

University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Aleksander Jankowski

Modeling transcription factor complex binding
to eukaryotic genomes

PhD dissertation

Supervisor

Professor Jerzy Tiuryn

Institute of Informatics
University of Warsaw

October 2014

Author's declaration:

aware of legal responsibility I hereby declare that I have written this dissertation myself and all the contents of the dissertation have been obtained by legal means.

October 14, 2014

date

.....

Aleksander Jankowski

Supervisor's declaration:

the dissertation is ready to be reviewed.

October 14, 2014

date

.....

Jerzy Tiuryn

Modeling transcription factor complex binding to eukaryotic genomes

ABSTRACT

The binding of transcription factors (TFs) to their specific motifs in genomic regulatory elements of eukaryotic organisms is commonly studied in isolation. However, in order to elucidate the mechanisms of transcriptional regulation, it is essential to determine which TFs bind DNA cooperatively as dimers or higher order complexes, and to infer the precise nature of these interactions. So far, only a small number of such cooperative complexes are known.

In this thesis, we present a method for predicting cell-type-specific TF-TF dimerization on DNA on a large scale. We applied it to DNase I hypersensitivity data, representing the universe of possible TF complexes by their corresponding motif complexes, and analyzed their occurrence at cell-type-specific DNase I hypersensitive sites. This way, we predicted 603 highly significant cell-type-specific TF dimers, the vast majority of which are novel. Our predictions included 69% (20/29) of the known dimeric complexes manually compiled from the existing biochemical literature. The predictions were also independently supported by evolutionary conservation, as well as quantitative variation in DNase I digestion patterns.

Notably, the known and predicted TF dimers were almost always highly compact and rigidly spaced, suggesting that TFs dimerize in close proximity to their partners, which results in strict constraints on the structure of the DNA-bound complex. Applying our method to ChIP-seq data, we uncovered a general principle governing the structure of TF-TF-DNA ternary complexes, namely that the flexibility of the complex is correlated with, and most likely a consequence of, inter-motif spacing.

To allow for a broad adoption of our method, we developed TACO, a software tool that takes as input any genome-wide set of regulatory elements and predicts cell-type-specific TF dimers based on enrichment of motif complexes. This is the first tool of such kind that can accommodate motif complexes composed of overlapping motifs, which are a characteristic feature of many known TF dimers. Our method comprehensively outperforms the existing approaches, iTFs and SpaMo, when benchmarked on a reference set of 29 known dimers.

Finally, we propose MOCCA, a novel computational method to identify individual TF binding sites from genome sequence information and cell-type-specific experimental data, such as DNase-seq. We combine the strengths of its predecessors, CENTIPEDE and Wellington, while keeping the number of free parameters in the model reasonably low. Our method is unique in allowing for multiple binding modes for a single TF, differing in their cut profile and overall number of DNase I cuts.

Keywords: transcription factor complexes, cooperative binding, dimerization, DNA sequence motifs, regulatory elements

ACM Classification: J.3

Modelowanie wiązania kompleksów czynników transkrypcyjnych w genomach eukariotycznych

STRESZCZENIE

Wiązanie czynników transkrypcyjnych do specyficznych motywów w elementach regulatorowych znajdujących się w genomach organizmów eukariotycznych jest zazwyczaj badane dla każdego z czynników oddzielnie, bez uwzględniania wpływu pozostałych. Jednakże w celu pełnego poznania mechanizmów regulacji transkrypcji należy rozstrzygnąć, które czynniki transkrypcyjne wiążą się kooperatywnie do DNA jako dimery lub kompleksy wyższego rzędu, a także poznać dokładną naturę tych oddziaływań. Do tej pory znana jest tylko niewielka liczba takich kooperatywnych kompleksów.

W tej pracy przedstawiamy wielkoskalową i specyficzną dla poszczególnych typów komórek metodę przewidywania dimeryzacji czynników transkrypcyjnych na DNA. Stosujemy ją do danych o nadwrażliwości na DNazę I, reprezentując możliwe kompleksy czynników transkrypcyjnych przez odpowiednie kompleksy ich motywów. Następnie analizujemy występowanie tych kompleksów w regionach otwartej chromatyny specyficznych dla poszczególnych typów komórek. W ten sposób przewidzieliśmy 603 istotnie nadreprezentowanych dimerów czynników transkrypcyjnych, spośród których zdecydowana większość nie była wcześniej znana. Nasze przewidywania obejmują 69% (20 z 29) zbioru znanych dimerów, który zebraliśmy z istniejącej literatury biochemicznej. Przewidywania były również niezależnie potwierdzone przez ewolucyjną konserwację, a także przez ilościową zmienność w profilach cięcia DNazą I.

Co istotne, zarówno znane, jak i przewidywane dimery czynników transkrypcyjnych niemal zawsze były zwarte i sztywnie rozmieszczone. Sugeruje to, że dimeryzacja czynników transkrypcyjnych zachodzi w bezpośredniej ich bliskości, co narzuca ściśle ograniczenia w strukturze kompleksu związanego z DNA. Stosując naszą metodę do danych ChIP-seq, ustaliliśmy ogólną zasadę regulującą strukturę kompleksów złożonych z dwóch czynników transkrypcyjnych i DNA, mianowicie, że ich elastyczność jest skorelowana z odstępem między motywami tych czynników na sekwencji DNA, prawdopodobnie będąc jego konsekwencją.

Aby umożliwić upowszechnienie naszej metody, opracowaliśmy program TACO, który przyjmuje jako wejście dowolny zbiór genomowych elementów regulatorowych i przewiduje dimery czynników transkrypcyjnych specyficzne dla poszczególnych typów komórek w oparciu o nadreprezentację kompleksów motywów. Jest to pierwsze narzędzie tego typu, które obsługuje kompleksy złożone z nachodzących na siebie motywów. Takie kompleksy są cechą wielu znanych dimerów czynników transkrypcyjnych. Na podstawie porównania ze wspomnianym zbiorem 29 znanych dimerów, stwierdziliśmy że nasza metoda prześciga istniejące, iTFs i SpaMo.

Pod koniec pracy przedstawiamy program MOCCA, będący nowatorską obliczeniową metodą identyfikacji poszczególnych miejsc wiązania czynników transkrypcyjnych, na podstawie informacji o sekwencji genomu oraz danych eksperymental-

nych specyficznych dla typów komórek, takich jak dane DNase-seq. Łączymy zalety dwóch poprzednich podejść, CENTIPEDE i Wellingtona, zachowując przy tym rozsądnie ograniczoną liczbę wolnych parametrów w modelu. Nasza metoda jest wyjątkowa przez to, że dopuszcza by pojedynczy czynnik transkrypcyjny miał wiele różnych stanów wiązania z DNA, różniących się profilami cięcia DNazą I oraz ogólną liczbą tych cięć.

Słowa kluczowe: kompleksy czynników transkrypcyjnych, kooperatywne wiązanie, dimeryzacja, motywy sekwencji DNA, elementy regulatorowe
Klasyfikacja tematyczna ACM: J.3

Contents

1	INTRODUCTION	1
1.1	Motivation	1
1.2	Organization of the thesis	3
2	COMPUTATIONAL PREDICTION OF TRANSCRIPTION FACTOR DIMERS	5
2.1	Introduction	5
2.2	Methods	7
2.2.1	Overview of the method	7
2.2.2	Identifying hypersensitive sites in 78 ENCODE cell types	9
2.2.3	Clustering of cell types into 41 cell-type clusters	9
2.2.4	Calculating motif occurrence statistics	12
2.2.5	Limiting the set of cooperativity predictions	15
2.2.6	Basic clustering of cooperativity predictions	17
2.2.7	Expanded clustering of cooperativity predictions	19
2.3	Results	21
2.3.1	Top-ranked predictions include known instances of TF cooperativity	21
2.4	Discussion	23
3	VALIDATION AND CHARACTERIZATION OF PREDICTED TRANSCRIPTION FACTOR DIMERS	25
3.1	Introduction	25
3.2	Methods	27
3.2.1	Comparison with ChIP-seq-based approach of Whittington et al. (2011)	27
3.2.2	Calculating DNase I cut density score	27
3.2.3	Calculating evolutionary conservation score	27
3.3	Results	29
3.3.1	Predicted interactions significantly overlap previous approach of Whittington et al. (2011)	29
3.3.2	DNase I cut density independently supports predicted physical interactions	29

3.3.3	Evolutionary conservation supports predicted physical interactions	31
3.3.4	Consistency of DNase-seq-based TF dimer predictions	32
3.3.5	Expanding the cooperativity landscape with additional DNase-seq datasets	33
3.3.6	ChIP-seq data extend the scope of TF dimer predictions	35
3.4	Discussion	37
4	STRUCTURAL PROPERTIES OF PREDICTED TRANSCRIPTION FACTOR DIMERS	39
4.1	Introduction	39
4.2	Methods	39
4.2.1	Analysis of motif spacing flexibility	39
4.3	Results	40
4.3.1	Predicted cooperative interactions are rigid and compact	40
4.3.2	Association between rigidity and compactness of TF dimers	43
4.3.3	Dynamic landscape reveals low TF dimer reuse across cell types	47
4.3.4	Predicted cooperative interactions indicate key role of FOXA1 in prostate cancer cells	47
4.4	Discussion	50
5	SOFTWARE FRAMEWORK FOR PREDICTING TRANSCRIPTION FACTOR DIMERS	55
5.1	Introduction	55
5.2	Identification of dataset-specific predictions	56
5.3	Implementation and applicability	57
5.4	Execution time and output	58
5.5	Benchmarking the dimer prediction tools	59
5.5.1	Benchmarking parameters for TACO	60
5.5.2	Benchmarking parameters for SpaMo	60
5.5.3	Benchmarking parameters for iTFs	61
5.6	Comparison of dimer prediction tools	61
5.7	Specification file format	63
5.7.1	General conventions	63
5.7.2	Reference genome	64
5.7.3	Input datasets	64
5.7.4	Sequence motifs	66
5.7.5	Scope of the analysis	66
5.7.6	Various options	67
5.7.7	Output files	68
5.8	Discussion	69

6	BUILDING ON TRANSCRIPTION FACTOR FOOTPRINTS TO PREDICT INDIVIDUAL BINDING SITES	71
6.1	Introduction	71
6.2	Methods	74
6.2.1	DNase-seq data from multiple sources	74
6.2.2	ChIP-seq data as a golden standard of TF binding	74
6.2.3	Prior probabilities of TF binding	76
6.2.4	Modeling the number of DNase I cuts	78
6.2.5	Expectation-Maximization approach	81
6.3	Results	88
6.3.1	MOCCA systematically outperforms the other tools	88
6.3.2	Knowledge of TF dimerization modes does not improve the prediction of individual TF binding sites	95
6.4	Discussion	99
7	CONCLUSION	101
	APPENDIX A TABLE OF ENCODE CELL TYPES	103
	APPENDIX B EXAMPLE SPECIFICATION FILES FOR TACO	109
	REFERENCES	119

List of Figures

2.1	Q-Q plot of observed vs. expected \log_{10} p -values of motif complex enrichment	8
2.2	Cell type dendrogram of 78 ENCODE cell types.	11
2.3	Definition of motif offset and motif spacing.	12
2.4	Identification of overrepresented cell-type-specific motif complexes.	14
2.5	The effect of motif complex orientation.	16
2.6	Cluster of highly similar motif complexes.	18
2.7	Top 10 predicted motif complexes, ranked by p -value.	22
3.1	DNase I cut density near predicted and incorrectly spaced motif complexes.	28
3.2	Evolutionary constraint signatures of predicted motif complexes.	31
3.3	Data sources and comparison of TF dimer predictions.	32
3.4	Top 10 predicted motif dimers in Duke DNase-seq data, ranked by p -value.	34
3.5	Top 10 predicted motif dimers in K562 ChIP-seq peaks, ranked by p -value.	36
4.1	Rigidity and compactness of transcription factor dimers.	41
4.2	Rigidity and compactness of transcription factor dimers after motif trimming.	43
4.3	Wide range of motif spacings for TF dimers predicted in K562 cells.	44
4.4	Predicted long range motif dimers in K562 ChIP-seq data.	45
4.5	Positive association between average motif spacing and flexibility of motif dimers.	46
4.6	Dynamic landscape of predicted TF dimers across cell types, UW DNase-seq data.	48
4.7	Dynamic landscape of predicted TF dimers across cell types, Duke DNase-seq data.	48
4.8	Key role of FOXA1 in prostate cancer cells (LNCaP).	49
4.9	Converging FOXA1 homodimer 3D structure seen from a different perspective.	50
5.1	Strongly cell-type-specific and weakly cell-type-specific paradigms.	57

5.2	Comparison of dimer prediction algorithms, UW DNase-seq data.	62
5.3	Comparison of dimer prediction algorithms, Duke and combined (UW+Duke) DNase-seq data.	63
5.4	Robustness of TACO with respect to motif sensitivity threshold chosen.	64
6.1	Example models of TF footprints learned by CENTIPEDE and MOCCA.	89
6.2	Example Receiver Operating Characteristic curves for the three tools.	91
6.3	Example Precision-Recall curves for the three tools.	92
6.4	MOCCA outperforms the other tools in terms of area under ROC curve.	93
6.5	Aggregated areas under ROC curves compared between three tools and three DNase-seq data sources.	94
6.6	Difference between MOCCA and Wellington areas under ROC curve.	94
6.7	Receiver Operating Characteristic curves for the known dimers. . .	96
6.8	Precision-Recall curves for the known dimers.	97
6.9	Distribution of the number of DNase I cuts learned by MOCCA for FOXA1 and its dimers.	98
6.10	Multinomial components of the models learned by MOCCA for FOXA1 and its dimers.	98

List of Tables

2.1	Known dimeric DNA-binding transcription factor complexes, manually compiled from the existing biochemical literature.	10
3.1	Cell-type-specific statistics of our predictions.	26
4.1	Motif dimers underlying the known DNA-binding TF complexes.	42
6.1	Numbers of reads in DNase-seq datasets used.	74
6.2	ChIP-seq datasets used as a golden standard of TF binding.	75
A.1	ENCODE cell types referred to in this thesis.	103

Acknowledgments

I would like to express my deep gratitude to both of my supervisors, Jerzy Tiuryn and Shyam Prabhakar. Professor Jerzy Tiuryn introduced me to the field of computational biology and then patiently guided me during my research. I am very grateful for his patience over the last five years. My second supervisor, Dr. Shyam Prabhakar from Genome Institute of Singapore, only for formal reasons could not be named as such on the title page. I very much appreciate the enormous amount of knowledge and experience he shared with me during my two-year stay in Singapore.

I would like to thank my two other co-authors, Ewa Szczurek and Ralf Jauch, for their ideas and questions, and for their inquisitiveness that motivated me to work.

I would also like to thank my colleagues with whom I worked. I really appreciated the help and advice of Paweł Bednarz, Przemysław Biecek, Agata Charzyńska, Piotr Dittwald, Norbert Dojer, Janusz Dutkowski, Anna Gambin, Paweł Górecki, Julia Herman-Iżycka, Bogusław Kluge, Mateusz Łącki, Jarosław Paszek, Michał Startek, Maciej Sykulski, Bartek Wilczyński and Michał Woźniak.

While in Singapore, I greatly benefited from the discussions with Nirmala Arul Rayan, Denis Bertrand, Akshay Bhinge, Gireesh Kumar Bogu, Rajoshi Ghosh, Jonathan Göke, Swee Hoe, Xiaoming Hu, Elita Jauneikaite, Asif Javed, Vibhor Kumar, Massimo Nichane, Niranjan Nagarajan, Jeremie Poschmann, Ricardo Cruz-Herrera del Rosario, Sigrid Rouam, Wenjie Sun, Davide Verzotto and Andreas Wilm.

Last but not least, I would like to thank my wife, Kasia, for her unconditional support. She always believed in me.

This work was supported by the Agency for Science, Technology and Research (A*STAR), Singapore [JCOAG03-FG02-2009]; the Ministry of Science and Higher Education, Poland [N N301 065236]; and the National Science Centre, Poland [N N519 652740, 2011/03/N/NZ2/03177, 2012/05/B/NZ2/00567]. I would also like to acknowledge the ENCODE Project and ENCODE Data Coordination Center at the University of California, Santa Cruz for making the experimental datasets analyzed in this thesis publicly available.

1

Introduction

1.1 MOTIVATION

Already Plato and Aristotle have discussed the nature of complex systems. Around 350 BCE, Aristotle in his *Metaphysics* noted that

“In the case of all things which have several parts and in which the totality is not, as it were, a mere heap, but the whole is something beside the parts, there is a cause.” (Aristotle, 2009, Book VIII, Part 6)

When the entirety has different properties than its components, these components cannot be only studied in isolation. Around 360 BCE, Plato in *The Republic* exemplified it by discussing beauty:

“Suppose that we were painting a statue, and some one came up to us and said, Why do you not put the most beautiful colours on the most beautiful parts of the body – the eyes ought to be purple, but you have made them black – to him we might fairly answer, Sir, you would not surely have us beautify the eyes to such a degree that they are no longer

eyes; consider rather whether, by giving this and the other features their due proportion, we make the whole beautiful.” (Plato, 2009, Book IV)

In the case of biological systems, the complexity arises out of a multitude of individual interactions between biochemical compounds. While each of such interactions may be effectively studied on its own, certain properties of the whole system may can only be studied at a higher level. Understanding how such emergent properties arise has been a major challenge for philosophers and naturalists over the centuries.

The beginning of the incredibly rapid development of molecular biology can be dated to 1953, when concurrent, albeit not joint, efforts of Francis Crick, Rosalind Franklin, James Watson and Maurice Wilkins led to the discovery of the structure of DNA. The drafts versions of the complete human genome were published in 2001 by the International Human Genome Sequencing Consortium and independently by Celera Genomics. The publicly funded follow-up, Encyclopedia of DNA Elements (ENCODE) Project, aims to identify all functional elements in the human genome (ENCODE Project Consortium et al., 2012). The policy of making the generated datasets publicly available allowed us to embark on the studies presented here.

The synthesis of proteins based on genetic information requires two steps: the first one being transcription, in which the messenger RNA (mRNA) is synthesized from a DNA template, and the second being translation, in which proteins are synthesized based on the matured mRNA transcript. The process of transcription is far more complex in eukaryotes, i.e. organisms whose cells contain a nucleus where the DNA is stored, than in prokaryotes, where the genetic material is not enclosed in any cellular compartment. Eukaryotes include all the multicellular and many unicellular organisms. In particular, all the animals, plants and fungi are eukaryotes.

Eukaryotic transcription is an extremely complex process taking place in the cell nucleus. The packaging of DNA around nucleosomes and subsequent higher order chromatin structure allows for a great level of regulatory control. The expression of individual genes is regulated by a multitude of processes affecting the corresponding DNA and RNA fragments. Here, we focus on the binding of transcription factors, i.e. proteins that recognize specific DNA sequence fragments and affect the rate of transcription by binding to the DNA.

1.2 ORGANIZATION OF THE THESIS

This thesis does not follow the traditional structure of Introduction, Methods, Results, Discussion and Conclusion chapters. Instead, I have decided to organize the thesis in five problem-based chapters, preceded by a general Introduction and followed by the Conclusion. Four of these five chapters are split into Introduction, Methods, Results and Discussion sections. This way, the results are put in a direct context of the previous work on a particular problem, and are presented immediately after describing the relevant methods.

In Chapter 2, we propose a novel computational method for predicting cell-type-specific TF-TF dimerization on DNA on a large scale. We apply it to DNase I hypersensitivity data and predict 603 highly significant TF dimers, the vast majority of which are novel. Chapter 3 is dedicated to validation and characterization of these predictions. They are independently supported by evolutionary conservation, as well as quantitative variation in DNase I digestion patterns. We also expand the cooperativity landscape by combining DNase-seq datasets from two sources, and by applying our method to ChIP-seq data.

In Chapter 4, we discuss the structural properties of predicted TF dimers. In particular, we observe a strong link between their rigidity and compactness, suggesting that TFs dimerize in close proximity to their partners. We also uncover a general principle governing the structure of TF-TF-DNA ternary complexes, namely that the flexibility of the complex is correlated with, and most likely a consequence of, inter-motif spacing.

Two last chapters are dedicated to software tools. In Chapter 5 we discuss TACO, a program that takes as input any genome-wide set of regulatory elements and predicts cell-type-specific TF dimers based on enrichment of motif complexes. We show that TACO comprehensively outperforms the existing approaches, iTFs and SpaMo. A dual problem is considered in Chapter 6. There, we propose MOCCA, a software tool that can leverage the information about dimerization partners while identifying individual TF binding sites from genome sequence information and cell-type-specific experimental data, such as DNase-seq.

In Appendix A, we list all the cell types referred to in this thesis. We suggest to refer to it to decipher the cell line acronyms and to understand their role in the human organisms. Finally, in Appendix B we give example specification files for TACO, the software tool discussed in Chapter 5.

Most of the work presented in this thesis is already published. Almost all results of Chapter 2, as well as parts of Chapter 3 and 4, were published in:

Jankowski, A., Szczurek, E., Jauch, R., Tiuryn, J., & Prabhakar, S. (2013). Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Research*, 23(8), 1307–1318.

This co-authored paper presented the method developed by me and my supervisors, Shyam Prabhakar and Jerzy Tiuryn. Moreover, Ewa Szczurek has proposed an approach to systematically validate our cooperative binding predictions against protein-protein interaction databases, which gave a strong support for our results. This approach, presented in the paper, is not included in the thesis. Furthermore, Ralf Jauch has developed the structural models described in Subsection 4.3.4, and drawn the right part of Figure 4.8. The manuscript has been drafted by me, and all the other co-authors greatly contributed to give it the final form.

The software presented in Chapter 5, along with the other parts of the results of Chapter 3 and 4, was published in:

Jankowski, A., Prabhakar, S., & Tiuryn, J. (2014). TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics*, 15(1), 208.

The idea for the software was developed jointly by me and my supervisors. The manuscript has been drafted by me, and subsequently edited by all the co-authors. The software and its documentation has been written by me; they are available online at <http://bioputer.mimuw.edu.pl/taco/>. The source code of TACO is also available on GitHub at <https://github.com/ajank/taco>.

The software presented in Chapter 6, codenamed MOCCA, was also written by me. Its source code is available on GitHub at <https://github.com/ajank/mocca>. Finally, all the errors that are found in this thesis are mine alone.

2

Computational prediction of transcription factor dimers

2.1 INTRODUCTION

Transcription factors (TFs) typically bind the eukaryotic genomes in clusters to form regulatory complexes (Berman et al., 2002). However, not much is known about the precise biochemical determinants of clustered TF binding. Moreover, the ability of TFs that have relatively low sequence specificity *in vitro* to bind with high specificity *in vivo* is one of the long-standing paradoxes of regulatory genomics.

One explanation for the above observations is provided by focal chromatin openness at regulatory elements, which attracts multiple TFs to the same stretch of genomic DNA, and is further reinforced by their co-binding. Such *indirect cooperativity* between proximal binding sites is mostly non-specific, since it applies in principle to any pair of TFs (Adams & Workman, 1995). Moreover, such co-binding TFs are only subject to the “fuzzy” spacing constraint of proximity (Hannenhalli & Levy, 2002; Yu et al., 2006).

Another biochemical mechanism is *direct cooperativity*, as exemplified by homo-

or heterodimerization of specific pairs of TFs on DNA. Note that this mechanism also applies to higher-order complexes of three or more TFs. However, for simplicity, we will henceforth only refer to TF “dimers”. Intuitively, one would hypothesize that such dimeric complexes should bind DNA with rigid or semi-rigid spacing (as opposed to variable or fuzzy spacing), due to the steric constraints imposed by protein-protein interaction. However, the actual prevalence of spacing constraints *in vivo* remains unknown, due to lack of comprehensive data.

Important examples of direct cooperativity between human TFs include the p53 homotetramer (Friedman et al., 1993), the NF- κ B heterodimer (Chen et al., 1998a), various bHLH dimers (De Masi et al., 2011), SOX2-POU5F1 (OCT4) dimerization in embryonic stem cells (Chen et al., 2008) and AR-FOXA1 dimerization in prostate cancer cells (Wang et al., 2011). Clearly, binding of dimeric TF complexes to DNA is central to gene regulation in many well-studied biological contexts. In addition to its role in facilitating TF clusters, direct cooperativity provides a simple resolution to the paradox of binding specificity. However, little is known about the overall extent and tissue-specificity of TF dimers in the human genome.

Here we present a method for comprehensively predicting cell-type-specific TF dimerization based on DNA sequence motifs of individual TFs and DNase I hypersensitivity data. As a showcase example, we utilize DNase I hypersensitivity profiles in 78 human cell types (ENCODE Project Consortium et al., 2012), as described in Subsection 2.2.2. Uniquely, our approach can model the statistics of overlapping motifs. As we show in Chapter 4, motif overlap is a feature of most TF dimers, and this capability is therefore a major improvement over existing techniques. We confirm the accuracy of our predictions by multiple means, including the analysis of their evolutionary conservation. Based on our method, in subsequent chapters we obtain new insights into the prevalence and scope of direct TF cooperativity, and the rigidity and compactness of such interactions.

Our method is based on enrichment analysis of motif pairs at specific spacings in cell-type-specific hypersensitive sites. Thus, it differs from several existing bioinformatics approaches that aim to identify fuzzily spaced co-binding of TF pairs, i.e. indirect cooperativity (Myšičková & Vingron, 2012; Qian et al., 2005; He et al., 2009; Bais et al., 2011). Recently, Whittington et al. (2011) described a method that, similarly to ours, predicts TF-TF dimerization based on enrichment of rigidly spaced motif pairs. However, this approach requires ChIP-seq data for one of the potentially cooperating TFs. In contrast, our approach is more broadly applicable, since

it requires only one experimental data set per cell type. Consequently, our TF-TF dimer predictions exceeded those of [Whittington et al. \(2011\)](#) by over a factor of 10, and the number of predicted dimeric binding sites in regulatory elements was greater by over a factor of 100.

ChIP-seq data have also been used for TF cooperativity prediction by [Wang et al. \(2012\)](#), who tested for non-randomly spaced motif pairs within binding peaks. The latter method is most suited for detecting fuzzily spaced TF-TF interactions. Consequently, the resulting predictions are different in nature from, and complementary to, those we present here.

2.2 METHODS

2.2.1 OVERVIEW OF THE METHOD

To account for the intrinsic similarity of many of the cell types considered, we used a systematic method to cluster them into coherent cell type clusters, based on the similarity of their hypersensitivity profiles. We describe the clustering method in detail in Subsection 2.2.3. Briefly, we accounted for the similarities between some of the 78 human cell types from the ENCODE Project ([ENCODE Project Consortium et al., 2012](#)) by grouping them into 41 distinct clusters, which we will henceforth refer to as “cell types.”

The 964 vertebrate motifs in TRANSFAC Professional ([Wingender, 2008](#)) were used as models of TF binding specificity, yielding 465,130 potential motif pairs. The central assumption of our method is that dimeric TF complexes would be juxtaposed in a constrained fashion when cooperatively bound to DNA. Consequently, the genomic binding sites of cooperating TFs should form rigid *motif complexes*, which we define as pairs of motifs with fixed relative orientation and offset (displacement between left edges of motifs). We therefore tested all possible compact motif complexes (motif spacing ≤ 50 bp; see Subsection 2.2.5) of each motif pair for enrichment in open chromatin regions specific to each of the 41 cell types.

To quantify enrichment, we counted the number of motif complex instances in each set of cell-type-specific hypersensitive sites, and then compared against a background model based on the number of instances in the union set of hypersensitive sites from all cell types (Figure 2.4). The significance of enrichment was assessed using a binomial distribution, after correcting for differences in motif co-occurrence frequency between foreground and background sets (see Subsection 2.2.4). The va-

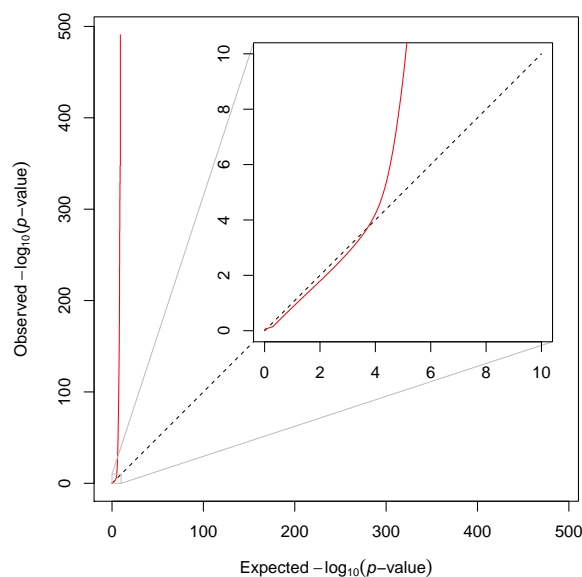


Figure 2.1: Q-Q plot of observed vs. expected $\log_{10} p$ -values of motif complex enrichment in all approximately 1.4 billion hypotheses tested. *Inset:* magnification of the first 10 decades of Q-Q plot. The calculated p -values fit the null expectation over the first 4 decades, indicating appropriate choice of statistical model.

lidity of our statistical approach is supported by the observation that our motif complex enrichment p -values fit the null expectation over four orders of magnitude and are, if anything, moderately conservative (Figure 2.1).

Motif complexes showing statistically significant enrichment ($p < 0.05$ after Bonferroni correction) were recognized as evidence of cell-type-specific TF cooperativity. Application of the approach across all approximately $1.39 \cdot 10^9$ motif and cell-type combinations yielded 5,233 significantly overrepresented motif complexes. For example, we found a highly significant AR-FOXA1 motif complex in the LNCaP prostate cancer cell line ($p = 8.1 \cdot 10^{-134}$; Figure 2.4), suggestive of widespread AR-FOXA1 dimerization at prostate cancer regulatory elements (Wang et al., 2011). Note that the motif complex was enriched only at one precise offset, indicating a rigid, strongly constrained heterodimeric structure.

Since the motif database frequently contains multiple motifs for a single TF, cooperative binding of one TF pair frequently resulted in enrichment of multiple equivalent motif complexes. We therefore clustered the 5,233 overrepresented motif complexes by similarity, so that each cluster constituted a distinct prediction of direct physical cooperativity in TF-DNA binding (see Subsection 2.2.6). Clustering yielded 603 distinct predictions, covering 30 of the 41 cell types (73%). Each cluster

was assigned the p -value of its most significant motif complex, which we refer to as the *signature motif complex*.

The number of known TF dimers is difficult to quantify, since the evidence is scattered over a large number of publications describing individual cases. We manually compiled a list of 29 known instances of direct cooperativity in DNA binding from the existing biochemical literature (Table 2.1). Although this list is possibly incomplete, it is nevertheless likely that our 603 predictions outnumber the known TF-TF-DNA complexes by over an order of magnitude.

2.2.2 IDENTIFYING HYPERSENSITIVE SITES IN 78 ENCODE CELL TYPES

We incorporated DNase I hypersensitivity datasets produced at the University of Washington as part of the ENCODE Project (UCSC Genome Browser track wgEncodeUwDnase). The 161 initially considered datasets covered 85 distinct cell types. We excluded some datasets with atypical GC-content spectra, reducing the number of datasets to 148, and the number of distinct cell types to 78 (data not shown). We relied on the hg19 read alignments provided by the ENCODE group. To identify hypersensitive regions, we used the F-Seq peak-calling algorithm (Boyle et al., 2008), treating each replicate separately.

We discarded hypersensitive regions, whose peak position lay within a repetitive region (union of RepeatMasker and Tandem Repeat Finder) and hard-masked repetitive basepairs in the remaining hypersensitive regions. We also hard-masked coding regions. To make the datasets obtained from different cell types comparable, we limited our analysis to the top 50,000 hypersensitive sites in each cell type. We also fixed the size of each hypersensitive region at 400 bp, centered on the F-Seq peak. Hypersensitivity calls from replicates were merged as described in the next subsection.

2.2.3 CLUSTERING OF CELL TYPES INTO 41 CELL-TYPE CLUSTERS

To avoid redundancy in our findings, we accounted for the similarities between some of the 78 human cell types by clustering them by their genome-wide DNase I hypersensitivity profiles (Figure 2.2). We represented the profiles of the 148 datasets as genome-wide binary vectors, with value 1 at positions within hypersensitive regions and value 0 elsewhere. We then calculated the dissimilarity between any two

TF complex	Sequence motif	PubMed ID
1. SOX–OCT (canonical)		22344693
2. SOX–OCT (compressed)		22344693
3. SOX–OCT (plus3)		22344693
4. HNF1–HNF1		2460858
5. p53–p53–p53–p53		8475074
6. SMAD–SMAD		21724602
7. TCF–RUNX		17158875
8. ETS–RUNX		20019798
9. AR–FOXA1		21572438
10. EBF1–EBF1		20876732
11. HNF4α–HNF4α		18829458
12. bHLH–bHLH		17148476
13. AR–AR, GR–GR or PR–PR steroid response elements (SREs)		10598584
14. p50–p65 (NF-κB)		9450761
15. ER–ER estrogen response element (ERE)		15036253
16. IRF–IRF interferon-stimulated response element (ISRE)		7687740
17. ETS–AP-1		16272134
18. ETS–IRF ETS–IRF composite element (EICE)		22992523
19. SOX9–SOX9		17264118
20. VD3R–VD3R vitamin D3 response element (VDRE)		1648450
21. TR–TR or RXR–TR thyroid hormone response elements (TRE)		1648450
22. RAR–RAR retinoic acid response element (RARE)		1648450
23. bHLH–GATA		9214632
24. STAT–STAT		7708771
25. AP-1–IRF AP-1–IRF composite element (AICE)		22992523
26. ETS-1–ETS-1		12034715
27. SOX2–PAX6		15558474
28. GATA–GATA GATApal		8628290
29. GABPα–CREB		23050235

Table 2.1: Known dimeric DNA-binding transcription factor complexes, manually compiled from the existing biochemical literature. For the complexes predicted in comprehensive analysis of UW DNase-seq data (Subsection 3.3.4, Figure 3.3b), their sequence motifs identified by TACO (see Chapter 5) are shown. The remaining motifs were compiled as spacing alterations of TACO predictions or juxtaposed TRANSFAC monomers.

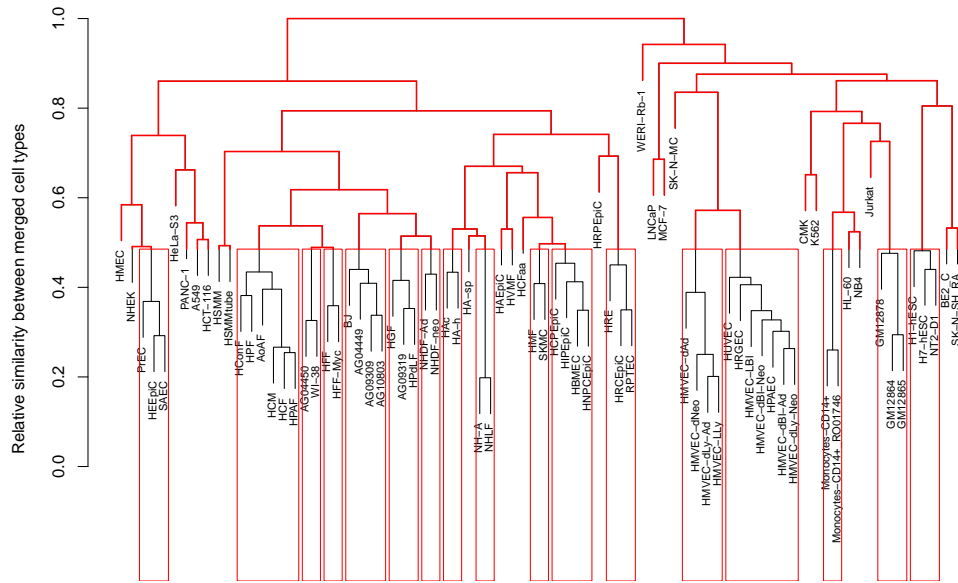


Figure 2.2: Cell type dendrogram of 78 ENCODE cell types. 78 ENCODE cell types were hierarchically clustered by the degree of overlap of their hypersensitive regions, resulting in 41 cell type clusters (see Subsection 2.2.3). The inner structure of the clusters is shown in black, whereas the relationships between the clusters are shown in red.

datasets as the Hamming distance between the respective binary vectors, scaled in such a way that the maximum dissimilarity across all comparisons equals 1.

Having calculated the dissimilarity matrix, we used the complete-linkage hierarchical clustering to collapse the 148 datasets from 78 cell types into cell type clusters. Before clustering, we first joined replicates from the same cell type at the lowest level of the dendrogram. The resulting dendrogram, along with the threshold defining the 41 cell type clusters, are presented in Figure 2.2. We then merged the sets of hypersensitive regions, obtained as described in the previous subsection, within each cell type cluster, combining overlapping regions into a single hypersensitive site.

Encouragingly, the resulting dendrogram recapitulated the expected developmental hierarchy. For example, blood cells formed a single supercluster, which split into lymphoid and myeloid branches. The lymphoid set further split into T-cell and B-cell subclusters, and the myeloid set into megakaryocytic leukemias (K562, CMK) and myeloblastoid cells (CD14+ monocytes and the promyelocytic leukemias, HL-60 and NB4). We manually thresholded the cell type dendrogram to define 41 distinct clusters, which we will henceforth refer to as “cell types.”

Cluster-specific hypersensitive regions were defined as genomic regions hypersensitive in a given cell type cluster, but not in any other cluster. In case of partial over-

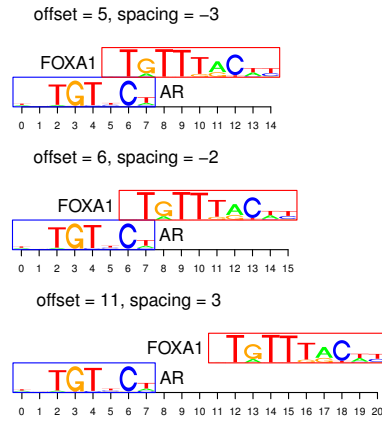


Figure 2.3: Definition of motif offset and motif spacing. Both are calculated between AR and FOXA1 motifs, in this order.

lap, the non-overlapping fragment was considered cluster-specific. For brevity, we will refer to the cluster-specific hypersensitive regions as “cell-type-specific hypersensitive regions.”

2.2.4 CALCULATING MOTIF OCCURRENCE STATISTICS

All 964 vertebrate motifs from TRANSFAC Professional 2011.2 were used as models of TF binding specificity. Given a pair of motifs, their *motif complex* was defined as a motif pair with a specified mutual orientation and offset. The *offset* was defined as the coordinate of the leftmost position of one motif in the coordinate system of the other motif (with zero-based start), whereas the *spacing* was defined as the number of intervening nucleotides between the edges of the two motifs (Figure 2.3). We allowed overlapping motif complexes, which were characterized by negative spacing. We considered only the motif complexes within up to 50 bp spacing between the two motifs. Let us denote by s the fixed orientation and offset of the motifs, and call it the *structure* of the motif complex.

For each combination of cell type, motif pair $(\mathcal{M}_1, \mathcal{M}_2)$ and its structure s , we calculated the significance of motif complex overrepresentation as follows. First, matches to individual motifs were identified within hypersensitive sites at a motif score threshold that provided at least 80% sensitivity (Rahmann et al., 2003). Pairs of motif matches that fit the specified structure s were taken as instances of the motif complex.

Let $C_{12}(s)$ and $c_{12}(s)$ be the number of observed motif complex occurrences in a

given set of cell-type-specific hypersensitive regions (foreground) and in the background set of all hypersensitive regions, respectively. Also, let $N_{12}(s)$ and $n_{12}(s)$ be the number of all possible complex occurrences in the foreground and in the background, respectively. By a possible occurrence of the motif complex we mean any occurrence such that the whole complex fits within the corresponding hypersensitive region. Then

$$f_{12}(s) = C_{12}(s)/N_{12}(s) \quad (2.1)$$

is the probability of observing the motif complex s in the foreground, and

$$b_{12}(s) = c_{12}(s)/n_{12}(s) \quad (2.2)$$

is the probability of observing the motif complex s in the background.

Let C_{12} be the total number of observed occurrences in the foreground of the pair of motifs $(\mathcal{M}_1, \mathcal{M}_2)$ with structure s ranging over spacings up to 50 bp and both orientations. In a similar way we define the numbers c_{12} , N_{12} , and n_{12} . Then $f_{12} = C_{12}/N_{12}$ is the probability of observing in the foreground the pair of motifs $(\mathcal{M}_1, \mathcal{M}_2)$ within a reasonable range of structures. Likewise, $b_{12} = c_{12}/n_{12}$ is the probability of observing in the background the pair of motifs $(\mathcal{M}_1, \mathcal{M}_2)$ within a reasonable range of structures.

The null hypothesis is that the conditional foreground probability $f_{12}(s)/f_{12}$ and conditional background probability $b_{12}(s)/b_{12}$ are the same. Consequently, the p -value of observing in the foreground at least $C_{12}(s)$ occurrences of the motif complex with a specified structure s can be calculated as the probability of observing at least $C_{12}(s)$ successes in $N_{12}(s)$ trials of the Bernoulli process with probability of success

$$f_{12} \cdot \frac{b_{12}(s)}{b_{12}}. \quad (2.3)$$

An intuition behind the success probability of the Bernoulli schema is that it is the background probability $b_{12}(s)$ of observing a given motif complex with structure s adjusted by the factor f_{12}/b_{12} , which reflects the relative motif pair densities in the foreground and in the background. Note that if we fix the pair of motifs and the structure s , then the background conditional probability stays the same and choice of cell type (foreground) affects the probability of success in the Bernoulli schema by the factor f_{12} .

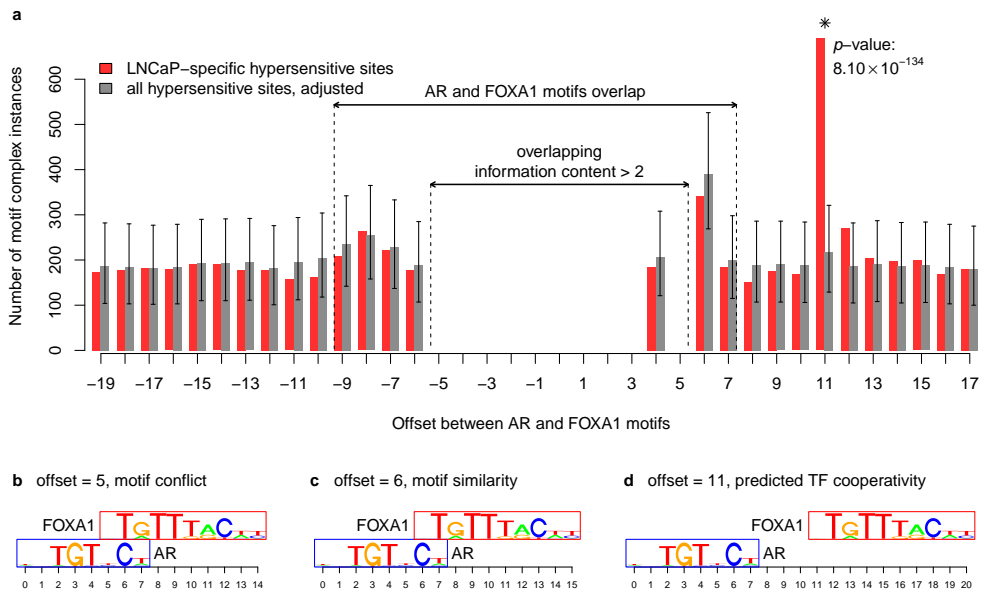


Figure 2.4: Identification of overrepresented cell-type-specific motif complexes. **(a)** Example of overrepresented motif complex specific to LNCaP (prostate cancer) cells. Number of instances of AR-FOXA1 motif complexes within LNCaP-specific hypersensitive sites (red bars) as a function of motif offset. Grey bars denote the expectation based on the background set of all hypersensitive sites (see Subsection 2.2.4). Offsets in the interval $[-9, 7]$ correspond to complexes with overlapping motifs. Offsets disallowed due to excessive motif overlap (see Subsection 2.2.5) are indicated. Error bars correspond to $p = 0.05$ after Bonferroni correction. The complex with offset 11, marked with an asterisk, was the only one overrepresented in LNCaP-specific hypersensitive sites; its Bonferroni-corrected p -value is indicated; **(b)** Examples of AR-FOXA1 motif complexes at 3 different offsets.

2.2.5 LIMITING THE SET OF COOPERATIVITY PREDICTIONS

We expected that transcription factors, which bind cooperatively in a particular cell type, should be also subject to individual overrepresentation in this cell type. To account for this expectation, we considered only pairs of motifs satisfying the condition $f_{12} \geq b_{12}$, i.e. pairs of motifs, which are at least as frequent in the foreground as in the background (within a reasonable range of structures).

Another constraint directly corresponded to steric hindrance between two TFs. Some approaches, e.g. [Whittington et al. \(2011\)](#) require that the motifs forming a motif complex must not overlap. However, many of the available motifs have redundant low-information positions at their ends, which would hinder the prediction of genuine TF cooperativities. Consequently, previous studies could not avoid trimming of low-information flanking regions of the motifs. We decided to apply a different approach, allowing minor motif overlaps, to retain all of the information contained in the binding affinity models. Our statistics accounts for possible over- or underrepresentation of motif complexes consisting of overlapping motifs (Figure 2.4). As explained below, excessive motif overlaps were disallowed as being highly unlikely; motif complexes dominated by one of the individual motifs were also disallowed.

To measure the degree of overlap, we introduced the concept of *overlapping information content*. For each overlapping motif position we define it as the minimum of the two information content values of the overlapping motifs. For the whole motif complex, we defined it as the sum of the overlapping information content values, ranging over all overlapping positions. We called an overlap *minor*, if the overlapping information content did not exceed 2 bits. We disallowed *major* (i.e. not minor) overlaps, because such colliding configurations are unlikely to correspond to direct TF cooperativity.

We also disallowed motif complexes, in which one of the individual motifs dominates the entire complex. To measure the share of an individual motif in a motif complex, we defined the *information contribution* of each motif. For a non-overlapping motif position it is simply equal to the information content of the individual motif at that position. For an overlapping motif position, if the two motifs differ in information content at that position, then the information contribution at that position of the more informative motif is equal to its information content at that position, and the information contribution at that position of the other motif

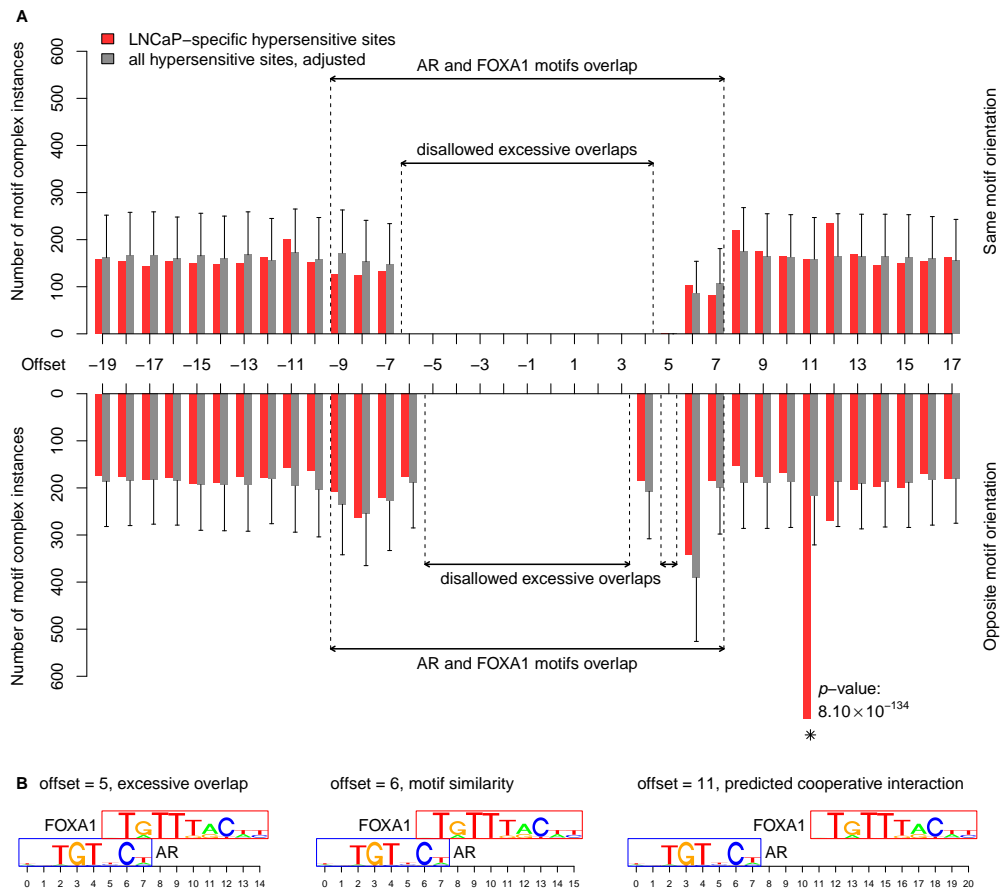


Figure 2.5: The effect of motif complex orientation. As in Figure 2.4, but indicating the number of motif complex occurrences for both of the possible motif orientations in a motif complex. The overrepresented complex characterized by offset 11 and opposite motif orientation manifests two mildly enriched shadow complexes, both of offset 12, one with opposite and one with same motif orientation.

is set to o . In case of equal information content, both of the motifs have the information contribution at that position set to half of their information content at this position. We defined the information contribution of a motif in the motif complex as the sum of its information contribution values, ranging over all positions. We considered only motif complexes in which both of the individual motifs had the information contribution of at least 6 bits.

To avoid artifacts arising from individual motifs that occur extremely rarely within hypersensitive sites, we considered only motif complexes that occurred at least 100 times within cell-type-specific hypersensitive regions (i.e. $C_{12}(s) \geq 100$). Moreover, we were aware that certain motifs are similar to themselves in a different layout. In particular, overrepresentation of a particular motif complex evokes possible overrep-

resentation of shadow motif complexes consisting of the same motifs, but with altered offset or orientation (Figure 2.5). We therefore allowed only one occurrence of each combination of motif pair and cell type, by incorporating only the motif complex with the smallest p -value. Finally, we considered only the overrepresented motif complexes with corrected p -value less than 0.05. The p -values were Bonferroni-corrected by multiplying by the total number of hypotheses tested, across all motif pairs, orientations, offsets and cell types (approximately 1.4 billion).

2.2.6 BASIC CLUSTERING OF COOPERATIVITY PREDICTIONS

Due to the redundancy of the motif database used, a single TF-TF cooperative interaction may be reported as multiple, mutually redundant, motif complexes (see, for example, Figure 2.6). We therefore clustered the 5,233 overrepresented motif complexes as described below. For each motif complex, we calculated its representative, called *dimer motif*, by counting nucleotide frequencies at all its instances, including a 5 bp margin on both sides.

As suggested by Gupta et al. (2007), we used the squared Euclidean distance (ED^2) as the dissimilarity measure of dimer motifs, assuming the clustering threshold of 2 for ED^2 . The overrepresented motif complexes were ranked by p -value in ascending order. We clustered them in a greedy manner, subsequently comparing each complex to already established clusters. The comparison was done by calculating ED^2 between the considered complex and the most significant motif complex in the considered cluster. If any ED^2 was less than 2, then the considered complex was merged with its counterpart with smallest p -value and discarded from further comparisons; in the other case, a new cluster was established. In this way we obtained the 603 clusters, which we refer to as *predicted dimers* or simply *predictions*. Each prediction was assigned the p -value of its most significant motif complex, which we refer to as the *signature motif complex*. Consequently, each prediction was characterized by the cell types in which its signature motif complex was predicted.

In rare cases, it may happen that longer monomer motif can be constructed by combining two short, degenerate motifs. To facilitate manual identification of such artifacts, we reported instances where the dimer motif closely matched ($ED^2 < 2.0$) a single motif from the database. Note that it would not be appropriate to automatically discard such dimer motifs, due to the contamination of motif databases with dimer motifs (e.g. SOX-OCT).

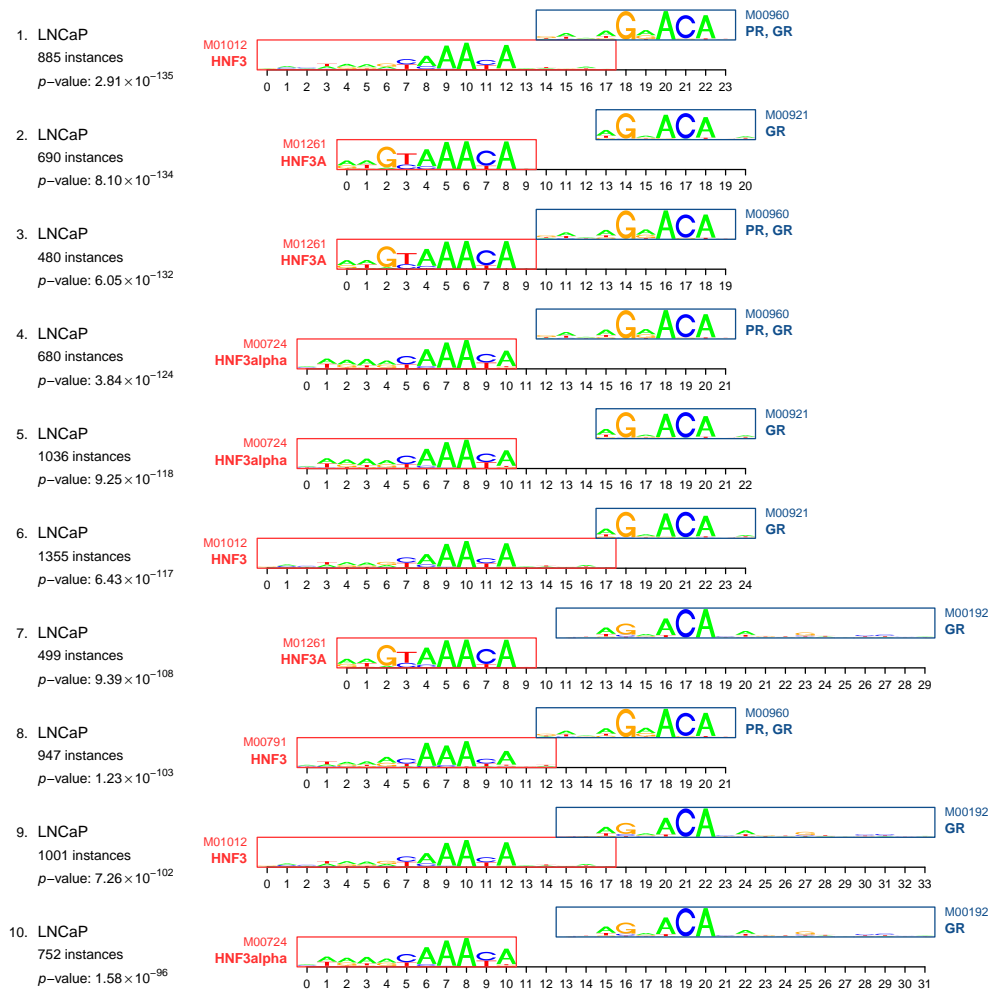


Figure 2.6: Cluster of highly similar motif complexes corresponding to AR-FOXA1 cooperativity. The top 10 overrepresented motif complexes comprising row 6 in Figure 2.7 are shown. The color of the motif bounding boxes indicates motif orientation relative to TRANSFAC motif database: blue – same orientation, red – opposite orientation.

2.2.7 EXPANDED CLUSTERING OF COOPERATIVITY PREDICTIONS

For the study presented in this chapter, we clustered the overrepresented motif complexes as described in the previous subsection. However, for subsequent studies we propose a more sophisticated algorithm, being a generalization of the one described above. The following expanded algorithm may be reduced to the basic one by setting $\alpha = 2.0$, $\beta = 0$ and $\gamma = +\infty$.

We rank the overrepresented motif complexes by p -value in ascending order (i.e. starting from the most highly enriched complex). Let us denote them by R_1, \dots, R_N . In order to cluster the complex R_n , we loop through $k = 1, \dots, n - 1$ and iteratively check if R_n is similar to R_k , as described below. If any of the comparisons yields a positive result, we immediately merge R_n into the cluster containing R_k . If the complex R_n cannot be incorporated into any of the existing clusters, a new cluster is created, with R_n as the *cluster seed*. In particular, the most enriched overrepresented motif complex, i.e. R_1 , gives rise to the first cluster.

To compare R_n to R_k , the following three tests are performed. If any of the three tests results in a positive outcome, the two complexes are deemed to be similar.

Test 1: motif complex identity. The first test is attempted only if R_k is the cluster seed of a previously established cluster. If R_n and R_k share the same motif complex, then R_n is *joined by motif complex identity* to the cluster of R_k . It occurs when the same motif complex is found overrepresented in different target datasets.

Test 2: dimer motif similarity. The second test is attempted only if R_k is a *signature motif complex*, i.e. the cluster seed or joined by motif complex identity to its cluster. Let $ED^2(R_n, R_k)$ be the squared Euclidean distance between the dimer motifs for complexes R_n and R_k . The simplest motif similarity criterion would be to impose a threshold on ED^2 . However, our approach allows highly specific motifs (those with high information content) to be further apart in Euclidean space, and still be considered similar. We therefore employ a distance threshold that is an affine function of the information content. If

$$ED^2(R_n, R_k) < \alpha \cdot IC(R_k) + \beta, \quad (2.4)$$

where α and β are user-provided parameters, and $IC(R_k)$ is the information content of the dimer motif for R_k , then R_n is *joined by dimer motif similarity*

to the cluster of R_k .

Test 3: overlap of genomic instances. The third test is attempted only if R_k is a signature motif complex or joined by dimer motif similarity. Let $C_{12}(R_n \cap R_k)$ be the number of their overlapping genomic instances (note that only overlaps conforming to the most common relative spatial arrangement of R_n and R_k are counted). Intuitively, we would like to capture the number of excess instances of R_n that are not also instances of R_k .

As described in detail in Subsection 2.2.4, the enrichment p -value of R_n is calculated as the probability of observing at least $C_{12}(R_n)$ successes in $N_{12}(R_n)$ trials of the Bernoulli process with probability of success $f_{12} \cdot (b_{12}(R_n)/b_{12})$, where $C_{12}(R_n)$ is the actual number of R_n instances in the target dataset, $N_{12}(R_n)$ is the number of all its possible occurrences in the target dataset, $b_{12}(R_n)$ is the probability of observing R_n in the control dataset, and f_{12} and b_{12} are the probabilities of observing the pair of motifs constituting R_n within a reasonable range of structures in the target and control dataset, respectively.

The success probability of this Bernoulli process combines two components: the “base” probability $b_{12}(R_n)$ of observing the motif complex R_n in the control dataset, and the factor f_{12}/b_{12} accounting for the enrichment of the underlying motif pair (i.e. motif complexes regardless of their spacing) in the target dataset.

Now we introduce

$$E_{12}(R_n) = N_{12}(R_n) \cdot f_{12} \cdot \frac{b_{12}(R_n)}{b_{12}} \quad (2.5)$$

as the expected number of instances of R_n following from the null model. Consequently, the number of excess instances over the null model now amounts to $C_{12}(R_n) - E_{12}(R_n)$. If

$$C_{12}(R_n \cap R_k) \geq \gamma \cdot (C_{12}(R_n) - E_{12}(R_n)), \quad (2.6)$$

where γ is a user-provided parameter, then R_n is *joined by overlap of genomic instances* to the cluster of R_k .

2.3 RESULTS

2.3.1 TOP-RANKED PREDICTIONS INCLUDE KNOWN INSTANCES OF TF COOPERATIVITY

All of the 10 most statistically significant cooperativity predictions matched known TF complexes (Figure 2.7). Moreover, the predicted cell type was also consistent with previous studies, in most of the cases. For example, the well-known cooperative interaction of POU5F1 (OCT4) with SOX2 (Ambrosetti et al., 1997; Chen et al., 2008), which is central to embryonic stem cell pluripotency, was ranked fourth and predicted in the correct cell type. Note that the OCT4–SOX2 heterodimer motif is sometimes mistakenly annotated in databases as an OCT4 or SOX2 monomer motif due to its high prevalence at OCT4 and SOX2 binding sites.

Note also that the monomers participating in cooperative binding are typically predicted only at the TF-family level, i.e. “OCT” or “SOX,” since TFs within a paralog family generally bind highly similar DNA sequences. Thus, additional domain knowledge or expression analysis is needed to determine exactly which representative of each TF family is involved in the DNA-bound complex (see, for example, (Carroll et al., 2005)). Occasionally, prior knowledge may alter the interpretation of TF identity within a dimeric complex. In most cases, this re-interpretation merely involves substituting one paralogous TF for another. However, in exceptional cases, such as the E-box motifs in Figure 2.7, the TFs implied by the predicted motif pairs are not paralogous to the actual TFs binding the motif (basic helix-loop-helix dimers).

Over all, 20 of the 29 known TF dimers (Table 2.1) were present among our predictions, suggesting that our method has 69% sensitivity. This number should be considered as a lower bound, since certain TFs from the set of known dimers may not be expressed in cell types considered in our study. Notably, our 36th ranked motif complex, NFAT–AP-1 ($p = 2.1 \cdot 10^{-40}$, <http://bioputer.mimuw.edu.pl/papers/tfdimers/>), matches the NFAT–FOS–JUN trimer that is known to synergistically regulate several immune-response genes (Chen et al., 1998b). This trimer was predicted by our algorithm because the sequence recognized by the FOS–JUN (AP-1) dimer was present as a single motif (accession number M00926) in TRANSFAC.

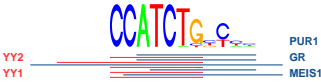

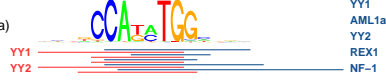
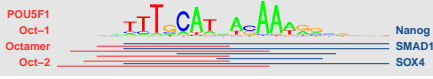
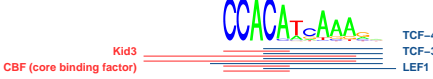
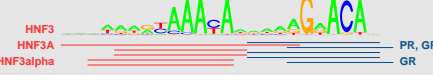

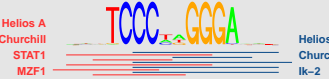

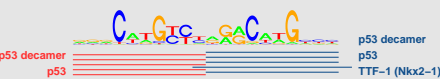
Prediction	Dimeric motif	Previous studies
1. WERI-Rb-1 (retinoblastoma) 3776 instances p -value: 2.01×10^{-482}		E-box dimer (ubiquitous) De Masi et al. 2011
2. HRCepiC:HRE:RPTEC (kidney epithelial cells) 1188 instances p -value: 7.83×10^{-320}		HNF1 homodimer (liver and kidney cells) Courtois et al. 1988 Cheret et al. 2002
3. WERI-Rb-1 + HFF:HFF-Myc + SK-N-SH_RA (retinoblastoma + fibroblast cells + neuroblastoma) 3306 + 218 + 858 = 4382 instances p -value: 5.55×10^{-268}		E-box dimer (ubiquitous) De Masi et al. 2011
4. H1-hESC:H7-hESC:NT2-D1 (embryonic stem cells) 5068 instances p -value: 1.26×10^{-198}		OCT-SOX heterodimer (embryonic stem cells) Ambrosetti et al. 1997 Chen et al. 2008
5. Jurkat (T lymphocytes) 1330 instances p -value: 7.15×10^{-167}		RUNX-TCF heterodimer (osteoblasts) Kahler and Westendorf 2003 Reinhold and Naski 2007
6. LNCaP (prostate adenocarcinoma) 885 instances p -value: 2.91×10^{-135}		FOXA1-AR heterodimer (prostate adenocarcinoma) Wang et al. 2011
7. GM12864:GM12865:GM12878 (B lymphocytes) 1218 instances p -value: 1.99×10^{-127}		IRF homotypic dimer (ubiquitous) Tanaka et al. 1993
8. BE2_C + NB4 (neuroblastoma) 464 + 152 = 616 instances p -value: 8.18×10^{-116}		EBF1 homodimer (B lymphocytes) Treiber et al. 2010
9. Jurkat (T lymphocytes) 712 instances p -value: 7.19×10^{-115}		ETS-RUNX heterodimer (T lymphocytes) Hollenhorst et al. 2009
10. HEEpiC:PREC:SAEC + NHEK (various epithelial cells) 740 + 444 = 1184 instances p -value: 5.03×10^{-97}		p53 homotetramer (ubiquitous) Friedman et al. 1993 McLure and Lee 1998

Figure 2.7: Top 10 predicted motif complexes, ranked by p -value. *Middle column:* below each motif complex, the locations of underlying individual motifs are indicated by red and blue lines. *Left column:* for each motif complex, the enriched cell types are separated by '+' symbols. The number of motif complex instances in hypersensitive sites specific to each cell type is also indicated. The p -value is given for the most significant prediction across the indicated cell types. *Right column:* TF dimer that binds the motif complex, with literature citations.

2.4 DISCUSSION

Genome-wide scans for DNase I hypersensitivity are a powerful tool for mapping cis-regulatory elements with high spatial precision in any given cell type (Crawford et al., 2006). One major advantage of this method is that, when combined with TF-DNA affinity models (motifs), DNase-seq can facilitate binding site predictions for a broad range of individual TFs (Pique-Regi et al., 2011) (Boyle et al., 2011). We have taken the latter approach one step further by using DNase-seq data to predict cooperatively bound TF complexes genome-wide. In all, we predicted cooperative binding of 603 signature motif complexes to 450,652 binding sites in regulatory regions specific to 28 different cell types. As a resource for future investigations, we provide these 603 motif complexes, along with exact genomic coordinates of their occurrences in cell-type-specific regulatory elements genome-wide (<http://bioputer.mimuw.edu.pl/papers/tfdimers/>).

The power of our method derives from the fact that it can in principle predict all TF complexes in a given cell type based on a single DNase-seq dataset. Additional datasets could be incorporated in the future to predict dimers in additional cell types. Judging from the set of 29 known cooperative dimers, our predictions have sensitivity of at least approximately 69%. The vast majority of the 603 predicted complexes are novel. Overall, our results suggest that TF dimerization is far more widespread than previously known. This provides at least a partial explanation to the paradox of TF-DNA binding specificity in large genomes. While TFs may individually possess low sequence selectivity, the complexes they form with other DNA-binding factors could be highly specific (Levine & Tjian, 2003). Thus, our results suggest that the current bioinformatics focus on predicting TF-DNA binding based on individual position weight matrices and chromatin openness data should be expanded.

3

Validation and characterization of predicted transcription factor dimers

3.1 INTRODUCTION

In the previous chapter, we have proposed a computational method for predicting transcription factor dimers. We have applied it to DNase-seq data for 41 human cell types, and found 5,233 significantly overrepresented motif complexes, which yielded 603 predicted complexes after clustering. Most of these predicted dimers are novel, hence they require a systematic validation and interpretation.

Most of the analyzed cell types was characterized by multiple TF dimers, with 15 cell types having at least 10 predictions after their clustering (Table 3.1). Not surprisingly, the number of individual overrepresented motif complexes and the number of predictions correlated well with the total length of cell-type-specific hypersensitive sites for a given cell type.

Apart from the above view, we also wanted to obtain a TF-centric perspective of our predictions from Chapter 2. Hence, we clustered all the individual 964 motifs using complete linkage hierarchical clustering, based on ED^2 between the motifs, to

Cell type cluster (individual cell types separated by colon, “:”)	Number of regions	Total length (base pairs)	Number of complexes	Number of clusters
A549	8 936	721 405	1	
AG04449:AG09309:AG10803:BJ	14 609	1 032 679	1	
AG04450:WI-38	7 700	569 630	1	1
AG09319:HGF:HPdLF	13 729	1 069 487	4	1
AoAF:HCF:HCM:HConF:HPAF:HPF	14 250	982 930		
BE2_C	17 217	2 298 836	199	17
CMK	16 575	2 199 464	59	9
GM12864:GM12865:GM12878	42 632	7 395 257	332	56
H1-hESC:H7-hESC:NT2-D1	44 857	8 661 450	395	62
HAc:HA-h	12 010	974 206		
HAepiC	7 414	701 085		
HA-sp	11 602	561 392		
HBMEC:HCPEpiC:HIPEpiC:HNPCEpiC	11 390	644 190		
HCFaa	6 184	322 649		
HCT-116	7 856	572 940		
HEepiC:PrEC:SAEC	19 322	1 474 439	28	10
HeLa-S3	15 086	1 661 709	10	1
HFF:HFF-Myc	6 992	433 603	47	3
HL-60	20 643	2 435 972	161	22
HMEC	18 037	1 125 606	2	1
HMF:SKMC	6 657	445 887	2	1
HMVEC-dAd:HMVEC-dLy-Ad:HMVEC-dNeo:HMVEC-LLy	21 941	2 399 603	13	5
HMVEC-dBl-Ad:HMVEC-dBl-Neo:HMVEC-dLy-Neo: HMVEC-LBl:HPAEC:HRGEC:HUVEC	26 516	2 659 135	462	15
HRCEpiC:HRE:RPTEC	21 572	3 004 717	775	53
HRPEpiC	14 025	2 131 177	28	12
HSMM	7 686	461 937	4	2
HSMMtube	12 334	1 079 127	13	7
HVMF	8 045	641 027	1	
Jurkat	22 054	3 675 475	397	35
K562	22 665	3 445 873	147	15
LNCaP	24 377	4 760 594	452	62
MCF-7	18 625	2 573 829	176	10
Monocytes-CD14+:Monocytes-CD14+_RO01746	21 556	2 585 081	581	36
NB4	11 025	1 042 263	38	5
NH-A:NHLF	6 451	317 333		
NHDF-Ad:NHDF-neo	14 627	1 654 633	18	6
NHEK	13 961	1 055 601	7	3
PANC-1	17 897	1 605 646	50	3
SK-N-MC	20 292	3 853 699	44	15
SK-N-SH_RA	22 308	2 818 340	28	6
WERI-Rb-1	33 411	7 400 569	757	162
control set (union of all hypersensitive sites from all cell types)	481 676	1 970 778 52		

Table 3.1: Cell-type-specific statistics of our predictions from Chapter 2. For each of the 41 cell types (i.e. cell type clusters) we indicate the number and total length of cell-type-specific hypersensitive sites, the number of overrepresented motif complexes and the number of predicted cooperative interactions.

obtain 350 motif clusters. The clustering threshold was set to 2.0, i.e. all the motifs in one motif cluster had their pairwise ED² not greater than 2.0. We observed that out of these 350 clusters of similar motifs, 129 participated in at least one prediction.

3.2 METHODS

3.2.1 COMPARISON WITH CHIP-SEQ-BASED APPROACH OF WHITINGTON ET AL. (2011)

We repeated our computational experiment from Chapter 2 using the motifs reported by Whittington et al. (2011). In case they used a custom motif, we applied the closest counterpart found in TRANSFAC, trimmed or extended respectively. We adjusted the motif sensitivity threshold in our method from 0.8 to 0.95, so that the number of individual motif occurrences in the genome was large enough for the overrepresentation statistics to be powerful.

3.2.2 CALCULATING DNASE I CUT DENSITY SCORE

We compared the number of DNase I cuts between the instances of a predicted signature motif complex and the instances of its slight alterations, which we refer to as *incorrectly spaced complexes*, consisting of the same two motifs, but with slightly increased spacing between them, by +1 up to +10 bp. Both the sets contained only the instances within hypersensitive sites specific to cell types for which the cooperativity prediction was made. Having fixed one prediction, we calculated the DNase I digestion patterns for both the predicted complex instances and incorrectly spaced complex instances, as shown in Figure 3.1. Our DNase I cut density score was the number of DNase I cuts in the ± 100 bp neighborhood of the motif complex instance, calculated with a triangular kernel and normalized within each prediction so that its average value for incorrectly spaced complexes equals 1. We then used the Mann–Whitney U test to assess whether the instances of predicted motif complex are more enriched in DNase I cuts than incorrectly spaced complex instances.

3.2.3 CALCULATING EVOLUTIONARY CONSERVATION SCORE

We followed a similar approach as for the DNase I cut density score, comparing the predicted and incorrectly spaced complexes. For each occurrence of the motif complex, we have calculated the weighted average of phyloP primate basepairwise cross-

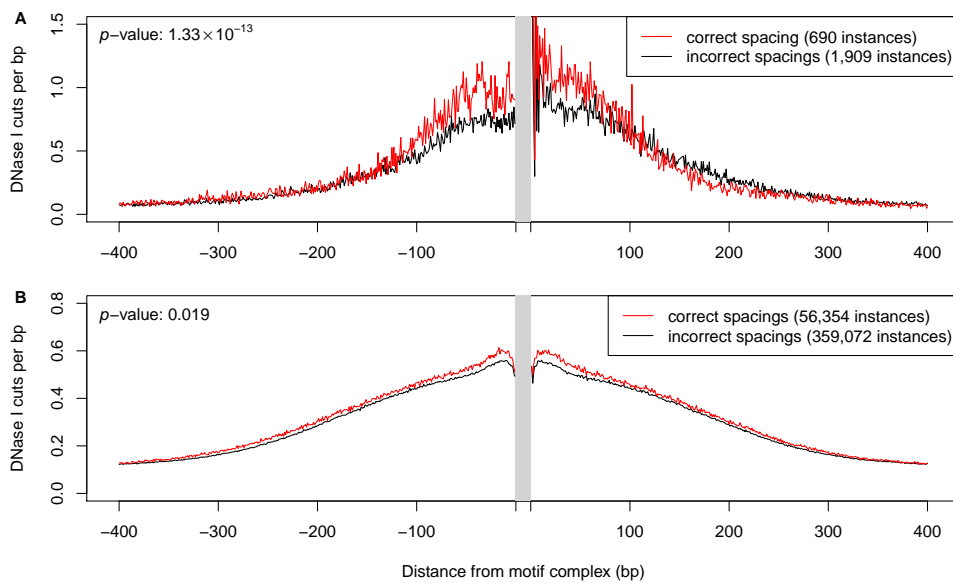


Figure 3.1: DNase I cut density near predicted and incorrectly spaced motif complexes.

(a) Example of AR-FOXA1. The average number of DNase I cuts in LNCaP-specific hypersensitive sites is shown in the vicinity of AR-FOXA1 motif complex instances. Red curve: DNase I cut density averaged over 690 instances of the predicted AR-FOXA1 motif complex (we predict that AR-FOXA1 heterodimer binds at these locations in LNCaP cells). Black curve: DNase I cut density averaged over 1,909 instances of incorrectly spaced AR-FOXA1 motif complexes (wider than the predicted spacing by 1 to 10 bp). The DNase I cut density is significantly higher within ± 100 bp of the predicted heterodimer binding sites;

(b) Similar to a: DNase I cut density averaged over the 54 predicted motif complexes that failed to show significant enrichment for DNase I cuts when analyzed individually.

species constraint scores (Pollard et al., 2010), where the weights were proportional to the information content at the corresponding nucleotide in the dimer motif. This weighting is justified by the fact that higher information content positions are likely to be more constrained. Again, we used the Mann–Whitney U test to assess whether the instances of predicted motif complex are more conserved than incorrectly spaced complex instances.

3.3 RESULTS

3.3.1 PREDICTED INTERACTIONS SIGNIFICANTLY OVERLAP PREVIOUS APPROACH OF WHITINGTON ET AL. (2011)

We compared our predictions with cooperative interactions inferred from motif analysis of ChIP-seq data (Whittington et al., 2011). We clustered the 59 human cell-type-specific motif complexes reported by Whittington et al. (2011) exactly as our complexes were clustered, and obtained 44 non-redundant predictions. Of these 44 predictions, 29 were reported in cell types for which we obtained DNase-seq data. We found that 9 of these 29 (31%) were also predicted by our method in at least one cell type, and 7/29 (24%) were predicted by our method in exactly the same cell type. Thus, there is a significant ($p = 2.6 \cdot 10^{-23}$), though incomplete, overlap between the two prediction sets. Apart from false positives and negatives in the two interaction sets, one possible reason for the limited overlap is that most of the TF-TF dimers predicted by Whittington et al. (2011) were predicted to bind at <30 locations in the genome. Our method, while more general, is only sensitive to TF-TF dimers with widespread binding, since it does not benefit from the precision of ChIP-seq data. This distinction is underlined by the observation that our 603 predicted TF dimers are estimated to bind at 450,652 locations genome-wide. In contrast, the human TF cooperativity predictions in Whittington et al. (2011) cover 1,821 genomic sites.

3.3.2 DNASE I CUT DENSITY INDEPENDENTLY SUPPORTS PREDICTED PHYSICAL INTERACTIONS

In predicting TF dimers, we did not use all of the information contained in the DNase-seq data. Specifically, we ignored variation in DNase-seq peak height – all hypersensitive sites were treated as equivalent. Consequently, we would expect false-positive motif complexes to be randomly distributed relative to peak height. In con-

trast, truly cooperative motif complexes should show a skew towards the “taller” hypersensitive peaks. This is because cooperativity would enhance TF-DNA binding and thereby enhance average chromatin openness (Pique-Regi et al., 2011; Boyle et al., 2011). This opens up another avenue for independently validating our predictions – we could test each predicted TF-TF dimer for bias towards taller hypersensitive peaks. Note that there is no circularity in this validation approach, since we are testing for peak-height skews within the set of DNase I hypersensitive sites, rather than between peaks and the rest of the genome.

For illustration, consider again the AR-FOXA1 motif complex. We predicted that AR-FOXA1 would bind cooperatively at 690 locations within LNCaP-specific hypersensitive sites, with the two individual motifs offset by 11 bp. We constructed the average density profile of DNase I cuts at cooperatively bound locations by aggregating over these 690 sites (see Subsection 3.2.2). For comparison, we considered 1,909 AR-FOXA1 motif complex instances with “incorrect” spacing (motif offset between 12 and 21 bp) within the same set of hypersensitive sites. If the two TFs did indeed bind cooperatively at the predicted motif offset, this cooperativity would result in stronger average TF-DNA binding at sites with the correct motif spacing, relative to sites with the incorrect spacing. Consequently, we would expect the cut density to be greater at the 690 correctly spaced sites, relative to the 1,909 incorrectly spaced sites. This is indeed the case, within the central 200 bp window (Figure 3.1a, $p = 1.3 \cdot 10^{-13}$). Our examination of the cut density profiles of other known TF dimers showed the same trend (data not shown).

We repeated the comparison of DNase I cut density profiles in Figure 3.1a for the entire set of 603 signature motif complexes, and found that, as a group, they collectively showed the expected cut density enrichment ($p < 10^{-300}$). At an individual level, 91% of the predicted cooperative interactions (549/603) showed statistically significant enrichment in DNase I cuts, after correcting for multiple testing ($FDR < 0.05$). Thus, most of our predicted dimers were independently supported by the cut-density test.

To obtain further insight into the remaining 54 (603 minus 549) predicted motif complexes that were rejected by this test, we averaged their collective DNase I cut profile, and compared it to the profile at the 540 corresponding incorrectly spaced complexes. Encouragingly, we again found significant local elevation of DNase I accessibility (Figure 3.1b, $p = 0.019$), suggesting that deeper sequencing of DNase-seq libraries could provide sufficient statistical power to validate several additional

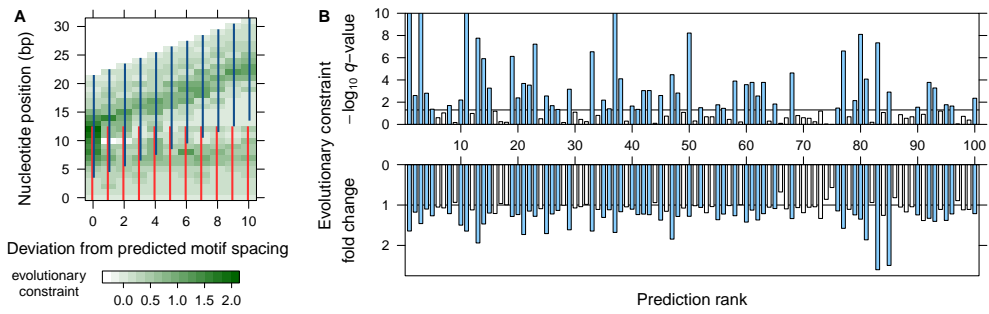


Figure 3.2: Evolutionary constraint signatures of predicted motif complexes.

(a) Example of FOXA1 (HNF3A) homodimer, ranked 11th and predicted in LNCaP (prostate cancer) cells. Again, we considered the predicted motif complex (*first column*) and its 10 incorrectly spaced variants. At each nucleotide position, color intensity indicates the average phyloP constraint score, weighted by information content at the corresponding motif position (see Subsection 3.2.3). Evolutionary constraint is highest at the predicted motif spacing;

(b) Evolutionary constraint q -values and fold change for the top 100 predicted motif complexes. Evolutionary constraint scores were calculated for each predicted motif complex and its 10 incorrectly spaced variants (see Subsection 3.2.3). For each prediction, we tested if the corresponding motif complex instances were enriched for evolutionary constraint relative to the remaining 10 spacings. We show the corresponding q -values (*top*) and fold changes (*bottom*) of evolutionary constraint scores between the predicted motif complex and its incorrectly spaced variants. Predictions with q -value below 0.05 are indicated by blue bars in both plots.

motif complexes.

3.3.3 EVOLUTIONARY CONSERVATION SUPPORTS PREDICTED PHYSICAL INTERACTIONS

Yet another approach to validate the predicted TF dimers would be to compare evolutionary conservation scores between predicted and incorrectly spaced motif complexes. This test has limited power, since TF binding sites are known to diverge very rapidly between species, and also because informative positions within motif complexes typically cover only approximately 5-10 bp. However, we still expected at least some of our predicted complexes to show a signal of evolutionary constraint; see, for example, the constraint profile of the FOXA1 (HNF3A) homodimer (Figure 3.2a). For this purpose, we used primate basepairwise conservation scores (Pollard et al., 2010), weighted by motif information content (see Subsection 3.2.3). For 23.7% of the predictions (143/603), we observed preferential evolutionary constraint (FDR < 0.05), further supporting the validity of our predictions (Figure 3.2b).

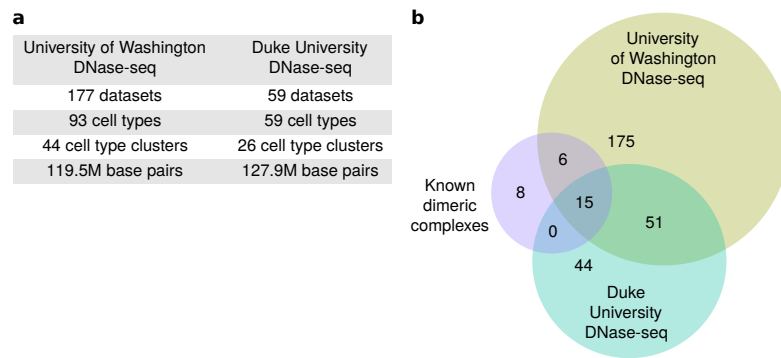


Figure 3.3: Data sources and comparison of TF dimer predictions.

(a) DNase-seq data sources.

(b) Comparison of TF dimer predictions obtained using UW and Duke DNase-seq data. The Venn diagram illustrates the overlap between the two sets and also the set of known DNA-binding TF dimers manually compiled from the existing biochemical literature (Table 2.1).

3.3.4 CONSISTENCY OF DNASE-SEQ-BASED TF DIMER PREDICTIONS

The ENCODE Project Consortium (ENCODE Project Consortium et al., 2012) provides multiple types of whole-genome open chromatin profiles, including data from DNase-seq experiments performed at the University of Washington (UW, track wgEncodeUwDnase) and Duke University (Duke, track wgEncodeOpenChromDnase). In order to obtain a comprehensive set of TF dimer predictions, and also assess the robustness and generality of our method, we ran it separately on both the UW and Duke collections.

For either of the data sources (UW or Duke), we considered all DNase-seq datasets from cell types under normal conditions (no treatment) that were not embargoed as of January 2013. We merged replicates and clustered cell types according to the similarity of their DNase-seq profiles, which resulted in 44 and 26 cell type clusters in UW and Duke, respectively (Figure 3.3a). Either of the data sources covered approximately 4% of the genome.

Application of our method to these two sets of genomic regulatory regions yielded 247 and 110 predicted TF dimers, respectively, of which 66 were shared (Figure 3.3b). Note that we did not expect complete overlap, since the 93 unclustered cell types from UW and the 59 from Duke shared only 15 cell types in common. After cell type clustering, the latter 15 contributed to 14 of the 44 UW cell types and 11 of the 26 Duke cell types. We also compare predicted TF dimers with a list of 29 known TF dimers manually compiled from the existing biochemical literature (Table 2.1).

Notably, we found that DNase-seq data from both UW and Duke were predictive of most of the known dimeric complexes.

3.3.5 EXPANDING THE COOPERATIVITY LANDSCAPE WITH ADDITIONAL DNASE-SEQ DATASETS

We expected that the known instances of direct TF cooperativity would tend to coincide with the most statistically significant TF dimer predictions, as was the case with our previous results based on UW DNase-seq data alone (Subsection 2.3.1). Focusing on the top 10 predictions derived from Duke data (Figure 3.4), we found 6 known interactions, the remaining 4 being novel predictions. Strikingly, while the known SOX9 homodimer (Genzer & Bridgewater, 2007) was detected as the 2nd ranked prediction, we also found two novel SOX homodimer motifs, ranked 5th and 10th respectively. The novel dimeric motifs are almost identical to the known SOX9 motif complex, except that the spacing between the monomer binding sites is increased or decreased by a single basepair. All three dimers were found to be specific to a cluster of melanoma (skin cancer) cell lines, consisting of Colo829 and Mel_2183. Interestingly, SOX9 is downregulated as melanocytes progress to melanoma, and its overexpression in melanoma cell lines inhibits tumorigenicity (Passeron et al., 2009). Our discovery of three distinct SOX9 homodimer binding modes in melanoma provides a single candidate molecular mechanism for the biological role of this TF in melanoma formation.

Another novel prediction, GATA-SMAD dimer ranked 6th, is in line with physical and functional interaction between GATA3 and SMAD3 reported by Blokzijl et al. (2002). However, we cannot rule out the alternative explanation, namely that this novel prediction is a variant of the known GATA-E-box dimer (Wadman et al., 1997), ranked 7th, with only a half-site of palindromic E-box motif being bound in this case.

The final novel prediction in Figure 3.4, GATA-GATA, ranked 8th in Figure 3.4, was found specific to K562 cell line. GATA is known to be a pioneer factor (Zaret & Carroll, 2011), and has been reported to bind cooperatively to a “GATApal” palindromic composite motif: ATCWGATAAG (Trainor et al., 1996). Our predicted dimer involves a converging pair of GATA motifs, as opposed to the diverging motifs in GATApal. By extension, we therefore call this prediction “GATAcpal”.

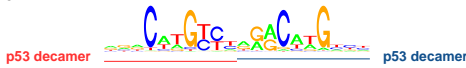


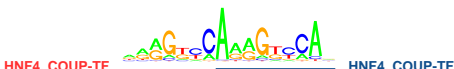
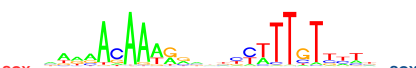


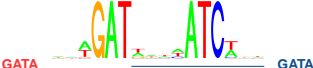

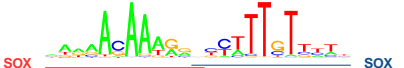
Prediction	Dimeric motif	Previous studies
1. HPDE6-E6E7;NHEK;pHTE:RWPE1, HMEC (various epithelial cells) 1602 + 220 = 1822 instances p -value: 4.98×10^{-254}		p53 homotetramer (ubiquitous) Friedman et al. 1993 McLure and Lee 1998
2. Colo829;MeI_2183 (skin cancer cells) 2110 instances p -value: 2.73×10^{-225}		SOX9-SOX9 homodimer (chondrocytic cells) Genzer and Bridgewater 2007
3. H1-hESC;H7-hESC;H9ES;iPS:iPS_CWRU1:iPS_NIH11:iPS_NIH7 (embryonic stem cells) 383 instances p -value: 3.23×10^{-101}		OCT-SOX heterodimer (embryonic stem cells) Ambrosetti et al. 1997 Chen et al. 2008
4. HepG2;Huh-7;Huh-7.5 (liver cancer cells) 3146 instances p -value: 4.99×10^{-92}		HNF4 α -HNF4 α homodimer (HeLa and pancreatic cells) Lu et al. 2008
5. Colo829;MeI_2183 (skin cancer cells) 1304 instances p -value: 3.22×10^{-88}		N/A
6. K562 (leukemia cells) 461 instances p -value: 1.90×10^{-64}		N/A
7. K562 (leukemia cells) 274 instances p -value: 1.61×10^{-62}		E-box-GATA heterodimer (leukemia cells) Wadman et al. 1997
8. MCF-7;T-47D (breast cancer cells) 568 instances p -value: 1.64×10^{-62}		N/A
9. GM12878;GM18507;GM19238;GM19239;GM19240 (B lymphocytes) 731 instances p -value: 2.17×10^{-62}		IRF homotypic dimer (ubiquitous) Tanaka et al. 1993
10. Colo829;MeI_2183 (skin cancer cells) 988 instances p -value: 9.74×10^{-54}		N/A

Figure 3.4: Top 10 predicted motif dimers in Duke DNase-seq data, ranked by p -value.

Left column: for each prediction, the enriched cell type, number of motif complex instances in cell-type-specific hypersensitive sites and p -value are indicated.

Middle column: below each dimer motif, binding sites for individual motifs are indicated. Only the structure of the cluster seed is shown. For clarity, we have manually interpreted the motif annotations.

Right column: literature citation on predicted TF dimer.

3.3.6 CHIP-SEQ DATA EXTEND THE SCOPE OF TF DIMER PREDICTIONS

To demonstrate the robustness of our method, we further applied it to 94 ChIP-seq datasets from K562 cells.

To demonstrate the ability to incorporate regulatory element annotations from multiple sources, we applied the algorithm to 127 replicates from 94 ChIP-seq experiments in K562 cells (ENCODE Project Consortium et al., 2012). For each experiment, we downloaded from Factorbook (Wang et al., 2012) the top 5 motifs found in ChIP-seq peaks using MEME (Bailey & Elkan, 1994).

We used our method to scan for motif complexes that contained at least one of the 5 motifs discovered in the respective dataset. The partner motif in the complex could be from the TRANSFAC database or from the entire set of motifs discovered in all K562 datasets. In total, our analysis yielded 81 predicted TF dimers, of which the top 10 are shown in 3.5. Ranked 1st is the known ETS-RUNX dimer (Hollenhorst et al., 2009), which was found in ChIP-seq peaks for PU.1, a transcription factor from the ETS family.

The 2nd ranked prediction, found in ChIP-seq peaks for NRSF (REST), actually represents a full-length, monomeric REST motif (Johnson et al., 2008). It was predicted by our method as a dimeric motif complex because “HudsonAlpha/NRSF: motif3”, the third-ranked motif discovered by MEME within REST ChIP-seq peaks, is actually only a fragment of the full-length REST motif, and the remaining fragment is very similar to the motif for nuclear receptors such as GR and PR.

The 4th ranked prediction is the known GATA-E-box motif complex (Wadman et al., 1997), which was also identified in the above-described analysis of Duke DNase-seq datasets (ranked 7th in Figure 3.4). Here, it is overrepresented in ChIP-seq peaks for the E-box-binding factor TAL1. Not surprisingly, among the top 5 motifs found in these ChIP-seq peaks, there is an E-box motif “Stanford/TAL1_(SC-12984): motif4”. The top 5 motifs also include the GATA motif “Stanford/TAL1_(SC-12984): motif2”. Such secondary TF motifs have been frequently reported in addition to the canonical ones (Wang et al., 2012). However, the biophysical interpretation of such secondary motifs is usually unclear. They could be a consequence of tethered binding, functional cooperativity or actual dimerization. These diverse mechanistic explanations can be distinguished more easily with the help of spacing analysis (Figure 2.4). In this case, it is clear that the secondary GATA motif at TAL1 ChIP-seq peaks is a consequence of GATA-TAL1 heterodimerization on DNA.

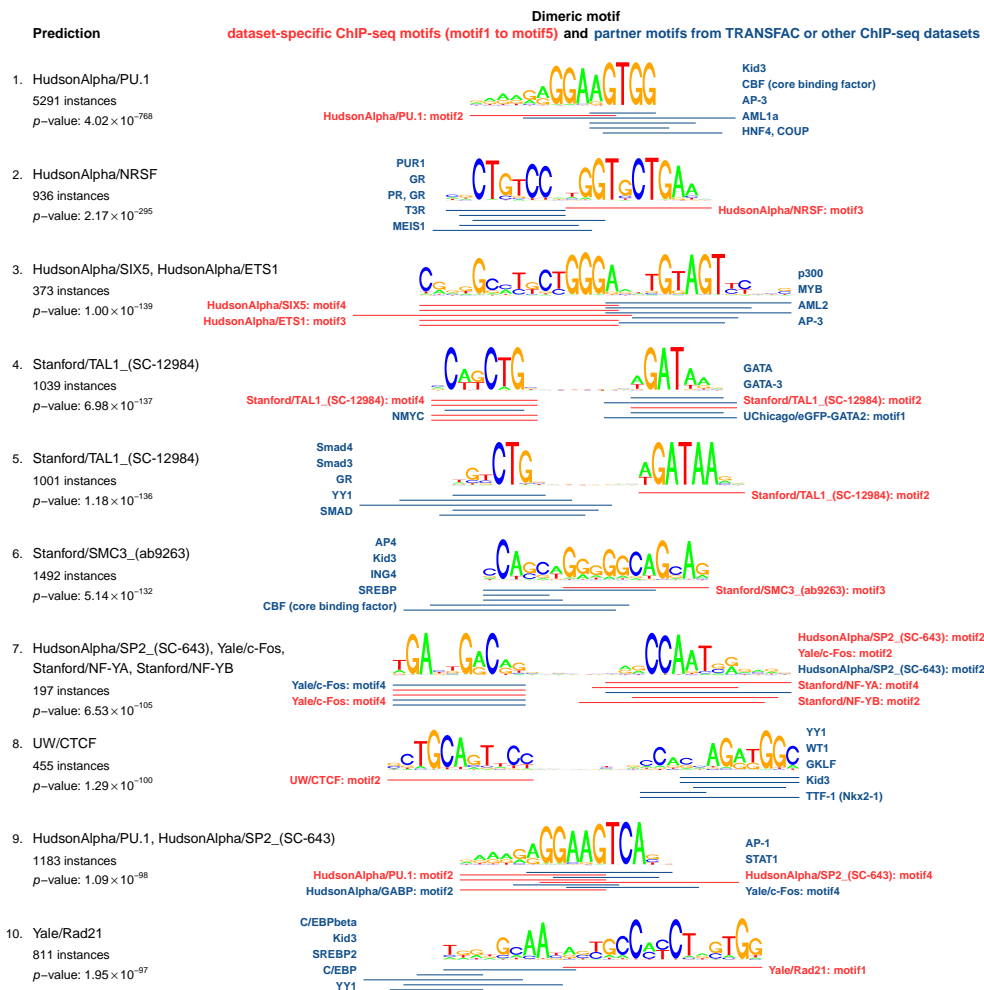


Figure 3.5: Top 10 predicted motif dimers in K562 ChIP-seq peaks, ranked by p -value.

Left column: for each prediction, the names of enriched ChIP-seq datasets, followed by the number of motif complex instances and p -value in most significantly enriched dataset.

Right column: below each dimer motif, the locations and names of underlying individual motifs are indicated for the top 5 overrepresented motif complexes. Red motifs correspond to the TF immunoprecipitated in an enriched ChIP-seq dataset, whereas blue motifs originate from TRANSFAC or other ChIP-seq datasets. For clarity, the red lines were drawn only once if the corresponding motif was shared across all 5 complexes.

3.4 DISCUSSION

We systematically validated our TF cooperativity predictions by comparing against a large-scale experimental database of protein-protein interactions and found highly significant overlap. This concordance is highly encouraging, given the profound differences between our computational method and experimental approaches. We also employed a novel statistical test to detect local elevation of the DNase-seq tag density, which validated 91% (549/603) predictions, and showed that at least some of the remaining 54 predictions would have also been validated if the corresponding DNase-seq libraries had been sequenced to greater depth. Another indication of functional relevance of the proposed complexes is the preferential evolutionary conservation of motif pairs with predicted structure. These findings independently support the accuracy of TF cooperativity predictions.

4

Structural properties of predicted transcription factor dimers

4.1 INTRODUCTION

We have so far confirmed the consistency and accuracy of the predicted TF dimers. This was done mostly by analyzing each of the predictions separately. In this chapter, we will discuss the structural properties of our predictions, arising from the general tendencies in the spacing of the overrepresented motif complexes, as well as from the degree of flexibility allowed in their structures.

4.2 METHODS

4.2.1 ANALYSIS OF MOTIF SPACING FLEXIBILITY

We defined motif spacing to be the number of intervening nucleotides between the proximal basepairs of the two motifs. In order to make the definition robust, we calculated motif spacing on the basis of trimmed motifs. Motif trimming was implemented as in [Whittington et al. \(2011\)](#), by eliminating flanking columns with in-

formation content less or equal 0.25 bit from both sides of the individual motifs. Note that motif trimming was only used to calculate motif spacing; our method did not require motif trimming.

To characterize the flexibility of TF-TF-DNA complexes, we grouped together the predictions that could have arisen from multiple spacings of the same TF dimer. In other words, we grouped together predicted motif complexes that shared the same pair of motifs in the same orientation, and varied only in their motif spacing. In the case of DNase-seq data, we only grouped predictions arising from the same dataset (for example, UW DNase-seq in GM12878 cells). Note that motif complexes within a group were constrained to all have the same left-right ordering of the individual motifs.

4.3 RESULTS

4.3.1 PREDICTED COOPERATIVE INTERACTIONS ARE RIGID AND COMPACT

There is some uncertainty in the literature about the spatial properties of motif pairs that are bound by TF dimers (Mirny, 2010; Biggin, 2011). Here, we define motif spacing as the number of intervening nucleotides between the edges of the two motifs (negative values indicate motif overlap). As noted above, numerous studies have tested for fuzzy motif spacing, and predicted TF-TF interactions with relatively large inter-motif distances (approximately tens of base pairs). In contrast, some biochemical analyses suggest that dimeric motif spacings should be rigid or semi-rigid, and also compact (<5 bp). Known TF complexes that fit this pattern include a number of SOX-OCT heterodimers (Ng et al., 2012) and several nuclear receptor dimers (Umesono et al., 1991). Our results clearly fit the latter model, as illustrated by the spatial pattern of motif complex enrichment scores corresponding to our top 100 predictions (Figure 4.1a). Note that most of the 603 predicted interactions require completely rigid spacing and the vast majority of the rest allow only 1 or 2 bp of variation in motif spacing (Figure 4.1b).

Interestingly, the vast majority (87.2%) of motif spacings among our 603 predictions were negative, indicating motif overlap (Figure 4.1c). It is possible that this high frequency of overlap merely represents an artifact of uninformative basepairs present at the flanks of TRANSFAC motifs. However, even after trimming potentially redundant motif positions (see Subsection 4.2.1), we still found that 67.8% of the motif pairs overlapped (Figure 4.2). Consistently, a high degree of overlap was

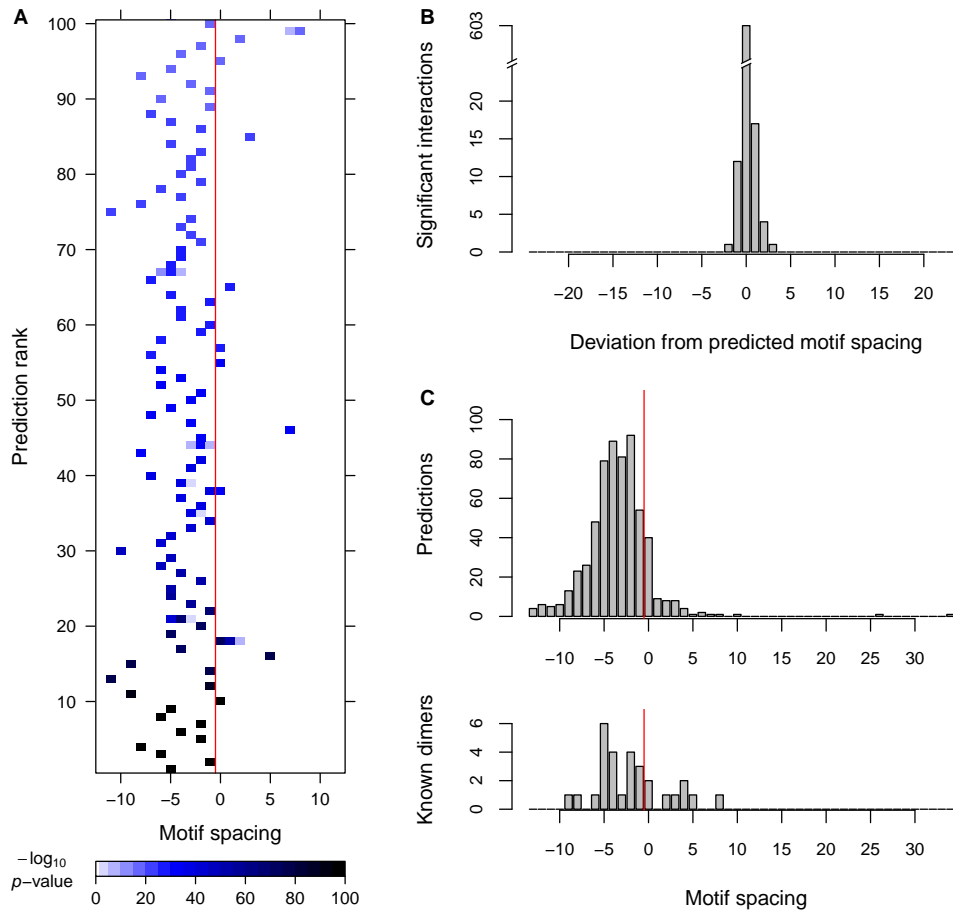


Figure 4.1: Rigidity and compactness of transcription factor dimers.

(a) For each of the top 100 predictions, we display the motif complex enrichment p -value as a function of motif spacing (see Subsection 2.2.4). Spacings to the left of the red line correspond to overlapping motifs;

(b) Very few of the 603 predicted motif complexes remain significantly enriched when motif spacing is altered, suggesting that cooperative motif complexes are rigidly spaced;

(c) Spacing distribution of 603 predicted motif dimers (top) and 29 known TF dimers (bottom; Table 4.1). Spacings to the left of the red line correspond to overlapping motifs. Predicted and known dimers are compact, i.e. tightly spaced.

Description	Motif 1	Offset	Motif 2	Spacing	Trimmed spacing	Cluster rank		
						Ch. 2	UW	Duke
SOX-OCT	M01247 →	12	M00795 →	-8	-2	4	1	3
SOX-OCT	M01247 →	11	M00795 →	-9	-3			
SOX-OCT	M01247 →	15	M00795 →	-5	1			
HNF1-HNF1	M01712 ←	5	M01712 →	-1	-1	2	3	22
p53-p53-p53-p53	M00761 →	10	M00761 ←	0	0	10	5	1
SMAD-SMAD	M01889 ←	3	M01889 →	-4	-2	10	5	
TCF-RUNX	M01705 ←	7	M01160 →	-2	-2	5	7	
ETS-RUNX	M00074 ←	8	M01160 →	-5	-3	9	8	89
AR-FOXA1	M00921 ←	12	M00724 ←	4	4	6	9	
EBF1-EBF1	M01003 →	5	M01003 ←	-6	-2	8	11	29
HNF4 α -HNF4 α	M00967 ←	7	M00967 ←	-2	0		12	4
bHLH-bHLH	M01808 ←	11	M01808 →	5	5	16	17	37
AR-AR, GR-GR or PR-PR	M00921 →	7	M00921 ←	-1	3	30	20	
p50-p65 (NF- κ B)	M01100 →	4	M00750 →	-5	-5	21	22	74
ER-ER	M00959 ←	6	M00959 →	-5	-5	25	25	
IRF-IRF	M01665 →	5	M01250 →	-2	0	7	31	9
ETS-AP-1	M01281 ←	4	M00926 →	-2	0	36	43	13
ETS-IRF	M00074 ←	9	M01250 →	-4	0	15	44	38
SOX9-SOX9	M01590 ←	12	M01590 →	0	2	293	50	2
VD3R-VD3R	M01270 ←	9	M01270 ←	2	3			
TR-TR or RXR-TR	M01270 ←	10	M01270 ←	3	4	179	58	
RAR-RAR	M01270 ←	11	M01270 ←	4	5			
bHLH-GATA	M01808 ←	14	M00789 ←	8	8	99	72	7
STAT-STAT	M00500 ←	7	M00500 →	-1	-1	63	109	11
AP-1-IRF	M00926 ←	4	M01881 →	-4	-1	98	210	68
ETS-1-ETS-1	M00074 ←	9	M00074 →	-4	0			
SOX2-PAX6	M01590 →	9	M00097 ←	-3	2			
GATA-GATA	M00462 →	5	M00462 ←	-5	1			
GABP α -CREB	M01660 →	1	M00113 ←	-5	-3			

Table 4.1: Motif dimers underlying the known DNA-binding TF complexes presented in Table 2.1. Cluster rank refers to: Ch. 2 - results in Chapter 2; UW and Duke - results in Subsection 3.3.4. Note that our method predicts p53-p53-p53-p53 homotetramer as a dimer of two homodimers, and SMAD-SMAD homodimer is found in the same cluster as p53 homotetramer.

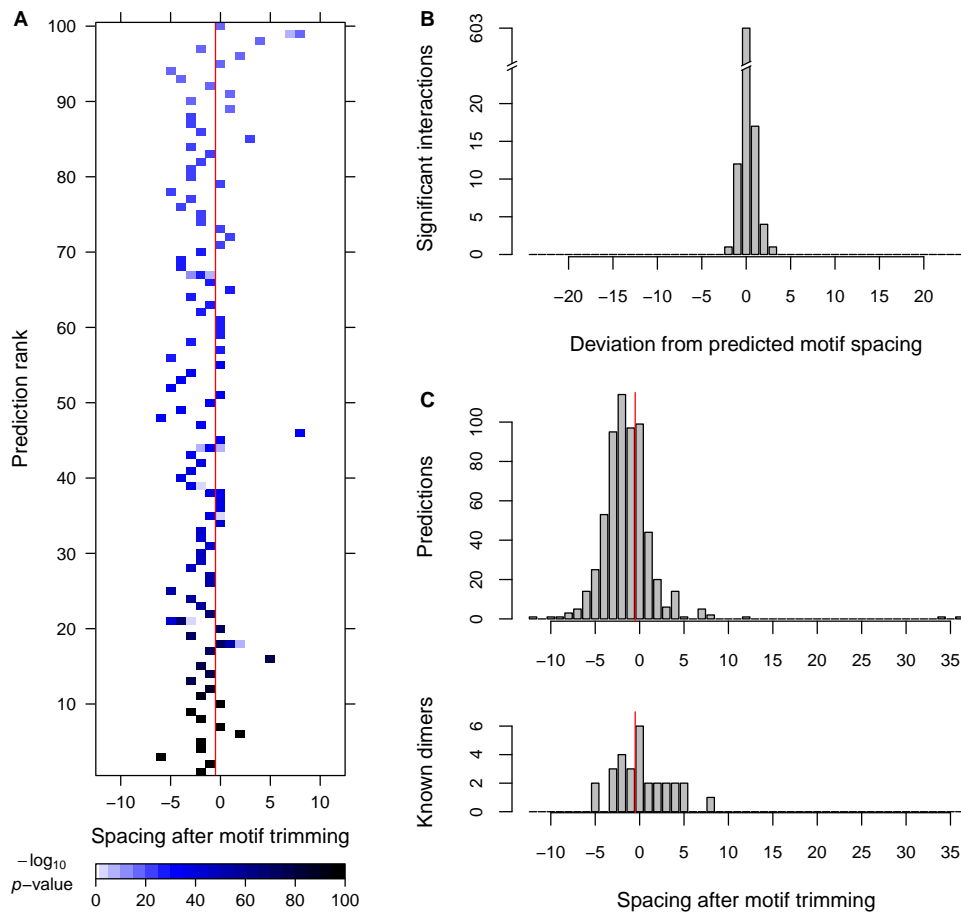


Figure 4.2: Rigidity and compactness of transcription factor dimers after motif trimming. Same as Figure 4.1, but after trimming potentially redundant motif positions (see Subsection 4.2.1).

observed even among the trimmed motif pairs corresponding to known TF dimers (Table 4.1, Figure 4.1c). Thus, 87.2% of the associations detected by our approach would be invisible to existing methods that do not allow motif overlap. Moreover, even after motif trimming, which is not necessarily advisable in all cases, 67.8% of our predictions would be undetectable by all existing approaches. Overall, our results indicate that TF dimers bind rigid and highly compact motif complexes.

4.3.2 ASSOCIATION BETWEEN RIGIDITY AND COMPACTNESS OF TF DIMERS

Notably, the analysis of overrepresented motif complexes in ChIP-seq peaks yielded multiple long-range interactions (spacing >15 bp), which were not discovered in our previous analyses of DNase-seq data (Figure 4.3). Most dramatically, we observed that in two such cases, ranked 40th and 41st, up to 5 motif spacings were signifi-

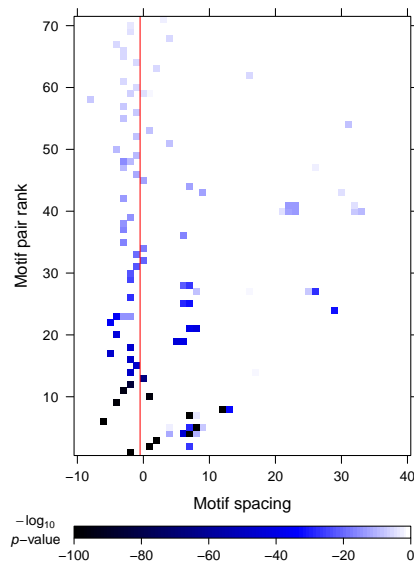


Figure 4.3: Wide range of motif spacings for TF dimers predicted in K562 cells. Predicted dimers that varied only in their spacing (same motif pair and orientation) were grouped together and ranked by the p -value of the most significant spacing. For each such group of dimer predictions in K562 ChIP-seq peaks, we show the motif complex enrichment p -value as a function of motif spacing. Spacings to the left of the red line correspond to overlapping motifs.

cantly overrepresented. Both of these predictions involved NF-Y homodimers, as did yet another of the predictions (Figure 4.4a). Of the 9 predicted NF-Y homodimers, 5 were direct repeats, 3 were divergent palindromes and 1 was a convergent palindrome. The 5 different spacings for the NF-Y direct repeat were broken up into two clusters one turn apart, and therefore phased to be on the same side of the DNA double helix. Another relatively widely spaced (>5 bp) interaction mentioned earlier, GATA-E-box, similarly permitted flexible spacing (Figure 4.4b).

In Subsection 4.3.1, we noted that TF dimers are mostly rigidly spaced and compact, and hypothesized that compactness explained rigidity. Here, we use the expanded set of dimer predictions to test this hypothesis. Consistently with this hypothesis, we uncovered a significant correlation between the rigidity and compactness of predicted TF dimers.

In order to quantify a potential association between rigidity and compactness of TF dimers, we aggregated our predictions derived from K562 ChIP-seq data into groups that varied only in their motif spacing (see Subsection 4.2.1), as in Figure 4.3. We then found Pearson correlation coefficient of $r = 0.51$ between the number of enriched complexes for a motif pair and their average motif spacing (Figure 4.5,

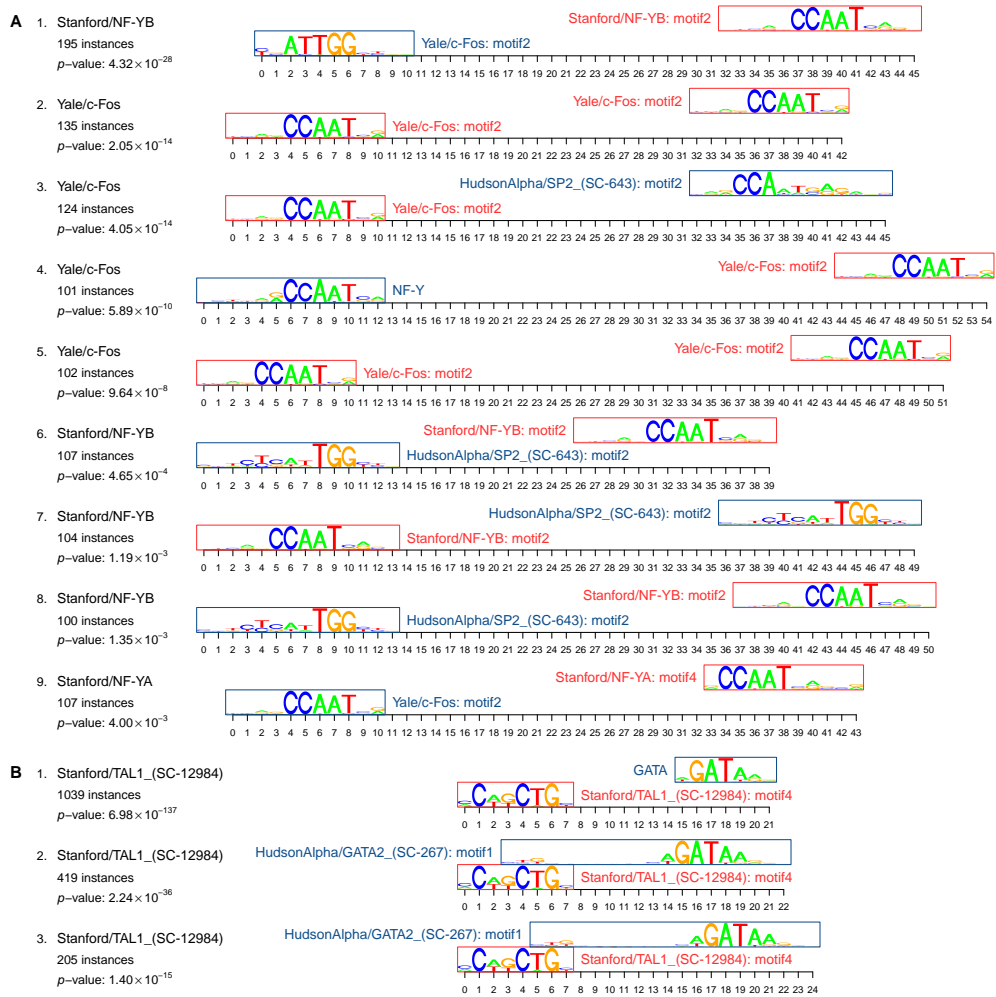


Figure 4.4: Predicted long range motif dimers in K562 ChIP-seq data. As in Figure 3.5, (a) NF-Y homotypic dimers and (b) GATA–E-box heterodimers predicted in K562 ChIP-seq data are shown in detail.

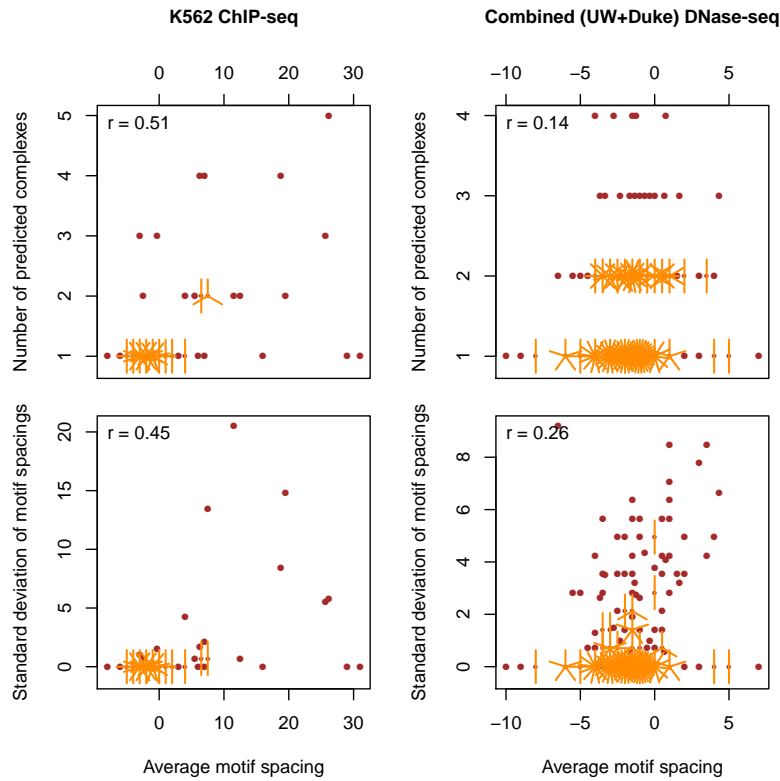


Figure 4.5: Positive association between average motif spacing and flexibility of motif dimers. *Left column:* predictions in K562 ChIP-seq peaks, *right column:* combined predictions from UW and Duke DNase-seq data. *Upper row:* sunflower plots show the number of predicted motif spacings for a group of dimer predictions as a function of the average of their motif spacings. In case of data points occurring more than once, their count is indicated by the number of petals (orange lines). *Lower row:* sunflower plots show the standard deviation of predicted motif spacings as a function of average motif spacing. The Pearson correlation coefficients are shown for all plots.

upper left). The difference in average motif spacing calculated within the prediction groups, compared between completely rigid motif complexes (single-spacing) and flexible complexes (more than one spacing) was found highly significant ($p = 4.07 \cdot 10^{-6}$, Mann-Whitney U test). Thus, we see a highly significant correlation between the rigidity and compactness of predicted TF dimers.

In order to test the generality of the abovementioned correlation, we applied the same approach to the combined set of DNase-seq dimer predictions, obtained using UW or Duke data. Again, we observed a positive Pearson correlation of $r = 0.53$ between the number of predicted complexes for a motif pair and their spacing. However, we noticed that four of the complexes in this case dominated the correlation coefficient by virtue of having outlier values for the motif spacing; their motif spacing was more than 5 interquartile ranges above the third quartile. When these four

data points were discarded, the correlation coefficient dropped to $r = 0.14$ (Figure 4.5, upper right). However, we still observed significantly larger average motif spacing among flexible complexes as compared to the completely rigid complexes ($p = 0.014$).

We further tested whether a more quantitative measure of dimer flexibility would also support the above findings on the structural properties of TF dimers. Consistently, we found that the average motif spacing also correlates with the standard deviation of motif spacings for a motif pair (Figure 4.5, lower left and right). In this case, the Pearson correlation coefficients were $r = 0.45$ for K562 ChIP-seq dimers and $r = 0.47$ for combined DNase-seq dimers ($r = 0.26$ after outlier removal). In summary, we found that the rigidity and compactness of motif complexes are consistently correlated, by multiple measures in two different data types.

4.3.3 DYNAMIC LANDSCAPE REVEALS LOW TF DIMER REUSE ACROSS CELL TYPES

The vast majority of TF dimers predicted in DNase-seq data were found specific to a single cell type only (87% or 215/247 in UW, 89% or 98/110 in Duke). Out of the 32 remaining dimers in UW, 29 were predicted in exactly two cell types (Figure 4.6) and usually found to be reused between related cell types (e.g. prostate cancer LNCaP and breast cancer MCF-7). Note that these predictions originated from disjoint sets of genomic regions (i.e. cell-type-specific hypersensitive sites), so the predictions in different cell types are independent. A similar trend of low TF dimer reuse was observed in Duke DNase-seq data (Figure 4.7).

4.3.4 PREDICTED COOPERATIVE INTERACTIONS INDICATE KEY ROLE OF FOXA1 IN PROSTATE CANCER CELLS

As noted above, all of the top 10 cooperativity predictions matched known TF dimers (Figure 2.7). However, the 11th-ranked prediction, which implies a FOXA1 (HNF3A) homodimer in prostate cancer cells ($p = 5.1 \cdot 10^{-93}$; Figure 4.8b), is, to the best of our knowledge, novel. This motif dimer also shows a very strong signal of preferential evolutionary constraint ($q = 5.2 \cdot 10^{-18}$; Figure 3.2a). Note that in the same prostate cancer cell line there already exists one well-known dimeric complex involving FOXA1, namely AR-FOXA1 (Wang et al., 2011), which ranked 6th amongst our predictions (Figure 4.8a). Inspired by these two cases, we searched for additional FOXA1 cooperative interactions among our predictions. Strikingly, we found a sec-

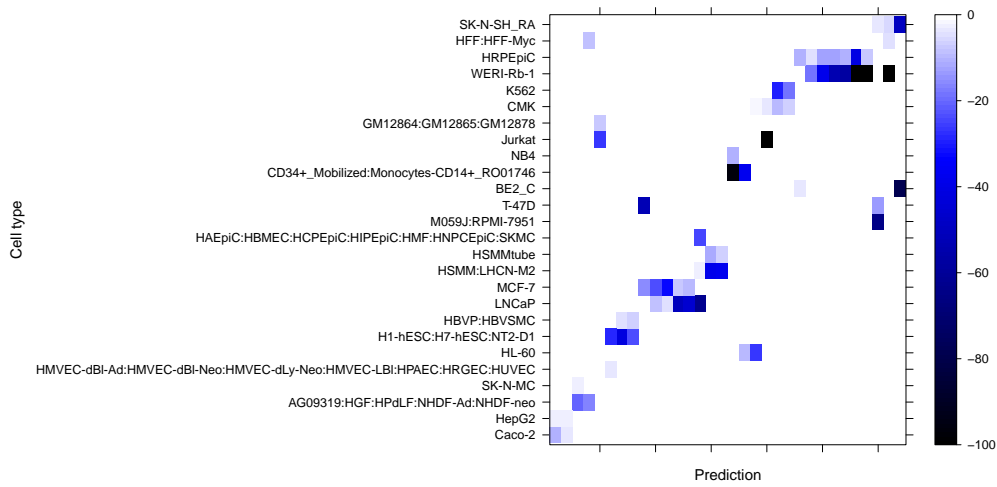


Figure 4.6: Dynamic landscape of predicted TF dimers across cell types, UW DNase-seq data. Each column of the heatmap represents a motif dimer predicted in UW DNase-seq data in more than one cell type. Dimers predicted only in a single cell type are not shown. Color intensity indicates the motif complex enrichment p -value in the given cell type. Rows and columns were clustered using complete linkage method with binary metric.

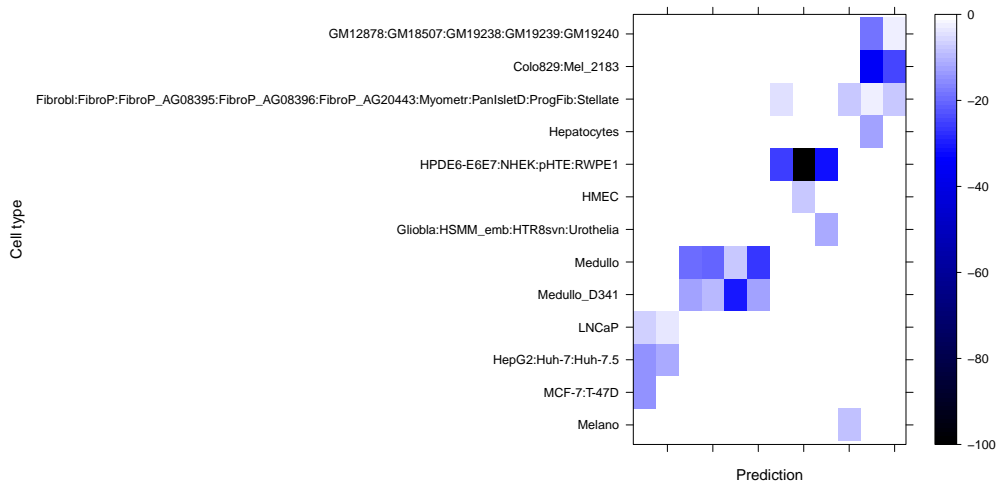


Figure 4.7: Dynamic landscape of predicted TF dimers across cell types, Duke DNase-seq data. As in Figure 4.6, but for motif dimers predicted in Duke DNase-seq data.

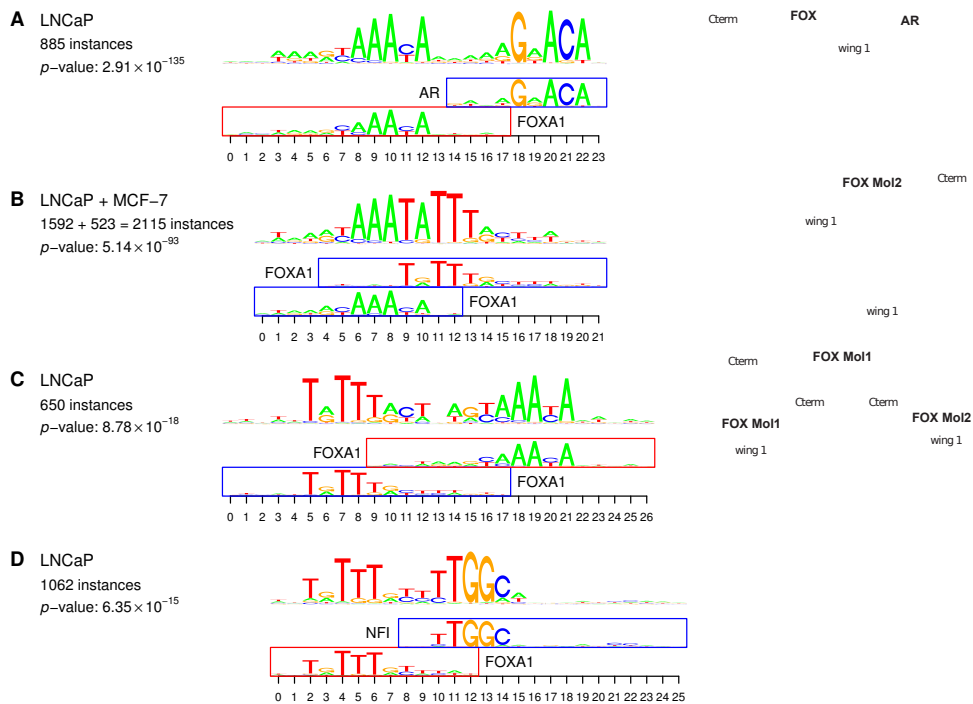


Figure 4.8: Key role of FOXA1 in prostate cancer cells (LNCaP).

Left: most significant cooperativity predictions involving FOXA1 and underlying overrepresented motif complexes. The number of instances and p -value are given as in Figure 2.7.

Right: predicted 3D structures of respective TF-TF-DNA complexes. (a) FOXA1-AR heterodimer; (b) diverging FOXA1 homodimer; (c) converging FOXA1 homodimer; (d) FOXA1-NFI heterodimer. Due to the lack of crystal structure for NFI in Protein Data Bank, no 3D structure is predicted in d.

and predicted FOXA1 homodimer, with a completely different structure (ranked 108th, $p = 8.8 \cdot 10^{-18}$; Figure 4.8c), as well as a predicted FOXA1-NFI heterodimer (ranked 139th, $p = 6.4 \cdot 10^{-15}$; Figure 4.8d). Thus, we predict that FOXA1 is involved in at least four strong cooperative dimeric binding modes in prostate cancer cells, only one of which was previously known.

To assess whether the four motif dimers involving FOXA1 topologically permit the assembly of dimeric TF complexes, we attempted to generate structural models. To this end, we first simulated ideal B-DNA structures containing the dimer motifs from Figure 4.8 using the w3DNA server (<http://w3dna.rutgers.edu/>). Next, we downloaded structural models from the Protein Data Bank (PDB) (Berman et al., 2000) containing androgen receptor (Shaffer et al., 2004) and FOX (Littler et al., 2010) DNA binding domains (PDB identifiers 1R4I and 3G73) when bound to DNA sequences that closely match the consensus of our composite motifs. Unfortunately, we found no PDB entries with reasonable sequence similarity to NFI. To assemble

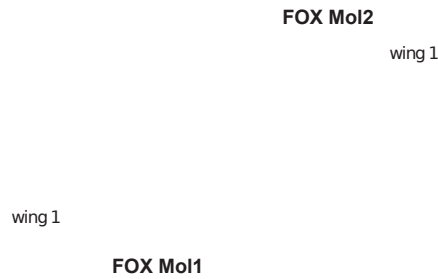


Figure 4.9: Converging FOXA1 homodimer 3D structure (Figure 4.8c) seen from a different perspective. Notably, C termini do not overlap.

hypothetical ternary TF-TF-DNA complexes, we superimposed DNA strands of the experimental crystal structures upon the simulated DNA with composite motifs using least-squares fitting in Coot (Emsley & Cowtan, 2004). We then visualized the resulting complexes using PyMOL (DeLano, 2002).

By analyzing the resulting models of TF dimers on DNA, we found that both homodimeric FOX complexes as well as the heterodimeric FOX-AR complex can assemble without any steric hindrance. Furthermore, the protein interfaces of the FOX-AR complex (Figure 4.8a) as well as the converging FOX homodimer (Figure 4.8c; Figure 4.9) are positioned favorably such that they could engage in direct protein-protein interactions. The diverging FOX homodimer (Figure 4.8b) is arranged on opposing faces of the DNA double helix, and direct protein-protein interactions between the DNA-binding domains are less likely in the present conformation, barring pronounced allosteric effects. It is possible that FOX-FOX binding cooperativity in this case is mediated by DNA conformational changes, as has been previously observed in multiple instances (Baburajendran et al., 2011).

4.4 DISCUSSION

We show that TF dimers were both rigid and compact, and hypothesize based on qualitative structural arguments that their rigidity is a consequence of their compactness. Such a causal relationship could arise for two reasons. Firstly, TF pairs binding widely spaced motifs are likely to form protein-protein contacts via their DNA-distal domains, or even via intervening cofactors. Such a configuration would in general be more flexible than direct physical contact between the DNA-binding domains. Secondly, a widely spaced complex might also gain flexibility from the greater deformability of the long stretch of intervening DNA. The widely spaced

complexes found in K562 cells provided us with an opportunity to test the above hypothesis. Our results indicate that TF dimers that bind widely spaced motif pairs are significantly more flexible in their spacing, thus providing statistical support for a causal relationship between compactness and rigidity (Figure 4.5). While our analysis provides the first evidence, further biochemical experiments are required to explore this relationship in greater detail.

In cases of very high inter-domain flexibility, as is perhaps true of NF-Y, even the relative orientation of individual motifs may vary. The NF-Y complex contains three proteins, NF-YA, NF-YB and NF-YC, of which only NF-YA forms specific contacts with DNA (Fleming et al., 2013). Thus, the NF-Y “dimer” motifs we identified are likely to be bound by pairs of such trimers, i.e. hexamers. It is possible that inter-trimer contacts are mediated not by the DNA-binding NF-YA subunit, but by the DNA-distal NF-YB or NF-YC subunits. Interestingly, the NF-Y motif was recently reported to form well-defined complexes of fixed spacing with E-box, E2F and TATA-box motifs at promoters genome-wide (Fleming et al., 2013), suggesting that the ternary complexes identified here are not the only cooperative interactions involving NF-Y. The same study also showed that NF-Y was unusually adept at binding genomic regions that showed no activating or repressive histone marks, suggesting that the TF acts as a pioneer factor. This is again consistent with our previous hypothesis that pioneer factors derive their DNA binding specificity from multiple dimeric binding modes.

Although the TF dimers predicted by our method are generally rigidly spaced, it is conceivable that this reflects to some extent an ascertainment bias of the algorithm. Dimers with highly flexible spacing would be harder to detect by this method, if they resulted in only weak enrichment of motif pairs at any given spacing. Similarly, the fact that all of the 29 known TF dimers we extracted from the literature are rigid or semi-rigid could also be questioned; one could hypothesize that existing biochemical assays for detecting cooperative dimerization on DNA are somehow biased against flexibly spaced dimers. However, we are not aware of any experimentally validated instances of TF dimers that can bind *cooperatively* with highly flexible motif spacing. Notably, in a recent study, even though the algorithm used to predict TF dimers permitted some flexibility in the spacing, all of the experimentally validated dimers turned out to be rigid, i.e. they bound with high affinity only at a single motif spacing (Kazemian et al., 2013). Thus, the evidence so far is strongly weighted towards rigid or semi-rigid TF dimers.

FOXA_I is well known to act as a pioneer factor in multiple cell types, including breast and prostate cancer cells (Zaret & Carroll, 2011). In other words, FOXA_I can initiate binding even at nucleosome-occluded DNA sites, and thereby potentiate subsequent binding of other factors. One would therefore imagine that FOXA_I should be able to bind all of its motif matches in the human genome. However, this is clearly not the case; in reality, FOXA_I binds only a small subset of its candidate sites (Lupien et al., 2008). Thus, there must be some other mechanism that compensates for the limited ability of chromatin openness to confer binding specificity upon pioneer TFs. Our results suggest that multiple homodimeric and heterodimeric binding modes could potentially contribute to the binding specificity of FOXA_I. Alternatively, one could hypothesize that dimerization may enhance the ability of this pioneer factor to compete with nucleosomes when the cognate DNA binding surface is not accessible. Interestingly, other known pioneer factors, such as GR and GATA (Zaret & Carroll, 2011), also appear among our top 40 predicted interactions, suggesting that dimerization could potentially represent a general specificity mechanism for pioneering TFs.

Previous studies have focused almost exclusively on fuzzily spaced co-binding of TFs, which is in general indicative of *functional* or *indirect* cooperativity. In contrast, biochemical studies suggest that only a single motif spacing, or at most 2-3 spacings, are compatible with *direct* cooperativity through TF dimerization (Cotnoir-White et al., 2011; Grove et al., 2009; Slattery et al., 2011). Moreover, even when TFs are seen to dimerize at a few different possible spacings, one spacing typically dominates in terms of binding affinity. For example, although OCT₄ and SOX₂ can dimerize at motif pairs separated by precisely three additional basepairs relative to the canonical OCT₄-SOX₂ motif spacing, the canonical spacing clearly provides greater binding affinity (Ng et al., 2012). Not surprisingly therefore, *in vivo* binding sites overwhelmingly favor the canonical spacing (Chen et al., 2008).

Our results indicate that there exists a large class of conformationally constrained TF dimers that bind rigidly-spaced motif complexes. The inflexibility of these motif complexes implies that dimerization on DNA frequently imposes strict constraints on the relative spatial conformation of the participating TFs. As in the case of OCT₄ and SOX₂, a small number of additional motif spacings may indeed provide alternate dimeric binding modes for the same factors, but these additional modes are likely to have lower affinity and also to contribute relatively few genomic binding sites. Finally, our predicted motif complexes are typically highly compact, perhaps

suggesting that TF dimerization is mediated by DNA-binding domains more commonly than by co-factors or DNA-distal domains.

5

Software framework for predicting transcription factor dimers

5.1 INTRODUCTION

The list of known DNA-binding TF dimers and multimers has expanded rapidly – we have compiled from the biochemical literature a list of 29 such complexes that have experimental support (Table 2.1). Concomitantly, numerous studies have used *in silico* analysis to computationally predict TF dimers. Since the goal of these studies was to predict specific ternary complexes of TFs with DNA, they scanned for pairs of TF-binding motifs enriched at a fixed relative orientation and spacing in regulatory regions. In Chapter 2 we have described such method that exploited the abundance of DNase-seq datasets available from the ENCODE consortium ([ENCODE Project Consortium et al., 2012](#)); subsequently, we also incorporate ChIP-seq data. Others have used DNase I hypersensitivity data on a smaller scale ([Kazemian et al., 2013](#)), as well as TF ChIP-seq data ([Whittington et al., 2011](#); [Hollenhorst et al., 2009](#)) and also sets of promoter or enhancer regions ([Chatterjee et al., 2012](#); [Fleming et al., 2013](#)) to define the regulatory elements of interest.

Up until recently, two software tools exist for performing the motif dimer enrichment analysis described above: SpaMo (Whittington et al., 2011) and iTFs (Kazemian et al., 2013). One important drawback of these tools is that they cannot assess enrichment of motif pairs that are so close that they overlap, even though such overlap is common (Figure 4.1). In Chapter 2, we propose a mathematical framework for TF dimer prediction that accommodates for motif overlap. Now, we introduce TACO (Transcription factor Association from Complex Overrepresentation), a software tool that generalizes this approach.

To allow for a broad adoption of our method, we encapsulate it into a configurable, publicly available standalone tool. We also compare TACO to SpaMo and iTFs, by benchmarking the three algorithms on the set of 29 known dimers.

5.2 IDENTIFICATION OF DATASET-SPECIFIC PREDICTIONS

We use DNA sequence motifs as models of TF binding specificity. In the default setting, we consider all possible pairs of the motifs provided. For each pair of motifs we test all possible compact motif complexes (all relative orientations and, by default, motif spacing of at most 50 bp) for enrichment in each of the target datasets. It should be noted that TACO can seamlessly handle the statistics of overlapping motif pairs, a property not shared by existing algorithms. As we will explain in detail in Chapter 4, this is an important feature, since a sizeable fraction of known TF dimers bind overlapping motif pairs.

To quantify enrichment, we count the number of motif complex instances in each target dataset, and compared it against the number of instances in the background model. The background model is based on the control dataset, defined as the union of all regulatory regions from all cell types. As described Subsection 2.2.4, the enrichment is calculated taking into account the difference in motif co-occurrence frequency between foreground (target) and background (control) datasets.

Motif databases very often contain multiple motifs for the same TF, or very similar motifs for different TFs. For this reason, a single underlying TF-TF interaction often results in the detection of multiple, highly similar motif complexes by TACO. We therefore cluster the overrepresented motif complexes, taking into account their similarity (measured by Euclidean distance) and overlap of their genomic instances, as described in Subsection 2.2.7.

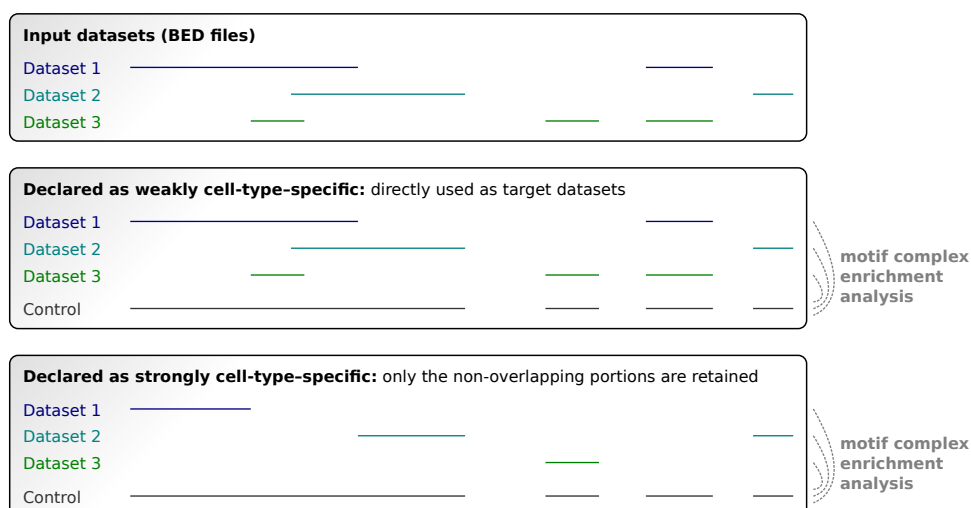


Figure 5.1: Strongly cell-type-specific and weakly cell-type-specific paradigms. In both of the paradigms, the control dataset is defined as the union of all input regions.

5.3 IMPLEMENTATION AND APPLICABILITY

Operating system(s): Unix-like, such as Linux and Mac OS X Other requirements: R or standalone R math library

TACO is a standalone C++ software tool. Its mandatory inputs are: reference genome sequence (FASTA format) and a list of TF motifs or a motif database. Accepted motif formats include TRANSFAC (Wingender, 2008), JASPAR (Bryne et al., 2008), SwissRegulon (Pachkov et al., 2013) and MEME (Bailey & Elkan, 1994) output. Moreover, a collection of genome-wide sets of regulatory regions should be provided (BED format). TACO can handle input regulatory region datasets of two kinds: strongly cell-type-specific or weakly cell-type-specific. Each input dataset should be declared as strongly or weakly specific (these two kinds can be provided simultaneously). In this thesis, all the DNase-seq datasets were processed according to the strongly specific paradigm. In contrast, ChIP-seq datasets considered here were treated as weakly specific.

Strongly and weakly cell-type-specific datasets are translated using different approaches into target datasets for TF dimer prediction (Figure 5.1). Regulatory regions of strongly specific datasets are intersected with each other, and only the non-overlapping (unique) portions are retained as target regions. In contrast, the weakly specific datasets are directly used as target datasets, without modification. The union of all input regulatory regions is used as a control dataset in order to build the null

model of motif complex occurrence.

The open chromatin datasets which could be used include publicly available DNase-seq data from the ENCODE Project (ENCODE Project Consortium et al., 2012). The input datasets can be provided as multiple replicates per cell type, to be merged by TACO within each cell type. In this way, closely related cell types, e.g. with similar genome-wide DNase I hypersensitivity profiles, may be merged as well.

The scope of the analysis may be narrowed down by screening for enrichment only in a subset of the target datasets. Moreover, instead of scanning for enrichment of all possible motif pairs, one or both of the motifs forming the motif complex can be fixed by the user. Below we provide three typical use cases for TACO.

Prediction of overrepresented motif complexes in a collection of DNase-seq datasets. All possible motif complexes are screened for enrichment in all cell-type-specific open chromatin regions. As stated, such analysis follows the concept presented in Chapter 2. Alternatively, only some of the datasets could be screened, with the remaining open chromatin datasets contributing only to the control set.

Prediction of overrepresented motif complexes in ChIP-seq peaks. The motifs of immunoprecipitated TFs are supplied, and all motif complexes with all possible partner motifs from the database are screened for enrichment in ChIP-seq peaks. This approach has previously been used by Whittington et al. (2011). The collection of ChIP-seq peaks should be large enough to provide a representative control set. For example, all publicly available ChIP-seq datasets from the ENCODE Project for a given cell type could be used.

Analysis of cooperative interactions between a given pair of TFs with known motifs. Some TF dimers allow for multiple spacings, and are overrepresented only in certain datasets (see Subsection 4.3.3). Given a pair of motifs of interest, all possible motif complexes are screened for enrichment in all datasets.

5.4 EXECUTION TIME AND OUTPUT

One of our priorities while developing TACO was to make the analyses computationally tractable. Comprehensive analyses using two sources of DNase-seq data, described in Subsection 3.3.4, where we took as input 964 vertebrate TF affinity motifs from TRANSFAC Professional (Wingender, 2008), requires the testing of 2.57 billion hypotheses. TACO completes this task in approximately 6 hours, using 16 cores of a 3.33 GHz machine and up to 11 GB of memory.

As output, TACO provides a multidimensional view of overrepresented cell-type-specific motif complexes. First, TACO clusters the enriched motif complexes as described in Subsection 2.2.7, and treats each cluster as a single predicted TF dimer. For each overrepresented motif complex within a cluster, the locations of all its genomic occurrences are reported. We also provide the position weight matrices inferred by counting nucleotide frequencies at each position within its genomic instances. Moreover, TACO also provides statistics that can be used to visualize the distribution of enrichment p -values using a Q-Q plot, and to generate spacing plots as in Figure 4.3.

The source code for TACO is freely available under the GNU GPL license, along with examples and documentation, at <http://bioputer.mimuw.edu.pl/taco/> and on GitHub at <https://github.com/ajank/taco>.

5.5 BENCHMARKING THE DIMER PREDICTION TOOLS

We compared TACO with the two other dimer prediction methods, SpaMo (13) and iTFs (12), by benchmarking the three algorithms on the set of 29 known TF dimers manually compiled from the existing biochemical literature (Table 2.1; Table 4.1).

Since the known dimers were used as a set of true positives, we tested 25 distinct motif pairs underlying the 29 known dimers. As a control, we added a set of 1000 random motif pairs, which were randomly chosen from all motif pairs which could be possibly formed using all 964 vertebrate motifs from TRANSFAC Professional 2011.2 (17). We also ensured that the set of 1000 random motif pairs does not overlap with the set of 25 positive motif pairs.

Each of the tools (TACO, SpaMo and iTFs) was applied to each of the 44 cell-type-specific DNase-seq datasets from University of Washington (UW) and each of the 26 cell-type-specific DNase-seq datasets from Duke University (Duke). In these datasets, we masked repetitive regions (as identified by RepeatMasker and Tandem Repeat Finder) and coding regions (extracted from Ensembl). Options specific to each of the tools are reported in the next subsections.

SpaMo and iTFs were evaluated both with and without trimming of uninformative positions at motif edges. Motif trimming was implemented externally and performed as in (13) and (12), by eliminating flanking columns with information content less or equal 0.25 bit from both sides of the individual motifs. Note that we did

not run TACO with trimmed motifs, since TACO is able to handle motif overlap.

Each of the tools was applied to each cell-type-specific dataset separately to calculate enrichment p -values for all motif complexes which could be formed from the abovementioned motif pairs. Note that iTFs uses binned spacing, so the p -values were provided for each spacing interval within each mutual motif orientation. Since we do not have complete a priori information on the cell-type-specificity of known dimers, we combined all the enrichment p -values across datasets by choosing the most significant p -value for a given motif complex. In the case of combined (UW+Duke) study, the enrichment p -values were combined across data sources as well.

Sensitivity was defined as the fraction of the 29 known dimers (i.e. known motif complexes) detected at any given p -value threshold. False-positive rate was defined as the fraction of the random motif dimers (i.e. all the other motif complexes) detected at the same threshold.

5.5.1 BENCHMARKING PARAMETERS FOR TACO

TACO was run with the default options (in particular, `MaxMotifSpacing = 50`). Individual motif matches were identified using the threshold criterion we recommend for TRANSFAC, i.e. `Sensitivity = 0.8`. In addition, to ensure that all p -values are reported, we specified `TargetInstancesThreshold = 0`, `FoldChangeThreshold = 0.0` and `PValueThreshold = Inf`.

We also minimized the motif complex clustering, which was irrelevant to the benchmarking, in order to adequately compare the execution time. Hence, we specified `ClusteringDistanceConstant = 0.0`, `ClusteringDistanceMultiplier = 0.0` and `ClusteringOverlapThreshold = Inf`.

5.5.2 BENCHMARKING PARAMETERS FOR SPAMO

SpaMo was run using `spamo -trim 0 -cutoff 1 -margin 50 -keepprimary -bgfile [background file] [dataset] [motif1].meme [motif2].meme`. The background file, containing the nucleotide frequencies, was generated by `fasta-get-markov` from the union set of all cell-type-specific DNase-seq datasets considered.

5.5.3 BENCHMARKING PARAMETERS FOR iTFS

iTFs was run with distance ranges as specified in [12], namely 0-10, 10-25, 25-50 and 50-100 bp. We noted that due to the spacing binning, three known hormone receptor homodimers (rows 20-22 in Table 2.1) cannot be easily distinguished. These motif complexes share the same motifs and orientation, and differ only by their spacing, which falls into the same distance range. To resolve this ambiguity, we referred to TACO and SpaMo predictions and found that these complexes were identified as enriched in different datasets.

In the case of UW data, TR-TR or RXR-TR (row 21) was most overrepresented in WERI-Rb-1 cell type (uncorrected TACO p -value = $1.53 \cdot 10^{-32}$), RAR-RAR (row 22) in SK-N-SH_RA ($p = 2.65 \cdot 10^{-4}$) and VD₃R-VD₃R (row 20) in WERI-Rb-1 ($p = 0.039$) and NB₄ ($p = 0.058$). Consequently, for iTFs analysis of the single motif pair yielding these complexes in UW data, we separated WERI-Rb-1, SK-N-SH_RA and NB₄ datasets, considering them as indicators of three different dimers, and excluded all the other datasets.

In the case of Duke data, we found no significant cell-type-specific overrepresentation of the motif complexes discussed above. However, to make the results comparable, we referred to the smallest (albeit insignificant) TACO p -values and found that TR-TR or RXR-TR was most overrepresented in HMEC cell type, RAR-RAR in LNCaP and VD₃R-VD₃R in 8988T. Consequently, for iTFs analysis of the single motif pair yielding these complexes in Duke data, we separated HMEC, LNCaP and 8988T datasets as above, and excluded all the other datasets.

In the case of combined (UW+Duke) study, for iTFs analysis of the single motif pair discussed above, we separated the three datasets as in the UW case, and excluded all the other datasets, including all Duke ones.

For all the other motif pairs, we combined all the datasets (cell types) as previously described.

5.6 COMPARISON OF DIMER PREDICTION TOOLS

We compared TACO with the two other dimer prediction methods, SpaMo (Whittington et al., 2011) and iTFs (Kazemian et al., 2013) using the 29 known dimers as a benchmark set of true positives (Table 2.1; Table 4.1). Henceforth, we tested 25 distinct motif pairs underlying the 29 known dimers, and as a control we included

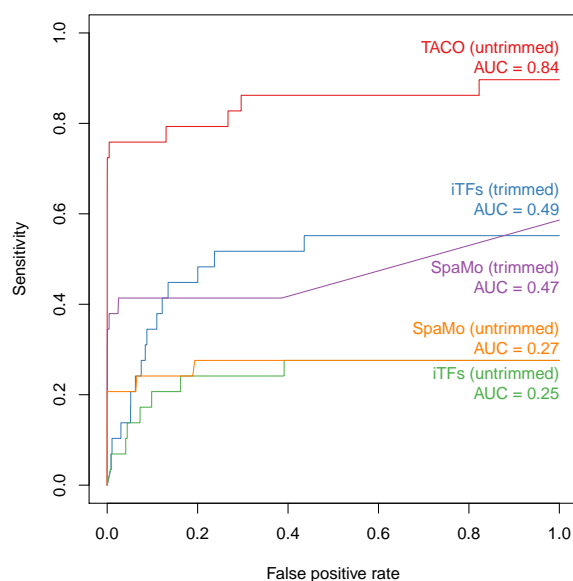


Figure 5.2: Comparison of dimer prediction algorithms, UW DNase-seq data. SpaMo and iTFs were evaluated both with and without motif trimming. Note that TACO does not require motif trimming. Sensitivity is shown as a function of false positive rate; Area Under Curve (AUC) is indicated.

a set of 1000 random motif pairs. All the tools were applied to each of the 44 cell-type-specific UW DNase-seq datasets. Sensitivity was defined as the fraction of the 29 known dimers detected at any given p -value threshold. False-positive rate was defined as the fraction of the random motif dimers detected at the same threshold (Figure 5.2).

SpaMo and iTFs were evaluated both with and without trimming of uninformative positions at motif edges. Motif trimming was performed as in [Whittington et al. \(2011\)](#) and [Kazemian et al. \(2013\)](#). As expected, both of these tools performed better with trimmed motifs. Notably, with motif trimming, iTFs performed marginally better than SpaMo (AUC = 0.49 vs. AUC = 0.47) despite the fact that it was not designed to predict rigidly spaced TF dimers ([Kazemian et al., 2013](#)). Ultimately, TACO (AUC = 0.84) clearly outperformed the other tools; note that we did not run TACO with trimmed motifs, since TACO is able to handle motif overlap. We also found that TACO is robust to the motif sensitivity threshold chosen (5.4). Notably, TACO and SpaMo completed the benchmarking analysis reasonably fast (2.7 and 6 hours on a single CPU machine, respectively; TACO may use multiple CPUs). However, iTFs could only complete the job in a feasible time when running on a cluster.

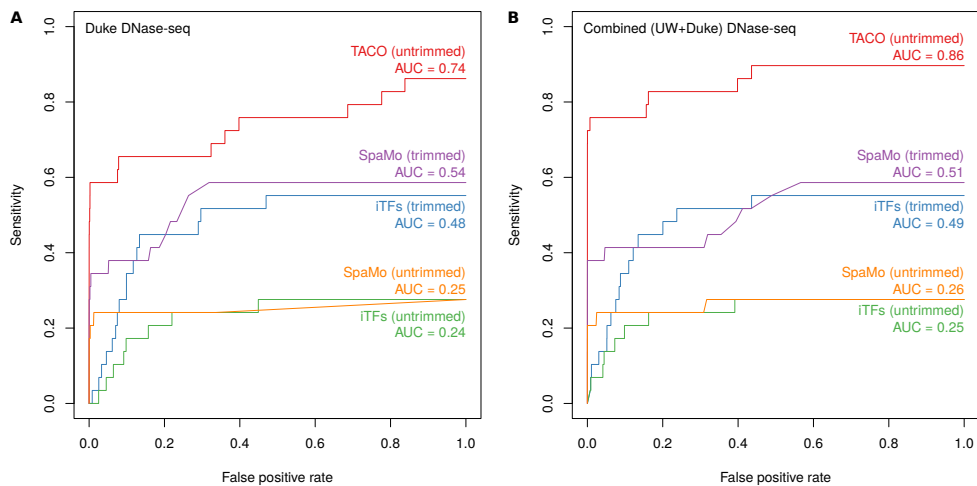


Figure 5.3: Comparison of dimer prediction algorithms, Duke and combined (UW+Duke) DNase-seq data. As in Figure 5.2, with algorithms evaluated using (a) Duke and (b) combined (UW+Duke) DNase-seq data.

Comparing the three tools by applying them to the 26 cell-type-specific Duke DNase-seq datasets yielded comparable results, with TACO (AUC = 0.74) again outperforming the two other tools (Figure 5.3a). Combining the predictions from both DNase-seq data sources gave even better performance (AUC = 0.86; Figure 5.3b).

5.7 SPECIFICATION FILE FORMAT

TACO, or Transcription factor Association from Complex Overrepresentation, is a program to predict overrepresented motif complexes in any genome-wide set of regulatory regions. TACO is a command line tool, and should be invoked with one argument: the name of the specification file to process.

5.7.1 GENERAL CONVENTIONS

A specification file, usually with `.spec` extension, has a HTML-like structure. A hash sign (`#`) begins a comment. Sequences of whitespace will collapse into a single whitespace. All the declarations have a form of `key=value` pairs. In case multiple values are allowed, they may come as multiple `key=value` pairs, as well as multiple values separated by whitespace. All filenames may contain wildcards. Each declaration has block scope, i.e. is applicable only to the declarations within `<section>...</section>`. All the applicable sections are described below.

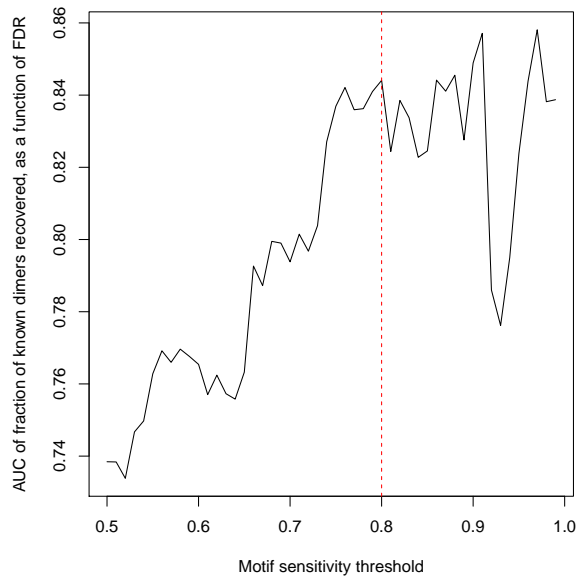


Figure 5.4: Robustness of TACO with respect to motif sensitivity threshold chosen. Area Under Curve (AUC) calculated as in Figure 5.2 in the main text. Red dotted line indicates the 0.8 sensitivity threshold used throughout this study.

5.7.2 REFERENCE GENOME

The <Genome> section specifies the reference genome sequence file(s), and possibly set(s) of genomic regions to be masked (excluded from the analysis). The following keys may be used:

FastaFile FASTA file(s) containing the genome sequences. Lowercase nucleotides (usually representing repetitive regions) are masked.

MaskedRegions BED file(s) listing regions to be masked, e.g. coding sequences.

5.7.3 INPUT DATASETS

Input regulatory region datasets of two kinds are handled: strongly cell-type-specific or weakly cell-type-specific. Each input dataset should be declared as strongly or weakly specific (these two kinds can be provided simultaneously). They are specified in the <StronglySpecificDatasets> section(s) and <WeaklySpecificDatasets> section(s), respectively. We recommend treating open chromatin datasets (such as DNase-seq datasets) as strongly specific, and ChIP-seq datasets as weakly specific. Each input dataset may consist of multiple replicates in narrowPeak or BED format.

Strongly and weakly cell-type-specific datasets are translated using different approaches into target datasets for TF dimer prediction. Regulatory regions of strongly specific datasets are intersected with each other, and then only the non-overlapping (unique) portions are retained as target regions.

In contrast, the weakly specific datasets are directly used as target datasets, without modification. The union of all replicates sharing a dataset identifier will be taken as a single input dataset. The union of all input datasets is used as a control dataset in order to build the null model of motif complex occurrence. The following keys may be used to specify input datasets:

Dataset Replicate filename, optionally preceded by the dataset identifier and whitespace. If no dataset identifier is provided, the replicate filename is taken.

DatasetList File(s) containing a list of datasets specified as above, in subsequent lines.

In addition, the following keys may be used to normalize the datasets (each replicate separately):

RegionSize Set each region size to a given value (in base pairs), centering at the peak. Default: 0 (leave it unchanged).

RegionMasking Specify how the regions overlapping masked fragments of the genome (see `MaskedRegions` in <Genome> section) should be treated. One of the following:

None no masking,

Peak exclude a region if its peak position (as specified in `narrowPeak` file) is masked in the genome

Majority exclude a region if most of the underlying genomic sequence is masked.

Default: None.

RegionCount Consider not more than the given number of regions with top `signalValue` (as specified in `narrowPeak` file). Default: 0 (consider all).

5.7.4 SEQUENCE MOTIFS

Motifs (position weight matrices) used in the analysis are specified in the `Motifs` section(s). Accepted formats include: TRANSFAC, JASPAR, SwissRegulon and MEME output (in PSPM format, including the header line starting with `letter-probability`). Although more than one motif database may be provided, it will be usually not necessary.

Motif identifiers are read along the position weight matrices. The uniqueness of the identifiers is enforced by suffixing them with underscore (`_`) and subsequent numbers if necessary. If no identifier is provided, motif filename is used instead. The following keys may be used:

Database Motif database file(s).

DatabaseSubset File(s) listing identifiers of motifs from the motif database(s) to be included in the analysis. If not provided, all the motifs from all motif databases are used.

Motif Motif filename, optionally preceded by the dataset identifier and whitespace.

Sensitivity Sensitivity value used for setting motif score threshold. Default: 0.9.

5.7.5 SCOPE OF THE ANALYSIS

The set of motif complexes or datasets considered in the analysis may be narrowed down in the `<Scope>` section(s). The following keys may be used:

Motif1 Motif identifiers for one of the motifs forming up the motif complex. Default: all motifs.

Motif2 Motif identifiers for the other motif forming up the motif complex. Default: all motifs.

Dataset Identifiers of the target datasets to consider. Default: all datasets.

5.7.6 VARIOUS OPTIONS

In the <Options> section, the following keys may be used:

NumberOfThreads Number of simultaneously running threads. Should be not more than the number of CPU cores of the machine. Default: 1.

MinMotifInformationContribution Minimal information contribution of each of the motifs forming up the motif complex. Default: 6.0.

MaxOverlappingInformationContent Maximal overlapping information content allowed in the motif complex. Default: 2.0.

MaxMotifSpacing Maximal spacing between the motifs forming up the motif complex. Motif spacing is defined as the number of intervening base pairs between the edges of the two contributing motifs; negative values indicate motif overlap. Default: 50.

ConsiderOrientationsSeparately Whether to calculate motif complex overrepresentation statistics separately for each of the two mutual motif orientations in a motif pair. One of: True, False. Default: True.

ConsiderMostSignificantComplexOnly Whether to consider only the most significant motif complex structure for a motif pair. One of: True, False. Default: False.

TargetInstancesThreshold Minimal number of instances in target dataset for an overrepresented motif complex not to be rejected. Default: 100.

FoldChangeThreshold Fold change threshold. Default: 1.

PValueThreshold *p*-value threshold, applied to *p*-values after Bonferroni correction. Default: 0.05.

DimerMotifFlanks Number of flanking basepairs for dimer motifs. Default: 5.

ClusteringAcrossDatasets Whether to allow prediction clusters to span across multiple datasets. One of: True, False. Default: True.

ClusteringDistanceConstant Constant in the affine function for joining motif complexes by dimer motif similarity. Default: 0.

ClusteringDistanceMultiplier Multiplier in the affine function for joining motif complexes by dimer motif similarity. Default: 0.15.

ClusteringOverlapThreshold Overlap threshold for joining motif complexes by overlap of genomic instances. Default: 0.2.

OutputPrefix Prefix for all output files. Default: name of the specification file, after truncating trailing .spec if possible.

OutputDetailedStats Whether to save detailed statistics for all motif complexes formed by pairs of motifs yielding an overrepresented motif complex. One of: None, Signature – only for signature motif complexes, All – for all overrepresented motif complexes. Default: All.

OutputDimerMotifs Whether to save dimer motifs. One of: None, Signature, All. Default: All.

OutputGenomicLocations Whether to save genomic locations of instances of overrepresented motif complexes. One of: None, Signature, All. Default: All.

GenomicLocationsMaxSpacingDeviation Number of incorrect motif spacings to be considered while saving genomic instances. If greater than 0, genomic instances of incorrectly spaced variants of overrepresented motif complexes will be saved, with spacing deviation ranging between 1 and the given value. Default: 0.

OutputPValueDistribution Whether to save p -value distribution, suitable for Q-Q plot. One of: True, False. Default: True.

5.7.7 OUTPUT FILES

The following files are created as output, subject to the options discussed above:

<OutputPrefix>.tab all overrepresented motif complexes, clustered into predictions

<OutputPrefix>.stats detailed statistics for all motif pairs yielding an overrepresented motif complex

<OutputPrefix>.pwms dimer motifs in TRANSFAC format

<OutputPrefix>.hits genomic locations of instances of overrepresented motif complexes, and possibly of their incorrectly spaced variants

<OutputPrefix>.pval p -value distribution.

5.8 DISCUSSION

We have demonstrated the generality and consistency of TF dimer predictions made by TACO by applying the algorithm to 152 DNase-seq datasets and 94 ChIP-seq datasets from the ENCODE Project. Moreover, we showed that TACO clearly outperforms existing dimer prediction tools when benchmarked on the set of 29 known dimers. Based on all TACO predictions, we found that TF dimers that bind widely spaced motif pairs are significantly more flexible in their spacing. Overall, we expect TACO to be widely applicable, since thousands of regulatory element datasets will be available in the near future. We also anticipate its application to regulatory annotations from assay types other than those discussed here, since the algorithm allows a great deal of flexibility in data type and mode of analysis.

6

Building on transcription factor footprints to predict individual binding sites

6.1 INTRODUCTION

In the previous chapters, we have focused on the identification of the putative structures of transcription factor complexes that bind cooperatively to DNA. The comprehensive knowledge of such functional structures is essential to fully understand the mechanisms of transcriptional regulation. In practice, while focusing on a certain mechanism, the dual problem is often faced: having focused on a particular set of TFs, one wish to identify the binding sites of these factors to the genome, in different cell types and conditions. Hence, the methods aimed at identification of TF complexes must be complemented with accurate tools to incorporate the knowledge about TF complexes to predict individual TF binding sites.

The traditional method of analyzing individual active regulatory elements in the genome involves the digestion by DNase I and subsequent identification of regions

where TFs are bound to the DNA fragment and protect the DNA from degradation by the enzyme. These protected sites, or TF footprints, can be identified on a large scale by a more recent protocol, DNase I digestion followed by high-throughput sequencing (DNase-seq).

CENTIPEDE ([Pique-Regi et al., 2011](#)) was the first algorithm aimed at combining sequence information with experimental data to identify the sites where a particular TF is bound in the genome. This method relied on the presence of a DNA sequence motif at the candidate binding sites considered. A logistic regression model allowed for multiple types of prior information to be incorporated, e.g. Position Weight Matrix (PWM) score, distance to the nearest Transcription Start Site and sequence evolutionary conservation. The posterior component consisted of a combination of negative binomial and multinomial positional models for each type of experimental data, such as DNase-seq or histone modification ChIP-seq. Overall, the main strength of CENTIPEDE is the ability to identify binding sites for multiple TFs from a single DNase-seq experiment.

MILLIPEDE ([Luo & Hartemink, 2013](#)), a method inspired by CENTIPEDE, also aims at identifying TF binding sites, and also combines DNase digestion data with TF binding specificity information. The method of MILLIPEDE is a dramatic simplification of the CENTIPEDE approach. Instead of a comprehensive combination of negative binomial and multinomial models to represent the positional distribution of DNase I cuts, these cuts were grouped into several (e.g. 5 or 12) bins. The log-transformed DNase I cut counts within these bins were incorporated in the logistic regression model, together with all the prior information. Overall, the number of parameters in MILLIPEDE is at least an order of magnitude smaller than in CENTIPEDE; hence the name MILLIPEDE. The authors showed that MILLIPEDE outperforms CENTIPEDE marginally in human but dramatically in yeast. This was attributed mostly to avoiding over-fitting of parameters by focusing on a more coarse-grained model, focused on the large-scale differences between the bound and unbound states.

Wellington ([Piper et al., 2013](#)) is another recent algorithm to predict occupied TF binding sites from DNase-seq data. This algorithm is based on a completely different approach, and does not require a DNA sequence motif to identify the prior set of candidate binding sites. Instead, Wellington quantifies an imbalance in the DNA strand-specific alignment information of DNase-seq data around virtually every location in the genome. The authors argue that in a DNase-seq experiment, most

of the DNA fragments captured for sequencing are in the order of 50 to 150 bp in length, and they are expected to originate from within the DNase I hypersensitive sites, as opposed to nucleosomal DNA. Since the length of DNase I hypersensitive sites is usually 200-250 bp, the captured fragments are likely to span the regions of DNA protected by bound TFs. These captured fragments manifest themselves after sequencing as 5' sequence tags, representing just one end of these fragments. Hence, a typical DNase I hypersensitive site should be enriched in forward strand tags upstream and in reverse strand tags downstream of bound TFs.

Wellington takes advantage of this strand imbalance criterion to greatly increase the specificity by reducing the number of false positives. The authors show that their method requires much fewer predictions than several previous approaches to recapitulate an equal amount of ChIP-seq data. For each base pair, Wellington tests the hypothesis that there are significantly more reads aligning to the forward strand in the upstream shoulder region with respect to the reads aligning to the forward strand in the footprint region. Moreover, a reverse complement hypothesis is tested, i.e. that there are significantly more reads aligning to the reverse strand in the downstream shoulder region with respect to the reads aligning to the reverse strand in the footprint region. The final Wellington p -value for a given genomic location is a product of the two for the aforementioned hypotheses.

Here, we propose MOCCA, a novel computational method to accurately identify TF footprints from genome sequence information and cell-type-specific experimental data, such as DNase-seq data. Our approach combines the strengths of CENTIPEDE and Wellington, while keeping the number of free parameters in the model reasonably low. For a given TF, we first identify candidate binding sites that have reasonable sequence affinity, using a position weight matrix. Then, like CENTIPEDE, we employ an Expectation-Maximization-based approach to simultaneously learn the DNase I cut profiles and classify the binding sites as bound or unbound.

Our method is unique in allowing for multiple bound states for a single TF, differing in their cut profile and overall number of DNase I cuts. To make the model robust, we employ a systematic approach to group the DNase I cuts, according to their location and strand. Inspired by Wellington, we take the forward strand DNase I cuts only upstream and within the cut site, while the reverse strand DNase I cuts – within the cut site and downstream. We model the total number of cuts as a negative binomial component, while the cut distribution (regularized by binning out-

Dataset		Number of reads		
Name	Genome Browser track	A549	HepG2	K562
Duke DNase	OpenChromDnase	51.6 M	13.6 M	80.8 M
UW DNase	UwDnase	33.3 M	22.1 M	35.8 M
UW DGF	UwDgf	350.6 M	168.9 M	180.0 M

Table 6.1: Numbers of reads in DNase-seq datasets used. Three ENCODE cell lines were considered: A549, HepG2 and K562. UW, University of Washington; DGF, Digital Genomic Footprinting.

side the cut site) is modeled as a multinomial component. Overall, MOCCA predictions agree well with experimental ChIP-seq measurements of TF binding at candidate motif sites. We also comprehensively compared the predictive performance of MOCCA, CENTIPEDE and Wellington, and show that MOCCA consistently outperformed CENTIPEDE and Wellington, especially when applied to DNase-seq datasets with lower sequencing depth.

6.2 METHODS

6.2.1 DNASE-SEQ DATA FROM MULTIPLE SOURCES

The ENCODE Project ([ENCODE Project Consortium et al., 2012](#)) provides three different tracks with DNase-seq data. Two of them follow the standard ENCODE DNase-seq protocol: wgEncodeOpenChromDnase from Duke University (Duke) and wgEncodeUwDnase from University of Washington (UW). The third track, wgEncodeUwDgf, follows the Digital Genomic Footprinting protocol, which yields much higher number of sequencing reads (Table 6.1).

6.2.2 CHIP-SEQ DATA AS A GOLDEN STANDARD OF TF BINDING

We have downloaded a collection of ChIP-seq datasets from ENCODE and used them as a golden standard for TF binding (Table 6.2). The same collection of datasets was used to assess the performance of Wellington in ([Piper et al., 2013](#)). The corresponding DNA sequence motifs for these TFs were taken from the HOMER (Hypergeometric Optimization of Motif EnRichment) suite ([Heinz et al., 2010](#)). The motif instances in the human genome of these motifs were downloaded from <http://homer.salk.edu/homer/> (HOMER Known Motifs track).

All the genomic motif instances were classified as either bound or unbound. The instances overlapping any ChIP-seq peak were classified as bound, and all the re-

Cell type	Transcription factor	ChIP-seq peaks			Motif instances		ENCODE narrowPeak filename
		total	with motif	without motif	inside ChIP-seq peaks	outside ChIP-seq peaks	
K562	ATF3	16 011	2 162	13 849	4 298	160 472	HaibK562Atf3V0416101
K562	c-Myc	5 023	2 098	2 925	4 331	509 454	SydhK562Cmyc
K562	CTCF	56 058	25 788	30 270	26 432	41 170	UtaK562Ctcf
K562	JunD	26 674	2 600	24 074	5 070	112 079	UchicagoK562Ejund
K562	Max	46 171	16 419	29 752	34 226	1 131 646	HaibK562MaxV0416102
K562	NFE2	2 637	1 619	1 018	1 750	50 360	SydhK562Nfe2
K562	NRF1	4 211	2 609	1 602	5 960	20 440	SydhK562Nrf1Iggrab
K562	NRSF	15 849	2 055	13 794	2 112	2 750	HaibK562NrsfV0416102
K562	PU.1	28 677	18 514	10 163	20 262	549 324	HaibK562Pu1Pcr1x
K562	Sp1	7 206	2 830	4 376	4 861	137 043	HaibK562Sp1Pcr1x
K562	USF1	18 521	12 431	6 090	23 808	524 887	HaibK562Usf1V0416101
A549	ATF3	6 580	308	6 272	636	164 134	HaibA549Atf3V0422111Etoh02
A549	bHLHE40	3 123	1 225	1 898	2 667	254 098	SydhA549Bhlhe40Iggrab
A549	CEBP	38 845	25 305	13 540	46 517	1 722 846	SydhA549CebpIggrab
A549	CTCF	45 732	23 536	22 196	24 289	43 313	UwA549Ctcf
A549	ELF1	8 611	5 075	3 536	6 937	348 641	HaibA549Elf1V0422111Etoh02
A549	ETS1	5 525	2 564	2 961	3 466	1 145 420	HaibA549Ets1V0422111Etoh02
A549	GABP	12 348	7 196	5 152	9 396	871 718	HaibA549GabpV0422111Etoh02
A549	Max	9 881	3 982	5 899	8 965	1 156 907	SydhA549MaxIggrab
A549	NRSF	11 970	1 938	10 032	1 861	3 001	HaibA549NrsfV0422111Etoh02
A549	USF1	8 004	4 710	3 294	9 452	539 243	HaibA549Usf1V0422111Etoh02
A549	YY1	10 259	2 148	8 111	2 079	52 873	HaibA549Yy1cV0422111Etoh02
A549	ZBTB33	7 152	626	6 526	1 052	14 443	HaibA549Zbtb33V0422111Etoh02
HepG2	ATF3	3 291	1 132	2 159	2 392	162 378	HaibHepg2Atf3V0416101
HepG2	c-Myc	4 413	1 762	2 651	3 558	510 227	UtaHepg2Cmyc
HepG2	CTCF	55 778	26 856	28 922	27 655	39 947	HaibHepg2Ctcfsc5916V0416101
HepG2	FOXA1	40 989	29 356	11 633	76 105	6 363 288	HaibHepg2Foxa2sc6554V0416101
HepG2	HNF4a	20 805	10 913	9 892	12 889	519 223	HaibHepg2Hnf4asc8987V0416101
HepG2	JunD	21 614	866	20 748	1 632	115 517	HaibHepg2JundPcr1x
HepG2	Max	11 854	4 707	7 147	10 726	1 155 146	SydhHepg2MaxIggrab
HepG2	MYB	17 898	8 016	9 882	10 306	2 389 507	HaibHepg2Mybl2sc81192V0422111
HepG2	NRF1	1 902	1 635	267	4 132	22 268	SydhHepg2Nrf1Iggrab
HepG2	NRSF	12 828	1 686	11 142	1 743	3 119	HaibHepg2NrsfV0416101
HepG2	RXR	17 063	6 976	10 087	9 044	1 265 842	HaibHepg2RrxraPcr1x
HepG2	Sp1	25 477	3 599	21 878	6 087	135 817	HaibHepg2Sp1Pcr1x
HepG2	Srebp1a	2 585	293	2 292	307	327 401	SydhHepg2Srebp1Insln
HepG2	TBP	13 806	2 490	11 316	3 798	3 136 778	SydhHepg2Tbplggrab
HepG2	TR4	2 953	660	2 293	836	88 251	SydhHepg2Tr4Ucd
HepG2	USF1	21 890	14 809	7 081	27 503	521 192	HaibHepg2Usf1Pcr1x
Total		670 214	283 494	386 720	449 140	26 312 163	
Percentage			42.3%	57.7%	1.7%	98.3%	

Table 6.2: ChIP-seq datasets used as a golden standard of TF binding. These datasets were generated by the ENCODE Analysis Working Group (AWG) using a uniform processing pipeline. The narrowPeak filenames follow the pattern "wgEncodeAwgTfbs...UniPk.narrowPeak.gz", where only the changing "... part is given above.

remaining ones were classified as unbound. We considered all the genomic motif instances as candidate binding sites in the analysis.

6.2.3 PRIOR PROBABILITIES OF TF BINDING

The prior component of the model captures the genomic sequence and other static (i.e. independent of cell type or conditions) characteristics of the candidate binding site for a TF of interest. Let us denote by i a particular genomic instance (motif match) of a motif of interest. Typically, the static characteristics assigned to motif instances are: the respective PWM score, average evolutionary conservation and so on.

To formalize the model, let us denote the value of the j -th static characteristics for genomic instance i by $x_i^{(j)}$, where $1 \leq j \leq J$. In the simplest case, where each motif instance can be either “bound” or “unbound”, we apply a logistic approach to model the ratio of the prior probabilities:

$$\frac{P(Z_i = 1)}{P(Z_i = 0)} = \exp \left(\beta_0 + \sum_j \beta_j x_i^{(j)} \right). \quad (6.1)$$

Here, $Z_i = 1$ indicates that the i -th motif instance is bound, whereas $Z_i = 0$ indicates that it remains unbound. Such a model has been used in CENTIPEDE (Pique-Regi et al., 2011).

Now we will generalize the above-described model. Let us consider a TF that manifests one or more cooperative binding modes, with well-defined structures of the underlying motif complexes. The cooperative binding modes, and the corresponding motif complexes, will be both denoted by $k = 2, \dots, K+1$. Each of these complexes imposes certain offset and orientation of the partner motif with respect to the primary motif.

Again, let us denote by i a particular genomic instance of a motif of interest. This genomic instance implies the corresponding locations for all partner motifs within all defined motif complexes. We will now include in the model the static characteristics for these partner motif instances. These characteristics are calculated no matter how unfavourable they are, and included in the sequence $x_i^{(j)}$, where $1 \leq j \leq J$. Note that in the homodimer case, some of these characteristics may be derived from the same PWM, however scored at a different genomic location.

Now let us focus on a particular cooperative binding mode k , where $2 \leq k \leq$

$K + 1$. We introduce the indicators $\gamma_j^{(k)} \in \{0, 1\}$, specifying whether the static characteristics $x_i^{(j)}$ should be taken into account in this cooperative binding mode. The values of these indicators ensure that only the characteristics specific to the primary motif instance and to the partner motif instances within k -th motif complex will be taken into account. Moreover, the monomer binding mode, denoted by $k = 1$, should be characterized only by the characteristics referring to the primary motif instance. Hence, $\gamma_j^{(1)} = 0$ for all the characteristics j referring to any of the partner motifs.

To model the prior probabilities, we now apply a logistic model against the unbound ‘‘pivot’’ case of $Z_i = 0$:

$$\frac{P(Z_i = k)}{P(Z_i = 0)} = \exp\left(\beta_0^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)}\right), \quad (6.2)$$

where $k = 0$ indicates no binding, $k = 1$ refers to binding as monomer, and $k = 2, \dots, K + 1$ refer to the respective cooperative binding modes. This way, we have $K + 1$ outcomes separately regressed against the pivot outcome $Z_i = 0$.

For clarity, we impose an additional constraint such that $\gamma_j^{(k)} = 0$ implies $\beta_j^{(k)} = 0$. In other words, $\beta_j^{(k)} = 0$ for the partner motifs not involved in k -th binding mode. We can now explicitly formulate $P(Z_i = 0)$ by summing up Equation 6.2 for $k = 1, \dots, K + 1$:

$$\frac{\sum_{k=1}^{K+1} P(Z_i = k)}{P(Z_i = 0)} = \sum_{k=1}^{K+1} \exp\left(\beta_0^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)}\right) \quad (6.3)$$

$$\frac{1 - P(Z_i = 0)}{P(Z_i = 0)} = \sum_{k=1}^{K+1} \exp\left(\beta_0^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)}\right) \quad (6.4)$$

$$P(Z_i = 0) = \frac{1}{1 + \sum_{k=1}^{K+1} \exp\left(\beta_0^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)}\right)}. \quad (6.5)$$

Applying the above to Equation 6.2, we obtain an explicit formulation for all the probabilities $P(Z_i = k)$ where $k > 0$:

$$P(Z_i = k) = \frac{\exp\left(\beta_0^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)}\right)}{1 + \sum_{l=1}^{K+1} \exp\left(\beta_0^{(l)} + \sum_j \beta_j^{(l)} \gamma_j^{(l)} x_i^{(j)}\right)}. \quad (6.6)$$

6.2.4 MODELING THE NUMBER OF DNASE I CUTS

Apart from the static characteristics of the individual motif instances, our model incorporates positional data about the chromatin state, e.g. as represented by chromatin openness or histone modification patterns. This cell-type and condition-specific information is used to derive the likelihood of being in a particular bound state for each of the candidate binding sites.

In this study, we use DNase-seq data as a measure of chromatin openness. We include the numbers of DNase I cuts at individual base pairs, $(\text{DNase}_{i,j})_j$, counted in the vicinity of the motif instance i in a strand-specific manner. The forward strand cuts are taken only upstream and within the cut site, while the reverse strand cuts – within the cut site and downstream. For example, we may consider primary motif of length L and 200 bp margin; in such a case, $\text{DNase}_{i,j}$ contains the numbers of forward strand DNase I cuts ($j = 1, \dots, 200 + L$, starting 200 bp upstream) and reverse strand ones ($j = 201 + L, \dots, 400 + 2L$, starting at the cut site).

We observed that all kinds of positional data based on short sequence reads, in particular all kinds of DNase-seq data, are prone to artifactual spikes of reads (above 100 reads) mapped to a single location and strand in the genome. These spikes may arise from the sequence fragments originating at repetitive regions with incomplete representation in the reference genome. Hence, we applied clipping to the number of reads mapped to a single location and strand, choosing the threshold as the value of 99.9% quantile of all $(\text{DNase}_{i,j})_{i,j}$. The values of $\text{DNase}_{i,j}$ above the threshold were set to be equal to the threshold itself. We have tried to use other quantiles apart from the 99.9% quantile, namely 99% and 99.99%; they all gave similar results (data not shown).

Let $X_i = ((\text{DNase}_{i,j})_j, \dots)$ denote all the positional data available to the model. As stated in the previous subsection, we introduce the latent variables Z_i such that $P(Z_i = 0 \mid X_i)$ is the probability of motif instance i to be unbound, $P(Z_i = 1 \mid X_i)$ is the probability of it being bound by monomer, $P(Z_i = 2 \mid X_i)$ is the probability of it being bound in the first cooperative binding mode, and so on. Our primary interest is

$$p_i = \sum_{k=1}^{K+1} P(Z_i = k \mid X_i) = 1 - P(Z_i = 0 \mid X_i), \quad (6.7)$$

i.e. the probability of the motif instance i to be bound in any binding mode.

Taking the complement and following the Bayes theorem, we get

$$1 - p_i = P(Z_i = 0 \mid X_i) = \frac{P(X_i \mid Z_i = 0)P(Z_i = 0)}{\sum_{k=0}^{K+1} P(X_i \mid Z_i = k)P(Z_i = k)} \quad (6.8)$$

$$\frac{1}{1 - p_i} = \sum_{k=0}^{K+1} \frac{P(X_i \mid Z_i = k)P(Z_i = k)}{P(X_i \mid Z_i = 0)P(Z_i = 0)} = 1 + \sum_{k=1}^{K+1} \frac{P(X_i \mid Z_i = k)P(Z_i = k)}{P(X_i \mid Z_i = 0)P(Z_i = 0)} \quad (6.9)$$

$$\frac{p_i}{1 - p_i} = \sum_{k=1}^{K+1} \frac{P(X_i \mid Z_i = k)P(Z_i = k)}{P(X_i \mid Z_i = 0)P(Z_i = 0)}. \quad (6.10)$$

We will now explain how the conditional probabilities $P(X_i \mid Z_i = k)$ are modeled. We make a simplifying assumption that the different positional data types (such as DNase I cuts or histone modifications) used in the model are independent, given its binding state (Z_i). Hence, the conditional probability $P(X_i \mid Z_i = k)$ is a product of the corresponding conditional probabilities for all the types of positional data included in the model:

$$P(X_i \mid Z_i = k) = P((\text{DNase}_{i,j})_j \mid Z_i = k) \cdot \dots \quad (6.11)$$

Each type of positional data is modeled separately, using a mixture model. The first component captures the total number of reads mapped in the vicinity of the motif instance i , using the negative binomial distribution. This way, the model is robust with respect to the dataset coverage. The second component captures the spatial distribution of the given number of reads, using the multinomial distribution.

In case of DNase-seq data, we actually consider the DNase I cuts as two separate types of positional data, according to their strandness, i.e. considering the forward and reverse strand cuts separately:

$$P((\text{DNase}_{i,j})_j \mid Z_i = k) = P((\text{DNase}_{i,j}^+) \mid Z_i = k) \cdot P((\text{DNase}_{i,j}^-) \mid Z_i = k). \quad (6.12)$$

Furthermore, both the negative binomial and multinomial components are calculated for each strand separately. For brevity, we discuss the formulas for the forward strand DNase I component only; they are analogous for the reverse strand. The negative binomial component in binding mode k quantifies the total number

of DNase I cuts on the forward strand

$$\text{DNaseSum}_i^+ = \sum_j \text{DNase}_{i,j}^+ \quad (6.13)$$

and is naturally parametrized by the success probability $p^{+(k)} \in (0, 1)$ and the real-valued number of failures $r^{+(k)} > 0$.

The multinomial component quantifies the probability of a particular spatial distribution of the total number of DNase I cuts on a given strand. As opposed to CENTIPEDE (Pique-Regi et al., 2011), we do not keep a separate free parameter for each position (relative to the motif location) and strand, but apply a more flexible approach. For each binding mode k and positional data type (e.g. DNase I cuts on forward strand), we divide the positions j into one or more bins. Note that the bins are considered separately for each binding mode.

Let us denote by $\text{DNaseBin}_j^{+(k)}$ the bin number for position j in binding mode k . For clarity, let us assume that the bins are numbered by positive integers. In this study, we take 20 bp long bins outside the motif binding site, and single-base-pair bins within motif binding site. Moreover, for the unbound mode ($k = 0$) we put all the positions in a single bin:

$$\text{DNaseBin}_j^{+(k)} = \begin{cases} 1 & \text{for } k = 0 \text{ and any } j \\ \lfloor j/20 \rfloor & \text{for } k > 0 \text{ and } j = 1, \dots, 200 \\ 190 - j & \text{for } k > 0 \text{ and } j = 201, \dots, 200 + L. \end{cases} \quad (6.14)$$

Note that binding modes may differ in the way the positions are split into bins.

For a given binding mode k , we associate a free parameter $\lambda_b^{+(k)}$ with each bin $b = 1, \dots, B^{+(k)}$. However, for the multinomial distribution we must provide a vector of probabilities covering every single position in the vicinity of the motif instance. Hence, we calculate the actual multinomial coefficients $\tilde{\lambda}_j^{+(k)}$ by taking $\lambda_b^{+(k)}$ for $b = \text{DNaseBin}_j^{+(k)}$ and normalizing $\lambda_b^{+(k)}$ so that $\sum_j \tilde{\lambda}_j^{+(k)} = 1$. By definition, the multinomial coefficients $\tilde{\lambda}_j^{+(0)}$ for the unbound state are equal, i.e. there is no positional preference for DNase I cuts in the null model.

The joint probability of the DNase I positional data is obtained by the superposition of the negative binomial and multinomial components:

$$\begin{aligned}
P((\text{DNase}_{i,j}^+)_j \mid Z_i = k) \\
&= \text{NegativeBinomial}(\text{DNaseSum}_i^+ \mid p^{+(k)}, r^{+(k)}) \\
&\quad \cdot \text{Multinomial}\left((\text{DNase}_{i,j}^+)_j \mid \text{DNaseSum}_i^+, (\lambda_b^{+(k)})_b\right). \quad (6.15)
\end{aligned}$$

Now we can explicitly formulate the probabilities:

$$\begin{aligned}
&\text{NegativeBinomial}(\text{DNaseSum}_i^+ \mid p^{+(k)}, r^{+(k)}) \\
&= \frac{\Gamma(r^{+(k)} + \text{DNaseSum}_i^+)}{\Gamma(\text{DNaseSum}_i^+ + 1) \Gamma(r^{+(k)})} (p^{+(k)})^{r^{+(k)}} (1 - p^{+(k)})^{\text{DNaseSum}_i^+} \quad (6.16)
\end{aligned}$$

$$\begin{aligned}
&\text{Multinomial}\left((\text{DNase}_{i,j}^+)_j \mid \text{DNaseSum}_i^+, (\lambda_b^{+(k)})_b\right) \\
&= \text{DNaseSum}_i^+! \prod_j \frac{\left(\tilde{\lambda}_j^{+(k)}\right)^{\text{DNase}_{i,j}^+}}{\text{DNase}_{i,j}^+!} \\
&= \Gamma(\text{DNaseSum}_i^+ + 1) \prod_j \frac{\left(\tilde{\lambda}_j^{+(k)}\right)^{\text{DNase}_{i,j}^+}}{\Gamma(\text{DNase}_{i,j}^+ + 1)}, \quad (6.17)
\end{aligned}$$

where Γ is the standard gamma function, i.e. a continuous extension of the factorial function.

6.2.5 EXPECTATION-MAXIMIZATION APPROACH

To estimate the model parameters

$$\Theta = \left((\beta_j^{(k)})_{j,k}, (p^{+(k)})_k, (p^{-(k)})_k, (r^{+(k)})_k, (r^{-(k)})_k, (\lambda_b^{+(k)})_{b,k}, (\lambda_b^{-(k)})_{b,k} \right), \quad (6.18)$$

we apply the Expectation-Maximization approach. We use a common technique: instead of maximizing the likelihood function

$$L(\Theta) = \prod_i P(X_i \mid \Theta) \quad (6.19)$$

with unknown latent state, we maximize the complete likelihood function

$$L_C(\Theta) = \prod_i P(X_i, Z_i | \Theta) = \prod_i P(X_i | Z_i, \Theta)P(Z_i | \Theta), \quad (6.20)$$

which is more tractable.

The complete likelihood function, as stated above, is defined only for $Z_i = 0, \dots, K + 1$. However, we may rewrite it using indicator functions $Z_i^{(k)}$ such that $Z_i^{(k)} = 1$ if $Z_i = k$ and $Z_i^{(k)} = 0$ otherwise:

$$L_C(\Theta) = \prod_i \prod_{k=0}^{K+1} P(X_i | Z_i = k, \Theta)^{Z_i^{(k)}} P(Z_i = k | \Theta)^{Z_i^{(k)}}. \quad (6.21)$$

Let us denote by $\langle Z_i^{(k)} \rangle$ the expected value of $Z_i^{(k)}$. It holds that $\langle Z_i^{(k)} \rangle = P(Z_i = k)$. Taking the expected value of $L_C(\Theta)$ with respect to all $Z_i^{(k)}$, we obtain a real-domain function of Θ :

$$\langle L_C(\Theta) \rangle = \prod_i \prod_{k=0}^{K+1} P(X_i | Z_i = k, \Theta)^{\langle Z_i^{(k)} \rangle} P(Z_i = k | \Theta)^{\langle Z_i^{(k)} \rangle}. \quad (6.22)$$

The formulas will be easier to manipulate after taking the logarithm:

$$\begin{aligned} \log \langle L_C(\Theta) \rangle &= \overbrace{\sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log P(X_i | Z_i = k, \Theta)}^{L_A(\Theta)} \\ &\quad + \underbrace{\sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log P(Z_i = k | \Theta)}_{L_B(\Theta)}. \end{aligned} \quad (6.23)$$

Our goal is to maximize the (log-transformed) expected value of the complete likelihood function L_C . Note that the value of the first component, L_A , depends on $(p^{+(k)})_k, (p^{-(k)})_k, (r^{+(k)})_k, (r^{-(k)})_k, (\lambda_b^{+(k)})_{b,k}$ and $(\lambda_b^{-(k)})_{b,k}$, while the value of the second component, L_B , depends only on the parameters in Θ not listed previously, namely on $(\beta_j^{(k)})_{j,k}$. Therefore, we can maximize L_A and L_B separately.

We found no closed-form solution for $\beta_j^{(k)}$ that maximizes $L_B(\Theta)$, hence we apply the Broyden-Fletcher-Goldfarb-Shanno (BFGS) numerical optimization procedure here (Broyden, 1969; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). This method

uses the function values and gradients to build up a representation of the surface to be maximized. Substituting Equation 6.6 to the definition of L_B , we get:

$$L_B(\Theta) = \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \left(\beta_o^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)} \right) - \sum_i \log \left(1 + \sum_{l=1}^{K+1} \exp \left(\beta_o^{(l)} + \sum_j \beta_j^{(l)} \gamma_j^{(l)} x_i^{(j)} \right) \right) \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle. \quad (6.24)$$

Note that the last factor, $\sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle$, is equal to 1 and may thus be omitted. Differentiating L_B with respect to $\beta_j^{(k)}$, we get:

$$\frac{\partial L_B}{\partial \beta_j^{(k)}} = \sum_i \langle Z_i^{(k)} \rangle \gamma_j^{(k)} x_i^{(j)} - \sum_i \frac{\exp \left(\beta_o^{(k)} + \sum_j \beta_j^{(k)} \gamma_j^{(k)} x_i^{(j)} \right) \gamma_j^{(k)} x_i^{(j)}}{1 + \sum_{l=1}^{K+1} \exp \left(\beta_o^{(l)} + \sum_j \beta_j^{(l)} \gamma_j^{(l)} x_i^{(j)} \right)}. \quad (6.25)$$

Now let us focus on the other component of $\log \langle L_C(\Theta) \rangle$, i.e. L_A . For clarity, let us assume that DNase-seq data is the only kind of positional data provided. The derivations follow analogously for any other independent positional datasets. Substituting Equations 6.11 and 6.12 to the definition of L_B in Equation 6.23, we get:

$$L_A(\Theta) = \underbrace{\sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log P((\text{DNase}_{i,j}^+) | Z_i = k)}_{L_A^+(\Theta)} + \underbrace{\sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log P((\text{DNase}_{i,j}^-) | Z_i = k)}_{L_A^-(\Theta)}. \quad (6.26)$$

The two components, L_A^+ and L_A^- , depend on distinct sets of parameters in the same manner. Hence, we can maximize them separately. Without loss of generality, we will discuss the optimization procedure for L_A^+ . From Equations 6.15 to 6.17, we have:

$$L_A^+(\Theta) = \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \left(\frac{\Gamma(r^{+(k)} + \text{DNaseSum}_i^+)}{\Gamma(\text{DNaseSum}_i^+ + 1) \Gamma(r^{+(k)})} \right)$$

$$\begin{aligned}
& \cdot (p^{+(k)})^{r^{+(k)}} (1 - p^{+(k)})^{\text{DNaseSum}_i^+} \Big) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \left(\Gamma(\text{DNaseSum}_i^+ + 1) \prod_j \frac{(\tilde{\lambda}_j^{+(k)})^{\text{DNase}_{i,j}^+}}{\Gamma(\text{DNase}_{i,j}^+ + 1)} \right) \\
& = \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(r^{+(k)} + \text{DNaseSum}_i^+) \\
& - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(\text{DNaseSum}_i^+ + 1) - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(r^{+(k)}) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle r^{+(k)} \log(p^{+(k)}) + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ \log(1 - p^{+(k)}) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(\text{DNaseSum}_i^+ + 1) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \sum_j \text{DNase}_{i,j}^+ \log(\tilde{\lambda}_j^{+(k)}) - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \sum_j \log \Gamma(\text{DNase}_{i,j}^+ + 1).
\end{aligned} \tag{6.27}$$

Eliminating the additive inverse terms and noting that $\sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle = 1$, we get:

$$\begin{aligned}
L_{\mathcal{A}}^+(\Theta) & = \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(r^{+(k)} + \text{DNaseSum}_i^+) - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(r^{+(k)}) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle r^{+(k)} \log(p^{+(k)}) + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ \log(1 - p^{+(k)}) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \sum_j \text{DNase}_{i,j}^+ \log(\tilde{\lambda}_j^{+(k)}) - \sum_i \sum_j \log \Gamma(\text{DNase}_{i,j}^+ + 1).
\end{aligned} \tag{6.28}$$

Note that only the first three summands depend on $(r^{+(k)})_k$, only the third and fourth depends on $(p^{+(k)})_k$, and only the fifth depends on the parameters $(\lambda_b^{+(k)})_{b,k}$, which give rise to $(\tilde{\lambda}_j^{+(k)})_{j,k}$. Hence, we may find the values of $(\lambda_b^{+(k)})_{b,k}$ that max-

imize $L_{\mathcal{A}}$ independently of the other parameters. We need to maximize

$$\sum_{k=0}^{K+1} \sum_j \log \left(\tilde{\lambda}_j^{+(k)} \right) \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+ \quad (6.29)$$

subject to the constraint $\sum_j \tilde{\lambda}_j^{+(k)} = 1$ for each k . Since k -th element in the sum above depends only on $\left(\tilde{\lambda}_j^{+(k)} \right)_j$ and consequently only on $(\lambda_j^{+(k)})_j$, we can maximize each element of the sum independently. We use a common technique, and for a given k maximize the expression

$$\sum_j \log \left(\tilde{\lambda}_j^{+(k)} \right) \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+ + L \cdot \left(1 - \sum_j \tilde{\lambda}_j^{+(k)} \right), \quad (6.30)$$

where L is called the Lagrange multiplier.

Let us recall that the multinomial coefficients $\tilde{\lambda}_j^{+(k)}$ are equal to the corresponding parameters $\lambda_b^{+(k)}$ such that $b = \text{DNaseBin}_j^{+(k)}$. Now let us fix the bin b and define the set J_b grouping all the positions j falling within bin b :

$$J_b = \left\{ j: \text{DNaseBin}_j^{+(k)} = b \right\}. \quad (6.31)$$

Differentiating Formula 6.30 with respect to $\lambda_b^{+(k)}$ and setting the derivative equal to zero, we get:

$$0 = \sum_{j \in J_b} \frac{1}{\lambda_b^{+(k)}} \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+ - L \cdot |J_b|. \quad (6.32)$$

Note that the above is a decreasing function of $\lambda_b^{+(k)}$, hence we capture a local maximum here. Hence,

$$L \cdot |J_b| \lambda_b^{+(k)} = \sum_{j \in J_b} \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+ \quad (6.33)$$

and summing this equation over all $b = 1, \dots, B^{+(k)}$, we get

$$L \cdot \sum_{b=1}^{B^{+(k)}} |J_b| \lambda_b^{+(k)} = \sum_{b=1}^{B^{+(k)}} \sum_{j \in J_b} \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+. \quad (6.34)$$

We should now note that

$$\mathbf{I} = \sum_j \tilde{\lambda}_j^{+(k)} = \sum_{b=1}^{B^{+(k)}} \sum_{j \in J_b} \tilde{\lambda}_j^{+(k)} = \sum_{b=1}^{B^{+(k)}} |J_b| \lambda_b^{+(k)}. \quad (6.35)$$

Now Equation 6.34 becomes

$$L = \sum_{b=1}^{B^{+(k)}} \sum_{j \in J_b} \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+, \quad (6.36)$$

and substituting the above into Equation 6.33, we obtain the desired solution:

$$\lambda_b^{+(k)} = \frac{\sum_{j \in J_b} \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+}{|J_b| \sum_{c=1}^{B^{+(k)}} \sum_{j \in J_c} \sum_i \langle Z_i^{(k)} \rangle \text{DNase}_{i,j}^+}. \quad (6.37)$$

To increase the robustness of the model, we added pseudocounts (formally speaking, applied an estimator shrinkage) while calculating the values of the parameters $(\lambda_b^{+(k)})_{b,k}$. For each b and k , we take the regularized value $\frac{1}{2} \lambda_b^{+(k)} + \frac{1}{2} |J_b| / \sum_b |J_b|$.

Now we will find the values of $(p^{+(k)})_k$ that maximize L_A independently of the other parameters. Differentiating Equation 6.28 with respect to $p^{+(k)}$ and setting the derivative equal to zero, we obtain the closed form for $p^{+(k)}$:

$$0 = \frac{\partial L_A^+}{\partial p^{+(k)}} = \sum_i \langle Z_i^{(k)} \rangle r^{+(k)} \frac{\mathbf{I}}{p^{+(k)}} - \sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ \frac{\mathbf{I}}{\mathbf{I} - p^{+(k)}} \quad (6.38)$$

$$p^{+(k)} \sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ = (\mathbf{I} - p^{+(k)}) \sum_i \langle Z_i^{(k)} \rangle r^{+(k)} \quad (6.39)$$

$$p^{+(k)} = \frac{\sum_i \langle Z_i^{(k)} \rangle r^{+(k)}}{\sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ + \sum_i \langle Z_i^{(k)} \rangle r^{+(k)}}. \quad (6.40)$$

Again, the above is a decreasing function of $p^{+(k)}$, indicating a local maximum here.

It remains to establish the values of $(r^{+(k)})_k$ that maximize L_A . Let us recall that only the first four summands in Equation 6.28 depend on $(r^{+(k)})_k$ or $(p^{+(k)})_k$. We start with substituting Equation 6.40 into these four summands:

$$L_A'^+(\Theta) = \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(r^{+(k)} + \text{DNaseSum}_i^+)$$

$$\begin{aligned}
& - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \log \Gamma(r^{+(k)}) + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle r^{+(k)} \log \left(\sum_l \langle Z_l^{(k)} \rangle r^{+(k)} \right) \\
& - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle r^{+(k)} \log \left(\sum_l \langle Z_l^{(k)} \rangle \text{DNaseSum}_l^+ + \sum_l \langle Z_l^{(k)} \rangle r^{+(k)} \right) \\
& + \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ \log \left(\sum_l \langle Z_l^{(k)} \rangle \text{DNaseSum}_l^+ \right) \\
& - \sum_i \sum_{k=0}^{K+1} \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ \log \left(\sum_l \langle Z_l^{(k)} \rangle \text{DNaseSum}_l^+ + \sum_l \langle Z_l^{(k)} \rangle r^{+(k)} \right).
\end{aligned} \tag{6.41}$$

Unfortunately, there seems to be no closed-form solution for $r^{+(k)}$ that maximizes $L_A^+(\Theta)$. Here we again apply the BFGS numerical optimization (Broyden, 1969; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). For brevity, let us introduce the digamma function, $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$. Differentiating L_A^+ with respect to $r^{+(k)}$, we get:

$$\begin{aligned}
\frac{\partial L_A^+}{\partial r^{+(k)}} &= \frac{\partial L_A^+}{\partial r^{+(k)}} = \sum_i \langle Z_i^{(k)} \rangle \psi(r^{+(k)} + \text{DNaseSum}_i^+) - \sum_i \langle Z_i^{(k)} \rangle \psi(r^{+(k)}) \\
& + \log \left(\sum_i \langle Z_i^{(k)} \rangle r^{+(k)} \right) \sum_i \langle Z_i^{(k)} \rangle + \sum_i \langle Z_i^{(k)} \rangle \\
& - \log \left(\sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ + \sum_i \langle Z_i^{(k)} \rangle r^{+(k)} \right) \sum_i \langle Z_i^{(k)} \rangle \\
& - \frac{\sum_i \langle Z_i^{(k)} \rangle}{\sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ + \sum_i \langle Z_i^{(k)} \rangle r^{+(k)}} \sum_i \langle Z_i^{(k)} \rangle r^{+(k)} \\
& - \frac{\sum_i \langle Z_i^{(k)} \rangle}{\sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+ + \sum_i \langle Z_i^{(k)} \rangle r^{+(k)}} \sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+. \tag{6.42}
\end{aligned}$$

Writing the above equation using $p^{+(k)}$ as defined in Equation 6.40, we get:

$$\begin{aligned}
\frac{\partial L_A^+}{\partial r^{+(k)}} &= \sum_i \langle Z_i^{(k)} \rangle \psi(r^{+(k)} + \text{DNaseSum}_i^+) - \psi(r^{+(k)}) \sum_i \langle Z_i^{(k)} \rangle \\
& + (\log(p^{+(k)}) + 1 - p^{+(k)}) \sum_i \langle Z_i^{(k)} \rangle - \frac{p^{+(k)}}{r^{+(k)}} \sum_i \langle Z_i^{(k)} \rangle \text{DNaseSum}_i^+. \tag{6.43}
\end{aligned}$$

Now the numerical optimization procedure referred to above is used to find the local maximum.

The Expectation-Maximization procedure was initialized by assigning the prior probabilities as follows. In the monomer binding mode, we put $P(Z_i = 1)/P(Z_i = 0) = 100$ for the top 10% of motif instances with highest total number of DNase-seq cuts. In a dimer binding mode k , we put $P(Z_i = k)/P(Z_i = 0) = 100$ for the motif instances satisfying both of the following criteria: being within the top 10% of motif instances with highest total number of DNase-seq cuts, and being within the top 10% of motif instances with highest dimerization partner motif score. In the cases not mentioned above for any bound mode k , we put $P(Z_i = k)/P(Z_i = 0) = 0.01$.

We then estimate the values for $(\beta_j^{(k)})_{j,k}$, and for the first Maximization step we take the prior probabilities as the posterior ones. We iterate the Expectation-Maximization procedure, in each iteration getting a revised vector of parameters Θ_t , until the posterior probabilities do not change by more than 0.001, i.e.

$$\max_{i,k} |P(Z_i = k | X_i, \Theta_{t+1}) - P(Z_i = k | X_i, \Theta_t)| < 0.001. \quad (6.44)$$

In most of the cases described here, the algorithm converged in less than 30 iterations.

6.3 RESULTS

6.3.1 MOCCA SYSTEMATICALLY OUTPERFORMS THE OTHER TOOLS

We systematically benchmarked MOCCA along with two other tools for accurate footprint identification, CENTIPEDE (Pique-Regi et al., 2011) and Wellington (Piper et al., 2013). As stated in Subsection 6.2.1, we applied all the methods to DNase-seq data from three tracks: Duke DNase, University of Washington (UW) DNase and UW Digital Genomic Footprinting (DGF). From each of the DNase-seq data sources, sequence tag profiles were fetched for three human cell types: A549 (lung adenocarcinoma epithelial cell line), HepG2 (hepatocellular carcinoma cell line) and K562 (leukemia cell line). As a reference, we used ChIP-seq data from ENCODE as a golden standard for TF binding (see Subsection 6.2.2).

Since both MOCCA and CENTIPEDE learn a model for TF footprints, we visualized these models by plotting their multinomial components (Figure 6.1). Note that the curves for MOCCA were smoothed by replacing the fixed-value bins by a piecewise linear function. The multinomial models for MOCCA were based on

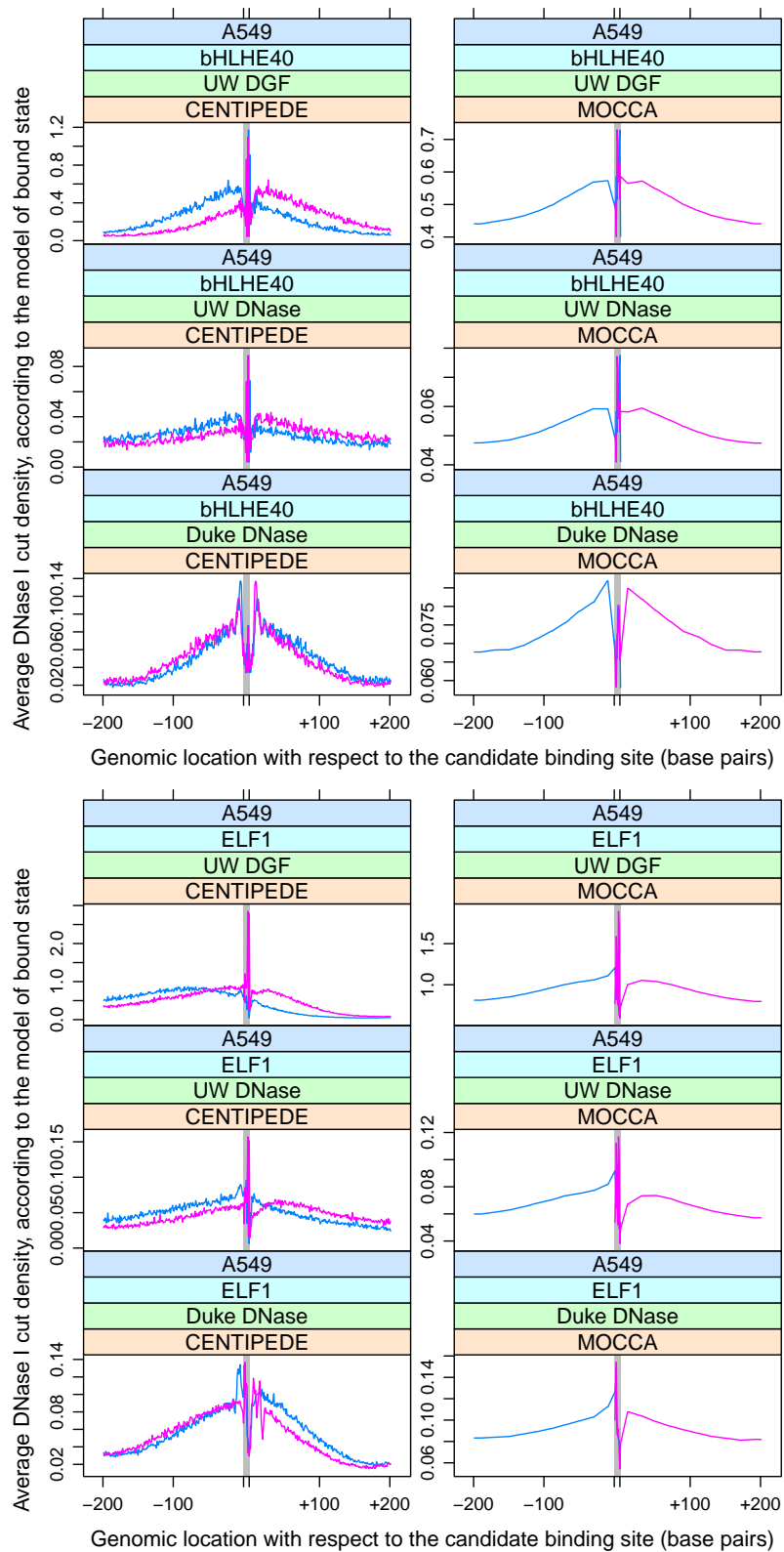


Figure 6.1: Example models of TF footprints learned by CENTIPEDE (*left*) and MOCCA (*right*) for two TFs: bHLHE40 (*top*) and ELF1 (*bottom*) in A549 cell line. Line colors indicate the strandness of DNase I cuts. In the case of MOCCA, the forward strand cuts (blue) are considered only upstream and within the cut site, while the reverse strand cuts (violet) are considered only within the cut site and downstream.

much fewer free parameters, hence they were much smoother than their CENTIPEDE counterparts.

To systematically benchmark the three tools, we have plotted Receiver Operating Characteristic (ROC) curves for each tool and each combination of cell type and TF for which we have reference ChIP-seq data. The ROC curves, showing the relationship between false positive rate (i.e. $1 - \text{specificity}$, x axis) and true positive rate (i.e. sensitivity , y axis), are the standard approach in assessing binary classifier performance. Example ROC curves for several TFs in K562 cells, shown in Figure 6.2, suggest a systematical supremacy of MOCCA over Wellington and CENTIPEDE.

Albeit routinely performed, application of ROC curves to assess the performance of TF binding site prediction is often criticized. These classifier performance assessments are characterized by relatively small number of true positives (motif instances actually bound by the TF) and large number of true negatives (motif instances that remain unbound). Hence, the shape of ROC curves and area under them is mostly affected by the ability of a particular tool to correctly predict unbound motif instances (Piper et al., 2013). To obtain a complementary view, we also plotted the Precision-Recall curves, showing the relationship between recall (i.e. sensitivity , x axis) and precision (i.e. $\text{positive predictive value}$, fraction of instances predicted as bound that is actually bound, y axis). Example Precision-Recall curves, shown in Figure 6.3, also indicate that MOCCA systematically outperforms the other tools.

A common single-dimensional measure in assessing classifier performance is the area under ROC curve, or AUROC. We have calculated this statistics for all the combinations of cell types and TFs considered, for each of the three DNase-seq data sources (Figure 6.4). We found that MOCCA consistently performed best in the case of UW and Duke DNase-seq data, while for UW DGF there was no clear supremacy of neither MOCCA nor Wellington. To obtain a summary view of the performance of each of the three tools considered (CENTIPEDE, Wellington and MOCCA) we have aggregated the AUROC statistics between different cell types and TFs in the form of boxplots (Figure 6.5). This aggregation suggests that the performance of CENTIPEDE is most affected by the choice of DNase-seq data source.

We also studied the difference between AUROC performance of MOCCA and Wellington, and found that for some TFs it is noticeably higher than for the other ones. We hypothesized that this may be related to the motif information content of these TFs. To test this hypothesis, we calculated the Pearson correlation of the difference between AUROC performance, as described above, and the motif informa-

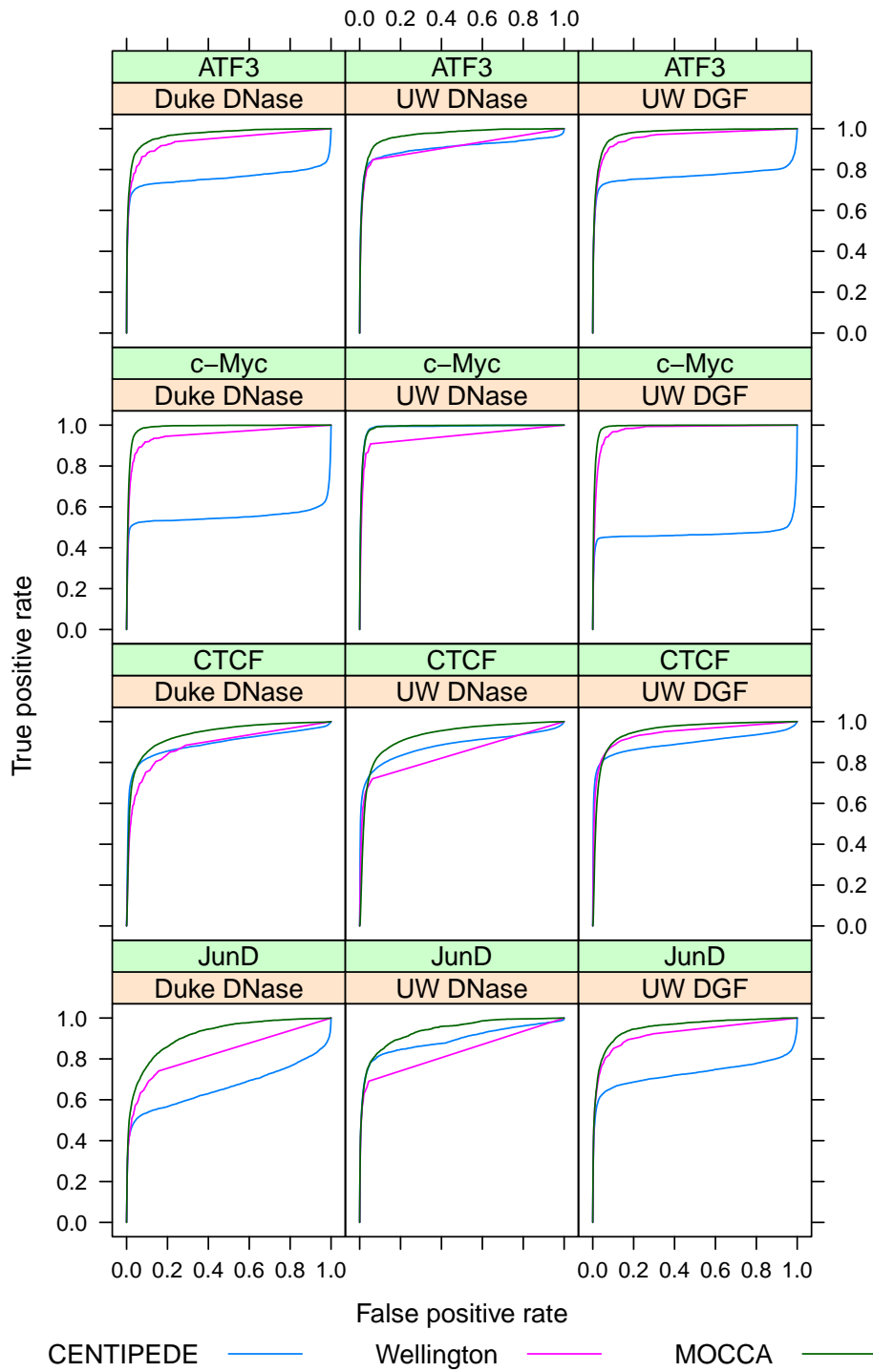


Figure 6.2: Example Receiver Operating Characteristic curves for the three tools. Prediction performance of CENTIPEDE, Wellington and MOCCA in K562 cells using three sources of DNase-seq data is shown. Only the results for a subset of 4 representative TFs are shown.

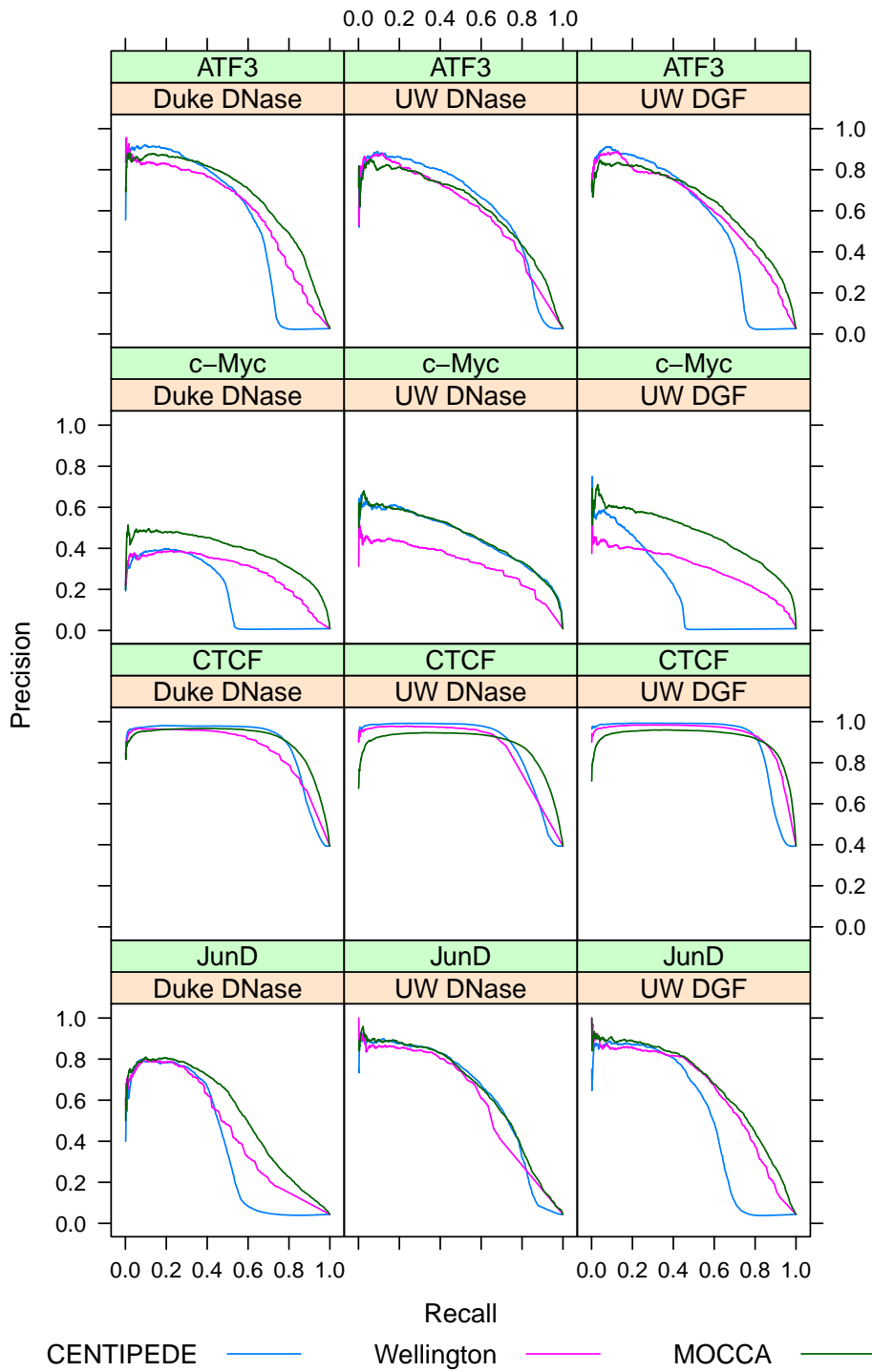


Figure 6.3: Example Precision-Recall curves for the three tools. Prediction performance of CENTIPEDE, Wellington and MOCCA in K562 cells using three sources of DNase-seq data is shown. Like in Figure 6.2, only the results for a subset of 4 representative TFs are shown.



Figure 6.4: Areas under Receiver Operating Characteristic curves for the three tools. Prediction performance of CENTIPEDE, Wellington and MOCCA in A549, HepG2 and K562 cells using three sources of DNase-seq data is shown. Apart from the values for individual TFs, the averages are indicated for each cell line.

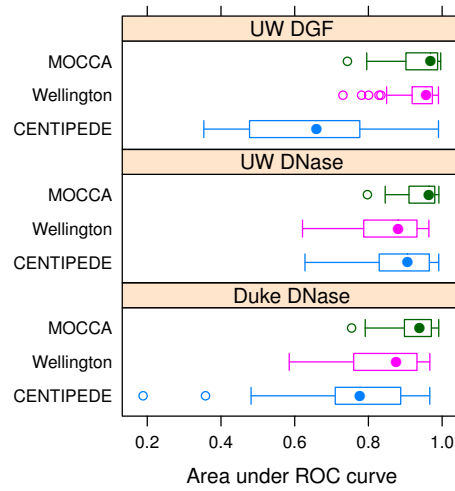


Figure 6.5: Aggregated areas under ROC curves compared between three tools and three DNase-seq data sources. Prediction performance of CENTIPEDE, Wellington and MOCCA using three sources of DNase-seq data is shown. All the cell lines (A549, HepG2 and K562) were considered jointly here.

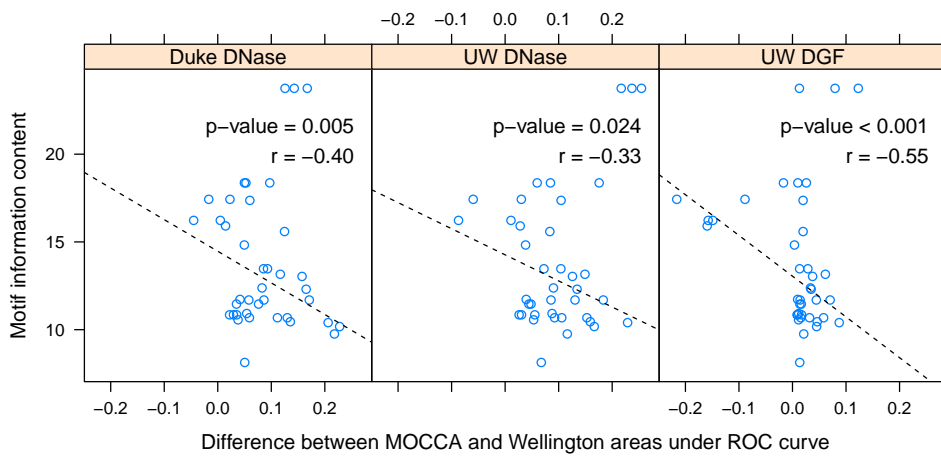


Figure 6.6: Difference between MOCCA and Wellington areas under ROC curve significantly correlates with motif information content. All the cell lines (A549, HepG2 and K562) were considered jointly here. Correlation values and p -values were calculated after excluding the outliers with information content above 20 bits.

tion content (Figure 6.6). We also permuted the motif information content between the classification tasks 1000 times to calculate the empirical p -value for the Pearson correlation coefficient r . For all the three DNase-seq data sources, we observed a statistically significant negative correlation (Figure 6.6).

6.3.2 KNOWLEDGE OF TF DIMERIZATION MODES DOES NOT IMPROVE THE PREDICTION OF INDIVIDUAL TF BINDING SITES

We benchmarked how well MOCCA predicted the binding sites for AR, FOXA1, SOX2 and OCT4, as identified by ChIP-seq datasets, when the information on the possible dimerization modes of these TFs is supplied. The choice of TFs for this study was guided by the multitude of dimeric binding modes known and predicted for AR and FOXA1. Apart from the dimers shown in Figure 4.8, we also included the AR-AR homodimer (row 13 in Table 2.1). Hence, the MOCCA model had 4 states for AR (i.e. AR monomer, AR-AR, AR-FOXA1 and unbound), and 5 states for FOXA1 (FOXA1 monomer, FOXA1-AR, FOXA1-FOXA1 divergent, FOXA1-FOXA1 convergent, unbound). The performance of TF binding site prediction was assessed in unstimulated, as well as androgen-stimulated, LNCaP cells.

For SOX2 and OCT4, we included the canonical SOX2-OCT4 heterodimer, as well as SOX9-SOX9 homodimer sharing the same SOX motif (rows 1 and 19 in Table 2.1). The MOCCA model had 4 states for SOX2 (i.e. SOX2 monomer, SOX2-OCT4, SOX9-SOX9 and unbound), and 3 states for OCT4 (i.e. OCT4 monomer, OCT4-SOX2 and unbound). The performance of TF binding site prediction was assessed in H1-hESC embryonic stem cells.

We expected that the additional binding modes would improve the overall predictive power, given that the prior information on partner motif score would allow to separate distinct dimer footprints. However, both in terms of Receiver Operating Characteristic curves (Figure 6.7) and Precision-Recall curves (Figure 6.8), we found no observable improvement. To verify whether the dimeric binding modes indeed have distinguishable profiles, we plotted the components of MOCCA model: the negative binomial component (Figure 6.9) and the multinomial component (Figure 6.10). We found that the models learned by MOCCA clearly differ between the binding modes, yet their inclusion does not improve the prediction of individual TF binding sites.

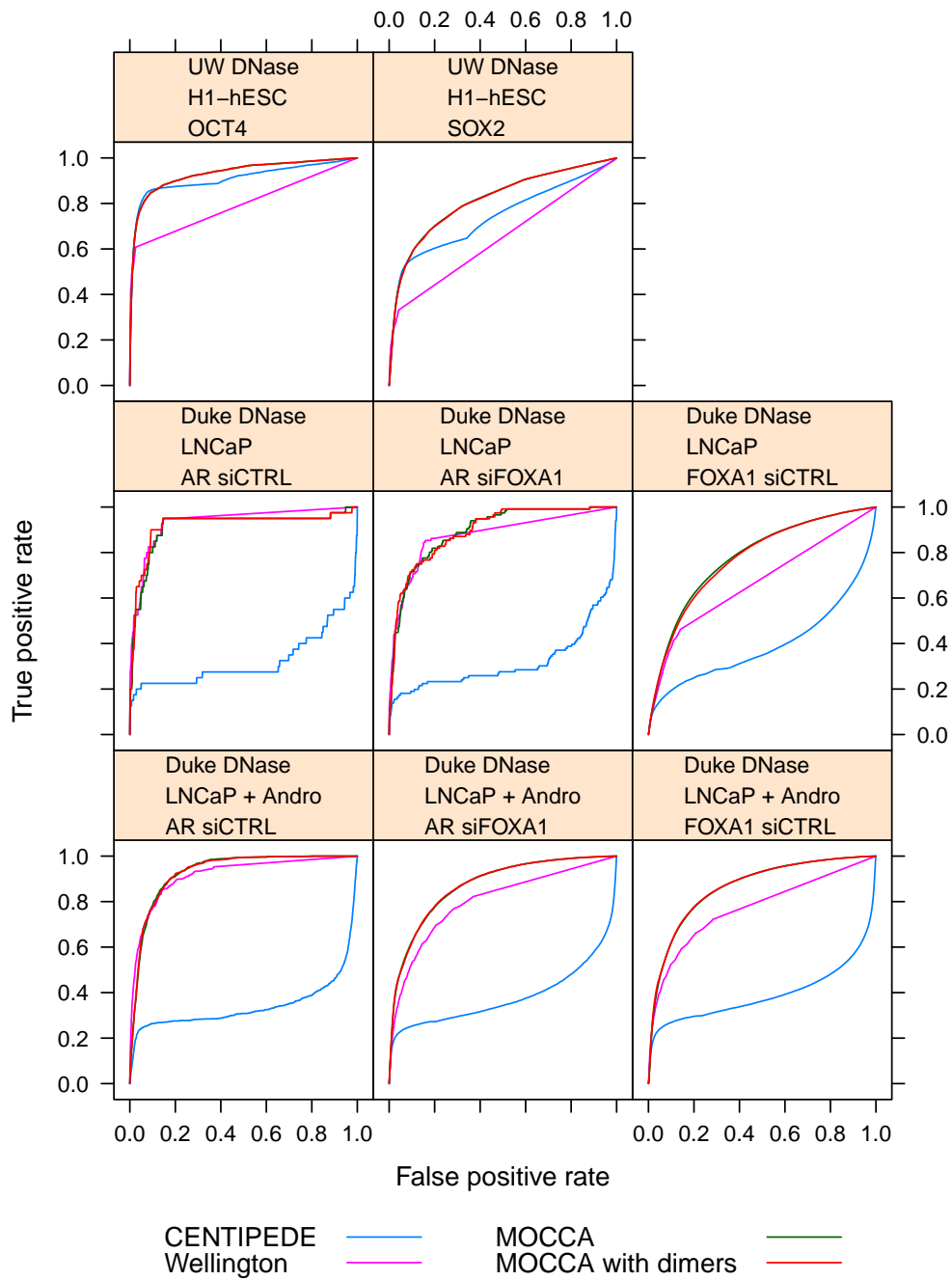


Figure 6.7: Receiver Operating Characteristic curves for the known dimers.

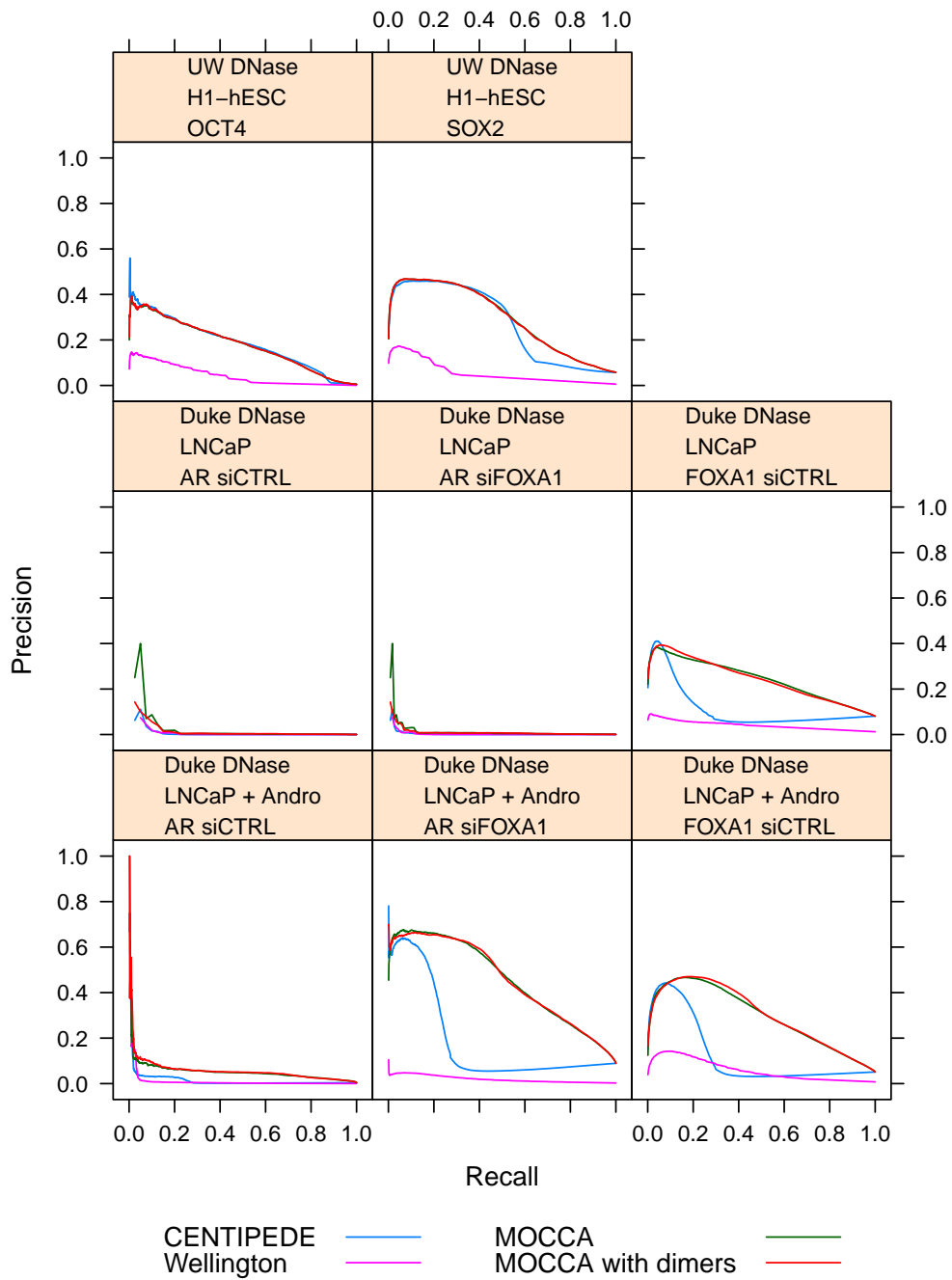


Figure 6.8: Precision-Recall curves for the known dimers.

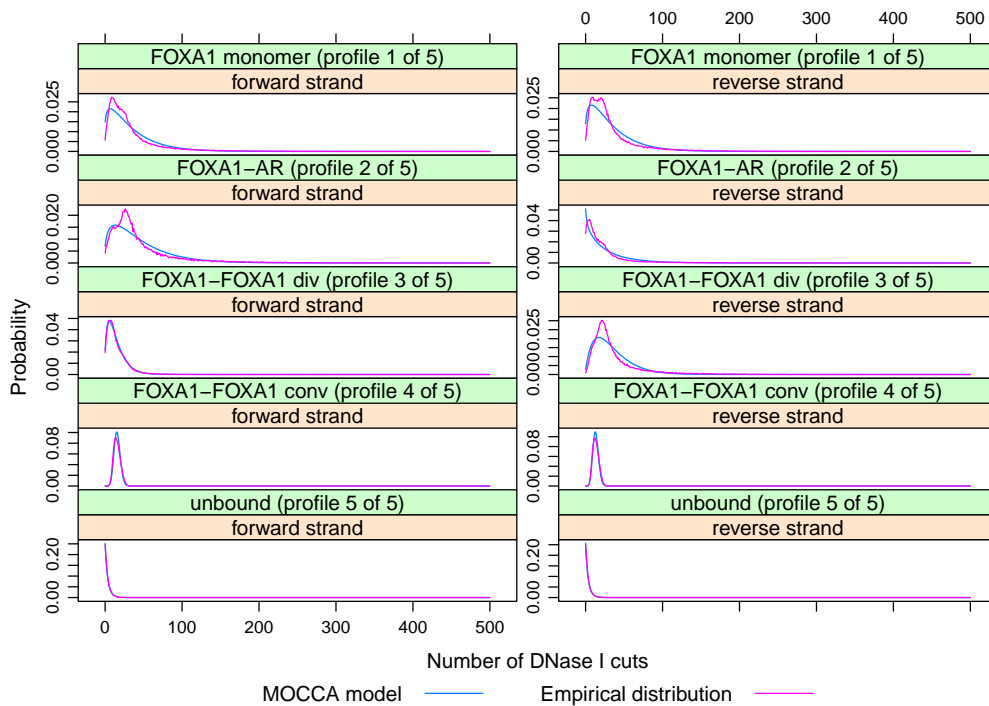


Figure 6.9: Distribution of the number of DNase I cuts learned by MOCCA for FOXA1 and its dimers in androgen-stimulated LNCaP cells. The curves show the MOCCA negative binomial model and the empirical distribution which is fitted to. *Left:* forward strand cuts, *right:* reverse strand cuts.

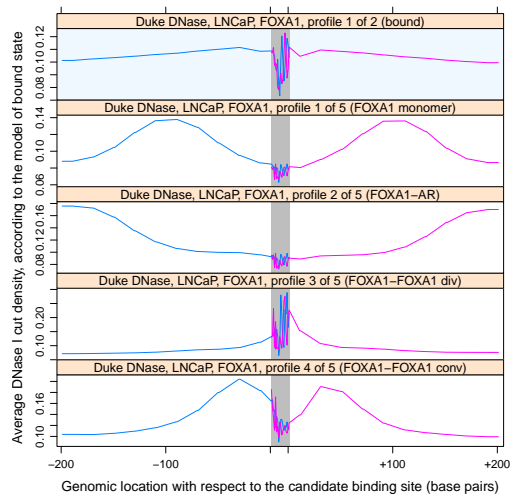


Figure 6.10: Multinomial components of the models learned by MOCCA for FOXA1 and its dimers in androgen-stimulated LNCaP cells. *The first panel* (blue background) corresponds to the model with no dimeric binding modes, and is shown here only for comparison. *The other panels* correspond to all the binding modes in the model allowing for dimerization. The unbound mode always follows the uniform distribution of cuts. The forward strand cuts (blue line) are considered only upstream and within the cut site, while the reverse strand cuts (violet line) are considered only within the cut site and downstream.

6.4 DISCUSSION

In this study, we comprehensively compared the predictive performance of three tools aimed at predicting TF binding sites from DNase-seq data. Two of them, CENTIPEDE and Wellington, used completely different approaches to address this problem. In our method proposed here, MOCCA, we combine the benefits of both of the two other tools, and showed that MOCCA consistently outperformed CENTIPEDE and Wellington, especially when applied to DNase-seq datasets with lower sequencing depth.

When allowing for more than one bound state in MOCCA, we found that the additional DNase I cut profiles can differ greatly. However, the inclusion of these additional states for the known TF dimers did not yield an increase in predictive power. We conclude that in the case of TF dimerization, it may happen that the bound TF dimers influence the chromatin state strongly enough to be detected even in a dimer-unaware manner.

7

Conclusion

Typically, TFs bind to only a very small fraction of their motif matches in the vast human genome. It is thought that the remaining motif matches remain unbound because they lie in closed chromatin ([ENCODE Project Consortium et al., 2012](#)). This model should not apply to pioneer factors, since they have the ability to bind closed chromatin. It is therefore not clear how do pioneer factors achieve binding specificity. We discovered multiple potential homo- and heterodimeric complexes involving FOXA1, and hypothesize that this pioneer factor could achieve binding specificity by exploiting a multiplicity of dimeric binding modes. The pioneer factor GATA may constitute yet another example of this phenomenon, given its multiple known and newly predicted dimeric binding modes (GATA-E-box, GATApal, GATAcpal).

We have so far assumed that the existence of a preferred motif spacing for a TF pair is indicative of dimeric binding. However, there is one other possible explanation that must be kept in mind. It has been shown that Smad4 dimers can bind cooperatively to DNA even in the absence of direct physical contacts ([Baburajendran et al., 2011](#)). The authors of this study suggested that DNA conformational changes induced by TF binding could be a mechanism for cooperative binding of

specific Smad4 homo- and heterodimers. It is conceivable that some of our predicted TF pairs might cooperate via allosteric changes in DNA structure rather than direct protein-protein contacts.

Throughout this thesis, we have almost exclusively focused on transcription factor dimers, despite the fact that higher order complexes may also play an important role. Actually we have been working on an extension of TACO from dimers towards higher order complexes. Given that even an analysis of all the possible motif trimers would be computationally too demanding, our approach was to combine overrepresented motif dimers found in the same cell type and sharing one of the motifs. We tested for enrichment of such a trimer as compared to the enrichment of underlying dimers. Unfortunately we found no strongly enriched motif trimer or any other strongly enriched higher order complex. An impeding factor here is that there are only a few higher order TF complexes that might be used as a reference set to assess the method performance. It may be also the case that possible higher order cooperative TF complexes are rather incidental, as opposed to TF dimers, which manifest themselves as motif dimers in multiple genomic locations.



Table of ENCODE cell types

Cell type	Tier	Description	Lineage	Tissue	Karyotype	Sex
8988T	3	pancreas adenocarcinoma (PA-TU-8988T)	endoderm	liver	cancer	F
A549	2	epithelial cell line derived from a lung carcinoma tissue	endoderm	epithelium	cancer	M
Adult_CD4_Th0	3	CD4+ cells isolated from human blood and enriched for Th0 populations	mesoderm	blood	normal	B
Adult_CD4_Th1	3	CD4+ cells isolated from human blood and enriched for Th1 populations	mesoderm	blood	normal	B
AG04449	3	fetal buttock/thigh fibroblast	ectoderm	skin	normal	M
AG04450	3	fetal lung fibroblast	endoderm	lung	normal	M
AG09309	3	adult toe fibroblast from apparently healthy 21 year old	ectoderm	skin		F
AG09319	3	gum tissue fibroblasts from apparently healthy 24 year old	ectoderm	gingiva	normal	F
AG10803	3	abdominal skin fibroblasts from apparently healthy 22 year old	ectoderm	skin		M
AoAF	3	aortic adventitial fibroblast cells	mesoderm	blood vessel	normal	F

Table A.1: ENCODE cell types referred to in this thesis. Table extracted and adapted from <http://genome.ucsc.edu/ENCODE/cellTypes.html>. Tier (1, 2 or 3) was assigned by ENCODE to prioritize the experiments on Tier 1 and Tier 2 first. Tier 1 consists of three most extensively studied cell lines: GM12878, H1-hESC and K562. Karyotype (normal or cancer) is indicated only if it is known. Sex of the sample donor(s) is indicated as follows: M – male, F – female, M – mixed, U – unknown.

Cell type	Tier	Description	Lineage	Tissue	Karyotype	Sex
BE2_C	3	neuroblastoma, BE-C is a clone of the SK-N-BE neuroblastoma cell line	ectoderm	brain	cancer	M
BJ	3	skin fibroblast	ectoderm	skin	normal	M
Caco-2	3	colorectal adenocarcinoma	endoderm	colon	cancer	M
CD20+_RO01778	2	B cells, caucasian, draw number 1, newly promoted to tier 2: not in 2011 analysis	mesoderm	blood	normal	F
CD34+_Mobilized	3	hematopoietic progenitor cells- mobilized, from donor RO01679.	mesoderm	blood		M
Chorion	3	chorion cells (outermost of two fetal membranes), fetal membranes were collected from women who underwent planned cesarean delivery at term, before labor and without rupture of membranes	extraembryonic mesoderm, trophecto-derm	fetal membrane		U
CLL	3	chronic lymphocytic leukemia cell, T-cell lymphocyte	mesoderm	blood	cancer	F
CMK	3	acute megakaryocytic leukemia cells	mesoderm	blood	cancer	M
Colo829	3	malignant melanoma	ectoderm	skin	cancer	M
Fibrobl	3	child fibroblast	ectoderm	skin	normal	F
FibroP	3	fibroblasts taken from individuals with Parkinson's disease, AG20443, AG08395 and AG08396 were pooled for this sample	ectoderm	skin	normal	U
FibroP_AG08395	3	fibroblasts taken from individuals with Parkinson's disease	ectoderm	skin fibroblast		F
FibroP_AG08396	3	fibroblasts taken from individuals with Parkinson's disease	endoderm	lung fibroblast		F
FibroP_AG20443	3	fibroblasts taken from individuals with Parkinson's disease	ectoderm	skin fibroblast		M
Gliobla	3	glioblastoma, these cells (aka H54 and D54) come from a surgical resection from a patient with glioblastoma multiforme (WHO Grade IV).	ectoderm	brain	cancer	U
GM04503	3	adult twin pair fibroblasts, monozygotic twin of GM04504	ectoderm	skin	normal	F
GM04504	3	adult twin pair fibroblasts, monozygotic twin of GM04503, 13% of the cells examined show random chromosome loss	ectoderm	skin	normal	F
GM06990	3	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah, treatment: Epstein-Barr Virus transformed	mesoderm	blood		F
GM12864	3	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah pedigree 1459, treatment: Epstein-Barr Virus transformed	mesoderm	blood		M
GM12865	3	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah pedigree 1459, treatment: Epstein-Barr Virus transformed	mesoderm	blood		F
GM12878	1	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah - European Caucasian, Epstein-Barr Virus	mesoderm	blood	normal	F
GM12891	3	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah pedigree 1463, treatment: Epstein-Barr Virus transformed	mesoderm	blood		M
GM12892	3	B-lymphocyte, lymphoblastoid, International HapMap Project, CEPH/Utah pedigree 1463, treatment: Epstein-Barr Virus transformed	mesoderm	blood		F

Cell type	Tier	Description	Lineage	Tissue	Karyotype	Sex
GM18507	3	B-lymphocyte, lymphoblastoid, International HapMap Project, Yoruba in Ibadan, Nigera, treatment: Epstein-Barr Virus transformed	mesoderm	blood		M
GM19238	3	B-lymphocyte, lymphoblastoid, International HapMap Project, Yoruba in Ibadan, Nigera, treatment: Epstein-Barr Virus transformed	mesoderm	blood		F
GM19239	3	B-lymphocyte, lymphoblastoid, International HapMap Project, Yoruba in Ibadan, Nigera, treatment: Epstein-Barr Virus transformed	mesoderm	blood		M
GM19240	3	B-lymphocyte, lymphoblastoid, International HapMap Project, Yoruba in Ibadan, Nigera, treatment: Epstein-Barr Virus transformed	mesoderm	blood		F
H1-hESC	1	embryonic stem cells	inner cell mass	embryonic stem cell	normal	M
H7-hESC	3	undifferentiated embryonic stem cells	inner cell mass	embryonic stem cell		U
H9ES	3	embryonic stem cells	inner cell mass	embryonic stem cell		F
HAc	3	astrocytes-cerebellar	ectoderm	cerebellar	normal	U
HA-EpiC	3	amniotic epithelial cells	pluripotent	epithelium	normal	U
HA-h	3	astrocytes-hippocampal	ectoderm	brain hippocampus	normal	U
HA-sp	3	astrocytes spinal cord	ectoderm	spinal cord	normal	U
HBMEC	3	brain microvascular endothelial cells	mesoderm	blood vessel	normal	U
HBVP	3	brain vascular pericytes	mesoderm	blood vessel	normal	U
HBVSMC	3	brain vascular smooth muscle cells.	mesoderm	blood vessel	normal	F
HCF	3	cardiac fibroblasts	mesoderm	heart	normal	U
HCFaa	3	cardiac fibroblasts- adult atrial	mesoderm	heart	normal	F
HCM	3	cardiac myocytes	mesoderm	heart	normal	U
HConF	3	conjunctival fibroblasts	ectoderm	eye		U
HCP-EpiC	3	choroid plexus epithelial cells	ectoderm	epithelium	normal	U
HCT-116	3	colorectal carcinoma	endoderm	colon	cancer	M
HEEpiC	3	esophageal epithelial cells	endoderm	epithelium	normal	U
HEK293T	3	embryonic kidney that expresses SV40 large T antigen	mesoderm	kidney		F
HeLa-S3	2	cervical carcinoma	ectoderm	cervix	cancer	F
Hepatocytes	3	primary hepatocytes, liver perfused by enzymes to generate single cell suspension	endoderm	liver	normal	B
HepG2	2	hepatocellular carcinoma	endoderm	liver	cancer	M
HFF	3	foreskin fibroblasts	mesoderm	foreskin	normal	M
HFF-Myc	3	foreskin fibroblast cells expressing canine cMyc	mesoderm	foreskin	normal	M
HGF	3	gingival fibroblasts	ectoderm	gingiva	normal	U
HIP-EpiC	3	iris pigment epithelial cells	ectoderm	epithelium	normal	U
HL-60	3	promyelocytic leukemia cells	mesoderm	blood	cancer	F
HMEC	3	mammary epithelial cells	ectoderm	breast	normal	U
HMF	3	mammary fibroblasts	ectoderm	mammary		F
HMVEC-dAd	3	adult dermal microvascular endothelial cells.	mesoderm	blood vessel	normal	F
HMVEC-dBl-Ad	3	adult blood microvascular endothelial cells, dermal-derived	mesoderm	blood vessel	normal	F
HMVEC-dBl-Neo	3	neonatal blood microvascular endothelial cells, dermal-derived	mesoderm	blood vessel	normal	M

Cell type	Tier	Description	Lineage	Tissue	Karyotype	Sex
HMVEC-dLy-Ad	3	adult lymphatic microvascular endothelial cells, dermal-derived	mesoderm	blood vessel	normal	F
HMVEC-dLy-Neo	3	neonatal lymphatic microvascular endothelial cells, dermal-derived	mesoderm	blood vessel	normal	M
HMVEC-dNeo	3	neonatal microvascular endothelial cells (single donor), dermal-derived	mesoderm	blood vessel	normal	M
HMVEC-LBl	3	blood microvascular endothelial cells, lung-derived	mesoderm	blood vessel	normal	F
HMVEC-LLy	3	lymphatic microvascular endothelial cells, lung-derived	mesoderm	blood vessel	normal	F
HNPEpiC	3	non-pigment ciliary epithelial cells	endoderm	epithelium	normal	U
HPAEC	3	pulmonary artery endothelial cells.	mesoderm	blood vessel	normal	F
HPAF	3	pulmonary artery fibroblasts	mesoderm	blood vessel	normal	U
HPDE6-E6E7	3	pancreatic duct cells immortalized with E6E7 gene of HPV	endoderm	pancreatic duct	normal	F
HPdLF	3	periodontal ligament fibroblasts	ectoderm	epithelium	normal	M
HPF	3	pulmonary fibroblasts isolated from lung tissue	endoderm	lung	normal	U
HRCEpiC	3	renal cortical epithelial cells	mesoderm	epithelium	normal	U
HRE	3	renal epithelial cells	mesoderm	epithelium	normal	U
HRGEC	3	renal glomerular endothelial cells	mesoderm	kidney	normal	U
HRPEpiC	3	retinal pigment epithelial cells	ectoderm	epithelium	normal	U
HSMM	3	skeletal muscle myoblasts	mesoderm	muscle	normal	U
HSMM_emb	3	embryonic myoblast	mesoderm	muscle		U
HSMM_FSHD	3	primary myoblast from Facioscapulohumeral Muscular Dystrophy (FSHD) patients, muscle needle biopsies	mesoderm	muscle		U
HSMMtube	3	skeletal muscle myotubes differentiated from the HSMM cell line	mesoderm	muscle	normal	U
HTR8svn	3	trophoblast (HTR-8/SVneo) cell line, a thin layer of ectoderm that forms the wall of many mammalian blastulas and functions in the nutrition and implantation of the embryo	ectoderm	blastula	normal	F
Huh-7	3	hepatocellular carcinoma	endoderm	liver	cancer	M
Huh-7.5	3	hepatocellular carcinoma, hepatocytes selected for high levels of hepatitis C replication	endoderm	liver	cancer	M
HUVEC	2	umbilical vein endothelial cells	mesoderm	blood vessel	normal	U
HVMF	3	villous mesenchymal fibroblast cells	mesoderm	connective	normal	U
iPS	3	induced pluripotent stem cell derived from skin fibroblast		induced pluripotent stem cell		B
iPS_CWRU1	3	iPS cells derived from MSC658 fibroblast		induced pluripotent stem cell		M
iPS_NIH11	3	iPS cells derived from AG20443 fibroblast		induced pluripotent stem cell		M
iPS_NIH7	3	iPS cells derived from AG08395 fibroblast		induced pluripotent stem cell		F
Jurkat	3	T lymphoblastoid derived from an acute T cell leukemia	mesoderm	blood	cancer	M
K562	1	leukemia	mesoderm	blood	cancer	F

Cell type	Tier	Description	Lineage	Tissue	Karyotype	Sex
LHCN-M2	2	skeletal myoblasts derived from satellite cells from the pectoralis major muscle of a 41 year old caucasian heart transplant donor, immortalized with lox-hygro-hTERT ("LH"), and Cdk4-neo ("CN")	mesoderm	skeletal muscle myoblast		M
LNCaP	3	prostate adenocarcinoma	endoderm	prostate	cancer	M
M059J	3	malignant glioblastoma, glioma, lack DNA-dependent protein kinase activity, deficient in repair of DNA double strand breaks, the cells are negative for glial fibrillary acidic protein (GFAP)	ectoderm	brain	cancer	M
MCF-7	2	mammary gland, adenocarcinoma	ectoderm	breast	cancer	F
Medullo	3	medulloblastoma (aka D721), surgical resection from a patient with medulloblastoma as described by Darrell Bigner (1997)	ectoderm	brain	cancer	U
Medullo_D341	3	Medulloblastoma cell line of neuron or neuron precursor origin	ectoderm	brain	cancer	U
Mel_2183	3	Melanoma Cell line	ectoderm	skin	cancer	U
Melano	3	epidermal melanocytes	ectoderm	skin	normal	U
Monocytes-CD14+	2	Monocytes-CD14+ are CD14-positive cells from human leukapheresis production, from donor RO 01826	mesoderm	monocytes	normal	F
Monocytes-CD14+_RO01746	2	Monocytes-CD14+ are CD14-positive cells from human leukapheresis production, from donor RO 01746	mesoderm	monocytes	normal	F
Myometr	3	myometrial cells	mesoderm	myometrium	normal	F
NB4	3	acute promyelocytic leukemia cell line	mesoderm	blood	cancer	U
NH-A	3	astrocytes (also called Astrocy)	ectoderm	brain	normal	U
NHBE_RA	3	bronchial epithelial cells with retinoic acid	endoderm	bronchial epithelium	normal	F
NHDF-Ad	3	adult dermal fibroblasts	mesoderm	skin	normal	F
NHDF-neo	3	neonatal dermal fibroblasts	mesoderm	skin	normal	U
NHEK	3	epidermal keratinocytes	ectoderm	skin	normal	U
NHLF	3	lung fibroblasts	endoderm	lung	normal	U
NT2-D1	3	malignant pluripotent embryonal carcinoma (NTera-2)	mesoderm	testis	cancer	M
Osteobl	3	osteoblasts (NH0st)	mesoderm	bone	normal	U
PANC-1	3	pancreatic carcinoma	endoderm	pancreas	cancer	M
PanIsletD	3	dedifferentiated human pancreatic islets from the National Disease Research Interchange (NDRI), same source as PanIslets	endoderm	pancreas		B
PanIslets	3	pancreatic islets from 2 donors, the sources of these primary cells are cadavers from National Disease Research Interchange (NDRI) and another sample isolated as in Bucher, P. et al., Assessment of a novel two-component enzyme preparation for human islet isolation and transplantation. Transplantation 79, 917 (2005)	endoderm	pancreas	normal	M
pHTE	3	primary tracheal epithelial cells	endoderm	epithelium		U
PrEC	3	prostate epithelial cell line	endoderm	prostate	normal	U
ProgFib	3	fibroblasts, Hutchinson-Gilford progeria syndrome (cell line HGPS, HGADFN167, progeria research foundation)	ectoderm	skin		M
RPMI-7951	3	human skin malignant melanoma cells	ectoderm	skin	cancer	F
RPTEC	3	renal proximal tubule epithelial cells	mesoderm	epithelium	normal	U
RWPE1	3	prostate epithelial	endoderm	prostate	normal	M

Cell type	Tier	Description	Lineage	Tissue	Karyotype	Sex
SAEC	3	small airway epithelial cells	endoderm	epithelium	normal	U
SKMC	3	skeletal muscle cells	mesoderm	muscle	normal	U
SK-N-MC	3	neuroepithelioma cell line derived from a metastatic supra-orbital human brain tumor	ectoderm	brain	cancer	F
SK-N-SH_RA	3	neuroblastoma cell line, treatment: differentiated with retinoic acid	ectoderm	brain	cancer	F
Stellate	3	hepatic stellate cells, liver that was perfused with collagenase and selected for hepatic stellate cells by density gradient	endoderm	liver	normal	F
T-47D	3	epithelial cell line derived from a mammary ductal carcinoma.	ectoderm	breast	cancer	F
Th1	3	primary Th1 T cells	mesoderm	blood		U
Th2	3	primary Th2 T cells	mesoderm	blood		U
Urothelia	3	primary ureter cell culture of urothelial cells derived from a 12 year-old girl and immortalized by transfection with a temperature-sensitive SV-40 large T antigen gene	mesoderm	urothelium	normal	F
WERI-Rb-1	3	retinoblastoma	ectoderm	eye	cancer	F
WI-38	3	embryonic lung fibroblast cells, hTERT immortalized, includes Raf1 construct	endoderm	embryonic lung	normal	F

B

Example specification files for TACO

The specification file listed below was used to benchmark TACO using University of Washington DNase-seq data in Section 5.5.

```
#
# Example specification file for TACO.
#
# Comprehensive prediction of transcription factor dimers in cell-type-specific
# open chromatin regions.

<Genome>
  FastaFile = hg19/*.fa
  MaskedRegions = coding_hg19.bed
</Genome>

#
# Open chromatin datasets, e.g. DNase-seq peaks
#
# One replicate per line, two fields are required: dataset name and BED filename.
# For each dataset, the union of all corresponding replicates will be taken.

<StronglySpecificDatasets>
  DatasetList = wgEncodeUwDnase_hg19.list

# Dataset normalization (each replicate separately):
```

```

# exclude a hypersensitive region if most of the underlying genomic sequence is masked
RegionMasking = Majority
# consider not more than the given number of regions with top signalValue
RegionCount = 50000
</StronglySpecificDatasets>

#
# Motif database, e.g. TRANSFAC, JASPAR, SwissRegulon -- preferably use only one of these
#

# TRANSFAC
<Motifs>
  Database = TRANSFAC/matrix.dat
  DatabaseSubset = TRANSFAC.vertebrata
  Sensitivity = 0.8
</Motifs>

# JASPAR
#<Motifs>
# Database = JASPAR/jaspar_CORE/non_redundant/by_tax_group/vertebrates/matrix_only/matrix_only.txt
# Sensitivity = 0.9
#</Motifs>

# SwissRegulon
#<Motifs>
# Database = SwissRegulon/weight_matrices
# Sensitivity = 0.95
#</Motifs>

#
# Various options, do not forget to adjust NumberOfThreads
#

<Options>
  NumberOfThreads = 16

  MinMotifInformationContribution = 6.0
  MaxOverlappingInformationContent = 2.0
  MaxMotifSpacing = 50
  ConsiderOrientationsSeparately = True
  ConsiderMostSignificantComplexOnly = False

  TargetInstancesThreshold = 100
  FoldChangeThreshold = 1.0
  PValueThreshold = 0.05

  DimerMotifFlanks = 5
  ClusteringAcrossDatasets = True
  ClusteringDistanceConstant = 0.0
  ClusteringDistanceMultiplier = 0.15
  ClusteringOverlapThreshold = 0.2

```

```
OutputDetailedStats = All
OutputDimerMotifs = All
OutputGenomicLocations = All
GenomicLocationsMaxSpacingDeviation = 0
OutputPValueDistribution = True
</Options>
```

In the case of Duke DNase-seq data, slightly different settings were used to enforce that all the DNase-seq peaks will have the same size. Below, only the fragment of the specification file that differs from the above one is listed.

```
<StronglySpecificDatasets>
  DatasetList = wgEncodeOpenChromDnase_hg19.list

  # Dataset normalization (each replicate separately):
  # make all hypersensitive regions the same size
  RegionSize = 300
  # exclude a hypersensitive region if most of the underlying genomic sequence is masked
  RegionMasking = Majority
  # consider not more than the given number of regions with top signalValue
  RegionCount = 50000
</StronglySpecificDatasets>
```


Bibliography

- Adams, C. C. & Workman, J. L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Molecular and cellular biology*, 15(3), 1405–1421.
- Ambrosetti, D. C., Basilico, C., & Dailey, L. (1997). Synergistic activation of the fibroblast growth factor 4 enhancer by sox2 and oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Molecular and cellular biology*, 17(11), 6321–6329.
- Aristotle (2009). *Metaphysics*. Translated by William D. Ross. The Internet Classics Archive, <http://classics.mit.edu/Aristotle/metaphysics.html>.
- Baburajendran, N., Jauch, R., Tan, C. Y. Z., Narasimhan, K., & Kolatkar, P. R. (2011). Structural basis for the cooperative DNA recognition by smad4 MHI dimers. *Nucleic acids research*, 39(18), 8213–8222.
- Bailey, T. L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2, 28–36.
- Bais, A. S., Kaminski, N., & Benos, P. V. (2011). Finding subtypes of transcription factor motif pairs with distinct regulatory roles. *Nucleic acids research*, 39(11), e76.
- Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., & Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2), 757–762.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1), 235–242.
- Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. *Developmental cell*, 21(4), 611–626.

- Blokzijl, A., ten Dijke, P., & Ibáñez, C. F. (2002). Physical and functional interaction between GATA-3 and smad3 allows TGF-beta regulation of GATA target genes. *Current biology: CB*, 12(1), 35–45.
- Boyle, A. P., Guinney, J., Crawford, G. E., & Furey, T. S. (2008). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21), 2537–2538.
- Boyle, A. P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., & Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research*, 21(3), 456–464.
- Broyden, C. G. (1969). A new double-rank minimization algorithm. *Notices of the American Mathematical Society*, 16, 670.
- Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., & Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36, D102–106.
- Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V., Geistlinger, T. R., Fox, E. A., Silver, P. A., & Brown, M. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, 122(1), 33–43.
- Chatterjee, R., Zhao, J., He, X., Shlyakhtenko, A., Mann, I., Waterfall, J. J., Meltzer, P., Sathyanarayana, B. K., FitzGerald, P. C., & Vinson, C. (2012). Overlapping ETS and CRE motifs ((g/c)CGGAAGTGACGTCA) preferentially bound by GABP α and CREB proteins. *G3 (Bethesda, Md.)*, 2(10), 1243–1256.
- Chen, F. E., Huang, D. B., Chen, Y. Q., & Ghosh, G. (1998a). Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA. *Nature*, 391(6665), 410–413.
- Chen, L., Glover, J. N., Hogan, P. G., Rao, A., & Harrison, S. C. (1998b). Structure of the DNA-binding domains from NFAT, fos and jun bound specifically to DNA. *Nature*, 392(6671), 42–48.
- Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L., & Ng, H.-H. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6), 1106–1117.

- Cheret, C., Doyen, A., Yaniv, M., & Pontoglio, M. (2002). Hepatocyte nuclear factor 1 alpha controls renal expression of the npt1-npt4 anionic transporter locus. *Journal of molecular biology*, 322(5), 929–941.
- Cotnoir-White, D., Laperrière, D., & Mader, S. (2011). Evolution of the repertoire of nuclear receptor binding sites in genomes. *Molecular and cellular endocrinology*, 334(1), 76–82.
- Courtois, G., Baumhueter, S., & Crabtree, G. R. (1988). Purified hepatocyte nuclear factor 1 interacts with a family of hepatocyte-specific promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 85(21), 7937–7941.
- Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G., & Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome research*, 16(1), 123–131.
- De Masi, F., Grove, C. A., Vedenko, A., Alibés, A., Gisselbrecht, S. S., Serrano, L., Bulyk, M. L., & Walhout, A. J. M. (2011). Using a structural and logics systems approach to infer bHLH-DNA binding specificity determinants. *Nucleic acids research*, 39(11), 4553–4563.
- DeLano, W. L. (2002). The PyMOL molecular graphics system.
- Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta crystallographica. Section D, Biological crystallography*, 60, 2126–2132.
- ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.
- Fleming, J. D., Pavesi, G., Benatti, P., Imbriano, C., Mantovani, R., & Struhl, K. (2013). NF- γ coassociates with FOS at promoters, enhancers, repetitive elements, and inactive chromatin regions, and is stereo-positioned with growth-controlling transcription factors. *Genome research*, 23(8), 1195–1209.
- Fletcher, R. (1970). A new approach to variable metric methods. *The Computer Journal*, 13, 317–322.
- Friedman, P. N., Chen, X., Bargonetti, J., & Prives, C. (1993). The p53 protein is an unusually shaped tetramer that binds directly to DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 90(8), 3319–3323.
- Genzer, M. A. & Bridgewater, L. C. (2007). A col9a1 enhancer element activated by two interdependent SOX9 dimers. *Nucleic acids research*, 35(4), 1178–1186.

- Goldfarb, D. (1970). A family of variable metric methods derived by variational means. *Mathematics of Computation*, 24, 23–26.
- Grove, C. A., De Masi, F., Barrasa, M. I., Newburger, D. E., Alkema, M. J., Bulyk, M. L., & Walhout, A. J. M. (2009). A multiparameter network reveals extensive divergence between *c. elegans* bHLH transcription factors. *Cell*, 138(2), 314–327.
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, 8(2), R24.
- Hannenhalli, S. & Levy, S. (2002). Predicting transcription factor synergism. *Nucleic acids research*, 30(19), 4278–4284.
- He, X., Chen, C.-C., Hong, F., Fang, F., Sinha, S., Ng, H.-H., & Zhong, S. (2009). A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data. *PLoS one*, 4(12), e8155.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4), 576–589.
- Hollenhorst, P. C., Chandler, K. J., Poulsen, R. L., Johnson, W. E., Speck, N. A., & Graves, B. J. (2009). DNA specificity determinants associate with distinct transcription factor functions. *PLoS genetics*, 5(12), e1000778.
- Johnson, R., Teh, C. H.-l., Kunarso, G., Wong, K. Y., Srinivasan, G., Cooper, M. L., Volta, M., Chan, S. S.-l., Lipovich, L., Pollard, S. M., Karuturi, R. K. M., Wei, C.-l., Buckley, N. J., & Stanton, L. W. (2008). REST regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS biology*, 6(10), e256.
- Kahler, R. A. & Westendorf, J. J. (2003). Lymphoid enhancer factor-1 and beta-catenin inhibit runx2-dependent transcriptional activation of the osteocalcin promoter. *The Journal of biological chemistry*, 278(14), 11937–11944.
- Kazemian, M., Pham, H., Wolfe, S. A., Brodsky, M. H., & Sinha, S. (2013). Widespread evidence of cooperative DNA binding by transcription factors in *drosophila* development. *Nucleic acids research*, 41(17), 8237–8252.
- Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945), 147–151.
- Littler, D. R., Alvarez-Fernández, M., Stein, A., Hibbert, R. G., Heidebrecht, T., Aloy, P., Medema, R. H., & Perrakis, A. (2010). Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. *Nucleic acids research*, 38(13), 4527–4538.

- Lu, P., Rha, G. B., Melikishvili, M., Wu, G., Adkins, B. C., Fried, M. G., & Chi, Y.-I. (2008). Structural basis of natural promoter recognition by a unique nuclear receptor, HNF4alpha. diabetes gene product. *The Journal of biological chemistry*, 283(48), 33685–33697.
- Luo, K. & Hartemink, A. J. (2013). Using DNase digestion data to accurately identify transcription factor binding sites. *Pacific Symposium on Biocomputing*, (pp. 80–91).
- Lupien, M., Eeckhoute, J., Meyer, C. A., Wang, Q., Zhang, Y., Li, W., Carroll, J. S., Liu, X. S., & Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6), 958–970.
- McLure, K. G. & Lee, P. W. (1998). How p53 binds DNA as a tetramer. *The EMBO journal*, 17(12), 3342–3350.
- Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52), 22534–22539.
- Myšičková, A. & Vingron, M. (2012). Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC genomics*, 13 Suppl 1, S2.
- Ng, C. K. L., Li, N. X., Chee, S., Prabhakar, S., Kolatkar, P. R., & Jauch, R. (2012). Deciphering the sox-oct partner code by quantitative cooperativity measurements. *Nucleic acids research*, 40(11), 4933–4941.
- Pachkov, M., Balwiercz, P. J., Arnold, P., Ozonov, E., & van Nimwegen, E. (2013). SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic acids research*, 41, D214–220.
- Passeron, T., Valencia, J. C., Namiki, T., Vieira, W. D., Passeron, H., Miyamura, Y., & Hearing, V. J. (2009). Upregulation of SOX9 inhibits the growth of human and mouse melanomas and restores their sensitivity to retinoic acid. *The Journal of clinical investigation*, 119(4), 954–963.
- Piper, J., Elze, M. C., Cauchy, P., Cockerill, P. N., Bonifer, C., & Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research*, 41(21), e201.
- Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3), 447–455.
- Plato (2009). *The Republic*. Translated by Benjamin Jowett. The Internet Classics Archive, <http://classics.mit.edu/Plato/republic.html>.

- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1), 110–121.
- Qian, J., Esumi, N., Chen, Y., Wang, Q., Chowers, I., & Zack, D. J. (2005). Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation. *Nucleic acids research*, 33(11), 3479–3491.
- Rahmann, S., Müller, T., & Vingron, M. (2003). On the power of profiles for transcription factor binding site detection. *Statistical applications in genetics and molecular biology*, 2, Article7.
- Reinhold, M. I. & Naski, M. C. (2007). Direct interactions of runx2 and canonical wnt signaling induce FGF18. *The Journal of biological chemistry*, 282(6), 3653–3663.
- Shaffer, P. L., Jivan, A., Dollins, D. E., Claessens, F., & Gewirth, D. T. (2004). Structural basis of androgen receptor binding to selective androgen response elements. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14), 4758–4763.
- Shanno, D. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24, 647–657.
- Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J., & Mann, R. S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*, 147(6), 1270–1282.
- Tanaka, N., Kawakami, T., & Taniguchi, T. (1993). Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system. *Molecular and cellular biology*, 13(8), 4531–4538.
- Trainor, C. D., Omichinski, J. G., Vandergon, T. L., Gronenborn, A. M., Clore, G. M., & Felsenfeld, G. (1996). A palindromic regulatory site within vertebrate GATA-1 promoters requires both zinc fingers of the GATA-1 DNA-binding domain for high-affinity interaction. *Molecular and cellular biology*, 16(5), 2238–2247.
- Treiber, N., Treiber, T., Zocher, G., & Grosschedl, R. (2010). Structure of an ebfi:DNA complex reveals unusual DNA recognition and structural homology with rel proteins. *Genes & development*, 24(20), 2270–2275.
- Umesono, K., Murakami, K. K., Thompson, C. C., & Evans, R. M. (1991). Direct repeats as selective response elements for the thyroid hormone, retinoic acid, and vitamin d3 receptors. *Cell*, 65(7), 1255–1266.

- Wadman, I. A., Osada, H., Grütz, G. G., Agulnick, A. D., Westphal, H., Forster, A., & Rabbitts, T. H. (1997). The LIM-only protein *lmo2* is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, *e47*, GATA-1 and *Id1*/NLI proteins. *The EMBO journal*, 16(11), 3145–3157.
- Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M. U., Ohgi, K. A., Glass, C. K., Rosenfeld, M. G., & Fu, X.-D. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature*, 474(7351), 390–394.
- Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M., & Weng, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9), 1798–1812.
- Whittington, T., Frith, M. C., Johnson, J., & Bailey, T. L. (2011). Inferring transcription factor complexes from ChIP-seq data. *Nucleic acids research*, 39(15), e98.
- Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in bioinformatics*, 9(4), 326–332.
- Yu, X., Lin, J., Zack, D. J., & Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic acids research*, 34(17), 4925–4936.
- Zaret, K. S. & Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25(21), 2227–2241.

THIS THESIS WAS TYPESET using \LaTeX , originally developed by Leslie Lamport and based on Donald Knuth's \TeX . The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.