

Mixture modelling to characterize diversity in DNA regions

A thesis
submitted in partial fulfillment of the requirements
of the degree of

Doctor of Philosophy

by

Anushua Biswas

ID: 20213301

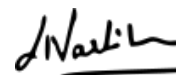


**INDIAN INSTITUTE OF SCIENCE EDUCATION AND
RESEARCH PUNE**

June 20, 2023

Certificate

Certified that the work incorporated in the thesis entitled “*Mixture modelling to characterize diversity in DNA regions*”, submitted by *Anushua Biswas* was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other university or institution.



Date: June 20, 2023

Dr. Leelavati Narlikar
Thesis Supervisor

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Anushua Biswas

Date: June 20, 2023

Anushua Biswas
Roll Number: 20213301

Acknowledgements

I had started my PhD journey from being a Project Assistant under Dr Leelavati Narlikar in National Chemical Laboratory (NCL) to joining the AcSIR PhD program there. When my advisor, Dr Narlikar moved to the new Data Science Department in IISER Pune, I was fortunate to get the chance to continue my PhD under her as a transfer student to IISER Pune, and I feel fortuitous to finally be able to submit my thesis here. Along this journey, I have many people to thank who helped me in different ways.

I would like to start by extending my heartfelt gratitude towards my advisor Dr Leelavati Narlikar for not only introducing me to this wonderful field of computational biology but also helping me at every step in this journey of PhD. It was her encouragement and support that gave me the motivation to learn from the basics of biology to the different, often times complicated methods, that are used to investigate the huge amounts of biological experimental data. She has been an excellent mentor - listening to my long, sometimes esoteric ramblings and helping me make sense of my ideas and results. She gave me plenty of opportunities to interact with and learn from people in the research community working in various interdisciplinary fields and present our work to them, for which I am ever grateful. Her insights and advice have played a major role in guiding me and helping me to learn how to be a better researcher. I am forever grateful to her for choosing me as her student!

As I had first joined the AcSIR PhD program in NCL, my first Doctoral Advisory Committee comprised of some of the distinguished scientists from there. I would like to thank Dr Chetan Gadgil, Dr Ramrup Sarkar and Dr Anuya Nisal for their invaluable inputs and advice regarding my work as well as the other aspects of working in research. I am thankful to Dr Gadgil for always coming up with some simple yet most important questions that never failed to give me a new perspec-

tive about my work. I will forever cherish the memories of the coffee sessions and the discussions, along with the numerous joint lab meetings which I had attended with Dr Gadgil and his students in the lab. I would like to thank Dr Sarkar for methodically evaluating my work and providing suggestions to make it more comprehensive. Also, I would like to thank Dr Nisal for her motivation and support in all the committee review meetings, which bolstered my spirits to give a clear and succinct presentation. When I transferred to the PhD program in IISER Pune, I had the good fortune of still having Dr Chetan Gadgil in my Research Advisory Committee along with Dr Girish Ratnaparkhi. They were both very supportive and understanding while giving me the push to finish my work and tying up the loose ends.

I would like to thank Dr Rahul Siddharthan from IMSc for the many discussions throughout my PhD work and even during the COVID-19 years. I am grateful to Dr Julia Zeitlinger and Melanie Weilert from Stowers Institute and Dr Shaun Mahony from Penn State University for their huge help with the ChIP-nexus and ChIP-exo datasets. Their questions and interest in my work helped make my experience of attending the online conferences delightful and interesting.

Next, I thank Dr Amit Apte for helping me during my term as a teaching assistant by tirelessly answering my endless questions and teaching me many new things along the way. I also thank Dr Pranay Goel for letting me be a part of some of their group meetings and giving some valuable suggestions to better understand my work.

I would next like to thank my colleagues, from both NCL and IISER. I am especially grateful to Sneha Mitra, who was my senior during my Masters course in Pune University as well as in Narlikar Lab in NCL. I had started my journey in computational biology by working on understanding the method DIVERSITY, which was written by Sneha. It was her generous help and unfaltering commitment towards building as bug free code as possible that gave me the initial impetus to persevere and get a thorough understanding of the method while trying to comprehend the biological connotations behind it. I thank Sucheta Gokhale, Bhagyashree Hote and Pooja Kumbhar, who were my colleagues in Narlikar Lab and from whom I learned a lot of things over the years. I also extend my heartfelt thanks to Pooja Dnyane who had joined the PhD program under Dr Gadgil at the same time as mine and was since my lab mate, batch mate and also a dear friend. Over the

years, we were joined by Aparna, Indrani and Ashley. I thank all of them for being great friends and for all the varied experiences we shared and for all the time when we discussed everything from science, culture to food. After I transferred to IISER to continue my PhD tenure, I got to share office space with some amazing people in the Mathematics Department. I would like to thank all of them, specially Visakh and Namrata, as they extended an unreserved welcome, taught me a variety of card games and for the uncountable small things that made my time there memorable. I also thank Ajit, Sandra, Somashree, Mukundan and Atreyi for being such wonderful people with whom I got a chance to work with.

I am grateful to all the staff members from the students' academic office, C & P division and Mr Raheja in CEPD office in NCL for their help and support in all admin related work throughout my tenure in NCL. I would also like to thank the staff in IISER offices: Sayalee, Suvarna, Yogesh, Tushar, Prabhakar and many others whose names I do not know, as they helped me make a seamless transition and answered all my questions with patience and grace.

Finally, I would like to thank my family for being there with me in all my trials, successes and failures and for their wholehearted love and support.

Abstract

The degrees of expression of the hundreds of genes in a eukaryotic cell influence its phenotype and functioning. The binding of proteins such as transcription factors to particular regulatory regions on the DNA play an important role in this process of regulation of gene expression. Mutations in these regulatory regions can affect the gene expression and can often lead to misregulation resulting in disorders and diseases. Therefore, to profile a wide range of regulation related biochemical activities, a variety of high-throughput experimental assays have been designed. They give a genome wide map of the regions having certain common characteristics for which they have been profiled. Some of the examples include STARR-seq which recognizes active enhancers, ATAC-seq, that detects accessible chromatin, ChIP-seq, which is used to identify TF binding sites. These assays report regions that are 200 to 1000 bases long, although the functional elements present in these regions are ≈ 15 bases in length. Current computational algorithms look for a common characteristic within these reported regions to identify these short sequence signatures. Evidence, however, suggests that these regions reported by the experiments have considerable heterogeneity in them. In fact, while these methods can pick up on the stronger signals, they can easily miss out on the weaker or less frequent ones.

In order to explicitly characterize the heterogeneity in these regions, I considered the question as a mixture modelling problem. Our first method, DIVERSITY, was developed to cluster regions from a ChIP-seq experiment into groups while simultaneously learning sequence signatures specific to each group in a *de novo* manner. DIVERSITY provides novel insights into the different ways in which a protein can bind DNA, including co-operative binding with other proteins. We next looked at regions identified by exonuclease-based ChIP experiments. They measure the exonuclease activity very close to the actual protein binding sites with

high precision, characterizing those regions with sharp read profile distributions. Our next method ExoDiversity models these regions by learning a joint probability distribution over the distinct ChIP-exo read signals in the forward and the reverse strands and the DNA sequences. It could resolve the binding footprints of the profiled TFs at nucleotide level resolution. These differences correlated with distinct DNA structure properties and sequence conservation profiles, implying that they are likely to have some functional importance. Finally, we investigated the varied mechanisms of TF binding at the regulatory regions. The combinatorial control exhibited by the TFs at these were modelled by our method CISDIVERSITY. Without requiring any prior knowledge of TF, cell type or organism CISDIVERSITY could determine discrete regulatory modules with particular combinations of TF binding sites when it was applied to multiple datasets across diverse types of species. We believe that our model-based approach of explaining the data in terms of various sequence components provides a comprehensive understanding of the regulatory information encoded in the data.

Contents

Abstract	xi
1 Introduction	17
1.1 Introduction	17
1.2 Regulation of gene expression	19
1.3 Transcriptional regulation	21
1.4 Cis-regulatory elements	23
1.5 Transcriptional misregulation	25
1.6 Motif model	26
1.7 High throughput sequencing experiments	28
1.7.1 Experiments to identify TF binding sites: Chromatin immunoprecipitation-based	28
1.7.2 Experiments to detect open chromatin regions	31
2 Mixture model for DNA regulatory regions	35
2.1 Problem Statement	35
2.1.1 Objective function	36
2.2 Gibbs sampling	37
2.3 Gibbs sampling applied to motif discovery	38
2.3.1 Collapsed Gibbs sampling	39
2.4 Mixture model	44
3 Identifying different modes of TF binding from ChIP data	47
3.1 Introduction	47
3.2 Problem statement	48
3.3 DIVERSITY model	49

3.3.1	Model and its parameters	50
3.3.2	Objective function	51
3.3.3	Model learning	52
3.3.4	Convergence of the sampler	53
3.3.5	Model selection	54
3.4	Results	55
3.4.1	CTCF in fly makes diverse contacts	55
3.4.2	Variations in Su(Hw) binding specificities	58
3.4.3	Pita interacts with CTCF and Su(Hw)	60
3.4.4	Binding activity of REST in neuronal and non-neuronal cells	61
3.4.5	Number of modes is a characteristic of the binding TF	66
3.4.6	Comparison with other motif discovery methods	67
3.5	Discussion	72
4	Resolving protein-DNA footprints from exonuclease-based ChIP experiments	77
4.1	Introduction	78
4.2	Approach	80
4.3	Model description	80
4.4	Objective function	84
4.5	Model learning	84
4.6	Model selection	87
4.7	Using models to predict	88
4.8	Using different distribution for the reads	89
4.9	Results	93
4.9.1	In silico mixed data	94
4.9.2	ExoDiversity results on ChIP-nexus datasets for Sox2 and Oct4	96
4.9.3	ExoDiversity models can distinguish between sets bound by TFs with similar database motifs	101
4.9.4	ExoDiversity finds diverse modes in FoxA1 and CTCF datasets	103
4.9.5	ExoDiversity results on the Twist dataset	106
4.10	Discussion	108

5	Detecting <i>cis</i>-regulatory diversity in DNA regulatory regions	113
5.1	Introduction	114
5.2	CISDIVERSITY model	115
5.2.1	Calculating the posterior probability	117
5.2.2	Model learning	118
5.3	A sequence model for representing <i>cis</i> -regulatory diversity	121
5.3.1	Input to CISDIVERSITY	121
5.3.2	CISDIVERSITY model	121
5.3.3	Output of CISDIVERSITY	122
5.4	Evaluating CISDIVERSITY on simulated datasets	123
5.4.1	Creating simulated datasets	123
5.4.2	Performance on simulated datasets	125
5.5	Results	127
5.5.1	Core promoter architectures retrieved from genome-wide TSS maps	128
5.5.2	CTCF bound regions display remarkable sequence level di- versity	131
5.5.3	GR binds to diverse regions after activation	135
5.5.4	Enhancers have a different structure based on the detection assay	141
5.5.5	Modules in open regions have differing future fates	146
5.5.6	Tissue specific modules discovered in open regions	150
5.6	Discussion	153
6	Software and datasets	159
6.1	Software for each of the methods	159
6.1.1	DIVERSITY	159
6.1.2	Input and output of DIVERSITY	161
6.1.3	ExoDiversity	164
6.1.4	CISDIVERSITY	165
6.2	Biological data sets	166
6.2.1	DIVERSITY	166
6.2.2	ExoDiversity	167
6.2.3	CISDIVERSITY	168

7	Conclusions and future directions	171
7.1	Mixture model-based probabilistic frameworks	172
7.1.1	DIVERSITY	173
7.1.2	ExoDiversity	174
7.1.3	CISDIVERSITY	178
7.2	Limitations of sampling-based methods	180
7.3	Deep learning to identify sequence signatures	180

List of Figures

1.1	Cis-regulatory elements (Reproduced from Maston et al. <i>Annu. Rev. Genomics Hum. Genet.</i> 2006)	23
1.2	Motif model	27
1.3	Enzyme-based experiments (Reproduced from Zentner et al. <i>Nature Reviews Genetics.</i> 2014)	32
2.1	Sequences with motif	39
3.1	In this example the red protein binds to DNA as (i) homodimer, (ii) monomer, (iii) indirectly to another DNA region via the blue protein or (iv) indirectly to yet another DNA region via the purple protein making no direct DNA contact this time. The ChIP experiment pulls down all the complexes that are associated with the red protein, and sequencing reports all the DNA regions participating in these complexes. DIVERSITY learns the positions of the motifs in each sequence while simultaneously splitting them into different subsets based on the motif it contains.	50

3.2	(a) 200 bp regions centered around the summit of ChIP peaks, input to DIVERSITY. (b) DIVERSITY finds eight modes which are aligned and reordered. (c) Motifs corresponding to modes. CTCF motifs from JASPAR and from high throughput SELEX [Jolma et al., 2013] are shown below. (d) Sequence conservation profile from phastCons, corresponding to nucleotides in b (e) The eight modes are displayed in decreasing ChIP score. (f) Violin plot of distance of each sequence from the closest transcription start site. (g) Violin plot of expression values of genes ($\log_2(1+\text{RPKM})$) with TSS within 2000 bp of the ChIP region. The red line shows the median value across all measured genes. (h) Overlaps with Su(Hw) and Pita ChIP experiments, respectively.	56
3.3	DIVERSITY results on 4 fly datasets for CTCF: 2180 sequences in <i>D. melanogaster</i> , 2191 sequences in <i>D. simulans</i> , 2984 sequences in <i>D. yakuba</i> , 2381 sequences in <i>D. pseudoobscura</i>	57
3.4	(a) Logos, (b) sequences, and (c) phastCons scores corresponding to the Su(Hw)-like motifs discovered by DIVERSITY.	59
3.5	(a) DIVERSITY finds the Pita motif as well as CTCF and Su(Hw) motifs. The literature motif is a result of conventional motif discovery on the same dataset [Zolotarev et al., 2016]. (b) Both Pita motifs have a significantly high ChIP score. (c) Table describes the various known direct binding motifs identified in each of the three datasets.	60
3.6	(a) The 12 modes are sorted on the basis on average ChIP score. (b) Sequences corresponding to RE-1 (mode 1), YY1 (modes 3 and 6), and RFX2 (mode 9) are more conserved. (c) While many modes are close to transcription start sites, modes 2, 4, and 9, are more variable in terms of their relative position. (d) All genes except those close to the RE-1 mode are significantly more expressed ($p < 10^{-5}$) than average.	62
3.7	(a,b) DIVERSITY identifies seven modes, all variants of RE-1 (modes sorted based on average ChIP score). Database motif shown below. (c) The full sites are more conserved (d) The full sites have well-organized chromatin structure.	63

3.8	DIVERSITY output on ChIP-seq datasets for REST in A549 and HepG2 cell lines	64
3.9	MEME output on ChIP-seq dataset for REST in GM12878 cell line.	69
3.10	DREME output on ChIP-seq dataset for REST in GM12878 cell line.	70
3.11	InMoDe output on ChIP-seq dataset for REST in GM1878 cell line	71
3.12	Run time of DIVERSITY is comparable to that of MEME. DIVERSITY and MEME were run on all 16 REST datasets as compiled by Rockowitz et al. (2014) [Rockowitz et al., 2014]. Both were run with the option of using 16 processors on Intel(R) Xeon(R) CPU E5-2630 v3.	73
4.1	The i th datapoint consisting of the DNA sequence $\mathbf{X}_i$, positive strand reads $\mathbf{R}_i^+$, negative strand reads $\mathbf{R}_i^-$ belonging to mode k is shown. The red and the blue boxes correspond to the positive and negative strand read windows having width w_k^+ and w_k^- respectively. They are modelled using Bernoulli distributions given in ϕ_k^+ and ϕ_k^- . The 1 and 0 are the binarized read count values. The pink box represent the DNA motif which is present at location Z_i , and it is modelled using PWM parameters ϕ_k . The positive and negative strand read windows are at a fixed distance τ_k^+ and τ_k^- from Z_i . Only $\mathbf{X}_i$, $\mathbf{R}_i^+$ and $\mathbf{R}_i^-$ are known; the rest are learned from the data	81
4.2	The red region shows all possible start positions for the positive strand read window and the blue region indicates the possible start positions for the negative strand read window. This assumption (without using the flag added later) was used to get the results presented here.	86
4.3	Motifs found in a mixed dataset with 5000 FoxA1 and 5000 CTCF regions in the MCF-7 cell line. Unlike in ExoDiversity, the numbers in the pie charts for ChExMix do not add up to 10000 since it finds multiple subtypes in some regions and none in others	95

4.4	ExoDiversity finds 12 modes in the Sox2 dataset. The DNA sequence and the reads in the positive and negative strands in a 50 bp neighbourhood are shown here. The central two lines in each plot mark the site of the corresponding motif. The last column indicates which of the regions overlap with the peaks reported in the ChIP-nexus experiment for Oct4.	98
4.5	The matrices show the probability distribution of a sequence over the 12 possible modes, averaged over sequences in each mode. The first matrix considers the full model, while the other two use only one of the components: either only the sequence part or only the reads part	99
4.6	ExoDiversity finds 30 modes in the ChIP-nexus dataset for Oct4. Of the many modes, it identified Oct4::Sox2 dimer motifs (modes 1-8), Oct4::Oct4 motif (mode 9), Oct4 motif (modes and 11), Sox2 motifs (modes 12, 13, 19, 20), Klf4 (modes 14, 15), Esrrb (mode 16), CTCF (mode 17). These plots were plotted in a 100 bp neighbourhood around the motifs, with the central two lines indicating the motif sites.	100
4.7	Performance of ExoDiversity-based classifier and gkm-SVM on distinguishing between Sox2-bound and Oct4-bound regions. The programs were trained on 80% of the dataset, with ExoDiversity being supplied the reads at those regions as well.	101
4.8	The average areas under the ROC and PR curves for classifying Sox2 versus Oct4 across five folds are shown. The BIC-based model performs the best with or without reads on the test data. On the training set, gkm-SVM performs the best.	102
4.9	ExoDiversity finds 13 modes in the FoxA1 dataset. The DNA sequence and the reads in the positive and negative strands in a 500 bp neighbourhood are shown. The central two lines in each plot mark the site corresponding to the FoxA1 motif.	104

- 4.10 ExoDiversity finds 10 modes in the CTCF dataset. The read profiles from both the positive (red) and negative (blue) strands over a 100 bp window about the motifs along with the motifs learned for the 8 modes resembling the canonical CTCF motif are shown in the left. The DNA sequence, the two sets of reads and the sequence conservation in terms of phastCons scores in the 100 bp neighbourhood are shown. The X-axes of the read profiles are aligned corresponding to the 19 bp JASPR CTCF motif which is also indicated by the two central lines in the four heatmap images and the two ticks in the read profiles. The read profiles are over a 100 bp region, while the motifs shown below them are over a 40 bp region. 105
- 4.11 ExoDiversity finds 25 modes in the Twist dataset. Modes 1-9 correspond to the Twist motifs, mode 10 matches a *Drosophila* core promoter motif and modes 17-20 are di/tri nucleotide repeat patterns. The DNA sequence, the two sets of reads and the DNA shape for propeller Twist in the 50 bp neighbourhood are shown with the central two lines marking the site corresponding to the Twist motif. Two representative modes for the two Twist variants are shown on the right. The mean positive and negative reads are shown in red and blue plots respectively (left Y-axis). The mean propeller twist values are shown in grey (right Y-axis). 107
- 4.12 In Sox2 dataset, ChExMix finds 2 motifs explaining 75% of the data, resembling the Sox2 and Oct4::Sox2 respectively, while in the Oct4 dataset, it finds a single motif that resembles a weak Oct4::Sox2 motif with low intensity reads explaining 76% of the data. In case of the FoxA1 dataset, ChExMix finds a single FoxA1 motif and an additional subtype characterized only by reads. In the Twist dataset, ChExMix finds a single weak E-box motif. 109
- 4.13 ExoDiversity was run with three different values of read widths on the mixed dataset. The result of using $w_k^{+/-} = 10$ is not very different from using a value of 5 (Fig: 4.3). However, for values of 50 and 100, the motifs become weaker. The average read plots are computed over a 200 bp neighbourhood around the motif: the ticks correspond to the motif positions shown below the plot in each case. 111

- 5.1 (A) DNA regions reported by the experiment are given as input to CISDIVERSITY. In this simulation, $n = 1000$ regions are a mixture of 3 types of regions. Each region is one of $r = 3$ regulatory modules. Each module can be represented in terms of probability of occurrence of $m = 5$ motifs. (B) CISDIVERSITY is run with upper bounds of $r \leq 10$ and $m \leq 20$ and it learns the planted structure of the dataset. The output has 3 components. First is the set of motifs that are learned, second, below is the $r \times m$ Bernoulli distributions describing the learned modules and the third is an image matrix of the data, where each DNA sequence is a row and the sites corresponding to each motif are represented in the column. If a site is absent, then those cells in the column are shown in black. CISDIVERSITY recovers the 5 motifs (motifs 1 and 3 are reverse complements of the planted motifs) and 3 modules to a great extent. The slight variability in the number of sites and sequences in each module is because of the stochastic nature of both the PWMs and the learning algorithm 123
- 5.2 Performance of CISDIVERSITY on 320 simulated datasets (A) Recovery of modules: Low values of β results in modules with more extreme (close to 0 or 1) probability distribution of motifs. This is where CISDIVERSITY performs better in recovering the planted modules. For $\beta = 0$ the performance with respect to recovery of modules is similar to that of a random clustering approach. (B) Recovery of motifs: Precision, recall and F-score of recovered motifs across the 320 datasets for 3 different programs. (C) Time taken: All programs were run on a single core. 126
- 5.3 When there is a single module ($r = 1$) or the modules are less separable ($\beta = 10$), the performance of CISDIVERSITY in terms of recovering the planted motifs is comparable to that of MEME or HOMER. 127

- 5.4 Promoter modules in fly (A) CISDIVERSITY identifies 24 motifs in fly promoters, but only 9 motifs that contribute to at least 20% of some modules are shown here. Core motifs' names are according to Ohler et al. (2002) [Ohler et al., 2002]. (B) 8 modules are learned. The fraction of sequences in each module that are focused promoters, i.e have a narrow peak are shown in green. (C) 3 representative modules are shown as sequence heat maps. They are sorted based on the position of the most frequently occurring motif. Module 1 contains DRE, which is visible owing to the manner in which the sequences were ordered but, Core 7 which is present in $> 40\%$ of the sequences is not visible because it has no clear positional relationship with DRE or the TSS. A weak signal corresponding to the AT-rich motif downstream from the TSS is visible. In contrast, module 3 displays a clear relationship between Core 1 and Core 6 motifs: Core 6 is present about 20 bp upstream of Core 1 and especially prominent when Core 1 is close to the TSS. Module 6 is largely composed of INR + DPE, but also contains TATA box especially when INR + DPE is a few bases downstream from the TSS. (D) Each motif has a distinct distribution about the TSS. 129
- 5.5 CISDIVERSITY identifies 25 motifs in pooled human promoter data, with motifs 21-25 contributing to one module with TSSs of 37 zinc finger genes. All the other motifs are strand invariant, except for motif 2 (TATA) and motif 6 (unknown). (B) Each motif here too has a distinct distribution about the TSS. 131
- 5.6 (A)-(C) CISDIVERSITY identifies 17 motifs and 13 modules in the fly CTCF data. Motif 1 shown in red box is the canonical fly CTCF motif. 132
- 5.7 (A)-(C) CISDIVERSITY identifies only 6 motifs and 6 modules in the human CTCF data. Motif 1 in the red box matches the canonical vertebrate CTCF motif. Motifs 2-5 resemble the vertebrate CTCF motif but differ at one of three parts, denoted with dotted lines. These motifs are shown with 10 bp flanking regions to ensure that they are genuine variants of the motif. 134

- 5.8 Diverse signals discovered in GR bound ChIP-seq regions. (A) CIS-DIVERSITY identifies 6 modules and 5 motifs. (B) The average DNase hypersensitive signal and the input subtracted ChIP-seq signal in reads per million (RPM) of 5 TFs: GR, JUNB, CEBPA, EP300 and CTCF before and after treatment at the GR bound regions are shown for each module. 138
- 5.9 Results of CISDIVERSITY on EP300 ChIP-seq data in A549 cells 1 hour after being treated with dex. These motifs were near identical to those learned on GR ChIP-seq data in the same cell type, but the distribution of the modules is different. 139
- 5.10 Results of CISDIVERSITY on a larger 400 bp neighbourhood of GR-bound summits. The modules and the motifs are recorded based on the motifs that are also found in the 200 bp neighbourhood (red box), those that are more spread out along the region and have low-complexity (green box) and those that resemble motifs of TFs active in the lung tissue (orange box)tissue. The histogram depicts the distribution of the binding sites of each motif along the 400 bp region. 140
- 5.11 CISDIVERSITY run on putative enhancers in K562 (A) On distant eRNAs, overlap with ChIP-seq data is significant (hypergeometric $p < 10^{-4}$; shown in bold) in modules that contain the matching TF motif. Note that in some cases the overlap looks large, but does not show up as significant, because the hypergeometric test corrects for the sizes of the overlaps and the modules. (B) Similar motifs are found in EP300 ChIP-seq data. 142
- 5.12 CISDIVERSITY run on putative enhancers in K562 (A) In STARR-seq peaks, YY1 is additionally discovered. CTCF and NFYA/NFYB are not enriched. (B) Distant enhancer like sequences and (C) proximal enhancer like sequences deduced from chromatin signatures have fewer motif-like signatures. 143
- 5.13 Sequences that are common across three assays that measure enhancer activity in K562. 145

- 5.14 CISDIVERSITY run on open regions: (A) A total of 13 modules and 28 motifs were learned on the ATAC-seq regions which are open in at least 110% of the cell 2-4 hour after egg laying. Only the 19 motifs that contribute to at least a quarter of the sequences in some modules are shown here for clarity. Modules are reordered: the red and green modules are significantly enriched (hypergeometric $p < 10^{-4}$) with promoters and depleted of them, respectively. (B) Gray indicates that there are only a few differences in the fraction of cells that are open within each module. Orange indicates modules 2, 3, 7, 8, 4 and 6 are significantly more open in the cells 6-8 hours after egg laying, whereas modules 12, 1 and 9 are closing at that time point. 148
- 5.15 CISDIVERSITY run on open regions: (A) A total of 18 modules and 21 motifs were learned on the DNase-seq regions in H1 ESCs. Again, only the motifs appearing in a quarter of the sequences in some modules are shown here. The red and green modules are significantly enriched (hypergeometric $p < 10^{-4}$) with promoters and depleted of them, respectively. (B) Gray indicates promoter modules have a higher DNase signal in general, but there are variations among them. Blue indicates the fraction of each module (and total below) that is also open in trophoblast, mesenchymal and neuronal stem cells (all derived from H1 ESCs) and GM12878 and shows considerable variation in modules and cell types. 149
- 5.16 CISDIVERSITY run on accessible regions in mouse tissue: 13 modules are learned from the ATAC-seq regions from the brown fat tissue. They are sorted according to median entropy (most brown-fat specific on top). The 15 discovered motifs can be split into 2 sets: those that are enriched in tissue specific regions (with median entropy value < 4.0) and those that are constitutively active. Box-plots on the right indicate the relationship between tissue specificity, accessibility and proximity to TSS 151

- 5.17 **Left panel:** shows modules with the lowest median entropy for each tissue. Many tissues have multiple modules with median values lower than 4.0: all such modules are shown. Motifs that are present in at least 10% of the sequences in any module are shown. **Right panel:** shows the top 2 modules based on highest median RPM (reads per million) values per tissue. Median entropy, RPM, and distance to closest TSS are displayed on the right for each module. TOMTOM matches were used to assign putative TF identities to motifs and combined across the CISDIVERSITY individual runs. . . . 152
- 5.18 Distribution of posterior scores of models learned on the simulated datasets are shown in black (100 datasets containing non-repetitive human genomic regions with no planted motifs) and red (320 datasets with planted motifs and modules as per the Methods section) . . . 156
- 6.1 DIVERSITY server 161
- 6.2 DIVERSITY bestModel.html file 162

List of Tables

3.1	Output of DIVERSITY on ChIP-seq data from 10 TFs in K562 cell-line	68
6.1	Datasets on which CISDIVERSITY was run in this study. Processed data or peaks were used from the original study in each case.	170
7.1	ExoDiversity was run on the in silico mixed datasets by varying the proportion of sequences from both FoxA1 and CTCF ChIP-exo data.	177

Chapter 1

Introduction

1.1 Introduction

Higher multicellular organisms are made up of millions of cells which have diverse functions and structure. They can be grouped into different types based on the functions they perform, and the morphological traits shared by the cells in a same group. Although all the cells contain the same genomic content, cells belonging to different cell types have different phenotypic features as well as some specific biological processes occurring within them. These differences arise out of variations in the expression levels of the thousands of genes present in the genome. Differential gene expression makes the stem cells capable of differentiating into a variety of specialized cell types without losing any genetic information.

Gene expression is the process by which the information encoded in the genes is used to form functional elements in the cell, like RNA molecules, proteins. Regulation of the amount of the genetic products formed is carried out by activating or suppressing the expression of one or more groups of genes. This forms the cornerstone of cell differentiation, morphogenesis and is also responsible for the different biochemical processes in a cell. This regulation can take place at the many stages of gene expression, for example, during chromatin remodeling, transcription initiation, post transcriptional modifications, translation etc.

At the first stage of gene expression, genetic information is transferred from the DNA to the mRNA. The production of mRNA is guided by the process of transcriptional regulation. This is a highly dynamic and complex process which involves participation by many entities and encompass a number of steps like spe-

cific protein-DNA interactions to form different protein complexes, assembly of the transcription initiation machinery, and participation of different types of non-coding RNAs that control the production of the final protein Casamassimi and Ciccodicola [2019]. The rate of transcription and thus in turn the amount of mRNA produced is primarily controlled by the transcription factors (TFs) that bind to the DNA. The regions on the DNA at which these TFs bind are called regulatory regions. The combinatorial binding of the TFs at these regulatory regions along with their interactions with other regulatory elements determine the mechanism of transcriptional regulation of the target genes. Variations in these regulatory regions, even small ones, can lead to changes in gene expression. The transcription factors binding to that region may no longer bind with the same affinity, and this can alter the intended functionality of the target gene. In such cases, even though the genes are perfectly normal, their transcription gets affected, which subsequently leads to diseases and disorders. Mutations and epigenomic aberrations in the regulatory regions of the DNA have been found to be one of the main causes for development of cancer Weinhold et al. [2014]. These variations also give rise to neurological disorders, diabetes, along with many other diseases Levine and Tjian [2003]. Thus, identifying these regulatory regions and understanding the role of the transcription factors associated with them is crucial in decoding the process of transcriptional regulation. These regulatory regions are mostly located in the non-coding parts of the DNA and are a lot more in number as compared to the genes. Also, unlike the coding regions or the genes, which have more defined structure, regulatory regions have variable structures and more sequence variation even at the transcription factor binding sites. They may be located thousands of base pairs away from the target gene, and thus even after being identified, it is difficult to confirm which are the genes a particular regulatory region is affecting. With the non-coding part of the DNA forming the major part of the genome (98% of the human genome), it is a challenging job to identify and annotate these regulatory regions Sheffield and Furey [2012].

Due to the technological improvements in sequencing and computation it is now possible to investigate the entire genome for particular biochemical activities and this has been a major step towards identifying and characterizing the regulatory regions based on their functions.

Next generation sequencing (NGS) technologies have rapidly evolved, allowing

us to sequence DNA and RNA with a high throughput and at low cost, producing large volumes of data from different experiments. Analyzing and interpreting this data is essential to develop an effective understanding of how different regulatory regions control gene expression. Different consortia like ENCyclopedia of DNA Elements (ENCODE), Genome Reference Consortium have put together datasets from various high throughput experiments along with maintaining and updating the reference genomes of a wide range of organisms. This data is made available to the scientific community to be analyzed using diverse techniques to form a comprehensive understanding from different perspectives.

In this chapter, we will describe the various stages of gene regulation and look into some of the details of the first step of transcriptional regulation. Then we will review the types of regulatory regions and their functions. This will be followed by understanding transcriptional misregulation and appreciating its importance by discussing some of the examples of the cases of misregulation. After that, we will describe the high throughput sequencing experiments that are used to identify and characterize the regulatory regions and the processes involved in the analyses of the experimental data.

1.2 Regulation of gene expression

Higher eukaryotic organisms have a large number of genes that code for an even greater number of proteins. The human genome is estimated to contain around 20,000 genes across the 23 pairs of chromosomes Salzberg [2018]. However, not all genes are active in any given cell at any point of time. The cell controls which genes are turned on and at what time, depending on the intended function. For example, the eye color in human is mostly determined by proteins formed from two adjacent genes *HERC2* and *OCA2* in chromosome 15 while a different set of genes, called *HOX* genes, regulate the formation of bones in developing limbs.

Gene expression comprises several steps, and the cell can regulate any of these to modulate the expression levels.

- i. **Transcriptional control:** The process of transcription copies the genetic information from the DNA to the primary RNA transcript. It is initiated by the binding of transcription factors to several regulatory regions across the DNA that ultimately help recruit the specific polymerase enzyme. The

interactions of the transcription factors with different regulatory regions determine under what conditions transcription should proceed and to what levels.

- ii. **Post transcriptional modifications:** It involves several processing steps by which a primary RNA transcript is modified into forming either a functional RNA molecule that can then move out of the nucleus to the cytoplasm to perform different functions or can be rendered non-functional. The steps involve capping, splicing, addition of a poly-A tail at the 3' end and edits to the poly-A tail; all of which can be controlled by the cell.
- iii. **RNA transport:** After the mRNA molecules undergo the above processing steps, they are associated with certain proteins that control the next phases in their life cycle. They move from the nucleus to the cytoplasm, close to the site where the final protein molecule is needed. The process of translation is repressed during their journey and is only activated when the mRNAs reach their final destination.
- iv. **Translation control:** This step involves the translation of mRNAs into proteins, and its regulation controls the final levels of protein synthesis. There are 2 general modes of control at the level of translation- the global level in which the translation of most mRNAs in a cell are regulated, and mRNA-specific control in which the translation of a specific group of mRNA is modulated. The regulatory sequences present in the mRNA like internal ribosome-entry sequences can initiate cap-independent translation, while sequences like upstream open reading frames (uORFs) mainly reduce translation from the main open reading frame. Also, the activities of the translation-initiation factors are mostly limited and changes in their activity results in changes in the global protein synthesis Gebauer and Hentze [2004].
- v. **mRNA degradation:** mRNAs get degraded in the cell cytoplasm, depending upon their stability. The poly-A tail of the mRNA is shortened by exonucleases which leads to the destabilization of the cap binding complex and the degradation machinery (RNase enzymes and protein complexes) then breaks down the mRNA molecule. The mRNAs which are actively translated undergo a slower degradation process than the inactive ones.

Thus we see that gene expression is a highly complex process, which can be regulated at every step and the cell can use a combination of these regulatory stages to determine its fate as well as its functions. Although the process of gene expression can be regulated at all the above stages, initiation of transcription is a key regulatory point for many genes.

1.3 Transcriptional regulation

Transcription of genes is strictly governed by the spatial and temporal organization of the chromatin (which is composed of the associated DNA and the histone proteins) and the transcription factors binding to the regulatory regions in the DNA. Transcription factors will not be able to bind to the DNA by being merely present in the cell, as their target sequences may be inaccessible due to the tightly wound structure of the chromatin. The chromatin has to be in an open state, facilitating the binding of the transcription factors to initiate transcription of a gene. Here, we outline the roles of the chromatin and the transcription factors in the process of transcription.

Chromatin accessibility: Chromatin is generally packaged as either highly condensed heterochromatin or the more open euchromatin. Heterochromatin renders the DNA inaccessible to transcription factor binding, while euchromatin consists of DNA which is accessible to the transcription machinery. The condensed form of the chromatin contains DNA wrapped around a core of an octamer of histone proteins, forming nucleosomes. The compact DNA is opened up by chromatin remodelers to change the chromatin structure dynamically. This is achieved by the remodeling or sliding of nucleosomes, by altering the histone-DNA contacts or by making histone modifications Tyagi et al. [2016]. It can bring about the formation of chromatin loops, which enables contacts between distal regulatory elements that help to govern the finer regulation of the transcription process.

Transcription factors Transcription and its regulation are driven by the binding of specific protein factors called transcription factors (TFs) to the regulatory regions of the genome. However, binding to the DNA alone is not sufficient for a transcription factor to affect transcription. It has to interact with other proteins called cofactors and in some cases also with the RNA polymerase in order to control the activation as well as rate of transcription. Transcription factors are

themselves regulated at many different levels that ranges from their synthesis to regulating their activity through signal transduction pathways as well as the interplay between various cofactors. Certain TFs are required for some of the common functionalities across all cell types and are thus ubiquitous, while other TFs are cell type specific.

Transcription factors contain DNA-binding domains that they use to bind to the specific regulatory regions like promoters and enhancers. The TFs that bind to the promoter regions near the transcription start site can act as activators and help form the transcription initiation complex, or they can act as repressors by preventing the binding of the RNA polymerase enzyme. TFs binding to the enhancer regions can either stimulate or repress the transcription of the related gene. These enhancer sequences can occur within *cis*-regulatory regions as well as thousands of base pairs away from the target gene, and mostly harbor more than a single TF-binding site. TFs participate in co-operative binding with other TFs and also in forming protein complexes wherein they might not directly bind the DNA but bring about interactions between promoter and enhancer regions through the other DNA bound TFs.

Thus, it is important for the transcription factors to recognize the correct DNA sequences and bind to that at the right point of time to perform their precise role in regulating gene expression. Variations in the sequence structure, either due to unavailability of the binding sites due to epigenetic reasons or because of mutations in the nucleotide sequence, can lead to misregulation. While the mutations in the protein-coding regions of the genes have been extensively studied and associated with human genetic diseases, interpreting the functional importance of the mutations in the non-coding regions of the genome have been challenging. The extensive non-coding regions contain the binding sites of several transcription factors, and changes sometimes as small as a single base pair mutation can lead to developmental disorders, neurological disorders and even cancer. We will now discuss the different types of regulatory regions that are present in the non-coding part of the genome and their roles in regulating transcription.

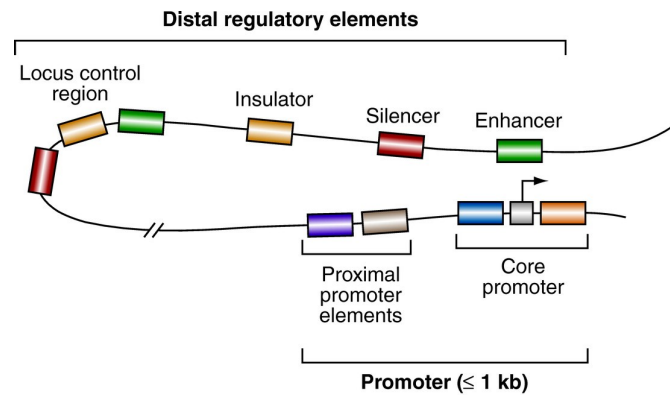


Figure 1.1: Cis-regulatory elements (Reproduced from Maston et al. *Annu. Rev. Genomics Hum. Genet.* 2006)

1.4 Cis-regulatory elements

Transcriptional regulatory units Fig. 1.1 in eukaryotic cells can be classified as core promoter elements, proximal promoter elements, enhancers, silencers, insulators and locus control regions Maston et al. [2006].

Core promoter The core promoter is a region located at the start of a gene and serves as a binding site for regulatory proteins which facilitates the organization and formation of the transcription pre-initiation complex. The core promoters can be broadly divided into two classes: those that have a single transcription start site (TSS) or a cluster of TSSs over a narrow region and those that have a broad range of potential TSSs over a 50-100 bp region Carninci et al. [2006], which often have an overrepresentation of CpG islands.

Proximal promoter It consists of the region immediately upstream (upto a few hundred base pairs) from the core promoter and can contain multiple binding sites for activator proteins. They also serve as tethering elements for active distant enhancers that bring about their interactions with the core promoter Su et al. [1991].

Enhancers They are regions on the genome that increase the transcription of a gene containing a promoter by up regulating the rate of formation and binding of the pre-initiation complex to the core promoter. Enhancers contain closely

grouped clusters of transcription factor binding sites, which act in a modular fashion such that a single promoter region can be acted upon by distinct enhancer elements under different conditions. These enhancer regions can be located up or downstream of the TSS, as far as 10 kb or 100kb away Consortium et al. [2007] and are brought into the proximity of the promoters by of DNA looping.

Silencers Silencers are sequence-specific elements that cause a down regulation in gene expression and serve as binding sites for negative transcription factors called repressors. They may be located far from the target gene, or in its intron or in its 3' untranslated region. Silencers can act over short range (within ≈ 100 bp) or have long range interactions via different cofactors Courey and Jia [2001].

Insulators Insulators, also called boundary elements, block the transcriptional activity of neighbouring genes. They have two main properties: enhancer blocking activity whereby they can block enhancer-promoter interaction and heterochromatin-barrier activity in which they prevent the spread of heterochromatin. Their functioning is position dependent but orientation independent.

Locus control regions They are groups of regulatory elements that are involved in regulating an entire locus or gene clusters Li et al. [2002]. They are composed of multiple *cis*-acting elements including enhancers, silencers, insulators and chromosome scaffold-attachment regions. Each of these components have their own effects on gene expression, and their collective activity defines the functionality of the locus control regions (LCR). They can regulate gene expression in a position-independent manner from a large distance and are characterized by strong, specific enhancer activity.

Disruption of the normal functioning of the genes can arise because of different types of events happening at these regulatory regions. Structural variations like chromosomal rearrangements can lead to an active enhancer being redirected to regulate genes that are not their original targets, resulting in activation of oncogenes. Enrichment of single nucleotide polymorphisms (SNPs) in the enhancer regions are also hallmarks of many diseases including diabetes, immune diseases among others Claringbould and Zaugg [2021]. Sequence variations can be at a single nucleotide level, or it may even span several bases and can occur at indi-

vidual sites of promoters or enhancers or in a group of regulatory regions [Cooper 1992].

1.5 Transcriptional misregulation

Mutations or sequence modifications that disrupt the functions of the regulatory elements lead to gene dysregulation, which in turn causes an imbalance in the gene products. This can give rise to abnormalities in phenotype along with different diseases like cancer, neurological disorders, developmental syndromes, diabetes, cardiovascular disease among others [Lee and Young 2013]. The DNA changes that can occur in these functional elements making them unfavourable involve translocations, deletions, duplications, insertions and point mutations. Here we list some of the examples of such misregulation.

The congenital abnormality preaxial polydactyly or extra digits in humans is an example of misregulation due to a point mutation in the regulatory region of the responsible gene. The expression of signalling molecule Sonic hedgehog (SHH) in developing limb bud is affected by changes in the regulatory region called ZRS of *SHH* gene, which is located 1 Mb upstream in another gene cluster. A single nucleotide variation in the ZRS region results in the loss of expression symmetry, which affects the development of digits in the limbs [Lettice et al. 2003].

Mutations in the proximal promoter of the *Gp1b β* gene causes reduced GATA-1 binding, which in turn reduces its gene expression. This leads to a disorder in blood clotting called the Bernard-Soulier Syndrome. Another case is that of mutations in the LCR of the α -globin or β -globin genes, which are responsible for production of hemoglobin in the body. Deletions in the DNA upstream of these genes lead to diseases α -thalassemia or β -thalassemia.

One example of neurological disorders is the increased susceptibility to Alzheimer's disease (AD) because of misregulation of the genes associated with the production and clearance of amyloid-B peptide in the brain. The expression levels of the AD genes must be very tightly controlled to keep a balanced amount of amyloid-B peptide in the brain under normal physiological conditions. Misregulation of these AD genes have been found to increase the risk of early onset of Alzheimer's disease or accelerate the disease progression [Chen et al. 2013].

The variations in the regulatory regions associated with diseases are most often

found in regions that are activated in a cell type specific manner, and the diseases are corresponding to that cell type or tissue. In order to understand which of these regulatory regions are active under different conditions in different cell types, we need to study which components of these regulatory regions are engaged in the regulational activity. One way to investigate this is by examining the transcription factor binding sites present in these regions and the mechanisms involved in the protein-DNA and the protein-protein interactions that ultimately affect the regulation of transcription. Determining the binding specificities of different transcription factors is a complex problem as it is influenced by which structural class the transcription factor protein belongs to, the length of the DNA sequence the TF binds to, whether the TF is involved in co-operative binding with other TFs at the binding sites. Here we describe a commonly used representation of the DNA sequence motif called the Position Weight Matrix (PWM) or Position Specific Scoring Matrix (PSSM) which capture the TF binding specificity to a certain extent.

1.6 Motif model

TFs generally bind to short DNA sequences, which are 10 to 20 bp in length. A single TF can bind to a set of DNA sequences that are similar but not exactly the same, with different binding affinities. These sequences with a similar repetitive pattern constitute the TF binding motif. Some of the positions in the motif can have variability in the nucleotide bases present there, while other positions might have a fixed base. Several models have been used to represent the binding specificity at each position in the motif. We use the standard model of PSSM, with the assumption that the bases present at each position in the motif contribute independently to the binding profile of the TF. The PSSM for a transcription factor motif is built from all the TF binding sequences where every column contains either the frequencies of the bases A,C,G,T or their probabilities at each position within the binding site. Thus, each column in the PSSM represents an independent categorical distribution over the 4 bases present at that position [Ambrosini et al. [2020], Boeva [2016]]. The PSSM matrices for TF motifs are usually generated by aligning the putative TF binding sites obtained from protein binding microarrays (PBMs) or from high throughput assays for genome wide mapping of TF binding

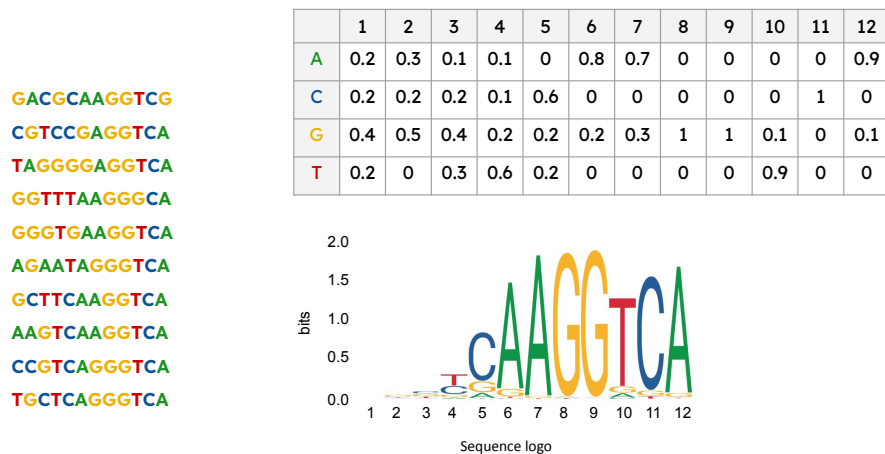


Figure 1.2: Motif model

sites like chromatin immunoprecipitation (ChIP)-based experiments. By systematically assimilating the data from different experiments the PWMs of thousands of TFs are now available across several motif databases such as JASPAR Khan et al. [2018] HOCOMOCO Kulakovskiy et al. [2018], TRANSFAC Matys et al. [2006], UniPROBE Hume et al. [2015]. Using the probabilities of the bases at each position within the binding site, we can score other nucleotide sequences to obtain the probability of the sequence being a binding site. This method can also be used to scan sequences with PWMs of known motifs to detect possible binding sites of a TF of interest. A graphical visualization of the TF motifs is the sequence logo Schneider and Stephens [1990]. The height of the stack of bases in each position of the logo is given in terms of the information content at that position in number of bits. If the length of the motif is w then the height of the stack at position i is given by $R_i = \text{maximum entropy} - \text{entropy at position } i (H_i)$. The height of each base at position i in the logo is proportional to the probability of that base

at position i .

$$R_i = \log_2(4) - H_i \quad \text{for } i = 1, \dots, w$$

$$H_i = - \sum_{a \in \{A, C, G, T\}} p_{a,i} \times \log_2(p_{a,i}) \quad p_{a,i} \text{ is probability of base } a \text{ at position } i$$

In the following section, we will discuss various high throughput sequencing methods to identify regulatory regions and the downstream processing of the data obtained from such experiments. To understand the functions of these regulatory regions, we examine the transcription factor binding sites present in them. These binding sites are modelled as DNA sequence motifs, and we use the approach of *de novo* motif discovery to learn the binding specificity of the TF at each of the binding sites.

1.7 High throughput sequencing experiments

With advancements in sequencing technologies, many high throughput sequencing-based experiments have been developed to identify different regulatory regions across the genome based on their distinct characteristics. Some of these experiments map entire open regions in the chromatin or regions that can be enhancers or promoters, while other experiments are focused on detecting the binding locations of a particular target protein. We would be going over a few specific experiments of these types in details next.

1.7.1 Experiments to identify TF binding sites: Chromatin immunoprecipitation-based

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq)

One of the most commonly used techniques to investigate the transcription factor binding sites in vivo, is chromatin immunoprecipitation Schmidt et al. [2009], Park [2009]. It involves cross linking the DNA binding proteins to the DNA by treating the cells with formaldehyde which is followed by cell lysis and shearing of the chromatin by sonication. The sheared chromatin is then immunoprecipitated using an antibody specific to the target protein, which could be a transcription factor

or a histone modification. Histone modifications are post translational modifications that include acetylation, methylation or phosphorylation of histone proteins. They are known to play important roles in transcriptional regulation by recruiting chromatin remodelling enzymes and other proteins with specific enzymatic activities. Multiple histone acetylations are known to be enriched at enhancer elements that facilitate TF access. The crosslinking in the precipitated fragments are then reversed, and the DNA is purified. After ligating known primers to the DNA, these fragments are amplified using Polymerase Chain Reaction (PCR). The PCR amplified library is sequenced as reads of length about 36 to 100 bp using high-throughput sequencing. The reads are then mapped to the genome using alignment tools Langmead et al. [2009], Li and Durbin [2009]. It has been observed that a large number of these reads form stacks or peaks near the binding site of the TF or the histone modification on both the forward and reverse strands of the DNA. Many methods have been developed to identify these peak regions (MACS2, GEM) Zhang et al. [2008], Guo et al. [2012] which are typically 200 to 1000 bp in length. However, the actual protein binding sites are shorter subsequences of length ≈ 20 bp, which are present somewhere within each peak region reported by the experiment. The next step in analyzing these peak regions involve finding the TF binding sequence components or motifs in them. This is done in broadly two different ways: de novo motif discovery from the DNA sequences corresponding to the peak regions; and scanning and ranking these sequences for the presence of known DNA motifs. Popular motif discovery methods look for a common sequence signature or an overrepresented motif within these DNA sequences that can provide the locations of the binding sites of the target TF. ChIP-seq has emerged as a robust and improved method with a higher resolution than its array-based predecessor, ChIP-chip, and has found applications in profiling different other features along the genome. It is used in profiling nucleosome level features like nucleosome positioning Schones et al. [2008] and also mapping histone variants Wang et al. [2008].

Although ChIP-seq has improved resolution over the previous methods, the variable sized DNA fragments generated during sonication still give a broad region as representative of the genomic TF binding site. Also, since it does not precisely delimit the binding sites, it is unable to resolve positional information between different motifs within binding sites in cases where the TF binds as a dimer or

participate in cooperative binding with another TF. To ameliorate the problems related to resolution and to reduce the noise in the experiment due to unbound DNA, modifications were made to the original ChIP-seq protocol. It involved the usage of lambda exonuclease enzyme to identify the genome wide set of TF binding sites at near single nucleotide resolution.

Exonuclease-based ChIP experiments

ChIP-exo and ChIP-nexus experiments are modifications of the ChIP-seq protocol that use the lambda exonuclease enzyme to improve the mapping resolution of binding sites. The first few steps are common to ChIP-seq procedure where DNA is crosslinked with associated proteins *in vivo*, chromatin is isolated from nuclei and subjected to sonication. Then, using an antibody against a specific protein, all the protein-DNA complexes with the target protein are immunoprecipitated. ChIP DNA is purified and is subjected to lambda exonuclease digestion. The enzyme degrades the DNA from 5' to 3' end on both the strands, until it encounters the bound protein. Free DNA is largely eliminated by exonuclease treatment. Then protein-DNA crosslinks are reversed and the extracted DNA is used for library preparation. The protocol was developed by Rhee et al. for use with Applied Biosystems SOLiD sequencing instruments Rhee and Pugh [2012]. ChIP-exo was adapted to the Illumina platform by Serandour et al. and they applied it to human cancer cell line experiments to reveal insights into the FoxA1-DNA interface in breast cancer cells and also in mouse liver Serandour et al. [2013]. In spite of having marked advantages over ChIP-seq like higher resolution and less noise, ChIP-exo had considerable technical challenges in its implementation. The extra wash and digestion steps decrease the quantity of DNA that can be recovered in comparison to conventional ChIP-seq experiments, which is crucial for the quality of the ChIP library. Also, the DNA fragments have to go through two ineffective ligation steps during library preparation to get adapters attached on both ends. Low starting DNA concentration frequently results in excessive amplification of artifacts during PCR leading to a lot of noise in the data. A more robust ChIP-exo protocol called ChIP-nexus (ChIP experiments with nucleotide resolution through exonuclease, unique barcode and single ligation) He et al. [2015] was designed for the Illumina platform by Zeitlinger et al. In this the standard ChIP-exo protocol is combined with the library preparation protocol from the iCLIP method for

mapping RNA-protein interactions. A unique randomized barcode was added to the adapter which is ligated to one end of the purified DNA and then the fragment is circularized, bringing the primer to the digested end. The libraries are then sequenced on Illumina platform with single-end sequencing. Systematic efforts to simplify the basic ChIP-exo protocol have been made that included the reduction in the number of enzymatic steps and alternative strategies for adapter ligation. Multiple alternative versions with their own advantages and limitations have been presented Rossi et al. [2018], which may be suitable for particular applications.

In these exonuclease-based ChIP experiments, the exonuclease enzyme degrades the free chromatin to the point where the crosslinking of the protein-DNA prevents it from going further. When these DNA fragments are sequenced and mapped to the genome they form sharp narrow peaks very near to the TF binding site on both the strands of the DNA. Aligning these peaks reveal the TF-DNA contact sites. This modification to the ChIP protocol not only increases the resolution of identifying TF binding sites, but it also minimizes the contamination caused by the presence of free DNA and can even capture the TF-DNA interactions that are crosslinked with low efficiency.

In the next section, we will discuss some of the high-throughput experiments to detect the complete set of open chromatin regions across the genome.

1.7.2 Experiments to detect open chromatin regions

Transcription factors gather on the DNA at the promoters and enhancer regions to form the transcription machinery. In this process, they destabilize and dislodge the nucleosomes present in these regions and open up the chromatin. Thus, open chromatin or accessible regions in the genome are considered as potential candidates for regulatory regions. Once the chromatin loses its condensed structure, then the open regions get susceptible to cleavage by DNase I enzyme. The functional sequences associated with these DNase I hypersensitive sites (HS) can be enhancers, silencers, promoters or locus control regions. They are also regions associated with histone modifications, with distinguishable patterns at promoters and distal sites (enhancers). Accurately finding and annotating these DNase I HS across the genome have become important in comprehending how gene expression is regulated in different cell types or in responses to stimuli under different circumstances.

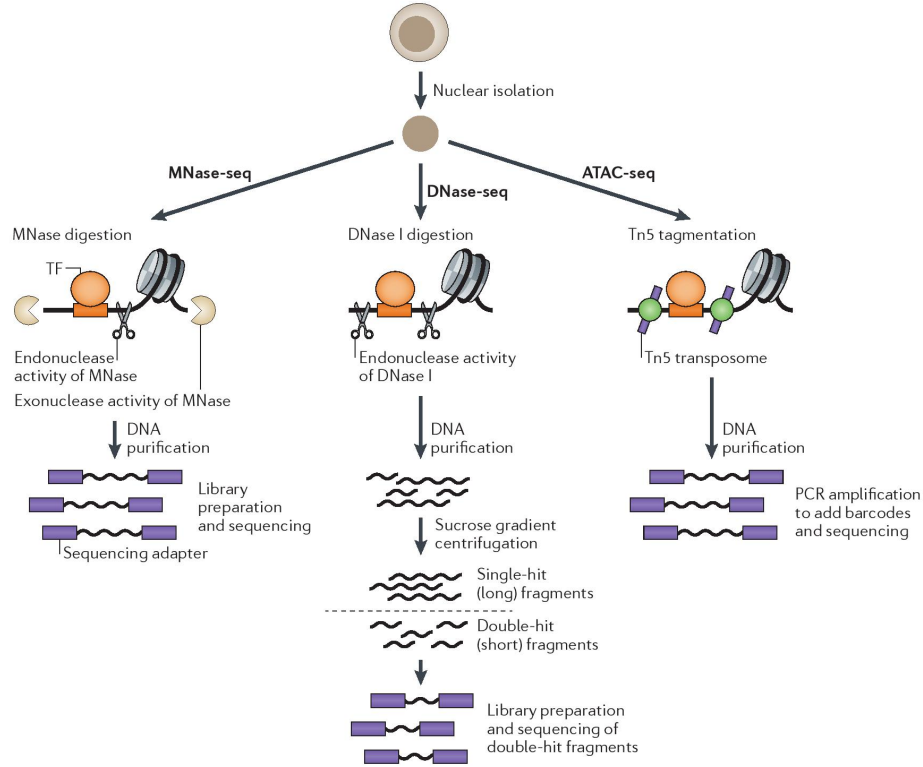


Figure 1.3: Enzyme-based experiments (Reproduced from Zentner et al. *Nature Reviews Genetics*. 2014)

DNase I hypersensitive sites sequencing (DNase-seq)

Traditionally DNase I HS were identified using Southern blot Wu [1980] which could only process a few regions of the genome at a time but with the advancements in high throughput sequencing, an improved protocol DNase-seq can now identify DNase I HS across the whole genome upto a single base pair resolution at the site of digestion Song and Crawford [2010]. In DNase-seq (Fig. 1.3), the nuclei are first digested by optimal concentrations of DNase I enzyme that preferentially cuts at the open chromatin sites. The DNase-digested ends are enriched and sequenced using next generation sequencing. However, DNase-seq does not provide any information about the function of the reported sequences. It needs to be followed up by experiments to precisely determine the function and the events

associated with those regions. Downstream data processing involves aligning and mapping of the sequence reads to the genome, followed by peak finding that would indicate regions where the most number of cuts have happened or where the DNase I HS are located.

Assay for Transposase Accessible Chromatin with high throughput sequencing (ATAC-seq)

ATAC-seq (Fig. 1.3) was developed as an alternative to the existing techniques to study open chromatin Buenrostro et al. [2015]. The existing methods were having limitations in many aspects, like the need for a large number of cells as starting material (1-50 million). They were time-consuming and had lengthy protocols. Also, they lacked the ability to assess open chromatin, nucleosome positioning and transcription factor occupancy in one assay. The ATAC-seq methodology uses the hyperactive Tn5 transposase that was already being used for preparing NGS libraries based on tagmentation technology. The transposase preferentially cleaves and inserts sequencing adaptors into the open regions of the DNA. The tagged DNA fragments are purified and PCR amplified, and then sequenced using high throughput sequencing techniques. After sequencing, approximately half the reads obtained are shorter than the canonical length of the DNA around a nucleosome and the rest half of the reads have length consistent to the approximate length of the DNA protected by a nucleosome. These different read concentrations provide information on both nucleosome-free regions (which may be occupied by transcription factors or epigenetic markers) and also positions of the nucleosomes.

These high throughput experiments report a set of genomic regions which share a common property. One way to analyze these regions is by identifying and understanding the roles of the transcription factors binding in them. These TF binding sites represented by motifs can be identified either by scanning these regions using known motifs from a database or by finding the motifs *de novo*. In case of using a motif database, it limits the possibility of presence of motifs to the ones present in the database. Also, not all the motifs are likely to be present in a particular set of genomic regions and scanning for them would not only take up resources but might give weak matches or false positives which would require further analysis to obtain some meaningful results. *De novo* motif discovery provides ways to finding the motifs from scratch and has the advantage of discovering novel motifs. The

algorithms can be built without any prior assumptions about the motifs or with provisions for using other sources of information to guide the algorithm in detecting the likely candidates for the motifs. However, the problem of motif discovery involves multiple challenges, beginning with a fair representation of the motif to identifying various motifs present in the data and lastly also the validation of the discovered motifs.

Traditional motif discovery methods focus on finding the most enriched motif in a dataset. In case of a ChIP experiment for a target protein, the overrepresented motif obtained by these algorithms is used to explain the binding specificity of the TF. However, the TF that has been profiled by the ChIP experiment could be binding directly to the DNA in which case its motif would be present in the DNA sequences, or it could be a part of a complex with many cofactors, without making direct DNA contact. In such a case, since the DNA bound to any complex containing the target protein gets immunoprecipitated in these experiments, we would get the motifs of those binding proteins as well, albeit in substantially smaller subsets of the ChIP sequences. Experiments that detect different regulatory regions in the chromatin report broad regions that contain a mixture of different TF binding sites.

We model this heterogeneity in these reported regions using the approach of mixture modelling, building model-based Bayesian frameworks that can characterize the data obtained from different high throughput experiments in terms of the TF motifs found in them.

Chapter 2

Mixture model for DNA regulatory regions

Here we will first look at the problem of motif discovery in a given set of DNA sequences and then define the motif model and its parameters. To estimate the parameters of the model we use one of the Markov Chain Monte Carlo methods: collapsed Gibbs sampling algorithm. We next show how Gibbs sampling has been used to learn the model parameters and select the best model describing our data.

2.1 Problem Statement

Assume we have n DNA regions $\mathbf{X} = \{X_1, X_2 \dots X_n\}$ with $X_{i,j} \in \{A, C, G, T\}$ all bound by a common TF. All these sequences will have the motif for the TF present within them at some location. Let w be the width of the TF binding site or the motif. The motif is modeled using a PSSM ϕ . $\phi_{a,b}$ denotes the probability of base b at position a within the motif, when $b \in \{A, C, G, T\}$ and $1 \leq a \leq w$. Each ϕ_a represents a categorical distribution over $\{A, C, G, T\}$.

The aim is to find the values for the matrix ϕ and the location of the motif within each sequence in $\mathbf{X}$.

We build the model M_1 for our data X containing a single motif. Let θ be the set of parameters for our model M_1 . $\theta = \{\mathbf{Z}, \phi, \phi_0\}$.

For our DNA sequences $\mathbf{X}$, let the length of sequence X_i be L_i for $1 \leq i \leq n$. The vector $\mathbf{Z}$ of length n denotes the starting positions of the motif in each sequence X_i . Thus, $Z_i = j$ would mean that the motif is present at location j in

the sequence X_i where $j \in \{1, 2, \dots, L_i - w + 1\}$ when width of the motif is w .

While the PSSM for the motif is denoted by ϕ , ϕ_0 is used to model the probabilities of the nucleotides $\{A, C, G, T\}$ that are not part of the motif, i.e., form the background. Here we model probabilities of the nucleotides in the background by using a second order Markov Model (which is learnt for every sequence).

The likelihood of a sequence X_i given the parameters θ of the model, can thus be computed as:

$$P(X_i | \mathbf{Z}, \phi, \phi_0) = P(X_{i,1}, X_{i,2} \dots X_{i,Z_i-1} | \phi_0) \times \prod_{a=1}^w \phi_{a, X_{i,Z_i+a-1}} \\ \times P(X_{i,Z_i+w} \dots X_{i,L_i} | \phi_0)$$

The likelihood of the full data will be:

$$P(\mathbf{X} | \mathbf{Z}, \phi, \phi_0) = \prod_{i=1}^n P(X_i | \mathbf{Z}, \phi, \phi_0) \\ = \prod_{i=1}^n \left(P(X_{i,1}, X_{i,2} \dots X_{i,Z_i-1} | \phi_0) \times \prod_{a=1}^w \phi_{a, X_{i,Z_i+a-1}} \right. \\ \left. \times P(X_{i,Z_i+w} \dots X_{i,L_i} | \phi_0) \right)$$

We observe that each sequence X_i has two parts: the motif region and the background. Let us denote the motif part of sequence X_i by X_i^M and the background by $X_i^{\overline{M}}$. Thus we have

$$P(X_i^M | \mathbf{Z}, \phi, \phi_0) = \prod_{a=1}^w \phi_{a, X_{i,Z_i+a-1}} \\ P(X_i^{\overline{M}} | \mathbf{Z}, \phi, \phi_0) = P(X_{i,1}, X_{i,2} \dots X_{i,Z_i-1} | \phi_0) \times P(X_{i,Z_i+w} \dots X_{i,L_i} | \phi_0) \quad (2.1)$$

The likelihood of X_i and the whole data can be written follows:

$$P(X_i | \mathbf{Z}, \phi, \phi_0) = P(X_i^M | \mathbf{Z}, \phi, \phi_0) \times P(X_i^{\overline{M}} | \mathbf{Z}, \phi, \phi_0) \quad (2.2)$$

$$P(\mathbf{X} | \mathbf{Z}, \phi, \phi_0) = \prod_{i=1}^n P(X_i^M | \mathbf{Z}, \phi, \phi_0) \times P(X_i^{\overline{M}} | \mathbf{Z}, \phi, \phi_0) \quad (2.3)$$

2.1.1 Objective function

Since our goal is to find the values of ϕ and $\mathbf{Z}$ that best capture the motif present in the sequences, our objective will be to find the values of ϕ and $\mathbf{Z}$ that will maximize the posterior distribution conditional on the data.

The objective function is thus:

$$\begin{aligned}\arg \max_{\phi, \mathbf{Z}} P(\phi, \mathbf{Z} \mid \mathbf{X}, \phi_0) &= \arg \max_{\phi, \mathbf{Z}} P(\mathbf{X} \mid \phi, \mathbf{Z}, \phi_0) P(\phi, \mathbf{Z}, \phi_0) \\ &= \arg \max_{\phi, \mathbf{Z}} P(\mathbf{X} \mid \phi, \mathbf{Z}, \phi_0) P(\phi) P(\mathbf{Z})\end{aligned}$$

considering independent priors over ϕ and $\mathbf{Z}$.

2.2 Gibbs sampling

To calculate the joint posterior distribution for our set of parameters, we use the Markov Chain Monte Carlo (MCMC)-based method, Gibbs sampling, to obtain an approximation of the posterior distribution by drawing samples from individual conditional distributions.

Gibbs sampler constructs a Markov Chain whose stationary distribution is the target posterior distribution. The Markov Chain is started with an initial set of random values for the variables. The sampling algorithm visits each variable iteratively and updates its value by sampling from its conditional distribution, given the current values of all the other variables. The idea behind the Gibbs sampler is to use the univariate conditional distributions where all random variables have a fixed assigned value except one. Such distributions are usually easier to sample from as compared to the complex joint distributions. Thus, if we have k random variables, we will sample each of them sequentially from k univariate condition distribution instead of sampling a single k -dimensional vector from the joint distribution in a single pass Carlo [2004].

Let us assume $\Theta = (\theta_1, \theta_2 \dots \theta_k)$ be a random variable containing k components. If $\theta_{[-i]}$ denotes all components of Θ except the i th one for $1 \leq i \leq k$, then let the conditional distributions $p(\theta_i \mid \theta_{[-i]})$ be known and easy to draw samples from. The algorithm for Gibbs sampling can be stated as follows:

1. Initialize $\theta_1^0, \theta_2^0 \dots \theta_k^0$ randomly
2. For $t = 1, 2 \dots N$
 - For $i = 1, 2 \dots k$

- Draw sample $\theta_i^t \sim p(\theta_i^t \mid \theta_{[-i]}^{(t-1)})$
i.e
 $\theta_i^t \sim p(\theta_i \mid \theta_1^{(t-1)}, \dots, \theta_{(i-1)}^{(t-1)}, \theta_{(i+1)}^{(t-1)}, \dots, \theta_k^{(t-1)})$

The Gibbs sampler generates a Markov chain of random variables and for a large value of N , it converges to the marginal distribution of Θ which is also the stationary distribution of the Markov chain. The initial values in the chain depend on the starting values $\theta_1^0, \theta_2^0 \dots \theta_k^0$, however this dependence wears off after a few iterations (burn-in period) and by the time the sampler converges it is independent of the starting values of the variables.

The Gibbs sampler can thus be effectively used to draw samples from the univariate conditionals and use them to approximate the joint distribution of all the variables Casella and George [1992].

2.3 Gibbs sampling applied to motif discovery

Given the model M_1 with its parameters as θ , we use Gibbs sampling to approximate the posterior distribution and find the values of the parameters θ corresponding to the maximum posterior probability. This gives us the optimal values of $\mathbf{Z}$ and ϕ that best capture the motif present in the sequences $\mathbf{X}$. However, if we directly apply Gibbs sampling to the motif discovery problem, we end up with the situation where we have to sample the values of each Z_i and the high dimensional ϕ . Instead, we can collapse the component ϕ by integrating it out and sampling $\mathbf{Z}$ based on the marginal distribution. The Collapsed Gibbs sampling strategy developed by Liu et al. Liu [1994] not only reduces the number of random variables to be sampled, it also improves the convergence time and makes the algorithm more efficient.

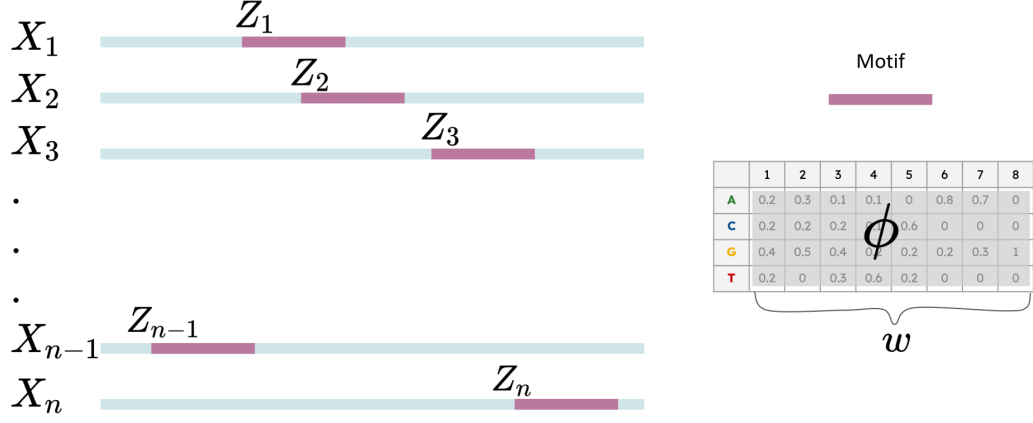


Figure 2.1: Sequences with motif

2.3.1 Collapsed Gibbs sampling

Calculating the posterior probability by collapsing ϕ :

$$\begin{aligned}
 P(\mathbf{Z} \mid \mathbf{X}, \phi_0) &\propto P(\mathbf{Z}, \mathbf{X} \mid \phi_0) \times P(\phi_0) \\
 &\propto P(\mathbf{Z}, \mathbf{X} \mid \phi_0) \\
 &= P(\mathbf{X} \mid \mathbf{Z}, \phi_0) \times P(\mathbf{Z}, \phi_0) \\
 &\propto P(\mathbf{X} \mid \mathbf{Z}, \phi_0) \times P(\mathbf{Z}) \\
 &= \int_{\phi} P(\mathbf{X}, \phi \mid \mathbf{Z}, \phi_0) \times P(\mathbf{Z}) d\phi \quad \text{Marginalizing over } \phi \\
 &= P(\mathbf{Z}) \times \int_{\phi} P(\mathbf{X} \mid \phi, \mathbf{Z}, \phi_0) \times P(\phi) d\phi \quad (2.4) \\
 &= P(\mathbf{Z}) \times \int_{\phi} P(\mathbf{X}^M \mid \phi, \mathbf{Z}) \times P(\mathbf{X}^{\overline{M}} \mid \mathbf{Z}, \phi_0) \times P(\phi) d\phi \quad \text{From (2.3)} \\
 &= P(\mathbf{Z}) \cdot P(\mathbf{X}^{\overline{M}} \mid \mathbf{Z}, \phi_0) \times \int_{\phi} P(\mathbf{X}^M \mid \phi, \mathbf{Z}) \times P(\phi) d\phi \\
 &= P(\mathbf{Z}) \cdot P(\mathbf{X}^{\overline{M}} \mid \mathbf{Z}, \phi_0) \times \int_{\phi} \prod_{i=1}^n P(X_i^M \mid \phi, \mathbf{Z}) \times P(\phi) d\phi
 \end{aligned}$$

The sequences in $\mathbf{X}^M$ consists of the motif instances X_{i, Z_i+a-1} for $a = 1, 2, \dots, w$

from all the sequences X_i for $i = 1, 2 \dots n$. For every motif instance in X_i , the nucleotides at each position a given by: X_{i,Z_i+a-1} follow categorical distribution ϕ_a for $a = 1, 2 \dots w$. For our likelihood with categorical distribution we assume a conjugate prior of Dirichlet distribution with prior hyperparameters as $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ corresponding to the categories $\{1, 2, 3, 4\}$.

$$\int_{\phi} \prod_{i=1}^n P(X_i^M \mid \phi, \mathbf{Z}) \times P(\phi) d\phi \quad (2.5)$$

$$= \int_{\phi_a} \prod_{i=1}^n \left(\prod_{a=1}^w P(X_{i,Z_i+a-1} \mid \phi_a, Z_i) \times P(\phi_a) d\phi_a \right)$$

$$= \prod_{a=1}^w \int_{\phi_a} \left(\prod_{i=1}^n P(X_{i,Z_i+a-1} \mid \phi_a, Z_i) \right) \times P(\phi_a) d\phi_a$$

$$= \prod_{a=1}^w \int_{\phi_a} \left[\left(\prod_{i=1}^n \text{cat-pmf}(X_{i,Z_i+a-1}, \phi_a) \right) \times \text{dir-pdf}(\phi_a, \alpha) \right] d\phi_a$$

$$= \prod_{a=1}^w \int_{\phi_a} \left[\left(\prod_{i=1}^n \prod_{b \in \{A,C,G,T\}} \phi_{a,b}^{[X_{i,Z_i+a-1}=b]} \right) \times \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \cdot \prod_{b \in \{A,C,G,T\}} (\phi_{a,b})^{\alpha_b-1} \right] d\phi_a$$

Here we use the notation of Iverson bracket $[y = b]$ to denote 1 when $y = b$ and 0 otherwise. Simplifying the integral on the RHS for any value of a

$$\begin{aligned}
& \int_{\phi_a} \left[\left(\prod_{i=1}^n \prod_{b \in \{A,C,G,T\}} \phi_{a,b}^{[X_{i,Z_i}=b]} \right) \times \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \cdot \prod_{b \in \{A,C,G,T\}} (\phi_{a,b})^{\alpha_b-1} \right] d\phi_a \\
&= \int_{\phi_a} \left[\left(\prod_{b \in \{A,C,G,T\}} \phi_{a,b}^{\sum_{i=1}^n [X_{i,Z_i}=b]} \right) \times \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \cdot \prod_{b \in \{A,C,G,T\}} (\phi_{a,b})^{\alpha_b-1} \right] d\phi_a \\
&= \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \int_{\phi_a} \left[\prod_{b \in \{A,C,G,T\}} \phi_{a,b}^{\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b - 1} \right] d\phi_a \\
&= \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \times \frac{\prod_{b \in \{A,C,G,T\}} \Gamma(\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b)}{\Gamma(\sum_{b \in \{A,C,G,T\}} (\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b))} \\
&\quad \times \int_{\phi_a} \left[\frac{\Gamma(\sum_{b \in \{A,C,G,T\}} (\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b))}{\prod_{b \in \{A,C,G,T\}} \Gamma(\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b)} \cdot \prod_{b \in \{A,C,G,T\}} \phi_{a,b}^{\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b - 1} \right] d\phi_a \\
&= \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \times \frac{\prod_{b \in \{A,C,G,T\}} \Gamma(\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b)}{\Gamma(\sum_{b \in \{A,C,G,T\}} (\sum_{i=1}^n [X_{i,Z_i}=b] + \alpha_b))} \times 1
\end{aligned}$$

$\sum_{i=1}^n [X_{i,Z_i}=b]$ is actually the count of category b in X_{i,Z_i} for $i = 1, 2 \dots n$ and can be denoted by $C_{a,b}$ for any any column a of the motif. Thus,

$$\begin{aligned}
& \int_{\phi_a} \left[\left(\prod_{i=1}^n \prod_{b \in \{A,C,G,T\}} \phi_{a,b}^{[X_{i,Z_i}=b]} \right) \times \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \cdot \prod_{b \in \{A,C,G,T\}} (\phi_{a,b})^{\alpha_b-1} \right] d\phi_a \\
&= \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha_b)}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha_b)} \times \frac{\prod_{b \in \{A,C,G,T\}} \Gamma(C_{a,b} + \alpha_b)}{\Gamma(\sum_{b \in \{A,C,G,T\}} (C_{a,b} + \alpha_b))}
\end{aligned}$$

Thus (2.5) becomes

$$\begin{aligned} & \int_{\phi} \prod_{i=1}^n P(X_i^M \mid \phi, \mathbf{Z}) \times P(\phi) d\phi \\ &= \prod_{a=1}^w \left[\frac{\Gamma(\sum_{b \in \{A, C, G, T\}} \alpha_b)}{\prod_{b \in \{A, C, G, T\}} \Gamma(\alpha_b)} \times \frac{\prod_{b \in \{A, C, G, T\}} \Gamma(C_{a,b} + \alpha_b)}{\Gamma(\sum_{b \in \{A, C, G, T\}} (C_{a,b} + \alpha_b))} \right] \end{aligned}$$

For the Gibbs sampler, we have to sample the value of Z_i . Let $\mathbf{Z}_{[-i]}$ denote the vector $\mathbf{Z}$ with all of its components except the i th one. Then the expression for sampling of Z_i is given as:

$$\begin{aligned} Z_i &\sim P(Z_i = j \mid \mathbf{Z}_{[-i]}, \mathbf{X}, \phi_0) \\ P(Z_i \mid \mathbf{X}, \mathbf{Z}_{[-i]}, \phi_0) &\propto P(\mathbf{X} \mid Z_i, \mathbf{Z}_{[-i]}, \phi_0) \cdot P(Z_i) \\ &= P(Z_i) \cdot \int P(\mathbf{X}, \phi \mid Z_i, \mathbf{Z}_{[-i]}, \phi_0) d\phi \\ &= P(Z_i) \cdot \int P(\mathbf{X} \mid \phi, Z_i, \mathbf{Z}_{[-i]}, \phi_0) \cdot P(\phi \mid Z_i, \mathbf{Z}_{[-i]}, \phi_0) d\phi \\ &= P(Z_i) \cdot \int P(\mathbf{X} \mid \phi, Z_i, \mathbf{Z}_{[-i]}, \phi_0) \cdot P(\phi \mid \mathbf{Z}_{[-i]}) d\phi \\ &= P(Z_i) \cdot \int P(X_i \mid \phi, Z_i, \phi_0) \cdot \left(P(\mathbf{X}_{[-i]} \mid \phi, \mathbf{Z}_{[-i]}, \phi_0) \cdot P(\phi \mid \mathbf{Z}_{[-i]}) \right) d\phi \end{aligned}$$

The first part of the integral $P(X_i \mid \phi, Z_i, \phi_0)$ is the likelihood of a single datapoint which can be resolved using (2.1) and (2.2).

The second part $P(\mathbf{X}_{[-i]} \mid \phi, \mathbf{Z}_{[-i]}, \phi_0) \cdot P(\phi \mid \mathbf{Z}_{[-i]})$ is the posterior for $\mathbf{Z}_{[-i]}$ (from (2.4)) and follows the Dirichlet distribution with parameters α'_a for $a = 1, 2 \dots w$ and $\alpha'_{a,b} = \sum_{\substack{p=1 \\ p \neq i}}^n [X_{p,Z_p+a-1} = b] + \alpha_b$.

Thus, if for some i , $Z_i = j$

$$\begin{aligned}
& P(Z_i = j \mid \mathbf{X}, \mathbf{Z}_{[-i]}, \phi_0) \\
& \propto P(Z_i) \cdot \int \prod_{a=1}^w \phi_{j, X_{i,j+a-1}} \times P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \text{Dir-pdf}(\phi_a, \alpha'_a) d\phi_a \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \int \prod_{a=1}^w \prod_{b \in \{A, C, G, T\}} \phi_{a,b}^{[X_{i,j+a-1}=b]} \cdot \text{Dir-pdf}(\phi_a, \alpha'_a) d\phi_a \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \cdot \prod_{a=1}^w \int \prod_{b \in \{A, C, G, T\}} \phi_{a,b}^{[X_{i,j+a-1}=b]} \cdot \text{Dir-pdf}(\phi_a, \alpha'_a) d\phi_a \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \\
& \quad \prod_{a=1}^w \int \prod_{b \in \{A, C, G, T\}} \phi_{a,b}^{[X_{i,j+a-1}=b]} \cdot \frac{\Gamma(\sum_{b \in \{A, C, G, T\}} \alpha'_{a,b})}{\prod_{b \in \{A, C, G, T\}} \Gamma(\alpha'_{a,b})} \prod_{b \in \{A, C, G, T\}} \phi_{a,b}^{\alpha'_{a,b}-1} d\phi_a \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \\
& \quad \prod_{a=1}^w \int \frac{\Gamma(\sum_{b \in \{A, C, G, T\}} \alpha'_{a,b})}{\prod_{b \in \{A, C, G, T\}} \Gamma(\alpha'_{a,b})} \prod_{b \in \{A, C, G, T\}} \phi_{a,b}^{[X_{i,j+a-1}=b] + \alpha'_{a,b} - 1} d\phi_a \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \\
& \quad \prod_{a=1}^w \left(\frac{\Gamma(\sum_{b \in \{A, C, G, T\}} \alpha'_{a,b})}{\prod_{b \in \{A, C, G, T\}} \Gamma(\alpha'_{a,b})} \right) \cdot \frac{\prod_{b \in \{A, C, G, T\}} \Gamma([X_{i,j+a-1} = b] + \alpha'_{a,b})}{\Gamma(\sum_{b \in \{A, C, G, T\}} [X_{i,j+a-1} = b] + \alpha'_{a,b})} \times \\
& \quad \int \frac{\Gamma(\sum_{b \in \{A, C, G, T\}} [X_{i,j+a-1} = b] + \alpha'_{a,b})}{\prod_{b \in \{A, C, G, T\}} \Gamma([X_{i,j+a-1} = b] + \alpha'_{a,b})} \cdot \prod_{b \in \{A, C, G, T\}} \phi_{a,b}^{[X_{i,j+a-1}=b] + \alpha'_{a,b} - 1} d\phi_a \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \\
& \quad \prod_{a=1}^w \left(\frac{\Gamma(\sum_{b \in \{A, C, G, T\}} \alpha'_{a,b})}{\prod_{b \in \{A, C, G, T\}} \Gamma(\alpha'_{a,b})} \right) \cdot \frac{\prod_{b \in \{A, C, G, T\}} \Gamma([X_{i,j+a-1} = b] + \alpha'_{a,b})}{\Gamma(\sum_{b \in \{A, C, G, T\}} [X_{i,j+a-1} = b] + \alpha'_{a,b})} \times 1 \\
& = P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \\
& \quad \prod_{a=1}^w \frac{\Gamma(\sum_{b \in \{A, C, G, T\}} \alpha'_{a,b})}{\prod_{b \in \{A, C, G, T\}} \Gamma(\alpha'_{a,b})} \cdot \frac{\prod_{b \in \{A, C, G, T\}} \Gamma([X_{i,j+a-1} = b] + \alpha'_{a,b})}{\Gamma(1 + \sum_{b \in \{A, C, G, T\}} \alpha'_{a,b})}
\end{aligned}$$

Using the fact that $\Gamma(x+1) = x\Gamma(x)$ we now simplify further

$$\begin{aligned}
& P(Z_i = j \mid \mathbf{X}, \mathbf{Z}_{[-i]}, \phi_0) \\
&= P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \\
&\quad \times \prod_{a=1}^w \frac{\Gamma(\sum_{b \in \{A,C,G,T\}} \alpha'_{a,b})}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha'_{a,b})} \cdot \frac{\prod_{b \in \{A,C,G,T\}} \Gamma([X_{i,j+a-1} = b] + \alpha'_{a,b})}{(\sum_{b \in \{A,C,G,T\}} \alpha'_{a,b}) \cdot \Gamma(\sum_{b \in \{A,C,G,T\}} \alpha'_{a,b})} \\
&= P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \prod_{a=1}^w \frac{\prod_{b \in \{A,C,G,T\}} \Gamma([X_{i,j+a-1} = b] + \alpha'_{a,b})}{\prod_{b \in \{A,C,G,T\}} \Gamma(\alpha'_{a,b}) \times \sum_{b \in \{A,C,G,T\}} \alpha'_{a,b}} \\
&= P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \prod_{a=1}^w \frac{\Gamma(1 + \alpha'_{a,X_{i,j+a-1}})}{\Gamma(\alpha'_{a,X_{i,j+a-1}}) \cdot \sum_{b \in \{A,C,G,T\}} \alpha'_{a,b}} \\
&= P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \prod_{a=1}^w \frac{\alpha'_{a,X_{i,j+a-1}}}{\sum_{b \in \{A,C,G,T\}} \alpha'_{a,b}} \\
&= P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \prod_{a=1}^w \frac{\sum_{\substack{p=1 \\ p \neq i}}^n [X_{p,Z_p+a-1} = X_{i,j+a-1}] + \alpha_{X_{i,j+a-1}}}{\sum_{b \in \{A,C,G,T\}} \left(\sum_{\substack{p=1 \\ p \neq i}}^n [X_{p,Z_p+a-1} = b] + \alpha_b \right)} \\
&= P(Z_i) \cdot P(X_i^{\overline{M}} \mid Z_i, \phi, \phi_0) \times \prod_{a=1}^w \hat{\phi}_{a,X_{i,j+a-1}}
\end{aligned}$$

Here $\hat{\phi}$ is the posterior mean of ϕ conditional on the current values of all the parameters $Z_{[-i]}$ and $X_{[-i]}$.

In this way we iterate over every $i = 1, 2 \dots n$ to sample the value for Z_i and update the other parameter values. This sequential updation of the parameter values based on their values until the previous iteration creates the Markov Chain and the Gibbs sampler ultimately converges to the stationary distribution of this Markov chain.

2.4 Mixture model

We extend this concept of collapsed Gibbs sampling to build mixture models that can partition the set of DNA sequences into disjoint clusters based on certain properties like the identity of the motif or the distribution over a set of motifs present in the sequences.

Our method DIVERSITY [Mitra et al., 2018] is a framework that uses collapsed Gibbs sampling to perform motif discovery in a set of DNA sequences. Instead of searching for a single motif it considers the sequences as a mixture of regions where subsets of the sequences might contain a different motif. DIVERSITY not only learns the different motifs in the sequences but also partitions the sequences into disjoint subsets based on the motifs present in them.

In our second method ExoDiversity [Biswas and Narlikar, 2021a] we modified the framework of DIVERSITY to learn a joint distribution over DNA sequences and their corresponding read count sequences to identify the modes of binding present in the data. ExoDiversity considers the set of sequences as a mixture of different binding modes where each mode is characterized by a sequence motif and its accompanying read distributions on both the positive and negative strands. We use collapsed Gibbs sampling in ExoDiversity to learn the sequence motifs and the read signals and their positioning relative to the sequence motifs while clustering the data into different modes of binding.

We developed our third method CISDIVERSITY [Biswas and Narlikar, 2021b], where the set of DNA sequences are considered as a mixture of distinct modules. Each module is defined by a different probability distribution over the set of motifs present in the data. In this method, collapsed Gibbs sampling is applied to learn the sequence motifs, their distributions over the modules while clustering the sequences based on their module identities.

We will go over the motivation behind each of these models and the biological characteristics which they manage to capture over different datasets. We will also look at the modifications made to the simple Gibbs sampling algorithm we saw in Section 2.3.1 while adapting it for the three methods.

Chapter 3

Identifying different modes of TF binding from ChIP data

In this chapter, first we look at the motivation for understanding the diverse ways in which a TF might bind to DNA using *de novo* motif discovery in ChIP-seq datasets. Then we look into the formulation of the problem statement and how it is different from the traditional motif discovery methods. After this, we go over the details of our mixture model-based framework DIVERSITY and its working. We show that applying DIVERSITY to ChIP data provides insights regarding different configurations of the target TF and together with other information like sequence conservation, gene ontology gives a more comprehensive idea about the regulatory activities the target TF is involved in.

The software for DIVERSITY was written and tested by Sneha Mitra. This work is published in Mitra et al. [2018], and much of the text and the images here, are taken from the paper.

3.1 Introduction

Transcriptional regulation is largely governed by the interactions between proteins called transcription factors and specific regions on the DNA. These regions on the DNA harboring the TF binding sites are called regulatory regions and play important roles in biological processes like differentiation, development, maintenance and apoptosis [Maston et al., 2006, Levine and Tjian, 2003]. In order to understand the roles of these regulatory regions under different circumstances, a wide range

of regulation-based biochemical activities have been measured and documented by various consortia like ENCyclopedia of DNA Elements (ENCODE) [Sloan et al., 2016] and other laboratories. These experiments and their analyses have provided evidence towards the fact that at any given time, there can be multiple regulatory mechanisms being at play [Struhl, 1991]. However, popular methods of analyzing these experimental datasets focus on finding a common characteristic in the data. For example, in the problem of motif finding, a common sequence signature typically represented in the form of a Position Weight Matrix (PWM) is learned from the protein-DNA binding data or promoters of co-regulated genes. This is done with the assumption that a particular motif for the specific TF must be over-represented in the data. But a TF can exert its influence on the genome in more than one way, by changing co-factors or through intermediate proteins without making a direct DNA contact [Farnham, 2009]. Thus, in a ChIP experiment for a target TF we can get a group of DNA regions that might contain the motifs or binding sites of different TF cofactors or various configurations of the target TF itself. Understanding these different types of binding configurations will help us to get a clearer picture of the role of the TF in gene regulation.

3.2 Problem statement

A ChIP experiment pulls down all regions that are associated with the target protein: whether directly bound to the protein or as a part of a protein complex in which the target protein is present. For example, consider a ChIP experiment targeting protein X, which binds DNA at its cognate sites and at the same time also interacts with four other proteins. Each of these proteins also binds to its own specific cognate site along the genome, which may be located at a farther distance from the binding locations of protein X. This scenario is consistent with the current view of transcriptional hubs and chromatin loops. In a ChIP experiment, when the cells are treated with formaldehyde, then protein-DNA as well as protein-protein interactions get cross-linked. Thus, in the immunoprecipitation step all the DNA regions associated with X will get pulled down, and the final data will be a union of five distinct sets of regions each containing binding sites of one of the five proteins. Characterizing all five different motifs and the corresponding sets of sequences is important to interpret the ChIP experiment and ultimately the role

of X in regulation. Our method DIVERSITY does this by partitioning the data into disjoint subsets based on the motif present in the sequences. The optimal number of motifs and the associated partitioning that together explain the entire dataset are both learned by DIVERSITY by using a Bayesian approach.

The goal in traditional motif discovery is to find a statistically overrepresented motif, typically one that appears in a large fraction of the data [Zambelli et al., 2013]. To identify more than one motif, the same approach is applied iteratively: occurrences of motifs identified in the previous passes are masked before searching for the next overrepresented motif. In doing so, they discover individually enriched motifs in the data, but those do not necessarily explain the entire dataset. This is conceptually different from DIVERSITY whose goal is to identify a set of motifs which together explain the entire dataset. We compared DIVERSITY results with the results from some of the other motif discovery methods: two standard approaches, namely MEME [Bailey et al., 1994] which is targeted for wide motifs corresponding to complexes and DREME [Bailey, 2011] which is usually used for finding shorter monomeric motifs likely to be cofactors. We also analyzed the results of another recent motif discovery method called InMoDe [Eggeling et al., 2017] that divides the data into a mixture of different ‘motif types’. Relaxing the inherent assumption of independence in PWMs, InMoDe identifies the dependencies within the binding sites by learning inhomogeneous parsimonious Markov models characterizing the motif types.

3.3 DIVERSITY model

For a given ChIP dataset, DIVERSITY assumes that the protein makes m different types of contacts, each of which is modelled with its own PWM. Formulating the problem as a mixture model, it splits the entire dataset into m disjoint subsets corresponding to the m contact types or modes of binding, with one PWM enriched in each subset. Neither the partitioning nor the PWMs are known a priori and are both inferred using a sampling-based approach. Several models with different values of m are learned, and the best model is identified using Bayesian model selection.

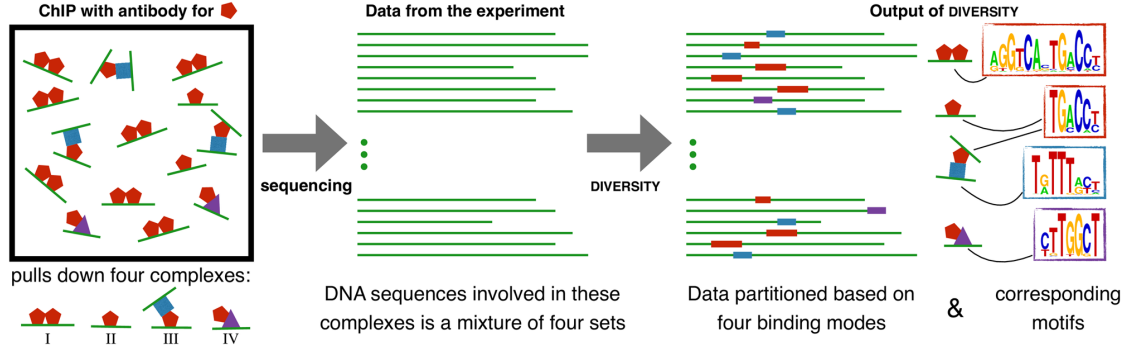


Figure 3.1: In this example the red protein binds to DNA as (i) homodimer, (ii) monomer, (iii) indirectly to another DNA region via the blue protein or (iv) indirectly to yet another DNA region via the purple protein making no direct DNA contact this time. The ChIP experiment pulls down all the complexes that are associated with the red protein, and sequencing reports all the DNA regions participating in these complexes. DIVERSITY learns the positions of the motifs in each sequence while simultaneously splitting them into different subsets based on the motif it contains.

3.3.1 Model and its parameters

Let us consider the case when the set of n DNA sequences $\mathbf{X}$ contains m different TF binding sites or motifs, we build a model M_m with parameters θ_m . Here, we consider each motif as a different mode of binding, and the cluster corresponding to each motif is thus called a mode. The set of parameters is given by $\theta_m = \{\mathbf{Z}, \mathbf{I}, \phi, \mathbf{w}, \gamma\}$ where:

- Z_i : The starting point of the motif in sequence $\mathbf{X}_i$, for $i = 1, 2, \dots, n$. $Z_i = j$ indicates that the binding site in sequence $\mathbf{X}_i$ is starting at location j .
- I_i : The cluster or mode of binding for sequence $\mathbf{X}_i$. $I_i = k$ would mean that the sequence $\mathbf{X}_i$ will belong to the cluster k .
- w_k : The width of the motif in mode k for $k = 1, 2, \dots, m$.
- ϕ_k : It represents the position-specific scoring matrix (PSSM) for the motif of mode k . $\phi_{k,a}(b)$ is the probability of finding nucleotide b at position a in the PSSM for mode k .

- γ_k : Gives the probability of a sequence belonging to mode k

We use ϕ_0 as a second order Markov model built from the input sequences $\mathbf{X}$ as the background model. To incorporate the heterogeneity in different genomic regions, we create the background model separately for every sequence. The motif occurrences in the sequences can be modelled as a ZOOPS model with zero or one occurrence per sequence or as OOPS which is one occurrence per sequence. DIVERSITY also has the provision where the motif occurrence per sequence can be determined based on a probability value given to the algorithm. In case of the ZOOPS model, the possibility of $\mathbf{X}_i$ having no motif is represented by having Z_i lie outside the sequence, i.e $Z_i = L_i + 1$, similar to the implementation in MuMoD [Narlikar, 2013] and PRIORITY [Narlikar et al., 2006]. In the OOPS model, we restrict the possible values of Z_i to $1, 2 \dots L_i - w_{I_i} + 1$.

We calculate the likelihood of the full data as:

$$P(\mathbf{X} \mid \boldsymbol{\theta}_m, \mathbf{M}_m) = \prod_{i=1}^n P(\mathbf{X}_i \mid \boldsymbol{\theta}_m, \mathbf{M}_m)$$

The likelihood of a single sequence $\mathbf{X}_i$ can be calculated as:

$$P(\mathbf{X}_i \mid \boldsymbol{\theta}_m, \mathbf{M}_m) = P(X_{i,1}, X_{i,2} \dots X_{i,Z_i-1} \mid \phi_0) \times \prod_{a=1}^w \phi_{I_i,a}(X_{i,Z_i+a-1}) \times P(X_{i,Z_i+w+1} \dots X_{i,L_i} \mid \phi_0)$$

Thus the posterior can be written as:

$$P(\boldsymbol{\theta}_m \mid \mathbf{X}, \mathbf{M}_m) \propto P(\mathbf{X} \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \times P(\boldsymbol{\theta}_m \mid \mathbf{M}_m)$$

3.3.2 Objective function

For a model with m modes, the aim is to find the optimal values of the parameters in $\boldsymbol{\theta}_m$ that maximize the posterior distribution

$$\begin{aligned} \hat{\boldsymbol{\theta}}_m &= \arg \max_{\boldsymbol{\theta}_m} P(\boldsymbol{\theta}_m \mid \mathbf{X}, \mathbf{M}_m) \\ &= \arg \max_{\boldsymbol{\theta}_m} P(\mathbf{X} \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \times P(\boldsymbol{\theta}_m \mid \mathbf{M}_m) \end{aligned} \tag{3.1}$$

3.3.3 Model learning

DIVERSITY uses Gibbs sampling to estimate the values of the parameters θ_m . It approximates sampling from the posterior distribution by drawing samples from individual conditional distributions. It applies collapsed Gibbs sampling to marginalize over ϕ and γ and integrate them out from the conditional distributions of $\mathbf{Z}$ and $\mathbf{I}$. All other components in θ_m : $\mathbf{Z}, \mathbf{I}, \mathbf{w}$ are learned by iteratively sampling from their conditionals. For each sequence $\mathbf{X}_i$ we first sample the motif start position Z_i in it and then its mode of binding I_i . Assuming conjugate priors for ease of integration, we use a Dirichlet prior over each position in ϕ_k having equal hyperparameters α_{motif} and another Dirichlet prior over γ with equal hyperparameters α_{mode} .

Using these priors, we construct the sampling expressions for Z_i and I_i .

Sampling for Z_i

If $Z_i = j$,

$$P(Z_i = j \mid \mathbf{X}, \mathbf{Z}_{[-i]}, \mathbf{I}, \mathbf{w}, \phi_0) = P(Z_i) \times P(X_{i,1}, X_{i,2} \dots X_{i,j-1} \mid \phi_0) \times \prod_{a=1}^{w_{I_i}} \hat{\phi}_{I_i,j}(X_{i,j+a-1}) \\ \times P(X_{i,j+w_{I_i}+1} \dots X_{i,L_i} \mid \phi_0)$$

$\hat{\phi}_{I_i}$ is the posterior mean of ϕ_{I_i} and it is calculated from the counts of the nucleotides from all the sequences except $\mathbf{X}_i$ having mode I_i .

$$\hat{\phi}_{I_i,a}(b) = \frac{\alpha_{\text{motif}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[X_{p,Z_p+a-1} = b, I_p = I_i]}{4\alpha_{\text{motif}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = I_i]}$$

$\delta[\text{condition}] = 1$ when *condition* is TRUE and 0 otherwise.

Sampling for I_i

Once the motif start position Z_i for $\mathbf{X}_i$ is sampled and assigned, we move on to sampling for its mode of binding. If $I_i = k$

$$P(I_i = k \mid \mathbf{X}, \mathbf{Z}, \mathbf{I}_{[-i]}, \mathbf{w}, \phi_0) = \hat{\gamma}_k \times P(X_{i,1}, X_{i,2} \dots X_{i,Z_i-1} \mid \phi_0) \times \prod_{j=1}^{w_k} \hat{\phi}_{k,j}(X_{i,Z_i+j-1}) \\ \times P(X_{i,Z_i+w_k+1} \dots X_{i,L_i} \mid \phi_0)$$

$\hat{\gamma}$ is the posterior mean of γ and is calculated from the counts of all sequences except $\mathbf{X}_i$ belonging to mode k .

$$\hat{\gamma}_k = \frac{\alpha_{\text{mode}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = k]}{m \cdot \alpha_{\text{mode}} + n - 1}$$

Sampling of w_k

We observed that the width of the motifs w_k for $k = 1, 2 \dots m$ do not change very often. Thus, we sample for the motif widths once every 10 passes of the sampler over all the sequences $\mathbf{X}$. For each mode, the width of the motif w_k is sampled from a set of values: where it is allowed to increase by one position or decrease by one position, or it remains the same on both the left and right ends of the motifs.

3.3.4 Convergence of the sampler

It is difficult to detect the convergence of the sampler to our target distribution when it is unknown and irregular. But we only need to determine the MAP estimate and do not really need to sample from the posterior. We fit a straight line through the last n iterations and stop sampling if the slope is close to 0. The maximum number of iterations for the sampler can be regulated by the user and DIVERSITY also has checks in place for the number of Z_i and I_i sampling iterations exceeding $2n^2$. This ensures that the algorithm will report a model in reasonable time in case of a bad initialization. After the sampler finishes, we use a Hill Climbing approach, starting from the model parameters corresponding to the maximum posterior probability. The final model parameters obtained at the end of Hill climbing are then reported by DIVERSITY.

3.3.5 Model selection

Since we do not know apriori the optimal number of modes that the data can be divided into, DIVERSITY learns models with different values of m in the range of $m_{\min}$ to $m_{\max}$ given to the algorithm. Theoretically, a model with more modes will have a better posterior score than a model with fewer modes. This will ultimately cause the dataset to be divided into n modes, with each sequence making up one mode, and this clearly is overfitting. In order to avoid this, we use Bayesian model selection and maximizing to determine which model best describes the data.

$$\begin{aligned}
 \arg \max P(\mathbf{M}_m \mid \mathbf{X}) &\propto \arg \max P(\mathbf{X} \mid \mathbf{M}_m) \cdot P(\mathbf{M}_m) \\
 &= \arg \max P(\mathbf{M}_m) \cdot \int_{\boldsymbol{\theta}_m} P(\mathbf{X}, \boldsymbol{\theta}_m \mid \mathbf{M}_m) d\boldsymbol{\theta}_m \\
 &= \arg \max P(\mathbf{M}_m) \cdot \int_{\boldsymbol{\theta}_m} P(\mathbf{X} \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \cdot P(\boldsymbol{\theta}_m \mid \mathbf{M}_m) d\boldsymbol{\theta}_m
 \end{aligned} \tag{3.2}$$

The prior on the model is exponential on the number of free parameters in the model. It thus penalizes the models with more parameters to a greater extent as compared to models with fewer parameters.

$$P(\mathbf{M}_m) \propto \exp(-\lambda |\mathbf{M}_m|)$$

Here λ is the hyperparameter that controls the magnitude of the penalty given to the models. The number of free parameters in $\mathbf{M}_m$ is the number of free parameters in the PSSMs used to represent the motifs.

$$|\mathbf{M}_m| = 3 \sum_{k=1}^m w_k$$

The integral in (3.2) is approximated by the maximum a posteriori (MAP) probability estimate, which has been obtained as $\hat{\boldsymbol{\theta}}_m$ in Equation (3.1) using Gibbs sampling. In DIVERSITY, we have used the value of λ as 5 to choose the best model, as it has been shown to give a reasonable clustering [Narlikar, 2013]. However, DIVERSITY reports models for all the values of m in the range $m_{\min}$ to $m_{\max}$ and it might be worthwhile for a Biologist to take a look at them.

DIVERSITY has a parallel implementation in Python and C and the software is available for download at <https://github.com/NarlikarLab/DIVERSITY/>

releases/tag/v1.0.0. A web server for DIVERSITY was also set up and can be accessed at <http://diversity.ncl.res.in/>.

We next look at some of the results of running DIVERSITY on various fly and human ChIP-seq datasets. For comparisons with MEME, DREME and InMoDe we used the following parameters for these methods:

- MEME parameters: `-nmotifs 20 -minw 6 -revcomp -p 16`
- DREME optional parameters: `-maxk 20 png`
- InMoDe was run in the flexible mode by setting the width to 20, motif orders to 0 (no dependencies) and number of modes as determined by DIVERSITY to be optimal when it too was run with a fixed motif width of 20. DIVERSITY was thus run twice: once when the motif width is allowed to vary (standard) and once when it was fixed to 20, only to compare with InMoDe.

3.4 Results

The sources for the biological datasets used to generate the following results are listed in Chapter: 6.

3.4.1 CTCF in fly makes diverse contacts

The CCCTC-binding factor, CTCF, is a highly conserved zinc finger protein that has been shown to have diverse roles in transcriptional regulation [Phillips and Corces, 2009]. Previous work has indicated several dependencies within its 20 bp binding site [Narlikar, 2013, Sharon et al., 2008] in mammals. To explore whether similar dependencies exist in invertebrates, we looked at data from the fruit fly. Ni et al. [Ni et al., 2012] have profiled CTCF in the white pre-pupa developmental stage across four related *Drosophila* species, separated by between ≈ 2.5 and 25 million years. They find a 9 bp core motif AGSKGGCGC to be enriched based on MEME in each species, implying that the binding specificity of CTCF has not evolved across those flies. This motif is found when MEME is supplied a motif width of 9 as a parameter, and it explains approximately half of the dataset. DIVERSITY was run on all four CTCF sets. Fig 3.2 (a) shows the *D. melanogaster* input to DIVERSITY and its output. It finds eight different

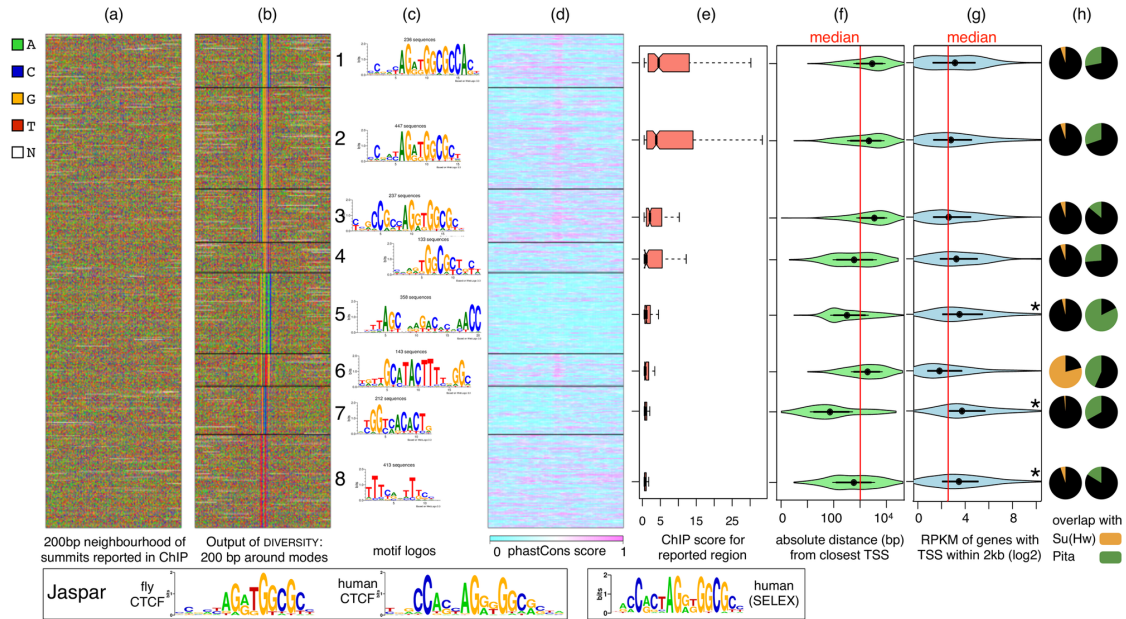


Figure 3.2: (a) 200 bp regions centered around the summit of ChIP peaks, input to DIVERSITY. (b) DIVERSITY finds eight modes which are aligned and re-ordered. (c) Motifs corresponding to modes. CTCF motifs from JASPAR and from high throughput SELEX [Jolma et al., 2013] are shown below. (d) Sequence conservation profile from phastCons, corresponding to nucleotides in b (e) The eight modes are displayed in decreasing ChIP score. (f) Violin plot of distance of each sequence from the closest transcription start site. (g) Violin plot of expression values of genes ($\log_2(1 + \text{RPKM})$) with TSS within 2000 bp of the ChIP region. The red line shows the median value across all measured genes. (h) Overlaps with Su(Hw) and Pita ChIP experiments, respectively.

modes, displayed based on the median ChIP binding score Fig 3.2 (b), (c) and (e).

Modes 1 and 2 have a similar ChIP score, although only mode 2 has been reported as the fly CTCF consensus. Mode 1 has an additional CAC at the 3' end. Interestingly, mode 3 resembles the human CTCF motif with the CC at the 5' end, but has a significantly lower ChIP score compared to modes 1 and 2. Unsurprisingly, modes 1–4 explain about half of the sequences. Modes 5, 6, 7 and 8 have a lower ChIP score than the first four modes and have no resemblance to any CTCF literature motif, suggesting these may not be direct binding sites. Instead, mode

5 matches the motif of a recently discovered insulator protein Pita [Maksimenko et al., 2015], mode 6 matches suppressor of hairy wing—Su(Hw)—a transcriptional repressor, and mode 7 matches a known fly promoter element [Ohler and Wassarman, 2010]. Mode 8 does not appear in any of the standard TF databases.

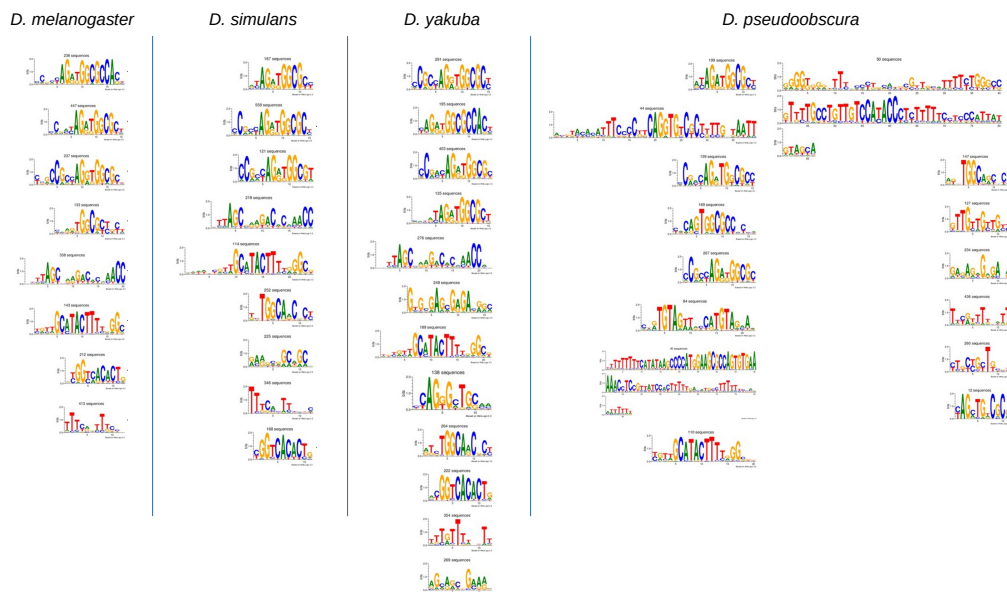


Figure 3.3: DIVERSITY results on 4 fly datasets for CTCF: 2180 sequences in *D. melanogaster*, 2191 sequences in *D. simulans*, 2984 sequences in *D. yakuba*, 2381 sequences in *D. pseudoobscura*

Surprisingly, the eight modes have different evolutionary profiles, as evident from phastCons scores Fig 3.2 (d). Modes 5 and 6 have opposite profiles, not only in terms of sequence conservation, but also in terms of functional conservation: mode 5 does not show up in *D. pseudoobscura*, which is farthest from *D. melanogaster* in the evolutionary tree, but is found in *D. simulans* and *D. yakuba*. In contrast, mode 6 appears in each of the four flies Fig 3.3. This suggests that the partnership of CTCF with Pita is specific to the melanogaster subgroup, while that with Su(Hw) is not.

The first three CTCF modes are far more variable in terms of where they bind along the genome with respect to gene Fig 3.2(f), which is typical of proteins exhibiting barrier or enhancer-binding function. Mode 4, on the other hand, is more proximal to promoters and is also less conserved across the four species,

suggesting this may be a non-functional artifact of more open regions being bound by the profiled protein and captured by ChIP. But additional evidence would be needed to be certain. Mode 7 is a well-established core-promoter motif, with almost half of the instances occurring within 100 bp of a TSS in the reported relative orientation [Narlikar, 2014]. Furthermore, the downstream genes are significantly more expressed Fig 3.2 (g) ($p\text{-value} < 10^{-15}$), suggesting that CTCF possibly activates transcription of these genes, by indirectly binding to the transcription initiation machinery assembling at these promoters. This could admittedly be a case of highly expressed promoters getting reported in the ChIP experiment, which are not specific to the profiled TF [Jain et al., 2015]. But even if that were true, it is still interesting that the CTCF ChIP-seq reports only those promoters that contain this particular element out of all the several different well-established fly promoter architectures [Narlikar, 2014].

3.4.2 Variations in Su(Hw) binding specificities

We next explored ChIP datasets of the co-factors of CTCF identified from Fig 3.2 (c), based on motif matches with the JASPAR database. Su(Hw), a zinc finger protein instrumental in chromatin organization [Vogelmann et al., 2011], has been profiled as part of modENCODE [Nègre et al., 2011] in the same developmental stage. Only sequences of CTCF mode 6 have a significant ($\approx 80\%$) overlap with this experiment (Fig 3.2 (h)). Further, the CTCF motif is not one of the 10 modes identified by DIVERSITY on the Su(Hw) set (Fig 3.4). This suggests that the Su(Hw)-CTCF contact might be like complex IV in Fig 3.1 where the hexagon is CTCF and the triangle is Su(Hw). Alternatively, CTCF may be bridging multiple Su(Hw)-DNA binding events, but not making direct DNA contact in the process. In any of these situations, all regions in the complex will be pulled down in both CTCF-ChIP and Su(Hw)-ChIP, but will only contain contacts of Su(Hw), not CTCF. However, the CTCF-ChIP will additionally report regions where CTCF does bind DNA directly, since the ChIP is against CTCF.

More surprisingly, DIVERSITY discovers six variants of the known Su(Hw) motif (Fig 3.4), of which mode 2 is most similar to the database motif. We can split the motif into 4 pieces, based on the places where the variations occur. Piece (i) is invariant across the modes, while piece (iv) is the most variable. Indeed, binding sites of zinc finger proteins are known to have interdependent effects within

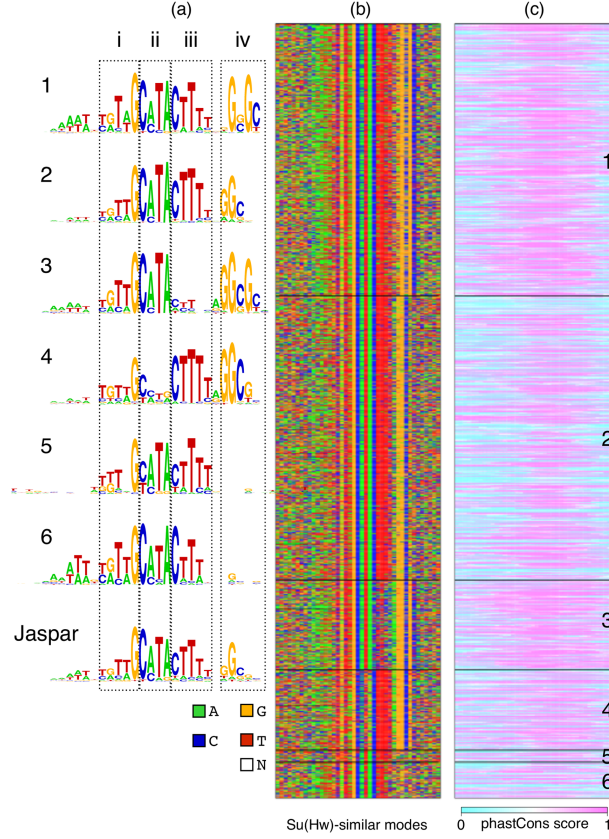


Figure 3.4: (a) Logos, (b) sequences, and (c) phastCons scores corresponding to the Su(Hw)-like motifs discovered by DIVERSITY.

positions [Bulyk et al., 2002]. Considering that individual zinc fingers interact with three or four consecutive nucleotides [Enuameh et al., 2013], variations in modes 3 and 4 are specially intriguing. We propose that the different zinc fingers of Su(Hw) bind two distant regions on the chromosome, one belonging to mode 3 and the other to mode 4. Three pieces of evidence support this. First, the two modes are complementary in terms of information content at pieces (ii) and (iii). Second, the number of sequences corresponding to the two modes is almost equal, suggesting that each region from mode 3 might have an interacting partner in mode 4. And finally and most importantly, if these were simply ‘weak’ or non-consensus binding sites for Su(Hw), the low information pieces would be under neutral selection. But that is not the case: the average conservation scores at the two pieces in both the modes is no different from the scores at the literature

consensus, all under negative selection. This suggests the organism prefers non-consensus pieces in these modes, possibly ensuring that some zinc fingers are free to make contact with the corresponding ‘missing’ piece at a different location.

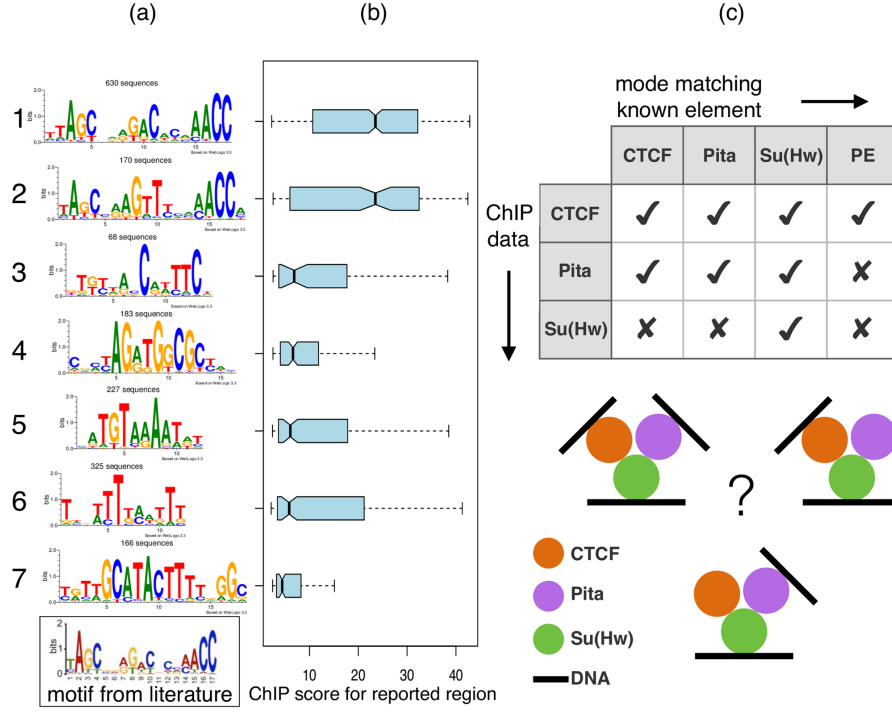


Figure 3.5: (a) DIVERSITY finds the Pita motif as well as CTCF and Su(Hw) motifs. The literature motif is a result of conventional motif discovery on the same dataset [Zolotarev et al., 2016]. (b) Both Pita motifs have a significantly high ChIP score. (c) Table describes the various known direct binding motifs identified in each of the three datasets.

3.4.3 Pita interacts with CTCF and Su(Hw)

We next explored the other potential CTCF co-factor, Pita, based on mode 5. A newly identified TF, also a C2H2-type zinc finger, it has not been profiled in the same very early stage of development, but in 0-12h embryos [Zolotarev et al., 2016]. DIVERSITY finds the literature Pita motif (mode 1), but also an additional variant, with the central piece differing in a fifth of sequences (Fig 3.5 (a)). The protein may have a different structural conformation at those regions. The ChIP score, however, is not different across the modes (Fig 3.5(b)).

Given the overlaps of this Pita set with the CTCF modes (Fig 3.2 (h)), it is not surprising that DIVERSITY finds CTCF and Su(Hw) motifs (modes 4 and 7, respectively). Taken together, this means, a CTCF ChIP pulls down direct sites of Pita and Su(Hw), a Pita ChIP pulls down direct sites of CTCF and Su(Hw), but Su(Hw) does not pull either sites of the other two (Fig 3.5 (a)). While additional experiments are needed to ascertain this, one possible explanation could be that Su(Hw), interacts with CTCF and Pita as part of one or more complexes, but those complexes do not make direct DNA contacts at CTCF or Pita binding sites.

3.4.4 Binding activity of REST in neuronal and non-neuronal cells

REST has many cofactors in neuronal cells

The RE-1 silencing transcription factor (REST) has been shown to repress neuronal genes in non-neuronal cell-types and play regulatory roles in differentiation and development of neuronal cells [Qureshi et al., 2010]. It binds directly to DNA, but it also interacts with a diverse set of co-factors, and the recruitment of specific complexes is believed to result in distinct transcription outcomes [Greenway et al., 2007]. We therefore consider it an apt TF for testing DIVERSITY. Rockowitz et al. [Rockowitz et al., 2014] have compiled and analysed REST binding in 15 non-neuronal human cell-types and differentiated human neurons. Here, we discuss detailed results only on the neurons (Fig 3.6) and one non-neuronal cell-type: the lymphoblastoid cell line GM12878 (Fig 3.7; results on a few other datasets are in Fig: 3.8). GM12878 was chosen due to availability of nucleosome occupancy data in this cell-type.

Rockowitz et al. applied the MAST tool of the MEME suite [Bailey et al., 2009], which scans sequences reported in neuronal cells on the basis of a user-supplied PWM corresponding to the 21 bp RE-1 motif (Fig 3.7 box). They showed there was only a marginal enrichment of RE-1, even in the top 600 sequences. Unsurprisingly, DIVERSITY also finds only a small fraction of sequences ($\approx 3.5\%$) contributing to a mode that looks like RE-1 (mode 1, Fig 3.6 (a)).

In addition to the RE-1 motif, DIVERSITY finds 11 other modes. For simplicity, only the top TF predicted to bind each mode by TOMTOM (with $p < 10^{-4}$) is listed on the left in Fig 3.6. But we note that in some cases a whole family of TFs

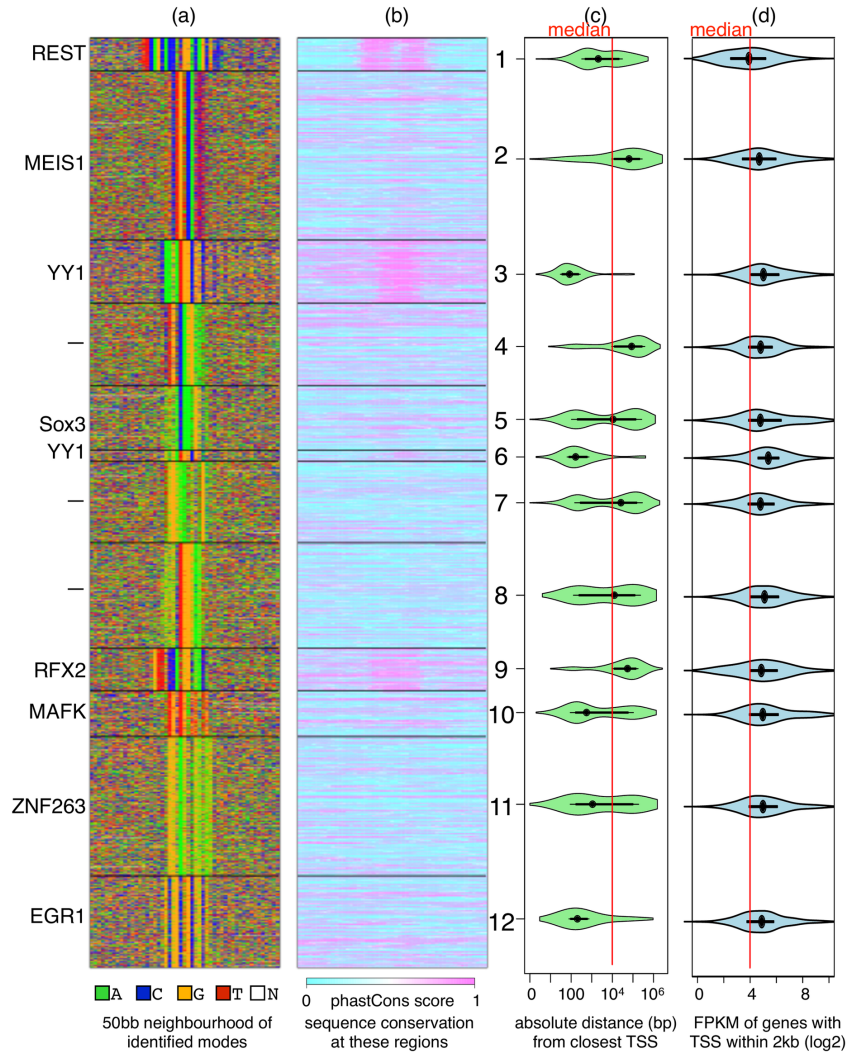


Figure 3.6: (a) The 12 modes are sorted on the basis on average ChIP score. (b) Sequences corresponding to RE-1 (mode 1), YY1 (modes 3 and 6), and RFX2 (mode 9) are more conserved. (c) While many modes are close to transcription start sites, modes 2, 4, and 9, are more variable in terms of their relative position. (d) All genes except those close to the RE-1 mode are significantly more expressed ($p < 10^{-5}$) than average.

have binding sites that match a mode, e.g., Sox2, Sox3, and Sox6 all have similar motifs and either one of them could be the factor in question. For all these modes, genes with transcription start sites within 2kb are significantly highly expressed (Fig 3.6 (d)). This supports the behavior of REST as an activator in neurons

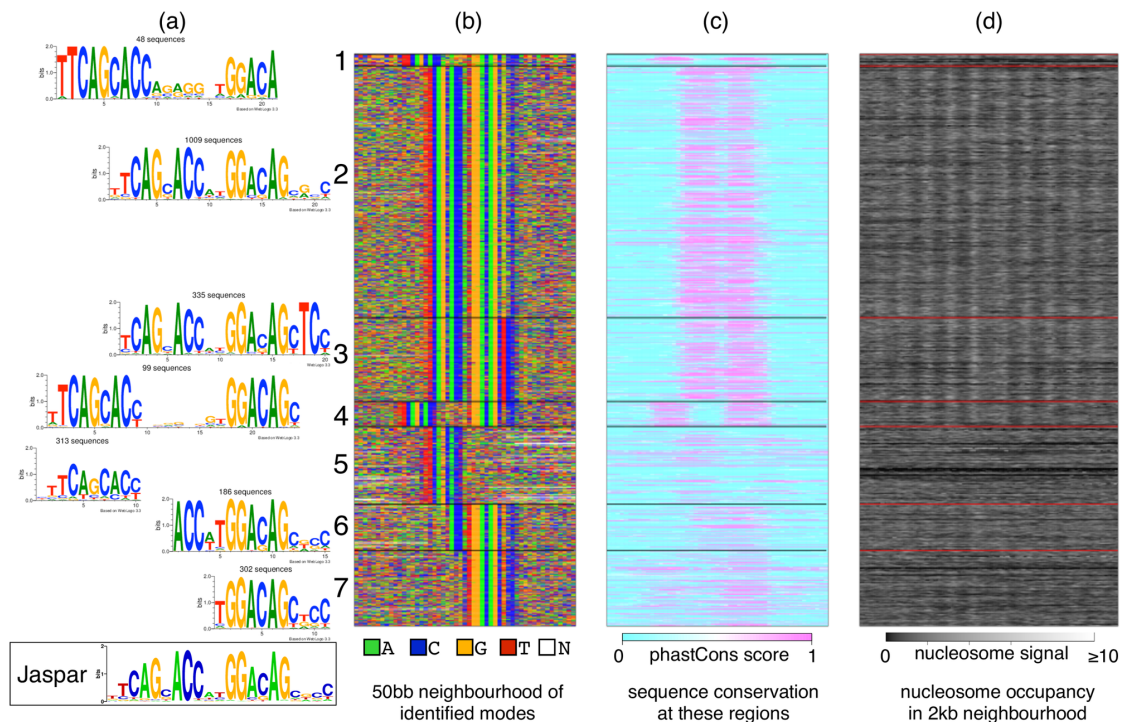


Figure 3.7: (a,b) DIVERSITY identifies seven modes, all variants of RE-1 (modes sorted based on average ChIP score). Database motif shown below. (c) The full sites are more conserved (d) The full sites have well-organized chromatin structure.

[Ooi and Wood, 2007], but suggests that this largely happens not by binding DNA directly. The top TF match from established vertebrate databases is listed on the right: these are most likely co-factors of REST. Although the modes are ranked in the order of ChIP score, there is no significant difference in the scores across the first 10 modes. This suggests that the binding of REST with DNA is as persistent as the binding of REST with the identified co-factors and of the co-factors with their own DNA recognition sites put together. This is, of course, based on the fairly reasonable assumption that REST cannot directly bind non-RE-1 motifs.

We now look at the co-factors in detail. MEIS1 is a homeobox encoding TF, known to be crucial for neuronal differentiation [Bouilloux et al., 2016]. REST has been shown to repress MEIS1 expression in non-neuronal cells via recruitment of Polycomb Repressor Complexes [Dietrich et al., 2012]. But how it interacts with MEIS1 in neuronal cells, where MEIS1 is expressed, is not yet known.

DIVERSITY finds two variants of YY1 motif (modes 3 and 6). YY1 plays a

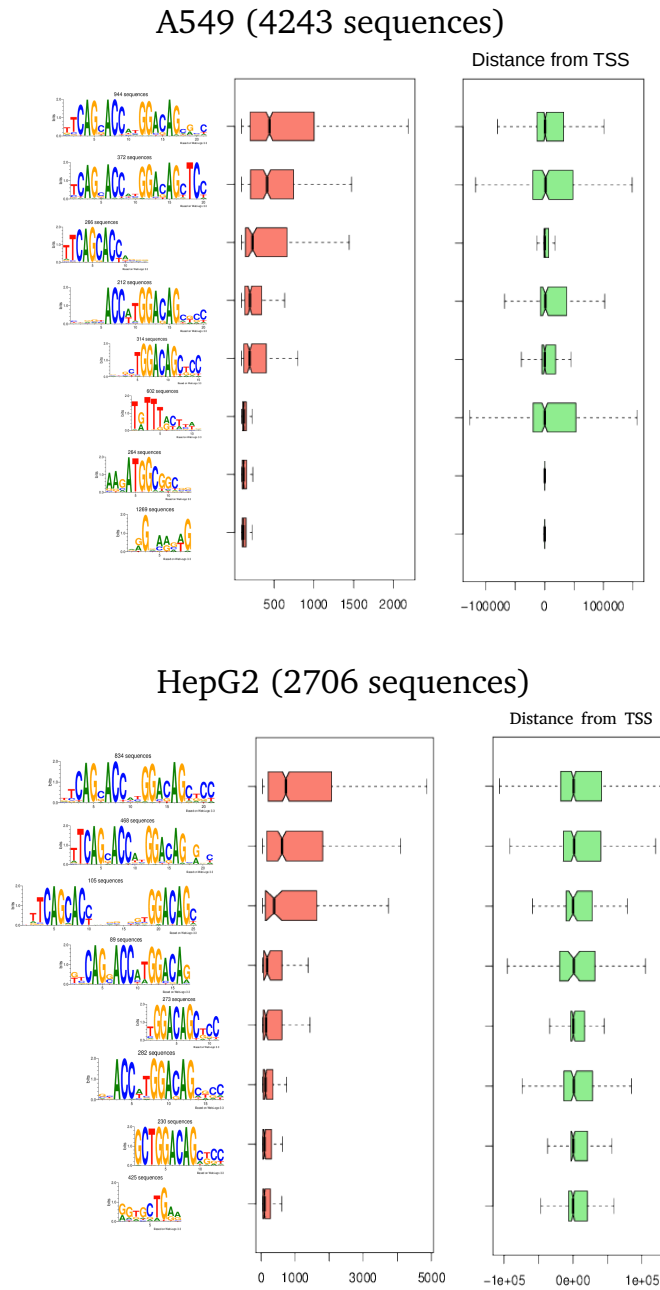


Figure 3.8: DIVERSITY output on ChIP-seq datasets for REST in A549 and HepG2 cell lines

key role in neuronal development and has been shown to positively regulate REST itself [He and Casaccia-Bonnel, 2008]. It is very likely that YY1 forms a complex with REST: both are zinc fingers of the same family and YY1 has been shown to bind with other zinc fingers [Brayer and Segal, 2008]. Both modes are highly conserved and are significantly closer ($p < 10^{-5}$) to the TSS.

Mode 9, which matches the winged helix RFX family, is the only other significantly conserved mode. All proteins in this family recognize a near-identical palindromic motif [Khan et al., 2018]. While the functions of RFX proteins are yet to be understood, at least one member, RFX1, has been shown to be critical for the development of the central nervous system in mouse [Feng et al., 2011]. DIVERSITY results suggest it also interacts with REST.

Mode 5 is a near perfect match to the SoxB1 proteins: Sox2 and Sox3, which have been shown to be critical for neuronal differentiation [Bergsland et al., 2011]. Mode 10 matches MAFK, a basic leucine zipper TF, involved in several functions including HDAC recruitment [Kannan et al., 2012], something REST is believed to do as well [Rockowitz et al., 2014].

Modes differ in terms of distance from the closest genes and expression of the downstream gene. MEIS1, RFX2, and mode 4 are far more variable, suggesting they might be distal regulatory regions such as enhancers or silencers. Without additional information such as ChIP of these proteins in the same set or chromatin structure information, we cannot say much more about what complexes must be forming along the genome. However, these results, from a single ChIP experiment, do point towards a rather complex TF-TF interaction landscape for REST in neuronal cells.

REST largely binds to RE-1 and variants in non-neuronal cells

In non-neuronal cell-types, the picture is dramatically different from neuronal cells. In GM12878 cell line (Fig 3.7), DIVERSITY finds only variants of the canonical 21 bp RE-1 motif, which contains two informative pieces separated by a small gap. These asymmetric half-sites of RE-1 have been shown to occur individually in REST-bound regions of non-neuronal cells [Johnson et al., 2008]: these are detected in modes 5 and 7. We find an additional mode 6 that contains the right half and a piece of the left half. All these modes: 5,6,7 are consistently found in the other 14 non-neuronal cell-types we examined as well (images are present in

the paper), but have a significantly lower ChIP score than the full site ($p < 10^{-5}$).

The full sites are more conserved than half-sites. Chromatin structure around full sites is strikingly different: nucleosomes are strongly phased and evenly spaced around the full sites, similar to what is known to happen with the insulator binding protein CTCF in mammals [Wiechens et al., 2016]. A weak nucleosome signal has been described in an earlier study, around computationally predicted RE-1 (not *de novo*) sites within REST binding peaks [Zheng et al., 2009].

3.4.5 Number of modes is a characteristic of the binding TF

Here we apply DIVERSITY to a collection of diverse TFs in the K562 cell-line (Table 3.1 to assess the generality of the method. The figures for these results are given in the Supplementary section of the paper. There is a clear variation in the number of modes detected across the TFs. As expected, the P300 protein, a general activator that does not bind DNA directly, has the most number of contact-types, which supports our current understanding about its function: it binds to several different DNA-binding TFs [Bedford et al., 2010] and is a marker for enhancers [Visel et al., 2009]. Indeed, in this dataset, diversity detects motifs resembling RUNX1, GATA, AP-1, SP1, and CTCF which are all active in this cell-type. In contrast, in the ChIP-seq of cell-type specific TFs FOXA1, GATA1, GATA2, and USF1, the number of modes is only two, one of which resembles the literature consensus of the respective TF. In the case of FOXA1, the second mode is the GATAA motif, while for the other three TFs, the second mode is a C-rich motif. This motif also occurs in some of the other TF sets, but in all cases the ChIP score at the sequences contributing to it is low. This may be a case of non-specific binding. RUNX1, also a cell-type specific protein, has a mode that resembles the RUNX1 motif, one that is a variant and a third that matches SP1, a known RUNX1 co-factor in leukemia [Pippa et al., 2017]. In all these cases, the sequences contributing to the known cognate motif of the profiled TFs have a significantly higher ChIP score than the other modes (S5 Fig), suggesting direct binding at those places.

FOSL1 from the Fos family and JUNB of the Jun family are part of the AP-1 complex, which is involved in multiple cellular processes. AP-1 complexes are known to be instrumental in looping DNA and involved in enhancer-promoter interactions [Phanstiel et al., 2017], which explains the multiple modes in these

TFs. But interestingly, there is no unique AP-1 complex: it can contain diverse combinations of TFs from both the Fos and Jun families [Andreucci et al. [2002]]. This explains why other than the characterized AP-1 motif of **TGA***s***TCA**, there are no common modes between the two TF datasets. In fact, in the case of FOSL1, the mode with the highest ChIP score is a strong motif but not recorded in the standard databases.

IRF2 is one of the interferon regulatory factors (IRFs), which bind to **AANNGAAA**. Variants of this motif are discovered as distinct modes by diversity. IRFs have different C-terminal regions, which help facilitate specific protein-protein interactions [Yanai et al., 2012]. This may explain the additional modes found in this set.

THAP1 is a zinc finger protein that is known to interact majorly with a general transcriptional regulator HCFC1. HCFC1 does not bind DNA directly but via interactions with other TFs such as YY1, E2F1, and THAP11 [Hollstein et al., 2017], all of which diversity detects as separate modes.

3.4.6 Comparison with other motif discovery methods

Here, we compare and contrast DIVERSITY’s output to that of the two standard motif finding approaches: MEME (for wide motifs) and DREME (for shorter motifs). We discuss the results on the 16 REST datasets, where the direct binding motif is well characterized. MEME detects the full RE-1 motif in 11 of the 16 sets (in spite of relaxing the definition of a ‘full RE-1 site’ to at least 14 bp containing cores of both half-sites). In the remaining 5 sets, MEME finds the two half-sites or variants as separate motifs (Figures are in the Supplementary section of the paper). This is because in these five sets MEME identifies the two half-sites first during its sequential motif discovery, and then masks them to find the next most enriched motifs, therefore missing the full site. In contrast, if the full site is more overrepresented than the individual variants, it gets detected first and the half-sites subsequently get detected if they are individually overrepresented in the full dataset. We see a similar picture with DREME: it never identifies the full site, possibly because of its bias to short ungapped motifs.

The number of motifs returned by each of the three programs is different, as well. In GM12878, for example, although neither MEME nor DREME finds the full RE-1 motif, DREME finds 8 non-RE-1 motifs (Fig: 3.10) and MEME finds 17 (Fig: 3.9), some of which are supported by only 4 sites. Indeed, multiple

TF	Known activity of TF	Number of contacts discovered
FOXA1	A Forkhead protein, binds DNA and interacts with chromatin	2 modes: DIVERSITY identifies the FOXA1 motif and the GATA motif. FOXA1 is believed to stabilize GATA complexes by changing the local chromatin landscape [Sekiya et al., 2009]
GATA1 GATA2	Members of the GATA family of zinc finger TFs	2 modes in each set: One resembles GATAA and the other is C-rich motif
USF1	A member of the basic helix-loop-helix leucine zipper family, recognizes the E-box	2 modes: The larger mode matches the E-box, while the other is a new motif
RUNX1	A heterodimeric TF that binds to a core element of many enhancers and promoters	3 modes: Two are variants of the RUNX1 motif, one resembles SP-1, which is a known cofactor of RUNX1 [Pippa et al., 2017]
JUNB	It is a part of the AP-1 complex	4 modes: Over half of the sequences are accounted by the AP-1 resembling mode, others are novel
FOSL1	Dimerizes with other leucine zipper proteins, is part of the AP-1 complex, activator	6 modes: In addition to variants of the TGAsTCA AP-1 motifs, DIVERSITY discovers GATAA motif and two unknown motifs
IRF2	An interferon regulatory factor, known to have both activating and repressing functions	10 modes: The multiple modes include promoter motifs, CTCF, and variants of IRF known motifs
THAP1	Contains a THAP domain, colocalizes with the apoptosis response protein PAWR/PAR-4 in leukemia	11 modes: Modes include motifs resembling SP1, AP1, YY1, E-box motif, THAP11 and many promoter elements
P300	A histone acetyltransferase that regulates transcription via chromatin remodelling	17 modes: CTCF, RUNX1, GATA, AP-1, SP-1 are among motifs that feature in these modes

Table 3.1: Output of DIVERSITY on ChIP-seq data from 10 TFs in K562 cell-line



user-defined/default parameter values such as the minimum number of sites for

GM12878

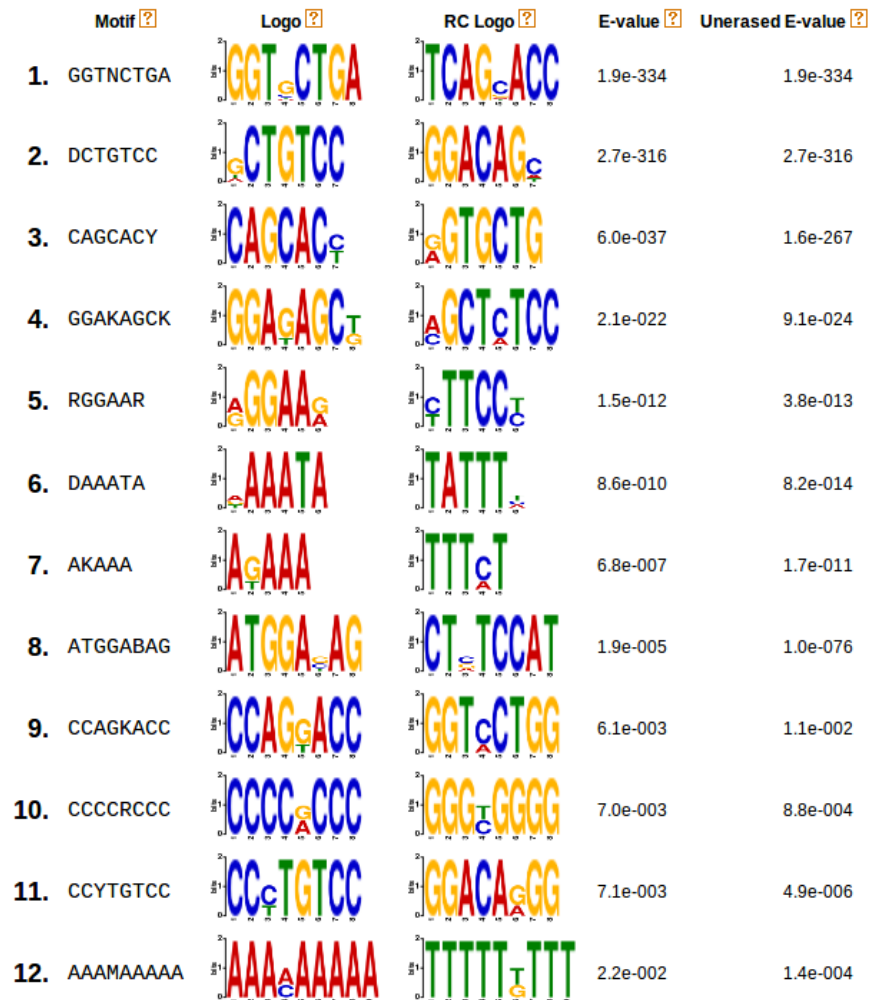


Figure 3.10: DREME output on ChIP-seq dataset for REST in GM12878 cell line.

a motif, the E-value cut-off for enrichment, etc. decide whether a motif will be reported in these methods. DIVERSITY, in its Bayesian formulation, uses one primary hyperparameter λ to determine how much to penalize models with more modes. Fig 3.7 suggests that variants of the RE-1 motif are probably enough to explain the REST-bound regions in GM12878: they cover the entirety of the set.

We of course cannot rule out the biological role, if any, of the motifs reported by MEME/DREME.

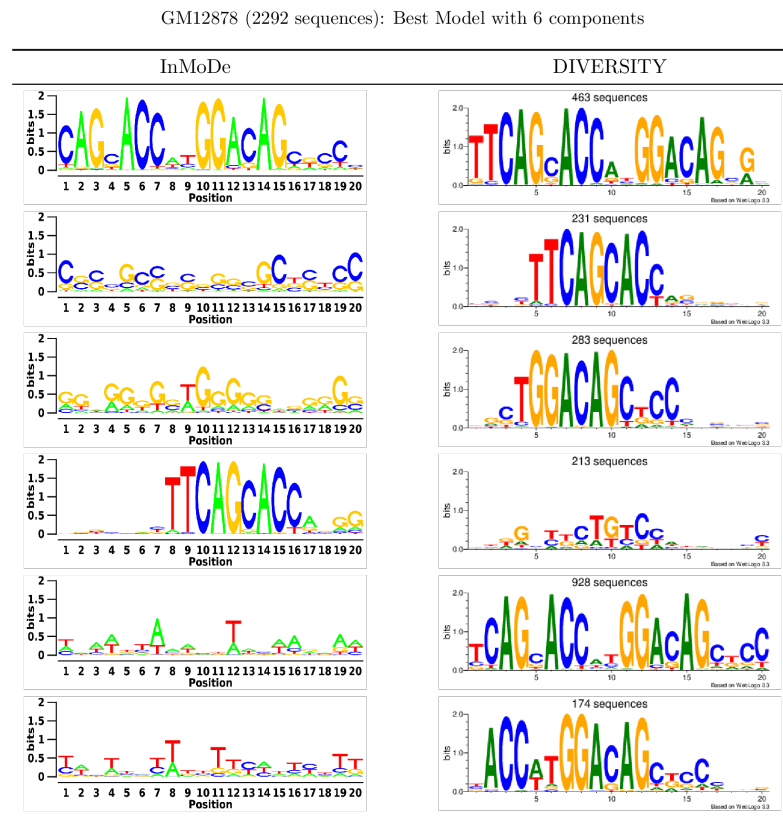


Figure 3.11: InMoDe output on ChIP-seq dataset for REST in GM1878 cell line

We also compared DIVERSITY results with the output of another motif discovery method, InMoDe that identifies different ‘motif types’ in a given dataset. However, InMoDe needs to be given both the width of the motifs and the number of modes to be identified by the user. Thus, we ran InMoDe on the REST datasets with the same number of modes that DIVERSITY found optimal (one such example is in Figs: 3.11). We kept the motif width as 20 to make sure that it has the chance to discover the full RE-1 motif. In 15 out of 16 REST datasets the full motif is one of the detected modes, but only in 5 are both half-sites detected as separate modes, which are understood to be prevalent across cell-types [Bruce et al., 2009] and are detected by DIVERSITY. Instead, InMoDe finds several modes with low information content, which may have biological significance, but at this point we cannot explain. We stress that this is not a fair comparison, since the motivation

behind InMoDe and therefore its objective function is very different from DIVERSITY's, implying that the number of optimal modes as determined by diversity may not be optimal for InMoDe. But there is no mechanism currently, in InMoDe, to identify the optimal width or number of modes. That said, InMoDe is significantly faster: it takes only 20 minutes on the neuronal REST dataset to find 12 modes on a single processor, compared with 85 minutes taken by DIVERSITY in parallel mode.

3.5 Discussion

A ChIP experiment reports all the regions that are associated with the target protein, whether the protein binds directly to that region or indirectly via intermediaries. Thus, we often get a mixture of regions that contain binding sites of the target protein along with binding sites of other TFs with which the target protein might be interacting. DIVERSITY discovers the various components present in the entire set of regions using no prior knowledge of the motif or the TF. The proportion of the sequences in each of these components vary across different datasets: for example, DIVERSITY could identify a small set of sequences containing the full RE-1 motif in the neuronal REST dataset.

Since our method uses a sampling-based approach, we have the option to start the algorithm from multiple initializations as well as use variable number of iterations after the burn in period to speed up the execution. DIVERSITY has a parallel implementation and is comparable in speed with standard motif discovery methods. The time required for convergence depends upon the dataset. If there are few clearly defined modes, then we get a faster convergence as compared to a large number of modes or when the modes are not distinctly separable. For example, the neuronal REST data which is one of the more diverse of our sets DIVERSITY takes an hour and 25 minutes to learn a model of 12 modes. When compared to a similar sized K562 REST data, DIVERSITY takes less than an hour there to identify 7 different modes. Ideally, DIVERSITY has to learn all models with number of modes in the range given by the user (1 to 20 in this case) before it can report the optimal model. Thus, the total time taken for the neuronal REST dataset is 27 hours. On the same machine, MEME when run in parallel mode takes over 40 hours to find 20 motifs (Fig: 3.12).

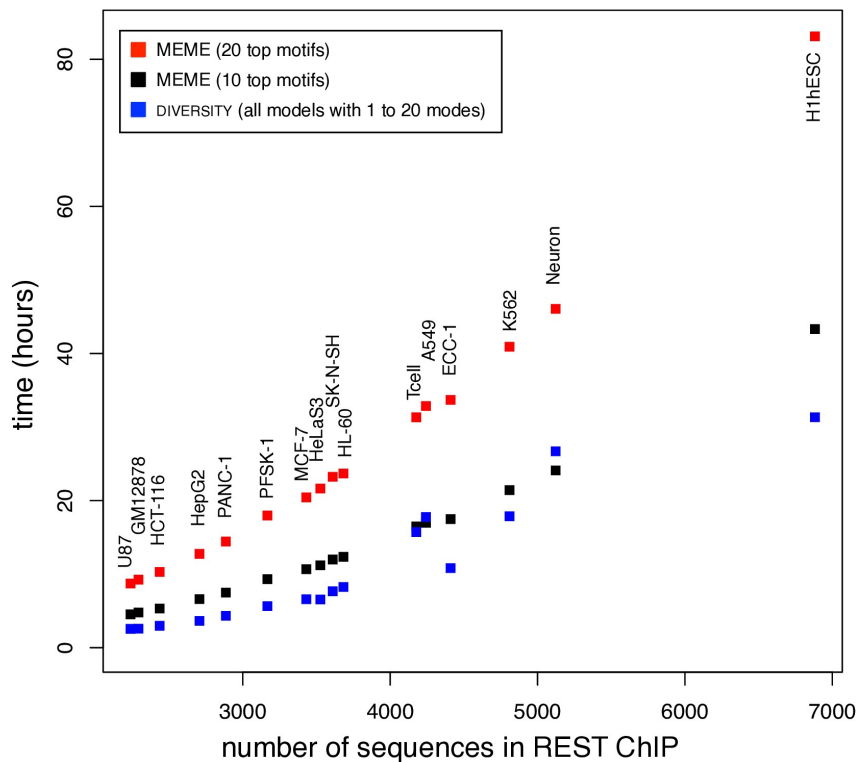


Figure 3.12: Run time of DIVERSITY is comparable to that of MEME. DIVERSITY and MEME were run on all 16 REST datasets as compiled by Rockowitz et al. (2014) [Rockowitz et al., 2014]. Both were run with the option of using 16 processors on Intel(R) Xeon(R) CPU E5-2630 v3.

From the results of DIVERSITY on various datasets, we can see supporting evidence towards the fact that diversity in regulation is driven in large part by diversity in sequence. The chromatin structure as well as sequence conservation correlates with different modes. DIVERSITY has also provided us with prospects of knowing the functional importance of each reported ChIP region by examining the different characteristics of the detected modes. It was observed that some of the modes played an important role in chromatin organization, some in activation while some in repression and so on. This can be learned by combining information from other sources such as histone/DNA modification, sequence con-

servation, gene ontology(GO) of downstream genes etc. In particular, we showed that DIVERSITY can give new insights into protein-DNA interactions even in widely studied TFs like the fly CTCF: it appears to bind human-like CTCF sites with lower efficiency; it interacts with specific promoter architectures; and that CTCF-Pita is likely a melanogaster subgroup-specific interaction, at least in the embryonic stage. Furthermore, our results suggest that one Su(Hw) molecule may be interacting with two distant DNA regions through its various zinc fingers. We hypothesize this as a similar event has been observed in the binding of Gata3. Gata3 has two zinc fingers and has been shown to bind to two GAT half sites separated by a long linker region [Chen et al., 2012]. Admittedly, such experiments for Su(Hw) are necessary to confirm our hypothesis.

DIVERSITY currently does not model the efficiency of the cross-linking or the immunoprecipitation step. Consider a situation where the profiled protein binds a DNA region through a chain of intermediaries. For the region to be reported, each interaction in the chain must get fixed during cross-linking and the antibody should be capable of recognizing the protein when it is part of this complex. Perhaps incorporating the accompanying ChIP binding score in the model will give further insights into the stability of the complexes. We also note that a ChIP experiment has its own limitations. Phantom peaks biased towards highly expressed regions have been reported in ChIP-seq experiments [Park et al., 2013]. Indeed, several promoter elements are identified as separate modes in many of the datasets studied here. Therefore we cannot discard the possibility that other modes are picked up perhaps because the regions are open and the TF “happens” to co-localise there, without its own cognate motif or is simply a result of an artifact of the ChIP experiment. One needs to be cautious when calling an identified mode the motif of a “co-factor”. Since the only information DIVERSITY uses is the DNA sequence at the ChIP regions, it can make no claim of the function of the identified components; that needs to be validated by separate means, with additional experiments/data.

The framework of DIVERSITY is conceptually different from standard motif discovery tools, since it asks and answers a very different question. Therefore, DIVERSITY does not seek to replace these tools but it can provide insights in cases where diverse configurations of the same TF are to be detected, especially in case of ChIP data. We note that it might be helpful to have such an analysis for other types of high-throughput experiments that identify regulatory regions

such as active enhancers [Arnold et al., 2013], accessible chromatin [Consortium et al., 2007], transcription initiation [Lizio et al., 2016] as there is even more likelihood of such data to be a mixture of multiple sequence components. For example, DNase I hypersensitive sites (DHSs) are accessible regions: they may be reported because they are active promoters, or enhancers, or insulators, or even matrix-attachment regions. There must be diverse sequence signatures that will explain the function of the DHSs. However, DIVERSITY limits each sequence to have only one binding site: in other words, one mode is defined by only one motif. For a ChIP-seq experiment that measures a specific TF-DNA interaction, this is a reasonable assumption, but to use DIVERSITY on the above mentioned high-throughput experiments the definition of a mode needs to be relaxed to a collection of motifs.

Chapter 4

Resolving protein-DNA footprints from exonuclease-based ChIP experiments

Chromatin immunoprecipitation followed by sequencing is the method of choice to identify the binding sites of a target protein throughout a genome. Although it is commonly used to investigate binding of a wide range of entities like TFs, modified histones, chromatin modifying enzymes, it suffers from a few limitations. Non-deterministic fragmentation of the DNA and contamination of the immunoprecipitation process due to unbound DNA produce false positives and leads to imprecision in mapping. To address these issues and to improve the resolution of the mapping of the TF binding sites, a modified version of the ChIP-seq protocol was developed called ChIP-exo [Rhee and Pugh, 2012]. It uses a lambda exonuclease enzyme to digest the unbound double-stranded DNA from 5' to 3' end until it encounters a bound protein. Free DNA is largely eliminated by the exonuclease digestion. It also ameliorates the mapping resolution by causing the 5' end cuts to be more concentrated near the binding sites. When these DNA read fragments are mapped to the genome, the profile of these 5' cuts can be used to precisely identify TF binding sites. Like ChIP-seq, ChIP-exo also captures both direct and indirect binding sites of the target protein, which can be characterized by the differences in not only the binding site but also in the 5' end read profiles. We show that using both the reads and sequence information available in the exonuclease-based ChIP experiments, we can resolve the mixture of ChIP-exo footprints into diverse modes

of binding of the target protein. To learn a joint distribution over the reads and the DNA sequences, we extended the model for DIVERSITY and built a framework ExoDiversity. We discuss the model and its working in details, followed by the results obtained by ExoDiversity on different datasets from exonuclease-based experiments.

This work is published in Biswas and Narlikar [2021a] and much of the work presented here follows from the paper.

4.1 Introduction

In a ChIP experiment, both protein-protein and protein-DNA interactions get crosslinked in vivo and thus genomic regions that are directly bound by the target TF or indirectly associated with the target TF as part of some protein complex, are equally likely to be immunoprecipitated. The identity of these DNA fragments are discerned by sequencing and then mapped to the reference genome. At this stage we get a set of genomic regions which are typically of length between 100 and 1000 bp. The actual TF binding site which is of length around 6 to 20 bp can be anywhere in this region.

To improve the positional resolution of these binding sites, several upgraded version of the ChIP method were developed that use the lambda exonuclease enzyme after the immunoprecipitation step to degrade the unbound DNA in the 5' to 3' direction. It leaves the DNA at the 3' end of the crosslinking event intact, and these read fragments once mapped to the genome produce precise directional footprints close to the complex on both the forward and the reverse strands. When these footprints are properly aligned, they are expected to reveal the DNA binding sites of the target protein. However, it has been observed that these footprints are not the same at all the locations and the variations in them depend upon multiple factors:

- i The actual protein binding to the DNA: it may be the target protein or some other protein which is a part of a complex or simply co-occurring with our target protein.
- ii The nucleotide sequence that is present at that binding site: a TF can normally bind to a set of distinct nucleotide sequences with varying binding

affinities.

- iii The precise location of the crosslink within the protein-DNA complex. The crosslinking creates a barrier for the exonuclease enzyme, resulting in the eventual distribution of the reads.

We need to categorize these footprints in order to identify the various protein-DNA contacts forming at the complexes. Two different approaches were pursued to understand and cluster the binding events in the reported set of regions. ExoProfiler by Starick et al. [Starick et al., 2015] is a pipeline that first scans the DNA regions with pre-established DNA motifs of interest and extracts the putative TF binding sites. Then the read footprints are plotted and analyzed separately at these binding sites to gain insights into the potential complexes and the binding mechanism involved. The second approach developed by Yamada et al. [Yamada et al., 2019] learns the different types of binding events present in the data by jointly considering the reads and the DNA sequences corresponding to the reported regions. Their method ChExMix first initializes potential binding events into subtypes either based on *de novo* motif finding or on the read distributions present in the top 500 to 1000 most enriched regions. Then these subtypes are refined over the complete dataset by using Expectation Maximization in the subsequent iterations. We present a new approach ExoDiversity that does not require the motifs to be known apriori and similar to ChExMix divides the ChIP-exo binding events into different binding modes based on both the read distributions and the sequence motifs. But, unlike ChExMix, ExoDiversity is built on a model-based Bayesian framework that uses both these sources of information simultaneously on the complete set of regions.

Using ExoDiversity on various datasets, we demonstrate that it can resolve the read footprints to a single nucleotide level, which correspond to differences at the base pair level in the sequence motifs. The binding modes identified by it relate to different properties of the genome like sequence conservation, DNA shape and helps us in understanding some underlying factors responsible for the exonuclease digestion activity.

4.2 Approach

ExoDiversity is built by extending our previous method DIVERSITY, incorporating the read distributions and some more parameters associated with them. It takes the bound regions reported by a ChIP-exo experiment as input along with the corresponding mapped reads in both directions and, in case of the availability of the control experiment, the corresponding reads from that as well. ExoDiversity goes on to partition the regions into different binding modes such that each mode is characterized by a DNA motif and the corresponding read footprints, which are learned as a joint distribution over the datasets. Multiple models with different number of modes are learned and finally using the Bayesian Information Criterion (BIC) the best model is selected.

4.3 Model description

From the input datasets we have the following:

- A set of n DNA sequences $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$. $X_{i,j}$ can be one of $\{A, C, G, T, N\}$ with $1 \leq j \leq L_i$ where L_i is the length of sequence $\mathbf{X}_i$ for $1 \leq i \leq n$.
- A set of n vectors $\mathbf{R}^+ = \{\mathbf{R}_1^+, \mathbf{R}_2^+, \dots, \mathbf{R}_n^+\}$ representing the read counts corresponding to $\mathbf{X}$ on the positive strand. $R_{i,j}$ denotes the number of reads whose 5' end maps to the genomic location corresponding to $X_{i,j}$ on the positive strand. These count values are discretized to binary values based on the median read count or a user-defined threshold.
- A similar set of n vectors $\mathbf{R}^- = \{\mathbf{R}_1^-, \mathbf{R}_2^-, \dots, \mathbf{R}_n^-\}$ corresponding to the negative strand.

A binding mode is defined by a joint probability distribution over the sequence motif and the read footprints in both the strands. The motif is modelled as a product of independent categorical distributions over the four nucleotides $\{A, C, G, T\}$ and the read footprints as a product of Bernoulli distributions for each of the strands at the two ‘read windows’ near the motif Fig: 4.1

Since exonuclease digestion produces very precise and sharp read distributions near the motif site, we consider a small window of width 5 on both the positive

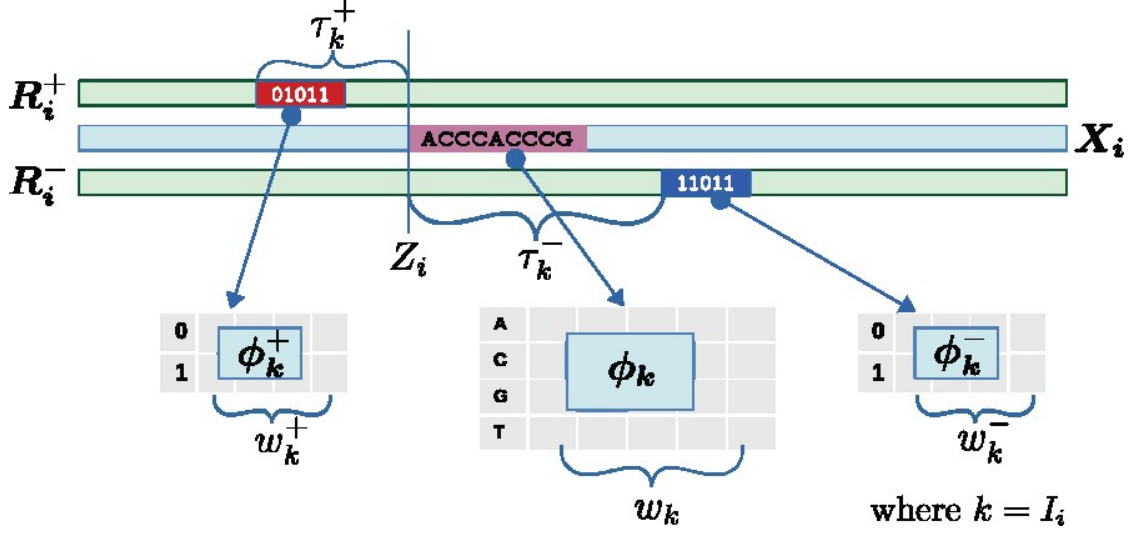


Figure 4.1: The i th datapoint consisting of the DNA sequence $\mathbf{X}_i$, positive strand reads $\mathbf{R}_i^+$, negative strand reads $\mathbf{R}_i^-$ belonging to mode k is shown. The red and the blue boxes correspond to the positive and negative strand read windows having width w_k^+ and w_k^- respectively. They are modelled using Bernoulli distributions given in ϕ_k^+ and ϕ_k^- . The 1 and 0 are the binarized read count values. The pink box represent the DNA motif which is present at location Z_i , and it is modelled using PWM parameters ϕ_k . The positive and negative strand read windows are at a fixed distance τ_k^+ and τ_k^- from Z_i . Only $\mathbf{X}_i$, $\mathbf{R}_i^+$ and $\mathbf{R}_i^-$ are known; the rest are learned from the data

and negative strands as the read windows. This width can be changed by the user, but it remains fixed for the read windows in both the strands. The positions of these read windows with respect to the motif, the motif locations within $\mathbf{X}$, the motif probabilities as well the Bernoulli probabilities are all unknown and are learned by ExoDiversity. Thus, ExoDiversity partitions the dataset into m binding modes while learning the values of the parameters θ_m of a model $\mathbf{M}_m$ which has six components $\theta_m = \{Z, \mathbf{I}, \mathbf{w}, \phi, \tau, \gamma\}$:

- Z_i : The starting position of the motif in $\mathbf{X}_i$
- I_i : The binding mode for $\mathbf{X}_i$; $1 \leq I_i \leq m$
- w_k : Width of the motif in mode k ; $1 \leq k \leq m$

- w_k^+, w_k^- : Width of the read window defining the positive strand footprint and negative strand footprint respectively in mode k . Here its default value is 5 but can be changed by the user.
- ϕ_k : PWM values for motif of mode k . $\phi_{k,a}(b)$ is the probability of finding nucleotide b at position a in the PWM of mode k ; $1 \leq a \leq w_k$
- ϕ_k^+ : Bernoulli probabilities for finding a read on the positive strand for mode k . $\phi_{k,a}^+(1)$ is the probability of having a positive strand read at position a within the read window of mode k , while $\phi_{k,a}^+(0) = 1 - \phi_{k,a}^+(1)$ is the probability of finding no read there; $1 \leq a \leq w_k^+$
- ϕ_k^- : Bernoulli probabilities for negative strand reads in mode k similar to ϕ_k^+
- τ_k^+ : distance of the start of the positive strand read window from the start of the motif in mode k (positive read window offset)
- τ_k^- : distance of the start of motif from the start of the negative strand read window in mode k (negative read window offset)
- γ : Categorical distribution over m modes. γ_k is the probability of a sequence belonging to mode k

The background distribution over the DNA sequences is considered as a second order Markov model learned separately for each sequence in $\mathbf{X}$ and its parameters are stored in ϕ_0 . Similarly ϕ_0^+ and ϕ_0^- have the parameters of the background read distributions for the positive and negative strand. These are considered as zeroth order Markov models learned directly from $\mathbf{R}^+$ and $\mathbf{R}^-$.

We assume that the DNA and the read are independent conditional on the mode information. Thus, the likelihood of a datapoint i consisting of $(\mathbf{X}_i, \mathbf{R}_i^+, \mathbf{R}_i^-)$ is given by:

$$P(\mathbf{X}_i, \mathbf{R}_i^+, \mathbf{R}_i^- \mid \theta_m, \mathbf{M}_m) = P(\mathbf{X}_i \mid \theta_m, \mathbf{M}_m) \times P(\mathbf{R}_i^+ \mid \theta_m, \mathbf{M}_m) \times P(\mathbf{R}_i^- \mid \theta_m, \mathbf{M}_m) \quad (4.1)$$

The first term in the above equation Eq: 4.1 is the likelihood of $\mathbf{X}_i$ considering it belongs to mode I_i . It thus contains an instance of the motif corresponding to I_i

at location Z_i and have the background nucleotide distribution elsewhere.

$$\begin{aligned}
P(\mathbf{X}_i \mid \boldsymbol{\theta}_m, \mathbf{M}_m) &= P(X_{i,1}, X_{i,2} \dots X_{i,Z_i-1} \mid \boldsymbol{\phi}_0) \times \prod_{a=1}^{w_{I_i}} \phi_{I_i,a}(X_{i,Z_i+a-1}) \times \\
&\quad P(X_{i,Z_i+w_{I_i}}, \dots X_{i,L_i} \mid \boldsymbol{\phi}_0) \\
&\quad \frac{\prod_{a=1}^{w_{I_i}} \phi_{I_i,a}(X_{i,Z_i+a-1})}{P(X_{i,Z_i} \dots X_{i,Z_i+w_{I_i}-1} \mid \boldsymbol{\phi}_0)}
\end{aligned} \tag{4.2}$$

This is obtained when we normalize the RHS by the background probability. In a similar way we can calculate the likelihood of $\mathbf{R}_i^+$ and $\mathbf{R}_i^-$ using the probabilities in the read windows in the positive and the negative strands and background probability at other places.

$$\begin{aligned}
P(\mathbf{R}_i^s \mid \boldsymbol{\theta}_m, \mathbf{M}_m) &= P(R_{i,1}^s, R_{i,2}^s \dots R_{i,Z_i^s-1}^s \mid \boldsymbol{\phi}_0^s) \times \prod_{a=1}^{w_{I_i}^s} \phi_{I_i,a}^s(X_{i,Z_i^s+a-1}) \times \\
&\quad P(R_{i,Z_i^s+w_{I_i}^s}, \dots R_{i,L_i}^s \mid \boldsymbol{\phi}_0) \\
&\quad \frac{\prod_{a=1}^{w_{I_i}^s} \phi_{I_i,a}^s(X_{i,Z_i^s+a-1})}{P(R_{i,Z_i^s}, \dots R_{i,Z_i^s+w_{I_i}^s-1}^s \mid \boldsymbol{\phi}_0)}
\end{aligned} \tag{4.3}$$

Here $s \in \{+, -\}$ and the position of the read window $Z_i^s = Z_i + \tau_{I_i}^s$.

The full likelihood of the dataset can thus be calculated as:

$$P(\mathbf{X}, \mathbf{R}^+, \mathbf{R}^- \mid \boldsymbol{\theta}_m, \mathbf{M}_m) = \prod_{i=1}^n P(\mathbf{X}_i, \mathbf{R}_i^+, \mathbf{R}_i^- \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \tag{4.4}$$

The motifs can be present on either strand and the read footprints might not be symmetric and as a measure to include the strandedness, the DNA sequences are appended with their reverse complements while taking care to reverse and append the reads appropriately.

4.4 Objective function

The goal is to learn the parameters of the model $\mathbf{M}_m$ to maximize the posterior distribution:

$$\widehat{\boldsymbol{\theta}}_m = \arg \max_{\boldsymbol{\theta}_m} P(\boldsymbol{\theta}_m \mid \mathbf{X}, \mathbf{R}^+, \mathbf{R}^-, \mathbf{M}_m) \quad (4.5)$$

$$= \arg \max_{\boldsymbol{\theta}_m} P(\mathbf{X}, \mathbf{R}^+, \mathbf{R}^- \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \times P(\boldsymbol{\theta}_m \mid \mathbf{M}_m) \quad (4.6)$$

4.5 Model learning

We use collapsed Gibbs sampling to marginalize over $\boldsymbol{\phi}$, $\boldsymbol{\phi}^+$, $\boldsymbol{\phi}^-$ and $\boldsymbol{\gamma}$ thus effectively bypassing the complexity of sampling every component of the high dimensional vector $\boldsymbol{\theta}_m$. While we integrate out these parameters from the conditionals, we iteratively sample for the other parameters: the motif start positions (Z_i) and the mode of binding (I_i) for each datapoint i , the read window offsets (τ_k^+, τ_k^-) and the width of the motifs (w_k) for each binding mode k . Conjugate priors of Dirichlet and Beta are chosen for convenience of integration. The Dirichlet prior with equal hyperparameters α_{motif} is taken over each position of $\boldsymbol{\phi}_k$ and a Beta prior with equal hyperparameters α_{read} is considered over each position of $\boldsymbol{\phi}_k^+$ and $\boldsymbol{\phi}_k^-$. For $\boldsymbol{\gamma}$, we again assume a Dirichlet prior with equal hyperparameters α_{mode} . The values of these hyperparameters can be changed by the user but by default we have the non-informative value of 1 for α_{read} and α_{mode} and a sparsity promoting value of 0.5 for α_{motif} . The sampling expressions for Z_i , I_i , and τ_k^+ or τ_k^- based on the conditionals are given below.

Sampling for Z_i

We use the notation $\mathbf{V}_{[-i]}$ to denote the vector $\mathbf{V}$ without the i th component. For any $j = \tau_k^+, \tau_{I_i}^+ + 1, \dots, L_i - \tau_{I_i}^- - w_{I_i}^-$ if $Z_i = j$, then the sampling expression for Z_i will be:

$$\begin{aligned} & P(Z_i = j \mid \mathbf{X}, \mathbf{R}^+, \mathbf{R}^-, \mathbf{Z}_{[-i]}, \mathbf{I}, \mathbf{w}, \boldsymbol{\phi}_0, \boldsymbol{\phi}_0^+, \boldsymbol{\phi}_0^-, \boldsymbol{\tau}^+, \boldsymbol{\tau}^-) \\ & \propto \frac{\prod_{a=1}^{w_{I_i}} \widehat{\phi}_{I_i, a}(X_{i, j+a-1})}{P(X_{i, j}, \dots, X_{i, j+w_{I_i}-1} \mid \boldsymbol{\phi}_0)} \times \prod_{s \in \{+, -\}} \frac{\prod_{a=1}^{w_{I_i}} \widehat{\phi}_{I_i, a}^s(R_{i, j+\tau_{I_i}^s+a-1})}{P(R_{i, j+\tau_{I_i}^s}^s \dots R_{i, j+\tau_{I_i}^s+w_{I_i}^s-1}^s \mid \boldsymbol{\phi}_0^s)} \end{aligned} \quad (4.7)$$

where $\hat{\phi}_{I_i,a}$ and $\hat{\phi}_{I_i,a}^s$ are the posterior means of $\phi_{I_i,a}$ and $\phi_{I_i,a}^s$ conditional on the current values of all the parameters except the i th datapoint.

$$\hat{\phi}_{I_i,a}(B) = \frac{\alpha_{\text{motif}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[X_{p,Z_p+a} = B, I_p = I_i]}{4\alpha_{\text{motif}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = I_i]} \quad (4.8)$$

$$\hat{\phi}_{I_i,a}^s(A) = \frac{\alpha_{\text{read}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[R_{p,Z_p+\tau_{I_i}^s+a}^s = A, I_p = I_i]}{2\alpha_{\text{read}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = I_i]} \quad (4.9)$$

$\delta[\text{condition}] = 1$ when *condition* is TRUE and 0 otherwise.

Sampling for I_i

The motif start position is fixed after it is sampled for X_i and then we sample for the binding mode of the sequence. For $I_i = k$

$$\begin{aligned} & P(I_i = k \mid \mathbf{X}, \mathbf{R}^+, \mathbf{R}^-, \mathbf{Z}, \mathbf{I}_{[-i]}, \mathbf{w}, \phi_0, \phi_0^+, \phi_0^-, \boldsymbol{\tau}^+, \boldsymbol{\tau}^-) \\ & \propto \hat{\gamma}_k \times \frac{\prod_{a=1}^{w_k} \hat{\phi}_{k,a}(X_{i,Z_i+a-1})}{P(X_{i,Z_i}, \dots, X_{i,Z_i+w_k-1} \mid \phi_0)} \times \prod_{s \in \{+, -\}} \frac{\prod_{a=1}^{w_k^s} \hat{\phi}_{k,a}^s(R_{i,Z_i+\tau_k^s+a-1}^s)}{P(R_{i,Z_i+\tau_k^s}^s \dots R_{i,Z_i+\tau_k^s+w_k^s-1}^s \mid \phi_0^s)} \end{aligned} \quad (4.10)$$

where $\hat{\gamma}$ is the posterior mean of γ conditional on the current values of $\mathbf{I}$ except for the i th datapoint

$$\hat{\gamma}_k = \frac{\alpha_{\text{mode}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = k]}{m \cdot \alpha_{\text{mode}} + n - 1}$$

Sampling for τ_k^+ and τ_k^-

Once the motif start position (Z_i) and the mode of binding (I_i) are sampled for all the sequences X_i for $i = 1, 2, \dots, n$, we iterate over the modes to sample the values for the read window offsets in both the positive and negative strands. τ_k^+

and τ_k^- determine the position of the read windows with respect to the start of the DNA motif for mode k . The value for τ_k^+ is sampled from all possible positions as long as the positive strand read window does not go beyond the motif's 3' end. Similarly, the value of τ_k^- is also sampled from all possible positions as long as the negative strand read window does not precede the motif. In a later version of the software, we have included a flag that allows for the positive strand read window to go beyond the 3' end of the motif, constraining only the start to lie before the 3' end. Similarly, in the case of the negative strand read window, the flag allows it to go beyond the 5' end of the motif with the constraint that the negative strand read window cannot lie entirely to the left of the 5' end of the motif. All the results presented here are obtained without using this flag.

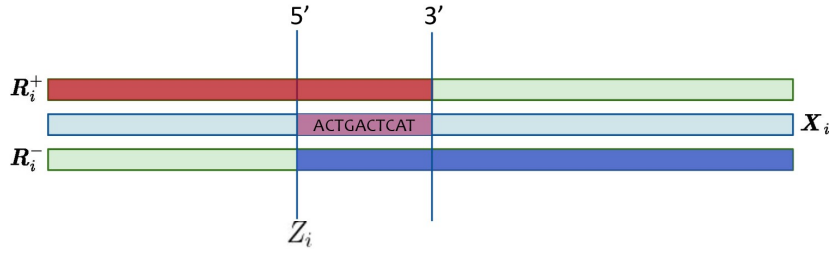


Figure 4.2: The red region shows all possible start positions for the positive strand read window and the blue region indicates the possible start positions for the negative strand read window. This assumption (without using the flag added later) was used to get the results presented here.

If l is one of the possible positions for τ_k^+ , then the conditional probability that $\tau_k^+ = l$ is given by:

$$P(\tau_k^+ = l \mid \mathbf{Z}, \mathbf{I}, \mathbf{X}, \mathbf{R}^+, \mathbf{R}^-) \propto \prod_{i=1}^n \prod_{I_i=k}^{w_{I_i}^+} \hat{\phi}_{k,a}^+(R_{i,Z_i-l+a}^+) \quad (4.11)$$

where $\hat{\phi}_{k,a}^+$ is similar to Eq: 4.9 but over all the sequence.

$$\hat{\phi}_{k,a}^+(A) = \frac{\alpha_{\text{read}} + \sum_{p=1}^n \delta[R_{p,Z_p+\tau_k^++a}^+ = A, I_p = k]}{2\alpha_{\text{read}} + \sum_{p=1}^n \delta[I_p = k]} \quad (4.12)$$

τ_k^- is sampled in a similar manner.

Sampling of w_k

The width w_k of the motif in mode k is sampled in a way similar to that of DIVERSITY. Instead of sampling from a large set of possible values in one iteration, the width is allowed to increase or decrease by 1 on either side of the motif, or stay the same. We noticed that the width does not change very often, so we sampled it after the other variables are sampled 10 times to speed up the entire sampling process. The initial width is set to a default value of 12, but it can be changed by the user.

Each sampling run is started from 10 different random initializations and this number can also be adjusted by the user.

4.6 Model selection

During sampling, for each sample from the Gibbs Sampler, we calculate the posterior score using the posterior distribution (from Eq: 4.5). Sampling is stopped when the value of the posterior score does not increase for $5 \times n$ iterations. At this point the sampler is assumed to have converged. After this, we perform a Hill climbing routine from the state of the model corresponding to the best posterior score obtained till now, similar to DIVERSITY. At the end of this, we report the final instance of the read distributions and the motifs as the MAP estimate of Eq: 4.5.

Since the optimal value of m is unknown, we learn models by varying the value of m . Then we calculate the Bayesian Information Criterion (BIC) score for each model, which is a function of the posterior score and the number of free parameters [Murphy, 2022]. The model with the minimum BIC score is reported as the best model.

$$\text{BIC}_{\mathbf{M}_m} = K \cdot \log_e n - 2 \cdot \log_e(\hat{P}) \quad (4.13)$$

Here, K is the number of free parameters, which for our model $\mathbf{M}_m$ includes the 3 free parameters for every column of the PWM, 1 parameter for every position of the read window and 2 parameters for the read window offsets for every mode. $\hat{P}$

is the posterior score of the model at the MAP estimate.

$$K = \sum_{i=1}^m (3.w_i + w_i^+ + w_i^- + 2)$$

4.7 Using models to predict

Given a model $\mathbf{M}_m$ learned by ExoDiversity and a new datapoint $\mathbf{D} = \{\mathbf{X}, \mathbf{R}^+, \mathbf{R}^-\}$ where the DNA sequence $\mathbf{X}$ is of length L with associated read counts sequences $\mathbf{R}^+$ and $\mathbf{R}^-$, we can calculate the probability of it belonging to mode k of model $\mathbf{M}_m$ by marginalizing over all possible motif positions $\mathbf{Z}$:

$$P(I = k \mid \mathbf{D}, \boldsymbol{\theta}_m, \mathbf{M}_m) \propto \gamma_k \sum_{j=1}^L P(\mathbf{D}, Z = j \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \quad (4.14)$$

where

$$\begin{aligned} P(\mathbf{D}, Z = j \mid \boldsymbol{\theta}_m, \mathbf{M}_m) &= \frac{1}{L} \cdot P(\mathbf{X} \mid Z = j, \boldsymbol{\phi}_k) \times \\ &\quad P(\mathbf{R}^+ \mid Z = j, \boldsymbol{\phi}_k^+, \tau_k^+) \times \\ &\quad P(\mathbf{R}^- \mid Z = j, \boldsymbol{\phi}_k^-, \tau_k^-) \end{aligned} \quad (4.15)$$

The three individual probability terms are computed using equations 4.2 and 4.3.

For a given model $\mathbf{M}_m$, we can also calculate the average probability of assigning a datapoint belonging to mode u to mode v . This can be represented as a confusion matrix where every entry (u, v) is computed as follows:

$$\frac{1}{\text{number of regions in mode } u} \sum_{\substack{i=1 \\ I_i=u}}^n P(I_i = v \mid D_i, \boldsymbol{\theta}_m, \mathbf{M}_m) \quad (4.16)$$

In the cases where only the motif or only the reads are used, the corresponding terms are dropped from Eq: 4.7.

We can similarly compute the probability of a datapoint given model $\mathbf{M}_m$ by marginalizing over $\mathbf{I}$ and $\mathbf{Z}$:

$$P(\mathbf{D} \mid \boldsymbol{\theta}_m, \mathbf{M}_m) = \sum_{k=1}^m \gamma_k \cdot \sum_{j=1}^L P(\mathbf{D}, Z = j \mid \boldsymbol{\theta}_m, \mathbf{M}_m) \quad (4.17)$$

assuming the starting position Z of the motif to be equally likely to appear anywhere within $\mathbf{X}$

Given two models $\mathbf{M}_a$ and $\mathbf{M}_b$ we can compute the log odds score for the datapoint:

$$S(\mathbf{D}) = \log \frac{P(\mathbf{D} \mid \boldsymbol{\theta}_a, \mathbf{M}_a)}{P(\mathbf{D} \mid \boldsymbol{\theta}_b, \mathbf{M}_b)} \quad (4.18)$$

which is used for classifying $\mathbf{D}$.

4.8 Using different distribution for the reads

Instead of discretizing the read count values to either 0 or 1 as is the case with the ExoDiversity model, we can take the original values and fit an appropriate distribution to them. We chose the Poisson distribution to model these discrete read count values at each position within the read windows. Thus, for the read distribution, we use $w_k +$ number of independent Poisson distributions for the positive strand read window and w_k^- number of independent Poisson distributions for the negative strand read window for each mode $k = 1, 2 \dots m$.

The background probabilities for the reads are Poisson probabilities with the parameter λ_{back}^i calculated as the mean of the read counts in each sequence $\mathbf{R}_i^s$. We observed that at certain locations the read count values are very large and thus taking them as it is leads to situation where these positions get a very low background probability which causes the posterior predictive score to be very high. Thus, we adopted a method of masking these outliers by the 98th percentile value of the dataset.

Here we look at the posterior calculation for $\mathbf{Z}$ when we learn a joint distribution over the sequence motifs and the reads and when the read counts at each position within the read window follow a Poisson distribution.

$$\begin{aligned}
& P(\mathbf{Z}, | \mathbf{X}, \mathbf{R}^+, \mathbf{R}^-, \phi_0, \phi_0^+, \phi_0^-) \\
& \propto P(\mathbf{Z}, \mathbf{X}, \mathbf{R}^+, \mathbf{R}^- | \phi_0, \phi_0^+, \phi_0^-) \\
& = P(\mathbf{X}, \mathbf{R}^+, \mathbf{R}^- | \phi_0, \phi_0^+, \phi_0^-, \mathbf{Z}) \times P(\mathbf{Z}) \\
& = P(\mathbf{X} | \phi_0, \mathbf{Z}) \times P(\mathbf{R}^+ | \phi_0^+, \mathbf{Z}) \times P(\mathbf{R}^- | \phi_0^-, \mathbf{Z}) \times P(\mathbf{Z}) \\
& = P(\mathbf{Z}) \int_{\phi} P(\mathbf{X}, \phi | \phi_0, \mathbf{Z}) d\phi \times \\
& \quad \int_{\phi^+} P(\mathbf{R}^+, \phi^+ | \phi_0^+, \mathbf{Z}) d\phi^+ \times \int_{\phi^-} P(\mathbf{R}^-, \phi^- | \phi_0^-, \mathbf{Z}) d\phi^- \\
& = P(\mathbf{Z}) \int_{\phi} P(\mathbf{X} | \phi, \phi_0, \mathbf{Z}) P(\phi) d\phi \\
& \quad \int_{\phi^+} P(\mathbf{R}^+ | \phi^+, \phi_0^+, \mathbf{Z}) P(\phi^+) d\phi^+ \times \int_{\phi^-} P(\mathbf{R}^- | \phi^-, \phi_0^-, \mathbf{Z}) P(\phi^-) d\phi^-
\end{aligned}$$

Consider reads on the +ve strand, and also let ϕ^+ be λ^+ . These are the parameters for the Poisson distributions in the positive strand read window

$$\begin{aligned}
\int_{\phi^+} P(\mathbf{R}^+ | \phi^+, \phi_0, \mathbf{Z}) P(\phi^+) d\phi^+ &= \int_{\lambda^+} P(\mathbf{R}^+ | \lambda^+, \phi_0^+, \mathbf{Z}) P(\lambda^+) d\lambda^+ \\
&= \int_{\lambda^+} P(\mathbf{R}^+ | \lambda^+, \mathbf{Z}) P(\lambda^+) d\lambda^+ \quad (4.19)
\end{aligned}$$

The above term gives the marginal likelihood of $\mathbf{R}^+$ where elements in every column in the reads window follow a Poisson distribution with parameter $\lambda_{k,j}^+$ for $j = 1, 2, \dots, w_k^+$ and $k = 1, 2, \dots, m$. We consider a conjugate prior of gamma distribution with hyperparameters $\alpha_{\text{read}}, \beta_{\text{read}}$ the values of which are assumed to be 1 by default. We use the following for the probability calculations:

$$\begin{aligned}
\text{Poisson-pdf}(x, \lambda) &= \frac{e^{-\lambda} \lambda^x}{x!} \\
\text{Gamma-pdf}(\lambda, \alpha, \beta) &= \frac{\beta^\alpha \lambda^{(\alpha-1)} e^{-\lambda\beta}}{\Gamma(\alpha)}
\end{aligned}$$

If C_k is the count of sequences in mode k and $S_{k,j}$ is the set of read counts at

position j within the read window belonging to mode k then,

$$\begin{aligned}
& \int_{\lambda^+} P(\mathbf{R}^+ \mid \lambda^+, \mathbf{Z}) P(\lambda^+) d\lambda^+ \\
&= \int \prod_{k=1}^m \prod_{j=1}^{w_k^+} \prod_{i=1}^n \frac{\exp(-\lambda_{k,j}) \lambda_{k,j}^{R_{i,j}^+}}{R_{i,j}^+!} \times \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{(\beta_{\text{read}})^{\alpha_{\text{read}}} \lambda_{k,j}^{\alpha_{\text{read}}-1} \exp(-\lambda_{k,j} \beta_{\text{read}})}{\Gamma(\alpha_{\text{read}})} d\lambda_{k,j} \\
&= \int \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{\exp(-C_k \cdot \lambda_{k,j}) \lambda_{k,j}^{\sum_{x \in S_{k,j}} x}}{\prod_{x \in S_{k,j}} x!} \times \frac{(\beta_{\text{read}})^{\alpha_{\text{read}}} \lambda_{k,j}^{\alpha_{\text{read}}-1} \exp(-\lambda_{k,j} \beta_{\text{read}})}{\Gamma(\alpha_{\text{read}})} d\lambda_{k,j} \\
&= \int \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{\exp(-C_k \cdot \lambda_{k,j} - \lambda_{k,j} \beta_{\text{read}}) \lambda_{k,j}^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}} - 1)}}{\prod_{x \in S_{k,j}} x! \cdot \Gamma(\alpha_{\text{read}})} (\beta_{\text{read}})^{\alpha_{\text{read}}} d\lambda_{k,j} \\
&= \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{(\beta_{\text{read}})^{\alpha_{\text{read}}} \Gamma(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}{\prod_{x \in S_{k,j}} x! \cdot \Gamma(\alpha_{\text{read}})} \times \\
&\int \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{\exp(-C_k \cdot \lambda_{k,j} + \lambda_{k,j} \beta_{\text{read}}) \lambda_{k,j}^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}} - 1)}}{(\beta_{\text{read}} + C_k)^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}} \frac{(\beta_{\text{read}} + C_k)^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}}{\Gamma(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})} d\lambda_{k,j} \\
&= \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{(\beta_{\text{read}})^{\alpha_{\text{read}}} \Gamma(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}{(\beta_{\text{read}} + C_k)^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}} \cdot \prod_{x \in S_{k,j}} x! \cdot \Gamma(\alpha_{\text{read}})} \times \\
&\prod_{k=1}^m \prod_{j=1}^{w_k^+} \int \frac{\exp(-(C_k + \beta_{\text{read}}) \cdot \lambda_{k,j}) \lambda_{k,j}^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}} - 1)}}{\Gamma(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})} (\beta_{\text{read}} + C_k)^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})} d\lambda_{k,j}
\end{aligned}$$

$$\begin{aligned}
&= \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{(\beta_{\text{read}})^{\alpha_{\text{read}}} \Gamma(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}{(\beta_{\text{read}} + C_k)^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}} \times \\
&\quad \cdot \prod_{x \in S_{k,j}} x! \cdot \Gamma(\alpha_{\text{read}}) \\
&\quad \prod_{k=1}^m \prod_{j=1}^{w_k^+} \int \text{Gamma-pdf}(\lambda_{k,j}, (\sum_{x \in S_{k,j}} x + \alpha_{\text{read}}), (\beta_{\text{read}} + C_k)) d\lambda_{k,j} \\
&= \prod_{k=1}^m \prod_{j=1}^{w_k^+} \frac{(\beta_{\text{read}})^{\alpha_{\text{read}}} \Gamma(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}{(\beta_{\text{read}} + C_k)^{(\sum_{x \in S_{k,j}} x + \alpha_{\text{read}})}} \times 1 \\
&\quad \cdot \prod_{x \in S_{k,j}} x! \cdot \Gamma(\alpha_{\text{read}})
\end{aligned}$$

Thus we see that the posterior probability for a position j within the read window for mode k follows the Gamma distribution with the following parameters:

$$\begin{aligned}
\alpha' &= \sum_{x \in S_{k,j}} x + \alpha_{\text{read}} \\
\beta' &= \beta_{\text{read}} + C_k
\end{aligned} \tag{4.20}$$

In order to score the read windows while sampling the start position of the motif we need to calculate the posterior predictive value for the read windows in both the strands. Consider a read value in the positive strand $x = R_{i,j}^+$ with $I_i = y$. If λ_z is the parameter for the Poisson distribution at a certain position z within the positive strand read window, then the posterior predictive score can be written as

$$\begin{aligned}
&\int_{\lambda_z} P(R_{i,j}^+ \mid \lambda_z, Z_i, \phi_0^+) P(\lambda_z) d\lambda_z \\
&= \int_{\lambda_z} P(x \mid \lambda_z, Z_i, \phi_0^+) P(\lambda_z) d\lambda_z \\
&= \int \frac{\exp(-\lambda_z) \cdot \lambda_z^x}{x!} \times \frac{(\beta')^{\alpha'} \cdot \lambda_z^{\alpha'-1} \cdot \exp(-\lambda_z \beta')}{\Gamma(\alpha')} d\lambda_z \\
&= \int \frac{\exp(-\lambda_z(1 + \beta')) \cdot \lambda_z^{(x+\alpha'-1)} \cdot (\beta')^{\alpha'}}{x! \Gamma(\alpha')} d\lambda_z \\
&= \frac{(\beta')^{\alpha'}}{x! \Gamma(\alpha')} \int \frac{\exp(-\lambda_z(1 + \beta')) \cdot \lambda_z^{(x+\alpha'-1)} \cdot (\beta' + 1)^{x+\alpha'} \cdot \Gamma(x + \alpha')}{(\beta' + 1)^{x+\alpha'} \cdot \Gamma(x + \alpha')} d\lambda_z
\end{aligned}$$

$$\begin{aligned}
&= \frac{(\beta')^{\alpha'}}{x! \Gamma(\alpha')} \times \frac{\Gamma(x + \alpha')}{(\beta' + 1)^{x + \alpha'}} \int \frac{\exp(-\lambda_z(1 + \beta')). \lambda_z^{(x + \alpha' - 1)}. (\beta' + 1)^{(x + \alpha')}}{\Gamma(x + \alpha')} d\lambda_z \\
&= \frac{(\beta')^{\alpha'}}{x! \Gamma(\alpha')} \times \frac{\Gamma(x + \alpha')}{(\beta' + 1)^{x + \alpha'}} \int \text{Gamma-pdf}(\lambda_z, x + \alpha', \beta' + 1) d\lambda_z \\
&= \frac{(\beta')^{\alpha'} \cdot \Gamma(x + \alpha')}{x! \Gamma(\alpha') \cdot (\beta' + 1)^{x + \alpha'}} \times 1 \\
&= \frac{\Gamma(x + \alpha')}{\Gamma(\alpha') \cdot x!} \times \frac{(\beta')^{\alpha'}}{(\beta' + 1)^{\alpha'}} \times \left(\frac{1}{1 + \beta'} \right)^x \\
&= \binom{x + \alpha' - 1}{x} \times \left(\frac{\beta'}{1 + \beta'} \right)^{\alpha'} \times \left(1 - \frac{\beta'}{1 + \beta'} \right)^x \\
&= \text{Neg-bin pdf}(x, r = \alpha', p = \frac{\beta'}{1 + \beta'})
\end{aligned}$$

Thus the second term for the reads in Eqns: 4.7 and 4.10 will become

$$\prod_{s \in \{+, -\}} \prod_{a=1}^{w_k^s} \frac{\text{Neg-bin pdf}(R_{Z_i + \tau_k^s + a - 1}^s, r = \alpha', p = \frac{\beta'}{1 + \beta'})}{\text{Poisson-pdf}(R_{Z_i + \tau_k^s + a - 1}^s, \lambda_{\text{back}}^i)}$$

Sampling the value of τ_k^s can be written as:

$$P(\tau_k^+ = l \mid \mathbf{Z}, \mathbf{I}, \mathbf{X}, \mathbf{R}^+, \mathbf{R}^-) \propto \prod_{a=1}^{w_{I_i}^+} \frac{\Gamma(x + \alpha_{\text{read}})}{\prod_{\substack{i=1 \\ I_i=k}} R_{i, Z_i - l + a}^s! (\beta_{\text{read}} + \sum_{i=1}^n \delta[I_i = k])^{(x + \alpha_{\text{read}})}}$$

where

$$x = \sum_{\substack{i=1 \\ I_i=k}}^n R_{Z_i - l + a}^s$$

$\delta[\text{condition}] = 1$ if $\text{condition} = \text{True}$, else 0

4.9 Results

The accession numbers and the sources of all the real datasets used to obtain the results in this section are listed in Chapter: 6.

4.9.1 In silico mixed data

To evaluate the performance of our method ExoDiversity, we should ideally run it on a simulated dataset where we can artificially plant the read and the motif footprints or on a ChIP-exo dataset where the true footprints are distinct and well established. However, the existence of such a ChIP-exo dataset is not currently known, and the simulated dataset will be constrained by the assumptions of the planter and may be far off from reality. Therefore, we followed a similar methodology like Yamada et al. [Yamada et al., 2019] where two different ChIP-exo datasets for TFs with well established direct binding motifs are mixed along with their read counts data. Here, we first describe the procedure that was followed to create the in silico mixed dataset.

Dataset creation

ChIP-exo datasets containing the mapped read fragments in BAM format for the two TFs CTCF and FoxA1, were mixed in equal proportions to create our dataset. MultiGPS [Mahony et al., 2014] was run on the BAM files for CTCF and FoxA1 separately with the corresponding control file to obtain the summits for each TF. Regions of length 240 bp around the summits were extracted and 5000 regions from each dataset were randomly chosen to make a dataset with a total of 10000 regions. While choosing these regions, it was checked that they did not contain repetitive DNA or overlapped with each other. These 10000 regions were mixed and the BAM files from both the experiments were merged. The mixed reads were control subtracted after scaling and along with the peak regions were given to ExoDiversity to run with its default parameters. While comparing with ChExMix, it was run with its default parameters with the regions file and the mixed and the control BAM files as input datasets.

Results from ExoDiversity

The peak regions of both the experiments were individually processed and 5000 such regions for the two experiments were mixed. The goal is to see if the CTCF and the FoxA1 footprints can be resolved considering this mixture of 10000 regions along with the combined reads. ExoDiversity finds a total of 6 modes. The first two resemble different variants of the CTCF motif, with over 85% of the sites

coming from the CTCF peak regions. The next two modes have variations of the FoxA1 motif, with almost 90% of the sites in the FoxA1 peaks. The last two modes contain motifs which do not resemble any known motif, but they account for the remaining 24% of the 10000 sequences. Among these last two modes, mode 5 is accompanied by a strong positive read signal five bases upstream in contrast to mode 6, which has no enrichment of reads.

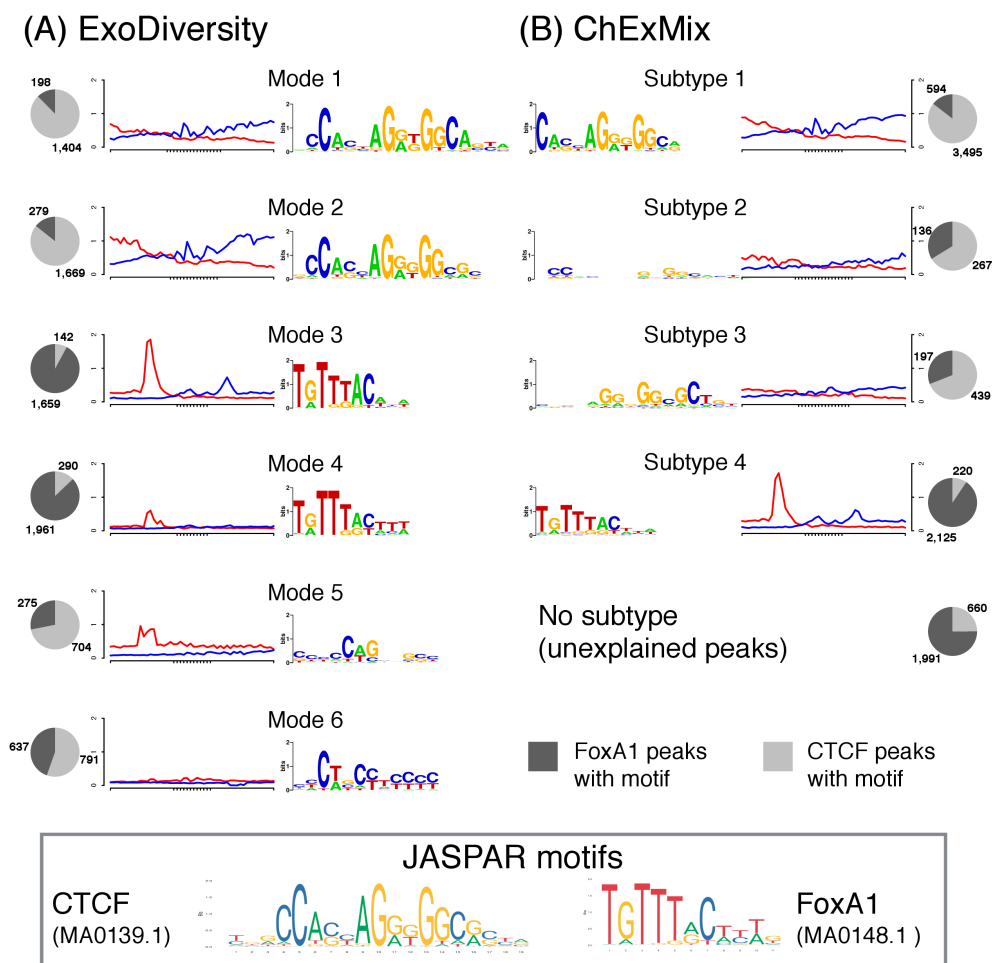


Figure 4.3: Motifs found in a mixed dataset with 5000 FoxA1 and 5000 CTCF regions in the MCF-7 cell line. Unlike in ExoDiversity, the numbers in the pie charts for ChExMix do not add up to 10000 since it finds multiple subtypes in some regions and none in others

We note that modes 1 and 2 differ in the 3' region of the motif: the presence

of these nucleotide dependencies have been shown before [Eggeling et al., 2014, Narlikar, 2013]. Interestingly, the read footprints are different in intensity within both the CTCF and FoxA1 variants: modes 2 and 3 have stronger read signals than their variants modes 1 and 4 respectively. The only other approach that detects subtypes of binding in ChIP-exo experiments using *de novo* motif finding is ChExMix, and it finds four subtypes. The biggest one is a clear match to CTCF and like the first two modes of ExoDiversity has over 85% of the sites in CTCF peaks. The next two motifs are weaker which although bears some resemblance to the CTCF motif but are less specific to CTCF peaks. The fourth subtype matches the FoxA1 motif and is enriched in the FoxA1 peaks, but it explains only 42.5% of the total FoxA1 peaks, in contrast to over 70% of the FoxA1 peaks that are explained by one of the modes 3 or 4 of ExoDiversity. Overall, the motifs learned by ChExMix are weaker in terms of information content than those learned by ExoDiversity. This might be because ChExMix uses the motif discovery method MEME in each subtype to find motifs in the top enriched 500 to 1000 regions, but when extended to the full set of regions the motif becomes weaker. Nevertheless, the total number of peaks not explained by a CTCF or a FoxA1 motif are only a little different between the two methods (2407 for ExoDiversity and 2651 for ChExMix)

4.9.2 ExoDiversity results on ChIP-nexus datasets for Sox2 and Oct4

We next applied ExoDiversity to 7465 ChIP-nexus regions for Sox2 binding in murine embryonic stem cells [Avsec et al., 2021]. Sox2 is a key factor that helps regulate pluripotency in embryonic stem cells (ESCs) and recognizes two types of motifs shown in the box in Fig: 4.4. ExoDiversity splits the data into 12 modes of binding, the first of which resemble the Sox2 motif. They differ in the 5' flanking DNA and these differences correspond to the variations in the read signals as well, which are shifted by one or two positions on either strand. This indicates that the Sox2-DNA binding at the motif variants is different structurally. Further, the ChIP enrichment scores as reported by MACS at these modes are also found to be different, with the first mode having a significantly higher Sox2 occupancy with a Wilcoxon p-value $\leq 10^{-10}$. Modes 8-10 match the Oct4::Sox2 dimer motif, and

these modes are also found to be significantly more occupied by Oct4 (grey plot at the right in Fig 4.4). They have differing Sox2 occupancy, with mode 8, like mode 1 having a significantly higher Sox2 occupancy (Wilcoxon p-value $\leq 10^{-10}$ on the ChIP scores) as compared to modes 9 and 10. An interesting feature to note here is that although the motifs in modes 8-10 when compared to motifs in modes 1-7 have an additional Oct4 motif on the 5' region, the positive strand read signals do not appear to be shifted. This suggests that even though Oct4 presumably does not bind to modes 1-7, the exonuclease digestion is nevertheless hindered similar to locations where Oct4 does bind. It might be the case that the position of the Sox2 protein is enough to obstruct the exonuclease from digesting that part of the DNA. Modes 11 and 12 do not match any database motifs, but the accompanying strong positive read signal for mode 12 suggests a potential experimental bias or artifact.

Looking at the partitioning given by ExoDiversity, we notice that visually, many modes appear similar. To quantify this similarity, we estimated the posterior probabilities in 3 different ways. First, we calculated the probability of every datapoint belonging to each mode based on its corresponding reads and sequence information using Eq: 4.7. Unsurprisingly, the ‘true mode’ is always the most probable. We did a similar exercise using only the sequence information ($\mathbf{X}$ and ϕ) and then using only the reads information ($\mathbf{R}^+, \mathbf{R}^-, \phi^+, \phi^-$). The mean probability of sequences of a particular mode predicted to belong to other modes using each of these scoring techniques is represented as three confusion matrices in Fig: 4.5. Modes 1-7 and 8-10, which are variations of Sox2 and Oct4::Sox2 motifs, respectively show only minor similarities when the full model is used. However, the similarity is more pronounced when only the motif information is used, as is expected. The reads contribute little by themselves, but in combination with the motif information they help make the modes distinctly different, as evident from the bold diagonal of the first matrix.

We also ran ExoDiversity on 19051 ChIP-nexus regions targeting Oct4. Here also it finds several variants of Oct4, Oct4::Sox2 dimer, and Sox2 motifs with different read footprints. Modes with the Oct4::Sox2 motifs are most highly enriched with Oct4 occupancy. Additionally, ExoDiversity finds motifs of Esrrb, Klf4, CTCF, (Fig: 4.6) all of which are active in the ESCs. All these modes have relatively weak read signals, suggesting indirect binding.

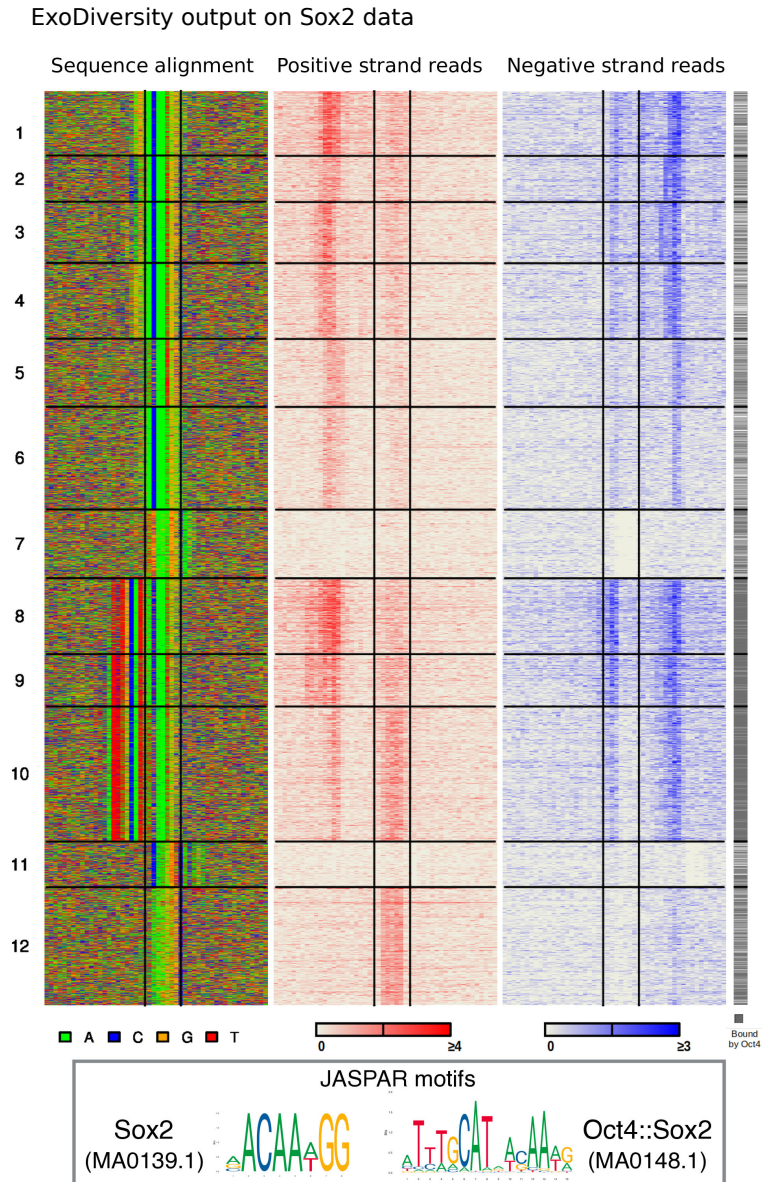


Figure 4.4: ExoDiversity finds 12 modes in the Sox2 dataset. The DNA sequence and the reads in the positive and negative strands in a 50 bp neighbourhood are shown here. The central two lines in each plot mark the site of the corresponding motif. The last column indicates which of the regions overlap with the peaks reported in the ChIP-nexus experiment for Oct4.

Confusion matrix over the modes

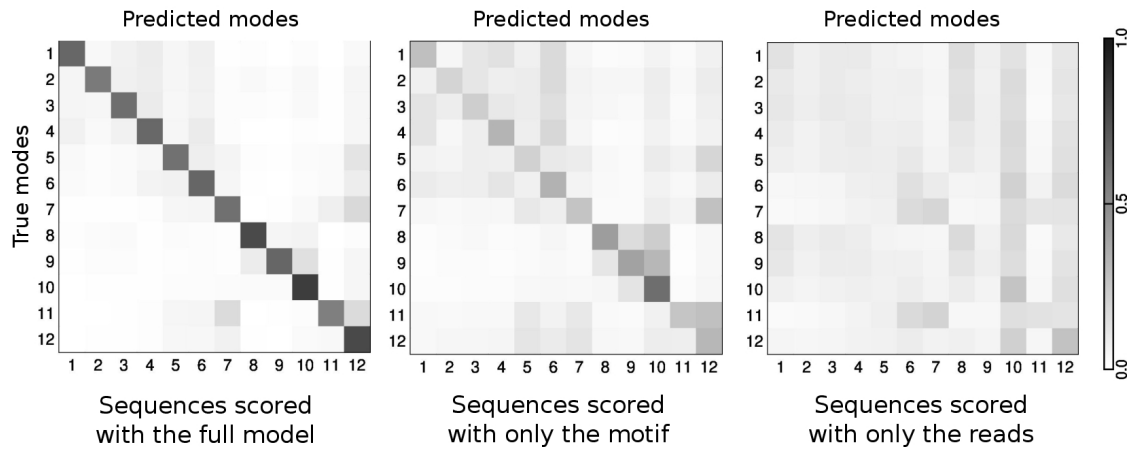


Figure 4.5: The matrices show the probability distribution of a sequence over the 12 possible modes, averaged over sequences in each mode. The first matrix considers the full model, while the other two use only one of the components: either only the sequence part or only the reads part

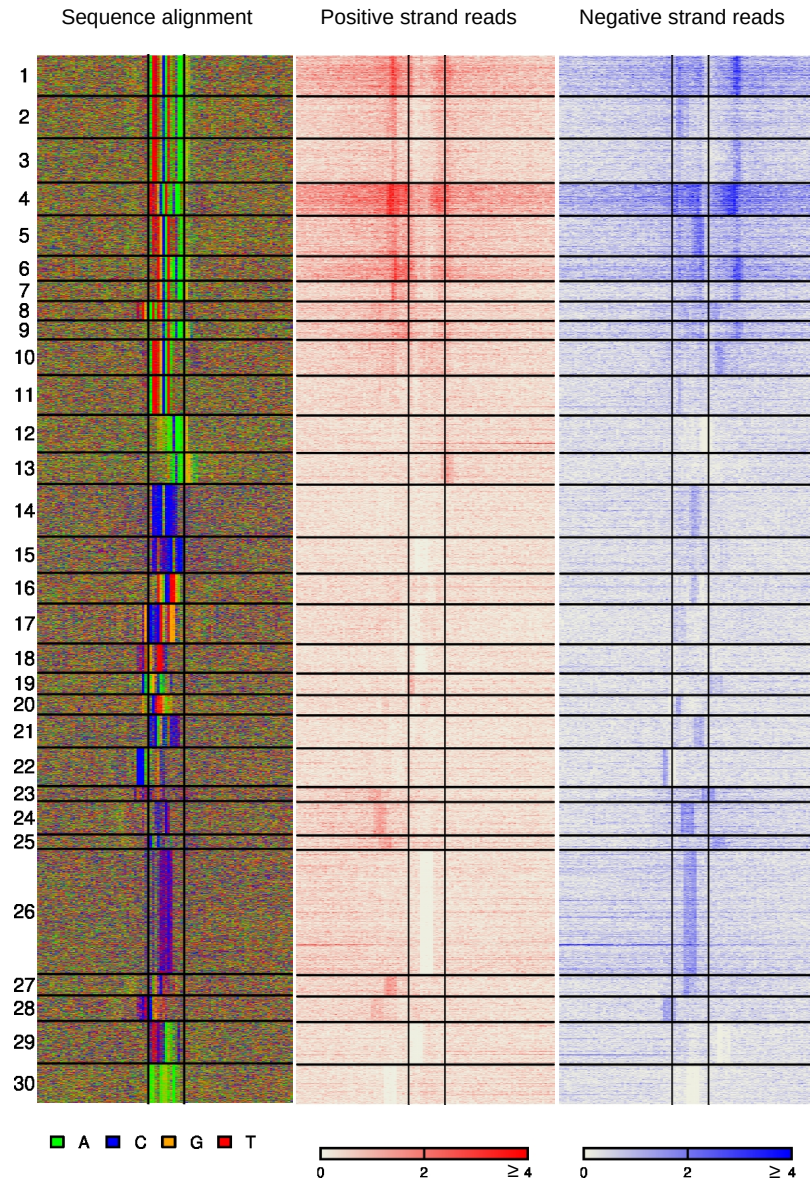


Figure 4.6: ExoDiversity finds 30 modes in the ChIP-nexus dataset for Oct4. Of the many modes, it identified Oct4::Sox2 dimer motifs (modes 1-8), Oct4::Oct4 motif (mode 9), Oct4 motif (modes and 11), Sox2 motifs (modes 12, 13, 19, 20), Klf4 (modes 14, 15), Esrrb (mode 16), CTCF (mode 17). These plots were plotted in a 100 bp neighbourhood around the motifs, with the central two lines indicating the motif sites.

4.9.3 ExoDiversity models can distinguish between sets bound by TFs with similar database motifs

We were confounded by the large number of modes captured by ExoDiversity in the Sox2 and Oct4 datasets. Although the two datasets overlap to a great extent, with the Oct4::Sox2 motif variant being enriched in the overlaps (Fig: 4.4), the individual modes are different in number and composition between the two sets. While the modes are indeed different in the probability space, we wanted to ensure they were not a result of over-fitting. One way to assess this is to check how general the model is in predicting unseen bound regions. The formulation of ExoDiversity enables us to find log likelihood ratios for a datapoint with respect to two models (Eq: 4.18).

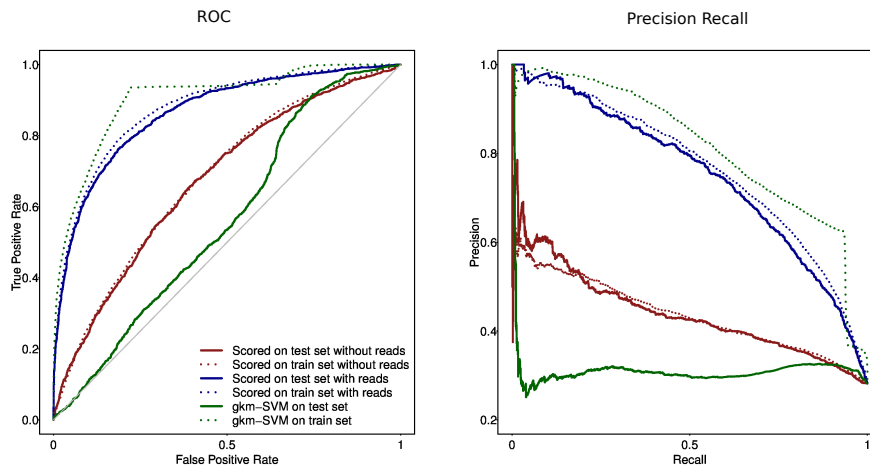


Figure 4.7: Performance of ExoDiversity-based classifier and gkm-SVM on distinguishing between Sox2-bound and Oct4-bound regions. The programs were trained on 80% of the dataset, with ExoDiversity being supplied the reads at those regions as well.

We set up binary classification using Sox2 and Oct4 as our two classes of data. Standard five-fold cross-validation was applied, where ExoDiversity was trained on each fold of each TF separately using their respective summit regions and reads. The held out sets for both the TFs were scored using the best models learned by ExoDiversity for Sox2 and Oct4. We used two scoring systems: one which used the reads ($\mathbf{R}^+$ and $\mathbf{R}^-$) in addition to the DNA $\mathbf{X}$, and the other used only $\mathbf{X}$. Considering Sox2 as the positive and Oct4 as the negative class, we plot the

ROC

Average auROC	Score with reads			Score without reads			gkm- SVM
	BIC model	1 mode	3 modes	BIC model	1 mode	3 modes	
On test set	0.87	0.82	0.81	0.67	0.62	0.64	0.56
On train set	0.88	0.83	0.82	0.68	0.62	0.65	0.90

Precision Recall

Average auPRC	Score with reads			Score without reads			gkm- SVM
	BIC model	1 mode	3 modes	BIC model	1 mode	3 modes	
On test set	0.76	0.67	0.67	0.43	0.39	0.41	0.31
On train set	0.78	0.67	0.68	0.44	0.39	0.41	0.82

Figure 4.8: The average areas under the ROC and PR curves for classifying Sox2 versus Oct4 across five folds are shown. The BIC-based model performs the best with or without reads on the test data. On the training set, gkm-SVM performs the best.

receiver operating characteristic (ROC) curve and the precision-recall curve for one representative fold (Fig: 4.7). In the test set, when both the reads and the DNA information are used to score the sequences, as expected, the performance in distinguishing the sets is higher. But, one may attribute that to the difference in the read counts, which would naturally be high for Sox2 at the Sox2 bound regions and vice versa. Indeed, when only the sequence component is used for scoring, the performance goes down. In spite of this, it is still higher than the state-of-the-art SVM-based classifier, gkm-SVM. It goes beyond single PWMs and using gapped k-mers can distinguish between functional regulatory elements to a remarkable degree [Ghandi et al., 2014]. This shows that, although we are finally using only the sequence features for classification, the reads data used in training by ExoDiversity is critical in identifying those features. To test whether the BIC selected model is not over-fitting by identifying excessive modes, we also scored the held out sequences with the models with fewer modes. Those models also do better than gkm-SVM but worse than the BIC selected model Fig: 4.8. We also note that the performance of gkm-SVM on the training data is significantly higher

than that on the test set, while the difference is lower in the models learned by ExoDiversity (dotted curves in Fig: 4.7). This further confirms that the model is not overfitting the training data.

4.9.4 ExoDiversity finds diverse modes in FoxA1 and CTCF datasets

ExoDiversity on FoxA1 data

We applied ExoDiversity to the complete FoxA1 dataset containing 40013 sequences in the MCF-7 cell line. It finds 13 modes (Fig: 4.9), out of which 11 of them are variants of the FoxA1 motif. The first two modes have an additional ‘C’ three (mode 1) or two (mode 2) bases upstream of the canonical motif. Of these, the second mode has been shown to be enriched in regions bound by both FoxA1 and its paralog FoxA2 [Bochkis et al., 2012]. The position of the reads on the positive strand also moves correspondingly upstream when compared to mode 5, which is the best match to the canonical motif (box in Fig: 4.9). Similarly, modes 7 and 9 lack a strong ‘A’ at the start and the corresponding cut in the positive strand reads goes a base deeper towards the motif. Notably having more informative positions in the motif, though does result in small shifts in the read patterns, does not necessarily correlate with more reads. This is counter-intuitive, because one expects to see higher occupancy of the TF at motifs with more information content. We surmise that these regions may also be bound by other proteins (example: FoxA2 in mode 2) and due to the resulting competition between FoxA1 and other TFs for the site of binding, fewer instances of FoxA1-DNA complex are pulled down causing fewer reads at these sites. The last two modes are a low complexity region (mode 13) and the CTCF motif (mode 12) both of which are associated with very few reads. CTCF is known to interact with FoxA1 and indirect binding between them has been suggested before [Yamada et al., 2019].

ExoDiversity on CTCF data

CTCF is a protein with 11 zinc fingers and while it binds to a core motif shown in Fig: 4.3 it uses its zinc fingers to bind to DNA in multiple ways. Two specific motifs on each side of the core motif have been shown to provide additional stability to the CTCF-DNA complex [Nakahashi et al., 2013]. In order to capture these, we

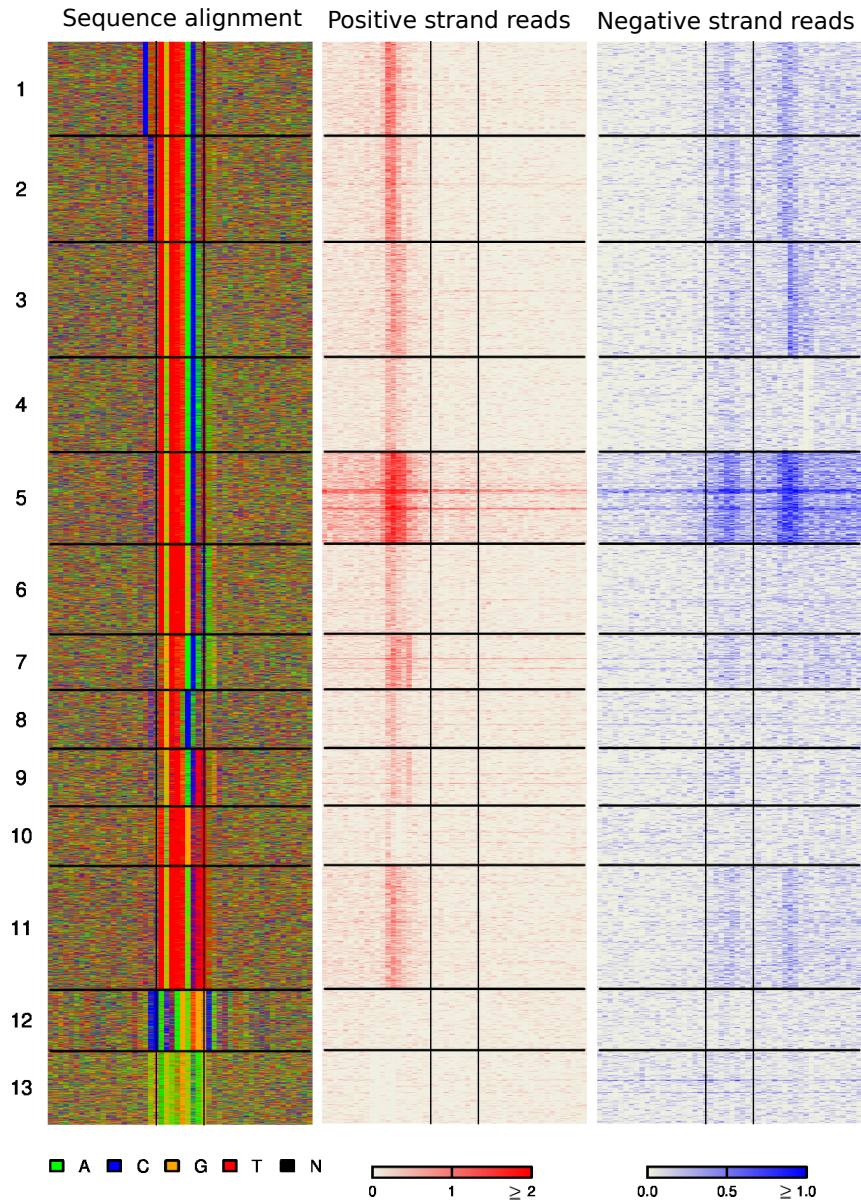


Figure 4.9: ExoDiversity finds 13 modes in the FoxA1 dataset. The DNA sequence and the reads in the positive and negative strands in a 500 bp neighbourhood are shown. The central two lines in each plot mark the site corresponding to the FoxA1 motif.

applied ExoDiversity to the CTCF set (55885 sequences) in the same MCF-7 cell line. Since we expected to identify wider motifs containing the core CTCF motif

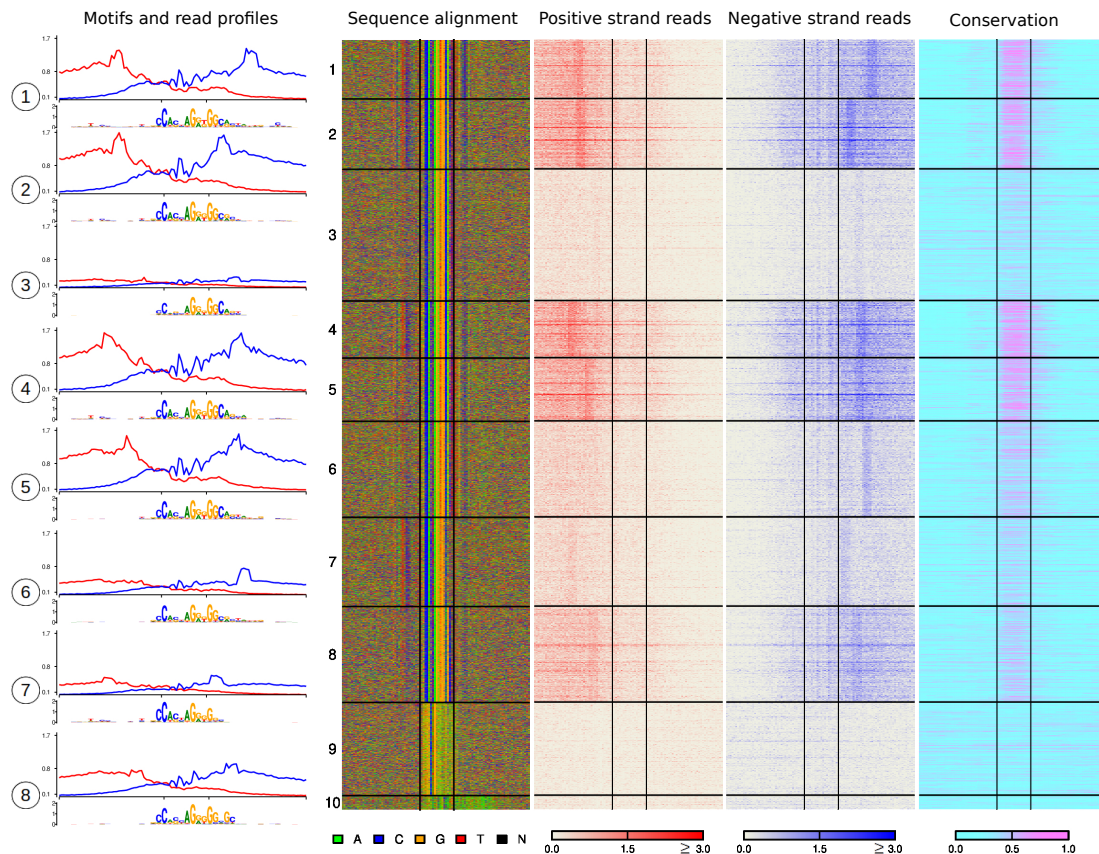


Figure 4.10: ExoDiversity finds 10 modes in the CTCF dataset. The read profiles from both the positive (red) and negative (blue) strands over a 100 bp window about the motifs along with the motifs learned for the 8 modes resembling the canonical CTCF motif are shown in the left. The DNA sequence, the two sets of reads and the sequence conservation in terms of phastCons scores in the 100 bp neighbourhood are shown. The X-axes of the read profiles are aligned corresponding to the 19 bp JASPR CTCF motif which is also indicated by the two central lines in the four heatmap images and the two ticks in the read profiles. The read profiles are over a 100 bp region, while the motifs shown below them are over a 40 bp region.

along with the additional motifs, we started ExoDiversity with an initial motif width of 40. It finds 8 modes that resembles the CTCF core motif with differing strengths of the additional flanking motifs Fig: 4.10. For example, modes 1 and

2 both have the upstream motif, but mode 2 has a much weaker downstream motif and the negative strand reads are also correspondingly shifted towards the core in this mode. Similarly, the core motif misses a few downstream bases in mode 7, but is associated with a band of reads on both strands, albeit weaker. In contrast, although the core motif is fairly conserved in mode 3, the bands of reads are missing. The only other difference between the two modes is the presence of the upstream motif in mode 7. Furthermore, sequence conservation varies across the modes: modes 1, 2, 4, 5 are significantly more conserved, while modes 3, 6-8 are less conserved (Wilcoxon p-value $\leq 10^{-10}$), which generally correlates with the number of reads associated with the modes. It does not necessarily correlate with the information content of the core motif.

4.9.5 ExoDiversity results on the Twist dataset

We next applied ExoDiversity to 29353 ChIP-nexus regions from the E-box binding TF, Twist, in early *Drosophila* embryos [He et al., 2015]. It finds a total of 25 modes out of which 9 of them match the E-box (Fig 4.11). The authors had shown that while overall, three peaks of reads are visible around the E-box, the individual peaks are specific to certain E-box core motifs: the first peak is enriched about 11 bp upstream of CAYATG and the second peak is enriched 2 bp inside CAKCTG. That analysis was done using the top 200 most enriched regions for each E-box variant. ExoDiversity also identifies similar correlations from the complete dataset without any prior information about the E-box. It finds additional variations of small shifts in the first peak, which co-occur with small differences within and around the E-box.

Genomic regions flanking the E-box have been known to possess specific local DNA shapes that correlate with binding [Gordân et al., 2013]. The propeller twist of the DNA, which denotes the rotation of one base with respect to its base-pair, has been shown to be important for other proteins recognizing E-boxes [He et al., 2015]. We therefore plotted computationally predicted propeller twist (proT) [Chiu et al., 2015] at the DNA for each mode. Motifs at modes 1, 2, 4, 6-8, which correspond to the CAYATG E-box have a lower propeller twist than the motifs at modes 3, 5, and 9 which correspond to the CAKCTG E-box. Since these DNA shape measures are based on the sequence, it is not surprising that the two variants have prominently different shape profiles. However, the upstream flanks,

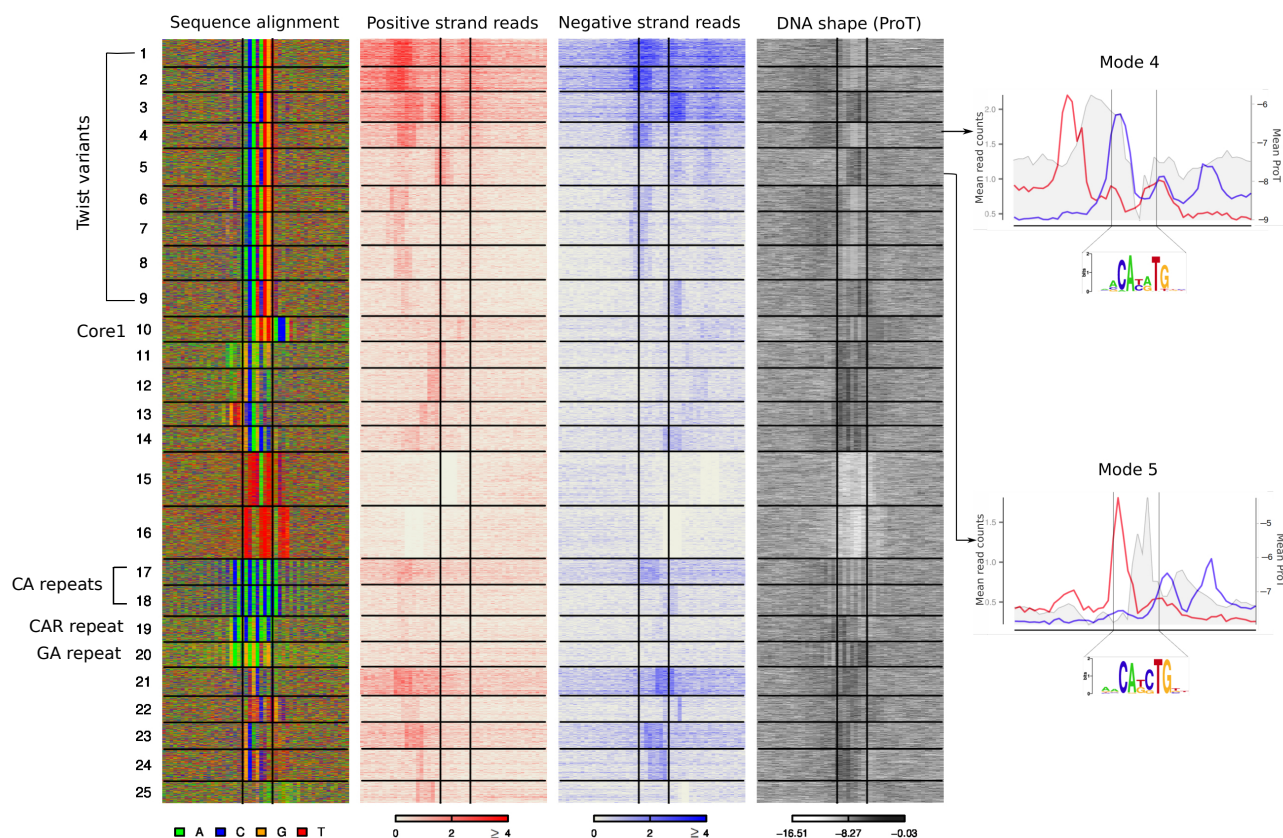


Figure 4.11: ExoDiversity finds 25 modes in the Twist dataset. Modes 1-9 correspond to the Twist motifs, mode 10 matches a *Drosophila* core promoter motif and modes 17-20 are di/tri nucleotide repeat patterns. The DNA sequence, the two sets of reads and the DNA shape for propeller Twist in the 50 bp neighbourhood are shown with the central two lines marking the site corresponding to the Twist motif. Two representative modes for the two Twist variants are shown on the right. The mean positive and negative reads are shown in red and blue plots respectively (left Y-axis). The mean propeller twist values are shown in grey (right Y-axis).

which do not have a strong motif signal, possess a higher propeller twist at the CAYATG modes than the CAKCTG modes. Curiously, the signal of reads on the positive strand aligns 7-10 bp upstream of the positions with high propeller twist across the E-box modes. Similarly, the signal on the negative strand aligns 7-10 bp downstream of the high propeller twist regions (Fig: 4.11). A high propeller

twist indicates DNA rigidity, which might explain the differing positions of the read signals and their intensities at the nine modes.

The Twist modes explain less than half of the total reported regions. ExoDiversity finds 16 modes that do not have E-box motifs, but many of which are well-studied fly motifs. For instance, mode 10 matches the *Drosophila* core promoter motif 1 [Ohler et al., 2002]. Indeed, the median distance of the sequences in that mode from the closest transcription start site is 25 bp (Wilcoxon p-value $\leq 10^{-10}$ when compared to the rest of the set). On the other hand, modes 17-20 contain specific di and tri-nucleotide patterns which are characteristic of fly enhancers [Brittain et al., 2014, Yáñez-Cuna et al., 2014]. Considering that these non-Twist motifs have far fewer reads overall, they are likely indirect binding events involving Twist.

4.10 Discussion

ExoDiversity learns a joint distribution over the DNA regions and the corresponding read counts obtained in an exonuclease-based ChIP experiment to resolve protein-DNA binding footprints. We applied ExoDiversity on a range of ChIP-exo and ChIP-nexus datasets and observed that it detects variations within the DNA motif which correlate with read signals in the immediate neighbourhood. We further noticed that some of the detected modes have distinct characteristics in terms of higher/lower sequence conservation, specific DNA shapes and proximity to transcription start sites. We note that while we cannot definitively establish whether the variations in all the detected modes are biologically relevant, these modes together are needed to describe the whole dataset. The ROC curves support the fact that the models are not over-fitting. Indeed, ExoDiversity is unique in its goal of finding an explanation for each reported region. It therefore identifies potential cofactors like CTCF and Esrrb in Oct4, which have relatively poor signals and make up only a small fraction of the total data.

In this implementation and subsequent application of ExoDiversity to datasets, we have used non-informative priors for all parameters in the model. However, the Bayesian formulation allows one to incorporate prior knowledge if available. For example, one could use motif-priors corresponding to its family-motifs if the structural family of the profiled TF is known. This might shed more light on

potential indirect binding modes. We also note that in its objective to explain the entire dataset ExoDiversity finds some modes which have low information content in the DNA motif but high read signals in either or both the strands. These motifs which are not obvious TF binding sites should be explored for possible experimental biases, especially when they are accompanied by strong read signals.

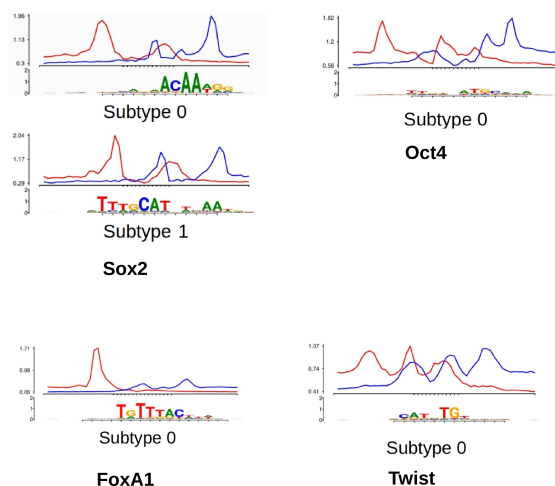


Figure 4.12: In Sox2 dataset, ChExMix finds 2 motifs explaining 75% of the data, resembling the Sox2 and Oct4::Sox2 respectively, while in the Oct4 dataset, it finds a single motif that resembles a weak Oct4::Sox2 motif with low intensity reads explaining 76% of the data. In case of the FoxA1 dataset, ChExMix finds a single FoxA1 motif and an additional subtype characterized only by reads. In the Twist dataset, ChExMix finds a single weak E-box motif.

ChExMix is another method that attempts to find subtypes based on motifs and reads. However, it is different from ExoDiversity in its problem formulation. Firstly, it is also a peak or event finder and therefore at times detects binding events well over 500 bp away from the input peak region. Secondly, it treats each subtype independently in such a way that it allows for multiple subtypes to occur in a region, as well as some regions to not have any binding event. Finally, motif discovery is conducted only in the top few regions of each subtype. Indeed, most downstream motif based analyses focus on the top X regions, either because of the time taken to analyze the full set or because those regions have a higher likelihood of containing the expected motif. ChExMix uses these strong motifs for subsequent scanning or as priors, but this causes the final motifs, those made

by all the final predicted binding events, to be weaker (Fig: 4.12), giving an average description of the dataset. In contrast, since ExoDiversity maximizes a joint distribution, its modes are not necessarily guided only by strong reads or enriched regions. This, we believe, is critical because regions with lower read counts are not necessarily uninformative. At times the motif slightly differs from the canonical one, but affects exonuclease activity, or it maybe a motif of a different protein, but captured due to its association with the target protein. We also note that unlike ChExMix we focus on a narrow window of reads (5 bp here) in the immediate neighbourhood of the motif. As a result, ExoDiversity’s modes are not only characterized by exonuclease cuts differing at the level of a few nucleotides, but also have stronger corresponding motifs when compared to ChExMix. We also ran ExoDiversity with larger read width values (10, 50, 100) and we noticed that while the results for read width 10 are not too different from read width 5, the motifs get progressively weaker for larger read widths (Fig: 4.13). This is because the read distribution start dominating over the motif information, effectively making a larger contribution to the joint distribution in Eq: 4.1. ExoDiversity gives the user the option of supplying a read width of their choice, and in the later version of ExoDiversity we have modified the framework to include the procedure for learning the read width in a manner similar to the motif width.

Avsec et al. [Avsec et al., 2021] have built a deep learning-based approach, BPNet, that trains on ChIP-nexus profiles and predicts profiles for unseen DNA regions to a remarkable degree. Their primary goal is to decipher the syntax or the architecture of TF binding sites within regulatory regions, and not specifically to resolve the footprints. We note that ExoDiversity is currently limited to identifying just one mode in each region. Changing the model to identify modules instead of modes of binding, where each module is characterized by a combination of binding modes [Biswas and Narlikar, 2021b], will be useful in investigating the binding activities at the regulatory regions. Further, ExoDiversity approximates the read distributions as a product of independent Bernoulli distributions conditional on the mode. The assumption here is that the cleaving activity of the exonuclease or the location where the cut happens, depends only on the type of complex forming at the DNA. This effectively results in an extension of the PWM model, which also assumes independence between positions of the binding site. However, other approximations as well as different distributions which overcome this assumption

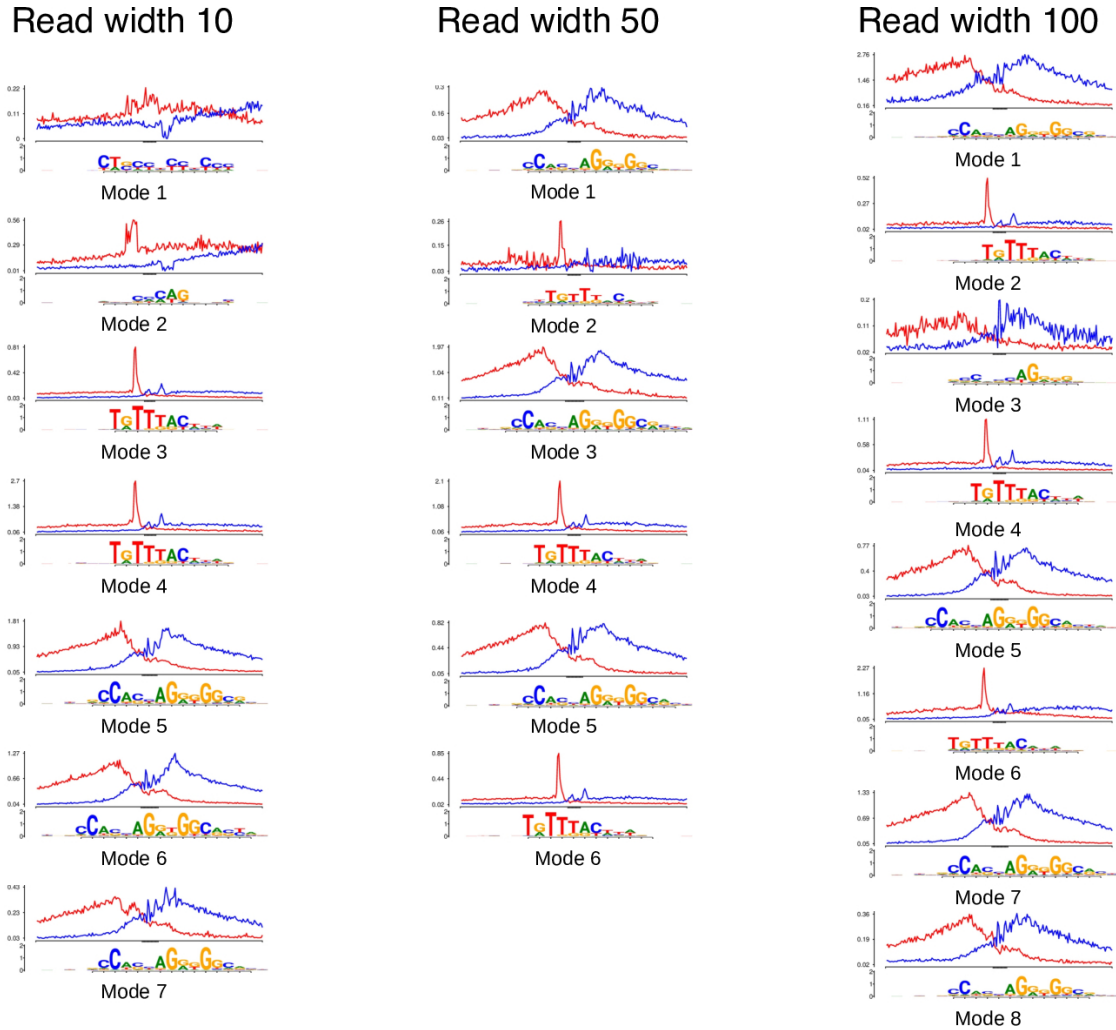


Figure 4.13: ExoDiversity was run with three different values of read widths on the mixed dataset. The result of using $w_k^{+/-} = 10$ is not very different from using a value of 5 (Fig: 4.3). However, for values of 50 and 100, the motifs become weaker. The average read plots are computed over a 200 bp neighbourhood around the motif: the ticks correspond to the motif positions shown below the plot in each case.

can be explored to study different aspects of TF binding. In fact, we have taken advantage of the structure of ExoDiversity's framework to model the discrete read count values as independent Poisson distributions at each position within the read window and have built a separate framework for that. Also, a noteworthy point in case of ExoDiversity, is that its formulation can be easily extended to other sources of information which are of high resolution and are indicative of DNA motifs in

the neighbourhood such as transcription start site data or sequence conservation.

We note that ExoDiversity relies heavily on the read information at the nucleotide level, and thus it is not applicable to the more widely used ChIP-seq protocol, as it reports wider read distributions. However, several modifications have been suggested recently to the original ChIP-exo protocol which may lead to a broader adoption of the technique as an alternative to ChIP-seq [Rossi et al., 2018]. In such high precision experiments, ExoDiversity can be easily used to gain more insights into the modes of binding of proteins of interest.

Chapter 5

Detecting *cis*-regulatory diversity in DNA regulatory regions

With advancements in sequencing technologies, a plethora of high throughput sequencing experiments have been developed to explore a variety of biological activities related to gene regulation in a genome wide manner. These biochemical activities include transcription factor (TF)-DNA binding, enhancer activity, 3D organization of the chromatin, open chromatin regions and many more. An important aspect of these investigations include understanding the underlying sequence components or motifs that are responsible for the measured activity. Typically, it is not just a single motif, but a combination of motifs bound by cooperatively acting proteins that determine the role of a particular genomic region. Moreover, as one can envisage, these regions can be diverse, with their diversity being governed by the different sets of TF binding sites present in them. Current approaches do not take into account this combinatorial activity of the TFs in defining the function of a region. We present here a statistical framework CISDIVERSITY which models regions as modules characterized by combinations of motifs. The model learns the set of motifs present in the data and their distributions over each module, while simultaneously clustering the regions into separate modules based on the motifs present in them.

Here we will first discuss the model for CISDIVERSITY followed by our evaluation of the model using simulated datasets. Then we will look at some of the results that we got by applying CISDIVERSITY to data obtained from multiple types of high throughput sequencing experiments. After this, we will briefly discuss some

of the advantages of CISDIVERSITY and a few additions that can be made to the existing algorithm.

The framework for CISDIVERSITY was written and tested by Leelavati Narlikar. This work is published in Biswas and Narlikar [2021b] and most of the text and images in this chapter are taken from the paper.

5.1 Introduction

Various biochemical activities occurring across the genome are mapped using high throughput sequencing experiments. Some of them include protein-DNA binding events([Johnson et al., 2007], [Vogel et al., 2007]), open chromatin regions([Boyle et al., 2008], [Giresi et al., 2007]), interacting chromatin domains([Fullwood et al., 2009], [Lieberman-Aiden et al., 2009]), active transcription start sites ([Shiraki et al., 2003]) and many more. One of the major goals of these efforts is to identify the sequence signatures that might be responsible for the specific activity which is being measured. Although, these methods detect regions that have are involved in a specific activity, the same sequence signatures may not be accountable for it at all the locations. For example in an assay for transposase accessible chromatin with sequencing (ATAC-seq) or DNase I hypersensitive sites sequencing (DNase-seq) which identify open chromatin regions, the reason behind a region to be open may be one of several: it can be an active promoter, or an enhancer or an insulator or even a matrix-attachment region. Therefore, the relevant sequence components at these different types of regions will also be different. Sometimes, this diversity is less pronounced, as in the case of chromatin immunoprecipitation (ChIP) experiments which are used to identify regions bound by a specific transcription factor. In fact, we have seen that these experiments report a miscellaneous set of genomic regions which might contain the direct binding site of the target TF or might contain the binding sites of other proteins that either interact with the target TF via an intermediate TF or might be cofactor or perhaps TFs that are in spatial proximity to the target TF.

Identification of these sequence components is generally carried out by finding overrepresented motifs in the data, either by sequentially scanning individual motifs from a known database or by using *de novo* motif finding. However, in this way we might miss out on motifs present in a small set of regions and are thereby

not statistically overrepresented in the entire set. Some of the recent methods have taken this into consideration and posed this as a mixture problem, with each component of the mixture being enriched with a potentially different motif ([Agrawal et al., 2018], [Eggeling et al., 2017], [Mittra et al., 2018]). Even so, these approaches do not take into account combinations of motifs that might be critical to drive the biochemical activity at the region. Further, there might multiple different combinations of motifs in the reported regions, with each type of motif-combination explaining a fraction of the regions.

Here we present our model CISDIVERSITY which aims to divide the dataset of reported regions in terms of motifs and their combinations, all of which is learned *de novo*. The idea is modelled based on the concept of topic modelling in computer science, where the goal is to cluster documents (which are DNA regions here) into different topics (which are various functions or modules) based on word frequencies (that correspond to DNA motifs here). In fact, our mixture model has an added layer of unknowns since unlike topic modelling where words are established a priori, here we also have to learn the motifs present in the dataset along with their distribution over different modules.

5.2 CISDIVERSITY model

Consider a set of regions reported in a high throughput experiment that share a common property in a specific context. The regions might display the property due to r different mechanisms, and each region contains one of these r mechanisms encoded in their DNA sequence. A ‘mechanism’ is represented as a regulatory module. Each of these r modules is characterized by the presence or absence of m different motifs. Each level of the model in terms of the motifs and modules is described below in a bottom-up manner:

1. Motif level:

Let there be m motifs present across the complete set of reported regions. These motifs are distributed across the modules with differing probabilities. The width of each motif is given by w_k and each motif is modelled as a product of w_k categorical distributions over the four nucleotides ‘A’, ‘C’, ‘G’, ‘T’ represented in the form of a position weight matrix (PWM). ϕ_k is used to denote these PWMs, where $\phi_{k,u}(B)$ gives the probability of finding nucleotide

B at position u within motif k , $B \in \{A, C, G, T\}$. Given any DNA sequence of length w_k : $d^1, d^2 \dots d^{w_k}$, the likelihood of it being an instance of motif k is $\prod_{u=1}^{w_k} \phi_{k,u}(d^u)$.

2. Module level:

We assume at least r modules are present in the entire set of regions. Each module is modelled as a product of Bernoulli distributions over the m motifs: module j has the probability $f_{j,k}$ of containing motif k and a probability of $(1 - f_{j,k})$ of not containing it.

3. Region level:

The set of regions reported by the experiment is considered as n DNA sequences which are denoted by $\mathbf{X}_1, \mathbf{X}_2 \dots \mathbf{X}_n$. Each sequence $\mathbf{X}_i$ is of length l_i with $X_{i,t} \in \{A, C, G, T\}$ and $1 \leq t \leq l_i$. Parts of each sequence $\mathbf{X}_i$ are instances of a subset of m motifs and the other parts which are not motif instances are considered as ‘background’ and modelled using a second order Markov chain over the nucleotides whose parameters are denoted by ϕ_0 . If motif k is present in sequence $\mathbf{X}_i$ then $Z_{i,k}$ denotes the starting point of the motif k in sequence $\mathbf{X}_i$ with $1 \leq Z_{i,k} \leq l_i - w_k + 1$ and if it is not present in $\mathbf{X}_i$ the $Z_{i,k} = -1$. Motifs do not overlap each other, and each motif can be present only once in a sequence. Each sequence $\mathbf{X}_i$ belongs to a module I_i and the distribution of the sequences across the r modules is modelled using a categorical distribution γ where γ_j is the probability of a sequence having a module identity j .

We are only given the set of regions $\mathbf{X}$ from which we first compute the background Markov model and the lengths of the sequences $\mathbf{l}$. CISDIVERSITY has an option where we can specify whether the motifs can occur on both the forward and reverse strand, or on the strand given in data only. The default being the former, we append the reverse complement of each sequence with itself, and their lengths are effectively doubled. We assume that the structure of the model is known and an upper bound to the values of r and m are given. The algorithm learns the unknown parameters of the model, which are: $\theta_m = \{\phi_{1,2,\dots,m}, \mathbf{w}, \mathbf{Z}, \mathbf{f}, \gamma, \mathbf{I}\}$. Considering

these assumptions, the likelihood of a sequence $\mathbf{X}_i$ can be computed as:

$$P(\mathbf{X}_i | \boldsymbol{\theta}_m) = \prod_{\substack{k=1 \\ Z_{i,k} \neq -1}}^m \prod_{u=0}^{w_k-1} \phi_{k,u}(X_{i,Z_{i,k}+u}) \times \prod_{\substack{t=1 \\ X_{i,t} \notin \text{any motif}}}^{l_i} \phi_0(X_{i,t} | X_{i,t-1}, X_{i,t-2}) \quad (5.1)$$

Here the first term in the above equation denotes the probability associated with all the motifs that are present in $\mathbf{X}_i$ while the second term is formed by the probability of those parts of the sequences that do not overlap any motif using the background Markov model. The default order of the Markov model is 2 in CISDIVERSITY which is built separately for each sequence to better capture the heterogeneity present in the eukaryotic sequences [Narlikar, 2013]. However, this is a parameter whose value can be set by the user as per convenience.

The likelihood of the entire data is thus calculated as:

$$P(\mathbf{X} | \boldsymbol{\theta}_m) = \prod_{i=1}^n P(\mathbf{X}_i | \boldsymbol{\theta}_m) \quad (5.2)$$

5.2.1 Calculating the posterior probability

Given a set of sequences $\mathbf{X}$, the goal is to find the optimal values of $\boldsymbol{\theta}_m$ that maximizes the posterior distribution. It can be calculated as follows:

$$\begin{aligned} P(\boldsymbol{\theta}_m | \mathbf{X}) &\propto P(\mathbf{X} | \boldsymbol{\theta}_m) \cdot P(\boldsymbol{\theta}_m) \\ &= P(\mathbf{X} | \boldsymbol{\theta}_m) \cdot P(\mathbf{Z}, \mathbf{I}, \boldsymbol{\gamma}, \mathbf{f}, \boldsymbol{\phi}, \mathbf{w}) \\ &= P(\mathbf{X} | \boldsymbol{\theta}_m) \cdot P(\mathbf{Z} | \mathbf{I}, \boldsymbol{\gamma}, \mathbf{f}, \boldsymbol{\phi}, \mathbf{w}) \cdot P(\mathbf{I} | \boldsymbol{\gamma}, \mathbf{f}, \boldsymbol{\phi}, \mathbf{w}) \\ &\quad P(\boldsymbol{\gamma}) \cdot P(\mathbf{f}) \cdot P(\boldsymbol{\phi}) \cdot P(\mathbf{w}) \end{aligned} \quad (5.3)$$

$$= P(\mathbf{X} | \boldsymbol{\theta}_m) \cdot P(\mathbf{Z} | \mathbf{I}, \mathbf{f}) \cdot P(\mathbf{I} | \boldsymbol{\gamma}) \cdot P(\boldsymbol{\gamma}) \cdot P(\mathbf{f}) \cdot P(\boldsymbol{\phi}) \cdot P(\mathbf{w}) \quad (5.4)$$

$$\begin{aligned} &= \left(\prod_{i=1}^n P(\mathbf{X}_i | \boldsymbol{\theta}_m) \left(\prod_{k=1}^m P(Z_{i,k} | f_{I_i,k}) \right) P(I_i | \boldsymbol{\gamma}) \right) \cdot \\ &\quad P(\boldsymbol{\gamma}) \cdot P(\mathbf{f}) \cdot P(\boldsymbol{\phi}) \cdot P(\mathbf{w}) \end{aligned} \quad (5.5)$$

$$\begin{aligned} &= \left(\prod_{i=1}^n P(\mathbf{X}_i | \boldsymbol{\theta}_m) \left(\prod_{k=1}^m f_{I_i,k} \delta[Z_{i,k} \neq -1] + (1 - f_{I_i,k}) \delta[Z_{i,k} = -1] \right) \right) \cdot \\ &\quad P(\boldsymbol{\gamma}) \cdot P(\mathbf{f}) \cdot P(\boldsymbol{\phi}) \cdot P(\mathbf{w}) \end{aligned} \quad (5.6)$$

Here, $\delta[\text{condition}] = 1$ when *condition* = True and 0 otherwise. The simplification of Eqn 5.3 to Eqn 5.4 can be done because of the structure of the model:

the motif start positions $\mathbf{Z}$ are independent of $\gamma, \phi, \mathbf{w}$ when conditional on the module identity $\mathbf{I}$ and its associated probabilities $\mathbf{f}$; $\mathbf{I}$ is independent of $\mathbf{f}, \phi, \mathbf{w}$ when its categorical distribution γ is known. The second term in Eqn: 5.5 is a product of Bernoulli probabilities, which are assumed to be independent. The third term is the categorical probability of the module identity. The rest of terms are independent priors over $\gamma, \mathbf{f}$ and ϕ . The prior over the widths is considered uniform.

5.2.2 Model learning

Appropriate priors over the components of θ_m are defined as follows:

We assume conjugate Dirichlet priors over all categorical parameters.

- Prior over γ , which characterizes the distribution of module identity $\mathbf{I}$, has r equal hyperparameters given by α_{module} whose value is considered as 1 by default.
- Prior over each $f_{j,k}$, which characterizes the distribution of presence or absence of motif k in module j , has hyperparameters defined by α_{YES} and α_{NO} which have default value of 0.1 here.
- Prior over $\phi_{k,u}$, which characterizes the distribution over the 4 nucleotides at position u in motif k , has 4 equal hyperparameters defined by α_{PWM} for all u and k and are all set to 0.1 by default.

Hyperparameters less than 1 assumes extreme final distributions, which fits in with our expectation of having informative motifs with nucleotide probabilities close to either 0 or 1. Likewise, we expect the modules to be distinctly defined by the presence or absence of motifs. However, we do not know the optimal number of modules in which our data can be divided. Thus, we take $\alpha_{\text{module}} = 1$ by default which assumes all distributions are equally likely a priori. Even so, all these values can be changed by the user.

We use Gibbs sampling to draw samples from the posterior distribution with the aim of learning the parameter values that maximize the posterior probability. To speed up the sampling and to reduce the search space, we use collapsed Gibbs sampling and marginalize over $\mathbf{f}, \phi$ and γ while sampling the other unknowns.

Objective function

After applying marginalization, our objective function reduces to

$$\begin{aligned}
 P(\boldsymbol{\theta}_m \mid \mathbf{X}) &= P(\mathbf{Z}, \mathbf{I}, \mathbf{w} \mid \mathbf{X}) \\
 &\propto P(\mathbf{X} \mid \mathbf{Z}, \mathbf{I}, \mathbf{w}) \times P(\mathbf{Z}, \mathbf{I}, \mathbf{w}) \\
 &= P(\mathbf{X} \mid \mathbf{Z}, \mathbf{I}, \mathbf{w}) \times P(\mathbf{Z} \mid \mathbf{I}, \mathbf{w}) \times P(\mathbf{I} \mid \mathbf{w}) \times P(\mathbf{w}) \\
 &= P(\mathbf{X} \mid \mathbf{Z}, \mathbf{I}, \mathbf{w}) P(\mathbf{Z} \mid \mathbf{I}) P(\mathbf{I}) P(\mathbf{w})
 \end{aligned} \tag{5.7}$$

We iteratively sample for the entities $Z_{i,k}$, I_i , w_k for each i and k .

Sampling of $Z_{i,k}$

Assuming that a motif is equally likely to be present anywhere in the sequence, we compute the score for binding of motif k in sequence $\mathbf{X}_i$ at position t by dividing the posterior probability by the background probability of the motif if it was present at that position. Thus, the sampling expressions for $Z_{i,k}$ can be written as:

$$P(Z_{i,k} = t \mid \mathbf{X}, \mathbf{Z}_{-[i,k]}, \mathbf{I}, \mathbf{w}) \propto \hat{f}_{I_i,k} \times \frac{1}{l_i} \prod_{u=0}^{w_k-1} \frac{\hat{\phi}_{k,u}(X_{i,t+u})}{\phi_0(X_{i,t+u} \mid X_{i,t+u-1}, X_{i,t+u-2})} \tag{5.8}$$

$$P(Z_{i,k} = -1 \mid \mathbf{X}, \mathbf{Z}_{-[i,k]}, \mathbf{I}, \mathbf{w}) \propto 1 - \hat{f}_{I_i,k} \tag{5.9}$$

The expression in Eqn: 5.8 gives the probability of a site of motif k present at position t in sequence $\mathbf{X}_i$ where $1 \leq t \leq l_i - w_k + 1$.

Similarly, the expression in Eqn: 5.9 gives the probability that no site of motif k is present in $\mathbf{X}_i$.

$\hat{\phi}_k$ and $\hat{f}_{I_i,k}$ are the posterior probabilities after integrating out ϕ_k and $f_{I_i,k}$ from Eqn: 5.7 when sampling $P(Z_{i,k} \mid \mathbf{X}, \boldsymbol{\theta}_m \setminus Z_{i,k})$ assuming Dirichlet prior over both:

$$\hat{\phi}_{k,u}(B) = \frac{\alpha_{\text{PWM}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[Z_{p,k} \neq 1] \cdot \delta[X_{p,Z_{p,k}+u} = B]}{4 \cdot \alpha_{\text{PWM}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[Z_{p,k} \neq -1]}$$

and similarly:

$$\hat{f}_{I_i,k} = \frac{\alpha_{\text{YES}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[Z_{p,k} \neq -1] \cdot \delta[I_p = I_i]}{\alpha_{\text{YES}} + \alpha_{\text{NO}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = I_i]}$$

The first term in Eqn: 5.8 is the contribution motif k makes to the current module I_i and the second term is similar to the standard collapsed Gibbs sampling applied to motif discovery with the counts computed only from those $\mathbf{X}_p$ that contain a motif k in the current iteration.

Sampling of I_i

The sampling expression for I_i can be derived by collapsing γ and $\mathbf{f}$ from Eqn: 5.7:

$$P(I_i = j \mid \mathbf{X}, \mathbf{Z}, \mathbf{I}_{[-i]}, \mathbf{w}) \propto \hat{\gamma}_j \times \prod_{k=1}^m \delta[Z_{i,k} \neq -1] \cdot \hat{f}_{j,k} + \delta[Z_{i,k} = -1] \cdot (1 - \hat{f}_{j,k}) \quad (5.10)$$

where $\hat{\gamma}$ is the posterior probability after integrating out γ

$$\hat{\gamma}_j = \frac{\alpha_{\text{module}} + \sum_{\substack{p=1 \\ p \neq i}}^n \delta[I_p = j]}{r \cdot \alpha_{\text{module}} + n - 1}$$

Sampling of w_k

The width of the motif k is sampled in a manner similar to that in [Mitra et al., 2018]. Instead of sampling from all possible values of w_k at one time, the width is allowed to either increase or decrease or remain the same on each side of the motif.

We sample the values of each of $Z_{i,k}$, I_i and w_k iteratively by using the above mentioned sampling expressions. The full posterior probability is computed after each sampling step and the values of the model parameters are stored if the posterior is found to be the maximum so far. If the posterior probability does not increase for a predetermined number of iterations (500 by default) the sampling is stopped and a Hill climbing approach is adopted. It starts from the sample with the highest posterior probability and instead of sampling, we store the parameter

values corresponding to the maximum score at each iteration. The final model is cleaned up: all those motifs that have less than 20 sites in the data are emptied out, and all the modules with less than 20 sequences are combined into one. These lower bound values can however be changed according to user preference. By default, CISDIVERSITY starts from 10 different initializations and reports the final model which scores the best across the 10 trials.

5.3 A sequence model for representing cis-regulatory diversity

5.3.1 Input to CISDIVERSITY

A high-throughput sequencing experiment generally reports a set of genomic regions that are enriched with a particular biochemical activity. We take the DNA sequences corresponding to these genomic regions as our data $\mathbf{X}$ which is the only input to CISDIVERSITY.

5.3.2 CISDIVERSITY model

We assume that r different regulatory modules might be present in the data that are responsible for the activity which was profiled by the experiment, and each module is characterized by the presence or absence of m motifs. The model defines a module as a product of m Bernoulli distributions corresponding to the contributions of each motif toward the module. Different modules will have different Bernoulli distributions over the same motifs, with each module describing a different functionality. A motif is modelled with the standard position weight matrix (PWM). In Fig: 5.1 A, it shows an instance of a simulated dataset of 1000 sequences, where 5 motifs were planted from the JASPAR database [Khan et al., 2018], with specific Bernoulli distributions across 3 modules. For example, motif 1 is present in all sequences of module 2 and in 1/5th of the sequences in module 1, and never in module 3. On the other hand, motif 5 is present in all sequences of module 3, but because the module consists of only about 40 sequences, overall, motif 5 occurs less frequently in the data. The aim of CISDIVERSITY is to learn the m motif parameters, the $r \times m$ Bernoulli distributions and the sequences which

belong to each of the r modules.

Gibbs sampling is used to iteratively sample each of these unknown values with the aim of finding the set of optimal values that maximizes the posterior distribution (Section: 5.2).

5.3.3 Output of CISDIVERSITY

CISDIVERSITY reports the output in 3 parts (Fig: 5.1 B). The first is the set of *de novo* motifs, ordered according to the number of sites that contribute to each motif. The second image in the form of the circle plot gives the general structure of the modules, denoting the contribution of each motif in every module. The color and size of the circles correspond to the proportion of sites for each motif in each module. The third part of the output consists of a heatmap plotting the sequences clustered together as per the identified modules. The image shows the sites that contribute to the PWMs in 4 colors for the nucleotides $\{A, C, G, T\}$. If a site is absent, those nucleotides are shown in black. The modules are ordered according to their size, with the largest one shown at the top. In the case in Fig: 5.1, CISDIVERSITY finds all the motifs and recovers the general module structure effectively.

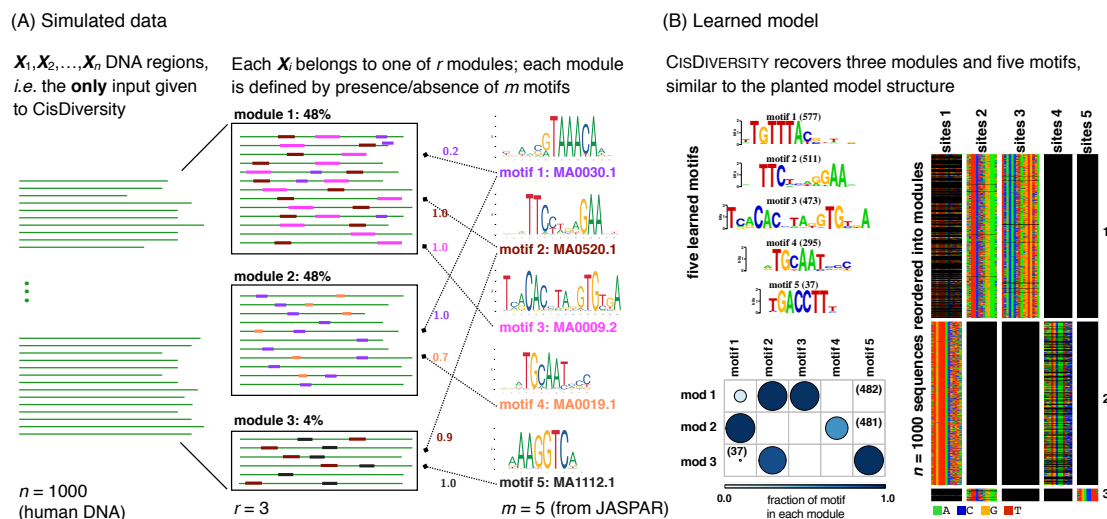


Figure 5.1: (A) DNA regions reported by the experiment are given as input to CISDIVERSITY. In this simulation, $n = 1000$ regions are a mixture of 3 types of regions. Each region is one of $r = 3$ regulatory modules. Each module can be represented in terms of probability of occurrence of $m = 5$ motifs. (B) CISDIVERSITY is run with upper bounds of $r \leq 10$ and $m \leq 20$ and it learns the planted structure of the dataset. The output has 3 components. First is the set of motifs that are learned, second, below is the $r \times m$ Bernoulli distributions describing the learned modules and the third is an image matrix of the data, where each DNA sequence is a row and the sites corresponding to each motif are represented in the column. If a site is absent, then those cells in the column are shown in black. CISDIVERSITY recovers the 5 motifs (motifs 1 and 3 are reverse complements of the planted motifs) and 3 modules to a great extent. The slight variability in the number of sites and sequences in each module is because of the stochastic nature of both the PWMs and the learning algorithm

5.4 Evaluating CISDIVERSITY on simulated datasets

5.4.1 Creating simulated datasets

In order to test the effectiveness of CISDIVERSITY a total of 320 datasets were generated. The JASPAR2018 CORE vertebrate motifs were utilized for this purpose, regardless of their information content. However many of these motifs were similar to each other, so a non-redundant set was created by removing motifs that

matched any earlier ones by using the motif comparison tool TOMTOM [Gupta et al., 2007]. This resulted in a total of 189 JASPAR motifs. Each simulated dataset had 1000 DNA sequences of 200 bp each, randomly selected from the non-repetitive part of the human genome. Each dataset was constructed as an instance of the previously described model. At the motif level, m randomly selected motifs from the set of non-redundant motifs were chosen. At the module level, the Bernoulli probabilities $f_{j,k}$ for each motif j in module k is sampled from Beta distributions with two symmetric or equal parameters α_{YES} and α_{NO} . At the sequence level, the categorical distribution γ which gives the probability of a sequence belonging to one of the r modules, was first sampled using the Dirichlet distribution with an uniform parameter value of $\alpha_{\text{module}} = 1$ and set for the model instance. Now to create the dataset from sequence to motif level, we first used γ to sample the module identity j for that sequence. Then for the m motifs, the $f_{j,k}$ were used to sample whether motif k was to be included in the sequence or not ($1 \leq k \leq m$). If it can be included without overlapping with any other motif, then a motif instance was sampled using the PWM k and randomly planted in the sequence on either strand with equal chance, ensuring that it does not extend into any other planted site.

Multiple datasets were created with the following possible values for m , r and α_{YES} : $m \in \{5, 10\}$, $r \in \{1, 2, 3, 5\}$ and $\alpha_{\text{YES}} \in \{0.01, 0.1, 1.0, 10.0\}$. For each parameter set $\{m, r, \alpha_{\text{YES}}\}$, we created 10 dataset instances resulting in a total of 320 different datasets on which CISDIVERSITY and other programs were evaluated.

We note that two modules can have very similar Bernoulli distributions over the m motifs, making them less separable. However the models were not screened for this possibility because (a) that would be rare with $m \geq 5$ and (b) this would only mean that we are underestimating CISDIVERSITY’s performance.

We ran CISDIVERSITY with $m = 20$ and $r = 10$ for all the datasets. The other parameters were set to default values. MEME ([Bailey et al., 1994]) was run on the datasets using the command line options: `-nmotif 20 -maxsize 100000000 -revcomp -dna -evt 10`. MEME was run with background Markov models of orders 0 through 5 separately, learned on the same set of non-repetitive human DNA sequences used to create the datasets here. Markov order of 3 gave the best F-scores and those results are reported here. HOMER ([Heinz et al., 2010]) was run using the command line options: `-noknown -nogo -basic`. HOMER was

also given the same background file of the non-repetitive human DNA sequences used here. TOMTOM was used to compare the learned motifs with the planted motifs from JASPAR and an E-value of 0.01 was considered a match.

5.4.2 Performance on simulated datasets

We can consider CISDIVERSITY as a joint clustering and de novo motif finding method. We generated many simulated datasets with different characteristics to thoroughly evaluate how well CISDIVERSITY can recover modules and motifs. Now, any clustering method's performance will depend on how easily the clusters can be separated whilst that of a motif finding method will depend on how informative the motifs are. The Bernoulli parameters which are the probabilities of the presence of the motifs in each regulatory module decide the separability of the modules. An extreme probability close to 0 or 1 for the motifs helps us to use them to precisely delineate the modules. A value closer to 0.5 for all $r \times m$ Bernoulli parameters will cause the modules to be less separable from each other. This variation is incorporated in the simulated datasets by sampling the Bernoulli parameters from a beta distribution with symmetric hyper-parameter $\beta \in \{0.01, 0.1, 1, 10\}$. A value much smaller than 1 results in extreme values of the Bernoulli probabilities while a value of 1 leads to uniform sampling between 0 and 1 for the Bernoulli parameters. A value much greater than 1, like 10, results in probabilities closer to 0.5 for the Bernoulli parameters and hence is used to create more "confusing" modules. In case of the motifs, to better emulate reality, we used real motifs from the JASPAR database.

For each combination of (m, r, β) we created 10 sets of data resulting in 320 simulated datasets and CISDIVERSITY was run on each of them. Module recovery was measured by Adjusted Rand Index (ARI) which is commonly used to compare the similarity of two clusterings of the same data. Two identical clusterings get an ARI value of 1, while two random clusterings get a value of 0. As expected, CISDIVERSITY's ability to identify the module structure goes down as the value of the β hyper-parameter increases (shown in Fig: 5.2 A). It is reassuring to observe that the correct npartitioning of the modules are identified although the m and r used in model learning (20 and 10 respectively) are larger than the true number of planted motifs and modules.

There are many techniques that address the problem of *de novo* motif discov-

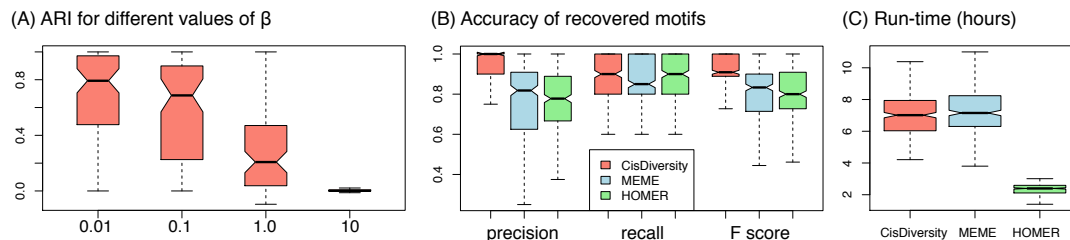


Figure 5.2: Performance of CISDIVERSITY on 320 simulated datasets (A) Recovery of modules: Low values of β results in modules with more extreme (close to 0 or 1) probability distribution of motifs. This is where CISDIVERSITY performs better in recovering the planted modules. For $\beta = 0$ the performance with respect to recovery of modules is similar to that of a random clustering approach. (B) Recovery of motifs: Precision, recall and F-score of recovered motifs across the 320 datasets for 3 different programs. (C) Time taken: All programs were run on a single core.

ery but none that address the issue of simultaneously finding modules, as far as we are aware. However, we utilized default parameters to run 2 widely used algorithms MEME ([Bailey et al., 1994]) and HOMER ([Heinz et al., 2010]) on these datasets to assess how CISDIVERSITY compares with the state-of-the-art *de novo* motif discovery methods. The recovery of a planted motif was evaluated using TOMTOM ([Gupta et al., 2007]) which is a part of the MEME suite. In the Fig: 5.2 B, we see that CISDIVERSITY performs better in terms of both precision and recall. Overall, of the total 2400 planted motifs across the 320 datasets CISDIVERSITY missed about 12.5%, MEME missed 15.2% and HOMER missed 15.1%. The differences in these methods were more pronounced in the specificity (True Negatives / (True Negatives + False Positives)): both MEME and HOMER falsely identified 23.5% and 27.8% additional incorrect motifs, compared to the 4.5% additional motifs identified by CISDIVERSITY. We acknowledge that this comparison is unfair because the data contain modules and no motif discovery tool takes this into consideration. That said, CISDIVERSITY is still very competitive (Fig: 5.3 A) with similar precision and recall rates when different modules are not a component of the data, i.e. all datasets where there is a single planted module ($r = 1$) or when the modules are almost indistinguishable ($\beta = 10$). This demonstrates that CISDIVERSITY is capable of discovering patterns with an accuracy which is

equivalent to the state-of-the-art techniques even when the data does not contain distinct modules. In a similar vein, we examined the approaches' false positive rate in 1000 randomly selected non-repetitive sections of the human genome that lacked any artificially added modules or motifs. CISDIVERSITY, once more identifies fewer motifs than any of the two motif discovery approaches, indicating it has a low false positive rate (Fig: 5.3 B). To compare the running time of the three programs, all of them were run in their serial mode using a single core on an Intel Xeon CPU E5-2630 v3 machine. HOMER is by far the fastest, with there being very less difference between MEME and CISDIVERSITY (Fig: 5.2 C).

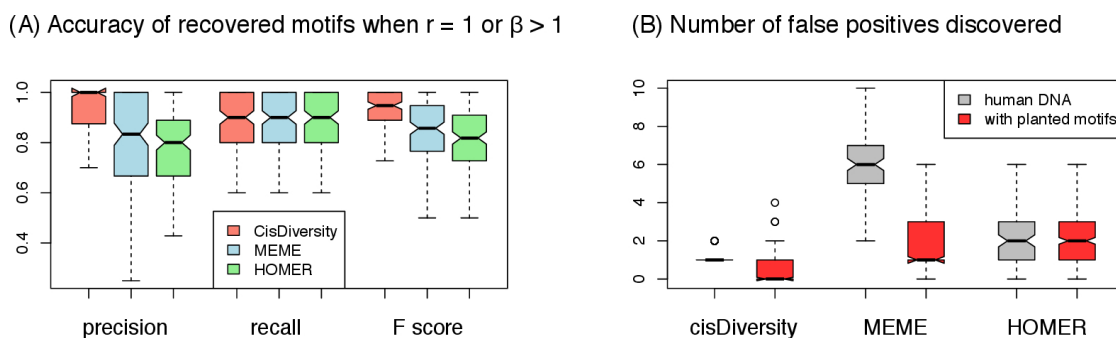


Figure 5.3: When there is a single module ($r = 1$) or the modules are less separable ($\beta = 10$), the performance of CISDIVERSITY in terms of recovering the planted motifs is comparable to that of MEME or HOMER.

5.5 Results

We applied CISDIVERSITY to datasets arising from a range of different types of high-throughput assays. Starting with modules discovered in core promoters, we then investigate ChIP-seq datasets, followed by diverse assays targeting enhancers and accessible regions. The sources for all the real datasets and the processing steps involved are given in details in Chapter: 6. The details of the complete results can be found in the Supplementary section of the paper.

Results guide

Each of the results will be described in the following way:

- In the beginning, we cite the study from which the data is taken. Here, we also state the results of the analysis which has already been done in the study and other information provided by the authors as available.
- This is followed by noting the motifs and modules discovered by CISDIVERSITY.
- Finally, we describe the modules and the motifs forming them and discuss their biological significance. This also includes our hypotheses and reasoning based on the results obtained by CISDIVERSITY.

5.5.1 Core promoter architectures retrieved from genome-wide TSS maps

CISDIVERSITY results on core promoters in fly

Source To test whether CISDIVERSITY can recover known promoter architecture and components, it was initially tested on core promoters. The input was a 200 bp area encompassing 4159 transcription start sites (TSSs) discovered by paired end analysis of fly embryos [Ni et al., 2010]. Ni et al. classified TSSs as narrow where reads map within a small window of 25 bases with a clear peak, broad with reads mapping to a larger window but still having a peak, and weak which would cover all other promoters. They analyzed each set separately for differential enrichment of established fly promoter motifs.

CISDIVERSITY output Since TSS data is strand oriented by nature, motif finding was done only on the given strand. A total of 8 modules and 24 motifs were found.

Discussion In Fig: 5.4A and B we show the 9 motifs that appear in at least a fifth of some module. The majority of these motifs, their module combinations and their spatial preferences in relation to the TSS have already been identified [Ohler et al., 2002]. TATA box is present ≈ 30 bases upstream, initiator (INR) at the TSSs, and downstream promoter element (DPE) is 20 bases downstream. Here, all regions are addressed collectively, and it is important to highlight that the identified modules contain data on both motif enrichment and transcriptional

Fly promoters

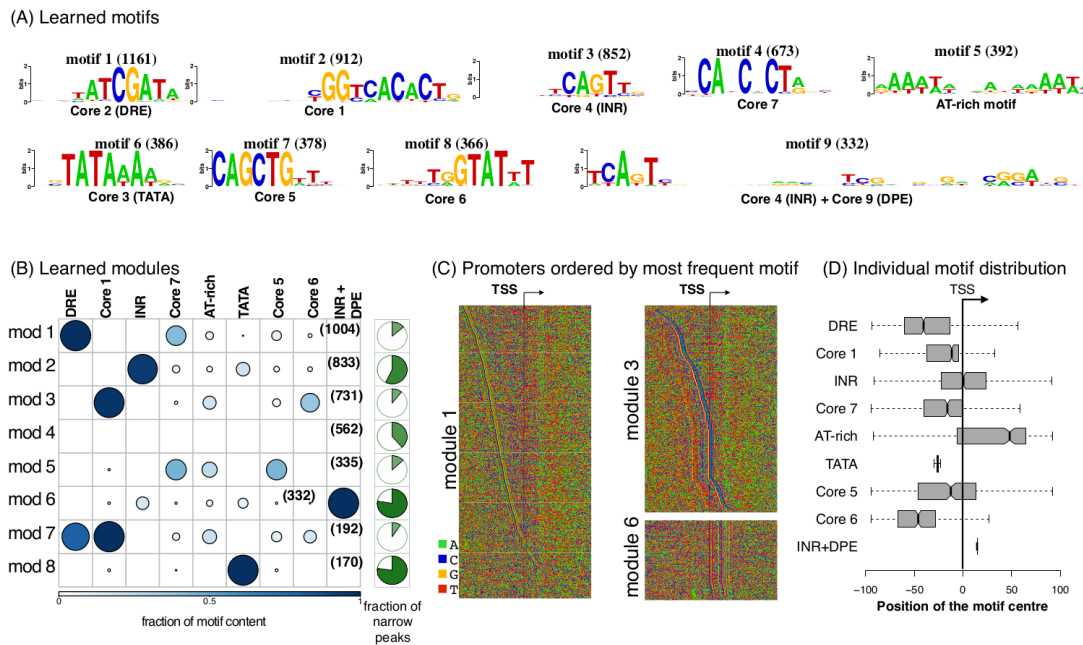


Figure 5.4: Promoter modules in fly (A) CISDIVERSITY identifies 24 motifs in fly promoters, but only 9 motifs that contribute to at least 20% of some modules are shown here. Core motifs' names are according to Ohler et al. (2002) [Ohler et al., 2002]. (B) 8 modules are learned. The fraction of sequences in each module that are focused promoters, i.e. have a narrow peak are shown in green. (C) 3 representative modules are shown as sequence heat maps. They are sorted based on the position of the most frequently occurring motif. Module 1 contains DRE, which is visible owing to the manner in which the sequences were ordered but, Core 7 which is present in > 40% of the sequences is not visible because it has no clear positional relationship with DRE or the TSS. A weak signal corresponding to the AT-rich motif downstream from the TSS is visible. In contrast, module 3 displays a clear relationship between Core 1 and Core 6 motifs: Core 6 is present about 20 bp upstream of Core 1 and especially prominent when Core 1 is close to the TSS. Module 6 is largely composed of INR + DPE, but also contains TATA box especially when INR + DPE is a few bases downstream from the TSS. (D) Each motif has a distinct distribution about the TSS.

activity. Narrow promoters are highly enriched in modules with the TATA box,

INR, and DPE while they are greatly depleted in modules 1, 3 and 5 ($p < 10^{-10}$). Motif 5 has a weak AT-rich sequence and is more prominent in these modules, but it has not been categorized as a promoter element earlier. However, it can be seen in the representative modules 1 and 3 in Fig: 5.4 C, specifically downstream of the TSSs. The INR-DPE motif dominates module 6, however it also contains a small number of promoters with a TATA box as well. The TATA box is enriched in promoters where the INR-DPE is shifted downstream from the TSS. The relevance of this intriguing observation is unknown to us. Narrow promoters are neither significantly enriched nor depleted in module 4. Additionally, it is the sole module with a notably lower tag count, indicating that its main constituents are weak promoters. This maybe the reason why CISDIVERSITY fails to detect any motifs in these locations.

CISDIVERSITY results on core promoters in human

Source We next applied CISDIVERSITY to 14408 non-overlapping human core promoters identified by pooling CAGE tags across (Firth and FANTOM consortium, 2014).

CISDIVERSITY output There are 25 motifs and 13 modules detected altogether (Fig: 5.5 A)

Discussion In contrast to the fly motifs, almost all motifs show both orientations or are palindromic. Motifs 8, 10, 13, 16 are examples of the latter. On the other hand, motifs 1 & 4, 3 & 17, 5 & 7, 9 & 20, 11 & 12, and 15, 18 and 19 are copies of the same motif in both orientations. Their positions with respect to the TSS does not differ (Fig: 5.5 B), suggesting these motifs function in a strand-invariant way. Amongst these, two motifs 3 & 9 do occur more frequently than their reverse complements 17 & 20. The only motifs seen in a single orientation are motif 6 which is present downstream and the TATA box which is present 29-35 bp upstream. The TATA box only appears in a single module that is rich in narrow promoters, although motif 6 is dispersed among multiple modules. It does not resemble the fly DPE, nor could we find evidence in literature supporting its existence as a human promoter element. The human donor splice site consensus sequence is the closest match, however we do not see how that may play a role

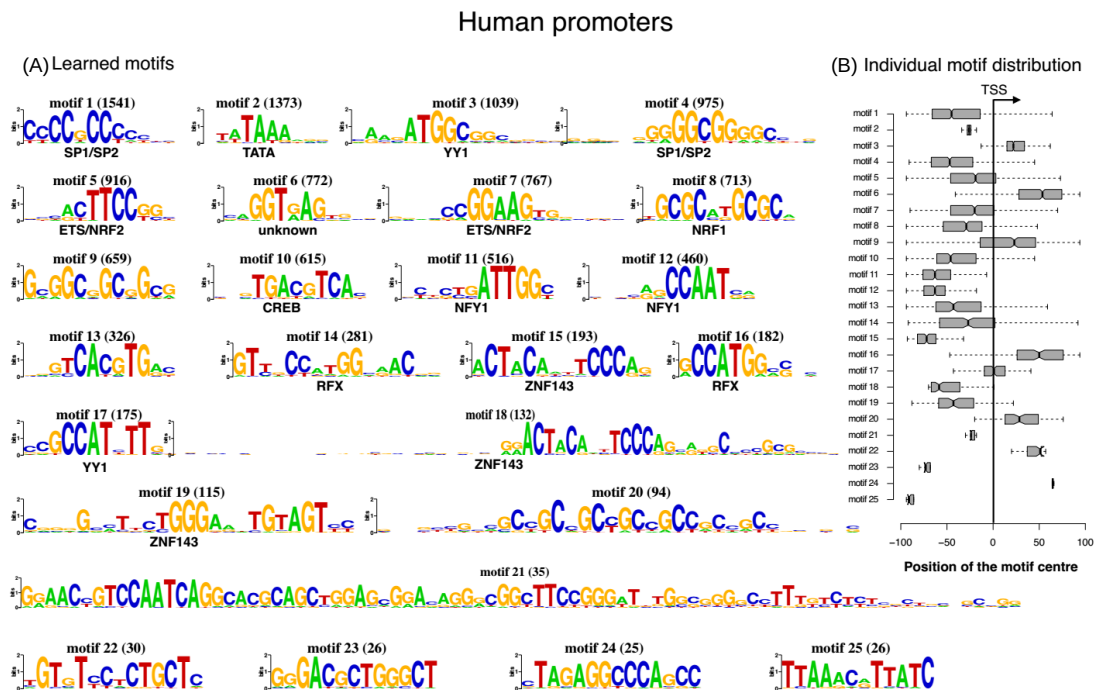


Figure 5.5: CISDIVERSITY identifies 25 motifs in pooled human promoter data, with motifs 21-25 contributing to one module with TSSs of 37 zinc finger genes. All the other motifs are strand invariant, except for motif 2 (TATA) and motif 6 (unknown). (B) Each motif here too has a distinct distribution about the TSS.

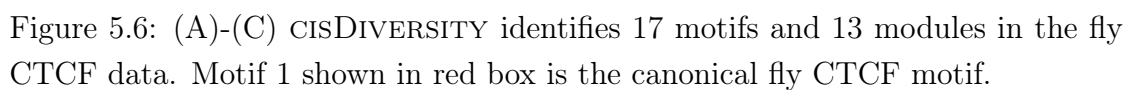
here. Similar to fly promoters, the identified modules exhibit unique properties in terms of enrichment or depletion of narrow promoters. Here also, we observe that modules with few binding sites have a considerably lower tag density.

5.5.2 CTCF bound regions display remarkable sequence level diversity

ChIP-seq for CTCF in fly

Source We next used CISDIVERSITY to investigate cis-regulatory diversity in ChIP-seq datasets. CTCF, a highly conserved TF plays a variety of crucial roles in regulation, from binding insulators to generating chromatin loops in various contexts [Phillips and Corces, 2009, Matthews and White, 2019]. Here, we apply

Fly CTCF ChIP-seq regions in white pre-pupa stage



Discussion The module with the highest number of sequences is dominated by the first motif which matches the reported CTCF motif. If we assume that CTCF binding is strongest in this module, then as expected it also has a significantly higher ($p < 10^{-10}$) ChIP signal. Motif 1 co-occurs with motif 5 in module 4 where it contributes only marginally (27%) to the module. Motif 5 is similar to the motif of another insulator binding protein suppressor of hairy wing, Su(Hw) [Khan et al., 2018]. Also, this module had the greatest degree of overlap with the ChIP-seq regions bound by Su(Hw) in the same developmental stage [Nègre et al., 2011], suggesting that sequences in this module may be bound by Su(Hw) directly.

According to multiple studies ([Ni et al., 2012, Smith et al., 2009]), fly promoters include the highest percentage of CTCF binding areas. In fact, if we take a look at the entire dataset, more than half of the bound areas are located within 1 kb of a TSS. However, the two modules mentioned above that include the CTCF motif, when considered individually, actually have fewer promoters ($p < 10^{-5}$) for both. Modules 5-8 on the other hand mostly overlap promoters (less than 100 bp from the TSS), and all but module 8 lack CTCF motifs. All of them are enriched with motif 3 which is a promoter element, Core 3 [Ohler et al., 2002], but have different co-factors. As we saw in the previous section, these are in fact several core promoter architectures. Given that the ChIP signal is also weak at these modules, CTCF most likely makes indirect interactions in these locations. Module 2 does not have a motif and has the lowest ChIP signal score compared to the other modules ($p < 10^{-5}$), indicative of poor or non-specific CTCF binding.

The motif of the zinc finger TF Pita is similar to motif 2 and motif 4 [Maksimenko et al., 2015]. Both these motifs appear in modules 3 and 7. This suggests that the sequences in those modules have 2 copies of the Pita motif and thus, both are necessary to explain the dataset. Now, the CISDIVERSITY model does not account for the spatial distribution of the motifs. But we see that in two thirds of the sequences the distance between the two copies of the Pita motif is less than 50 bp, suggesting cooperative binding.

In contrast to vertebrates, there is no evidence to support cohesin-CTCF mediated loop formations in fly [Matthews and White, 2019]. Instead, it is believed that looping may be mediated by interactions of CTCF with insulator binding proteins such as Su(Hw), BEAF-32, Ibf1/2 and Trl [Cuartero et al., 2014], all of which are recognized by CISDIVERSITY (motifs 5, 6, 9 in Fig: 5.6 A). We were

unable to identify any known TF motif in the literature that had any substantial similarities to other identified motifs. However, we noticed that many of these motif locations are conserved across fly species, which hints at the possibility that they may participate in CTCF-related regulation.

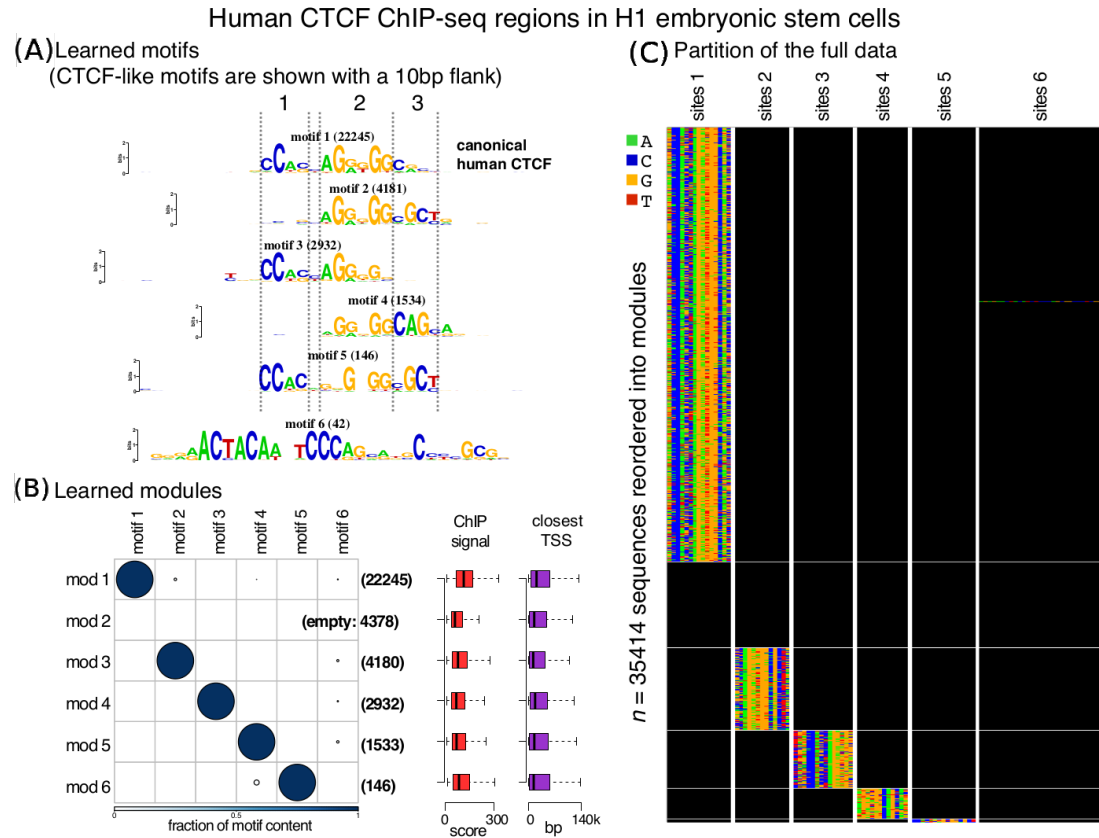


Figure 5.7: (A)-(C) CISDIVERSITY identifies only 6 motifs and 6 modules in the human CTCF data. Motif 1 in the red box matches the canonical vertebrate CTCF motif. Motifs 2-5 resemble the vertebrate CTCF motif but differ at one of three parts, denoted with dotted lines. These motifs are shown with 10 bp flanking regions to ensure that they are genuine variants of the motif.

ChIP-seq for CTCF in human

Source We ran CISDIVERSITY on ≈ 35000 ENCODE CTCF ChIP-seq regions from human H1 embryonic stem cells (ESCs).

CISDIVERSITY output Results obtained on human CTCF ChIP data are in stark contrast to those on fly data. Despite the fact that there are many more bound sequences in the human data than in fly, CISDIVERSITY only reports 6 modules and 6 motifs (Fig: 5.7 A-C).

Discussion The canonical CTCF motif is the top motif and the other 4 are variations with some of the components of the motif missing or showing nucleotide dependencies. Although these nucleotide level dependencies within vertebrate CTCF have been well documented ([Narlikar, 2013, Eggeling et al., 2014]), we have not previously found variants that are missing sections of the motif. These variations may be related to the varying utilization of CTCF zinc fingers [Phillips and Corces, 2009]. Furthermore, these 5 motifs account for about 85% of the sequences and are almost always found in different modules. The sole motif that is not CTCF matches to ZNF143, a transcriptional activator that binds at promoters associated with CTCF [Heidari et al., 2014], and in loop formations [Ye et al., 2020]. Nearly 90% of the sequences that contain ZNF143 also contain a variation of CTCF, suggesting that contact with CTCF at these sequences is not necessarily indirect or via ZNF143. The human CTCF motif itself is far more variable than the fly CTCF, and it does not seem to interact with DNA as frequently in a non-specific or indirect manner. The ChIP signal is significantly higher at the module with the canonical motif, indicating that CTCF is less strongly occupied by the variants in the other modules. CTCF is known to form loops and topological domains in mammals ([Phillips and Corces, 2009]). It is possible that the various zinc fingers of CTCF interact at different chromosomal locations, causing loop formation between them, but more research would be necessary to conclusively prove this idea.

5.5.3 GR binds to diverse regions after activation

Source The human glucocorticoid receptor (GR, encoded by NR3C1) binds to thousands of sites in response to exposure to the glucocorticoid (GC) hormone cortisol. According to studies by John et al. (2011) [John et al., 2011], GR primarily binds to DNase hypersensitive sites (DHSs) and often with other pioneer factors ([Biddie et al., 2011, Grøntved et al., 2013]). McDowell et al. (2018) [McDowell et al., 2018], have examined the binding of multiple factors including

GR, in A549 cells, before and after the synthetic GC dexamethasone treatment (dex).

CISDIVERSITY output The outcome of applying CISDIVERSITY to 6694 GR bound 200 bp regions identified 1 hour after dex treatment are shown in Fig: 5.8 A. It finds 5 motifs and 6 modules.

Discussion The largest module contains motif 1 which is a match to the GR motif. The other 4 motifs, which are motifs of TFs known to have a role in recruiting GR to its binding sites ([Biddie et al., 2011, Grøntved et al., 2013]), dominate modules 2-5. At these modules, we examined the ChIP-seq signals of the profiled TFs both before and 1 hour after dex treatment. As expected, GR signal is non-existent before dex treatment at any module (no GR bound regions were reported in this stage), but the GR signal increases at all modules after treatment (Fig: 5.8 B). The GR signal was found to be significantly higher at module 1 ($p < 10^{-10}$) which also contains the GR motif. ChIP-seq signal of TFs JUNB and CEBP is found to be stronger at the modules with their cognate sites. However, they seem to have been present in those modules even prior to dex treatment, albeit at a reduced level. After dex treatment, their signal at all modules, including module 1, increases. This indicates that GR may play a role in some form of cooperative binding at all modules.

As has been observed before [John et al., 2011], all modules exhibit some DNase hypersensitivity signal prior to the dex treatment, but the module with the GR motif shows the greatest difference in signal before and after the treatment. This difference is even more pronounced when we look at the binding signal for the enhancer mark: EP300. McDowell et al. [McDowell et al., 2018] used the GR motif to scan the GR bound regions and noticed that regions with initial EP300 binding (before treatment) had a weaker median GR motif strength than the regions which did not have initial EP300 binding. Our findings support this fact, but they also demonstrate that not only is the GR motif absent in the regions with initial EP300 binding, but those regions can be explained by the presence of other specific motifs. When the EP300 ChIP-seq regions are analyzed separately, CISDIVERSITY identifies nearly identical motifs but with different module distributions. The module with the GR motif is the fourth largest (Fig: 5.9). CTCF

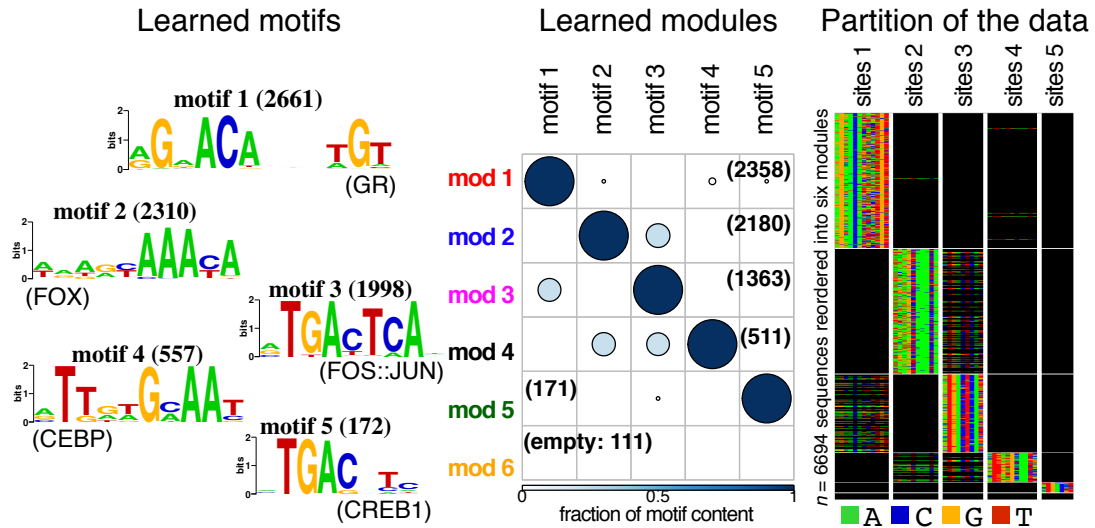
ChIP-seq signal is plotted as negative control, as there is no difference in the CTCF occupancy at these modules before and after dex treatment.

Running CISDIVERSITY on extended regions

Both the total number of motifs found and the number contributing to each module are less than the fly CTCF data we previously looked at. We re-ran CISDIVERSITY on a neighbourhood twice as long as before, of 400 bp size, to see whether there is more evidence of cooperativity in a bigger window surrounding the summit Fig: 5.10. Indeed, we get 13 motifs, which is more than twice as many we got earlier, and 9 modules in this data identified by CISDIVERSITY. The most prevalent motif is still the GR motif. Motifs resembling FOX, JUN, CEBP and CREB1 are discovered but with an additional copy of CEBP with both copies co-occurring in the data. Together these motifs continue to explain most of the regions. Three of the remaining 7 motifs, HNF4A, HNF1A and MAF::NFE2 resemble binding sites of TFs active in the lung. However, these being distributed across multiple modules do not contribute in a significant manner to any specific module. The remaining 4 motifs which are low-complexity di or trinucleotide repeats are distributed equally throughout the 400 bp regions, unlike the other motifs which are concentrated at the summits. These are probably the structural characteristics of the DNA [Yáñez-Cuna et al., 2014].

When these regions were lengthened further to 1000 bp, CISDIVERSITY found 21 motifs in them. All the prior motifs are still discernible, although GR is no longer the most prevalent motif. Instead, a number of additional FOX and JUN motifs, lung-specific TF binding sites and several low-complexity DNA motifs are found (Figure for this result is in the Supplementary section of the paper), indicating that GR motifs have less of an impact on the areas away from the ChIP summit.

(A) Human GR ChIP-seq regions 1 hr after dex treatment



(B) Signal within 500bp of GR peaks before and after dex treatment

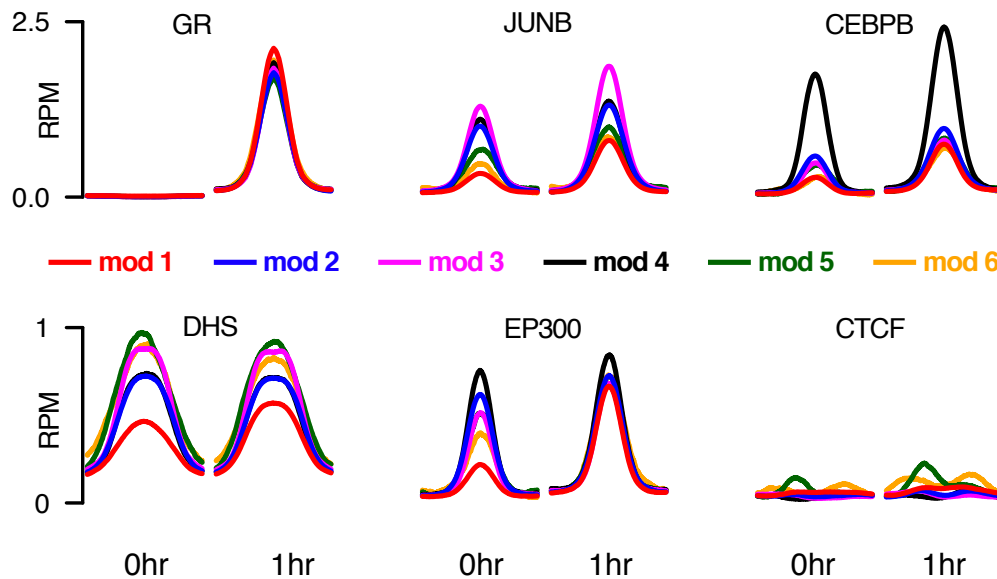


Figure 5.8: Diverse signals discovered in GR bound ChIP-seq regions. (A) CISDI-VERSITY identifies 6 modules and 5 motifs. (B) The average DNase hypersensitive signal and the input subtracted ChIP-seq signal in reads per million (RPM) of 5 TFs: GR, JUNB, CEBPA, EP300 and CTCF before and after treatment at the GR bound regions are shown for each module.

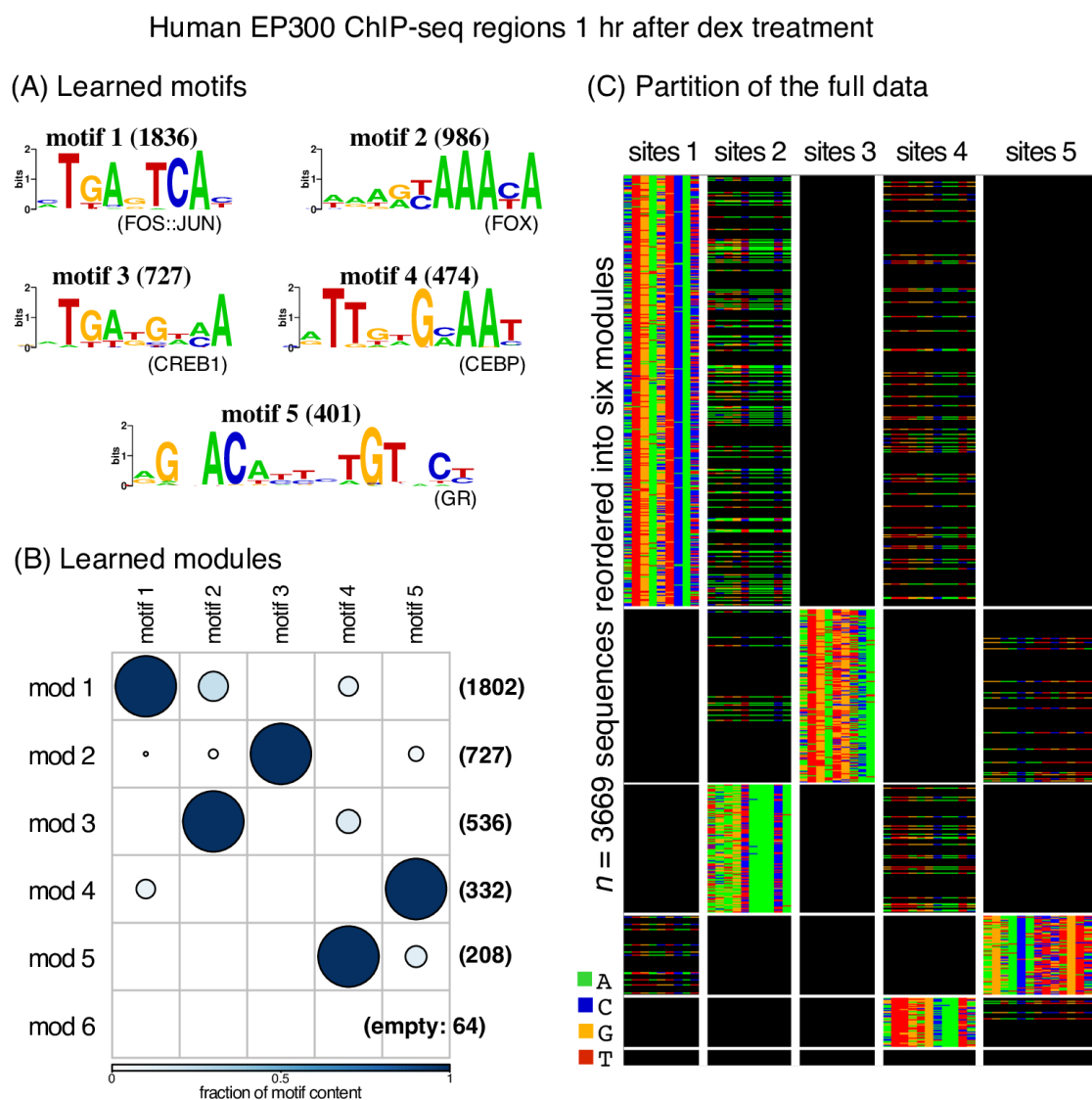


Figure 5.9: Results of CISDIVERSITY on EP300 ChIP-seq data in A549 cells 1 hour after being treated with dex. These motifs were near identical to those learned on GR ChIP-seq data in the same cell type, but the distribution of the modules is different.

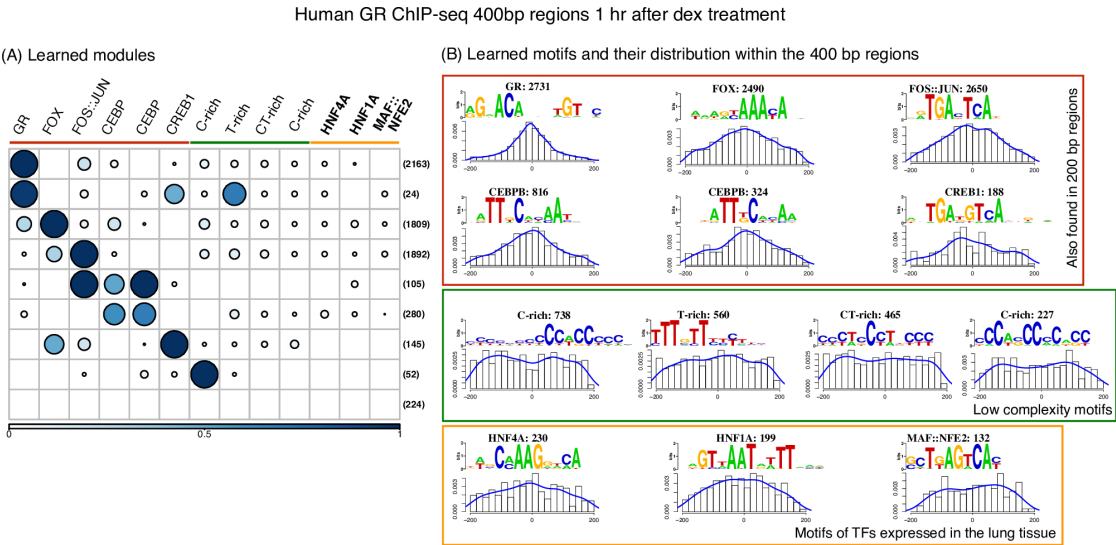


Figure 5.10: Results of CISDIVERSITY on a larger 400 bp neighbourhood of GR-bound summits. The modules and the motifs are recorded based on the motifs that are also found in the 200 bp neighbourhood (red box), those that are more spread out along the region and have low-complexity (green box) and those that resemble motifs of TFs active in the lung tissue (orange box)tissue. The histogram depicts the distribution of the binding sites of each motif along the 400 bp region.

5.5.4 Enhancers have a different structure based on the detection assay

Enhancers perform crucial roles in activating gene transcription, despite frequently being far from the target gene. There is no consensus sequence-based rule that can explain or describe these regions, despite the fact that several TF binding sites are enriched at the enhancers. Over the past few years, a number of fresh high-throughput assays have been created that profile a biochemical activity, which either directly or indirectly, indicates the presence of enhancers. For instance, short bidirectional and largely unstable transcripts have been shown to originate at active enhancers [Andersson et al., 2014]. The functions of these enhancer RNAs (eRNAs) are currently not fully known [Danko et al., 2015], but active TF binding at enhancers has been reported to increase corresponding eRNA levels.

CISDIVERSITY on distal eRNAs

Source We used CISDIVERSITY on 14300 distal eRNAs [Azofeifa et al., 2018] from the K562 erythroid cell line that were detected in the GRO-cap dataset [Core et al., 2014].

CISDIVERSITY output 9 motifs and 8 modules are discovered from this data shown in Fig: 5.11 A. There are actually 5 additional motifs that together describe the 9th module of 32 sequences. These sequences are classified as segmental duplications by the UCSC Genome Browser [Haeussler et al., 2019]. When module 9 was given as input to the multiple sequence alignment program CLUSTAL [Madeira et al., 2019], it could identify clear conserved structures in the sequences. We have not considered this module and the associated motifs in this section for the reason of clarity.

Discussion Unlike the models obtained from ChIP-seq regions, the largest module in this dataset which constitute more than half of the dataset, is largely devoid of any notable motif occurrences. One dominant motif is common to all other modules, which matches the motif of a TF active in K562 cell line. Thus, we next assessed overlaps between modules and ChIP-seq regions of all these TFs. As expected, the corresponding TF has a significant overlap (hypergeometric $p < 10^{-4}$),

although the overall overlap is minor with the entire set, for the majority of TFs. There is a higher degree of overlap between the active chromatin mark and the enhancer mark EP300, but there are differences in how much overlap there is between the modules as well. The essentially empty module in the distal eRNA does not significantly overlap with any of the external datasets. In fact, the overlaps with EP300 and other active chromatin marks are substantially lower than would be predicted by chance (hypergeometric $p < 10^{-4}$). This feature, along with the fact that no significant sequence motif is detected in this module, suggests that maybe these regions do not conform to the standard definition of enhancers, which are considered as regions enriched with TF binding and active chromatin marks.

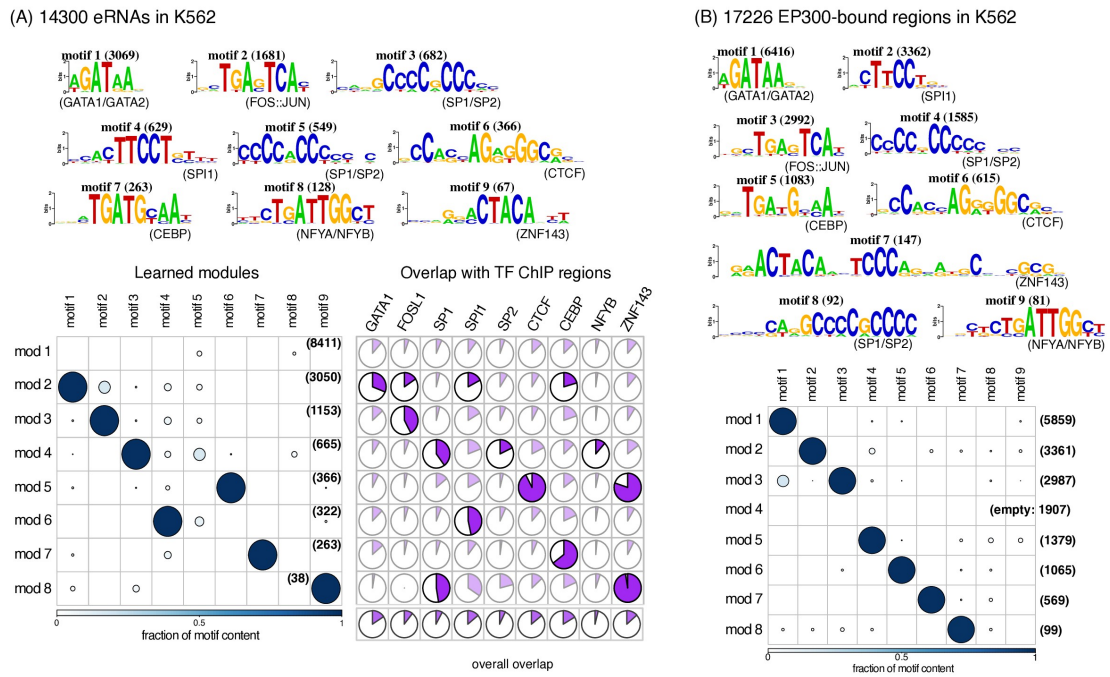


Figure 5.11: CISDIVERSITY run on putative enhancers in K562 (A) On distant eRNAs, overlap with ChIP-seq data is significant (hypergeometric $p < 10^{-4}$; shown in bold) in modules that contain the matching TF motif. Note that in some cases the overlap looks large, but does not show up as significant, because the hypergeometric test corrects for the sizes of the overlaps and the modules. (B) Similar motifs are found in EP300 ChIP-seq data.

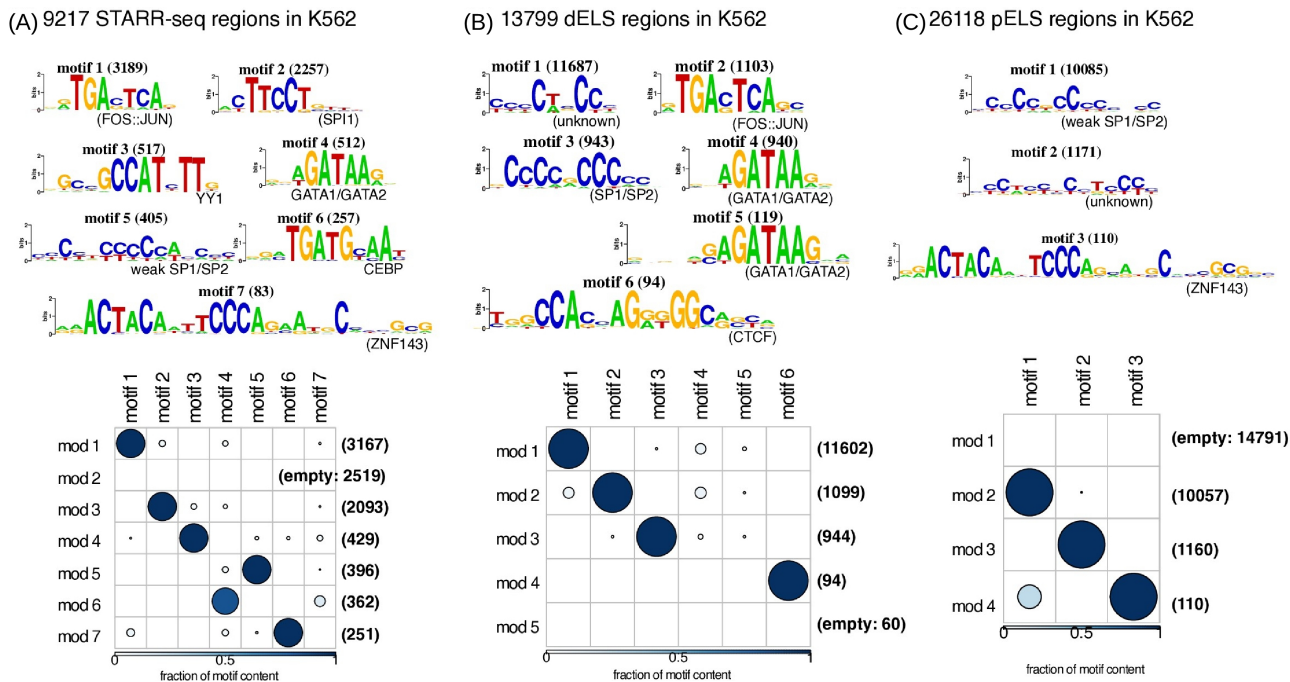


Figure 5.12: CISDIVERSITY run on putative enhancers in K562 (A) In STARR-seq peaks, YY1 is additionally discovered. CTCF and NFYA/ NFYB are not enriched. (B) Distant enhancer like sequences and (C) proximal enhancer like sequences deduced from chromatin signatures have fewer motif-like signatures.

CISDIVERSITY on proximal transcripts

Source Azofeifa et al. (2018) [Azofeifa et al., 2018] had eliminated the bidirectional transcripts that overlapped with annotated promoters to get the distal set of eRNAs. We applied CISDIVERSITY to these 8324 excluded proximal transcripts.

CISDIVERSITY output CISDIVERSITY retrieved a vastly different model with 10 modules and a much larger set of 25 motifs (Figure for this result can be found in the Supplementary section in the paper).

Discussion Both sets of distal eRNAs and the proximal transcripts share several motifs, such as those that match SP1, NFYA/B, and ZNF143. The proximal eRNAs are however enriched with a variety of other promoter motifs including YY1, RFX1, CREB1, NRF1 and others. There are more motifs per module compared to the modules in the distal set, indicating increased cooperative binding at

these locations, which is a characteristic of promoters [Maston et al., 2006]. The enrichment of the active chromatin marks also vary across these modules.

Except for H3K4me1, which is more frequently associated with enhancers [Heintzman et al., 2009] and less frequently with promoters [Cheng et al., 2014], these areas are more often accessible and have more overlaps with all active marks. This is consistent with the dataset being divided based on overlaps with annotated promoters. However, even after classifying eRNAs as proximal and distal, there are differences between the marks even within the modules in both datasets, implying that not all eRNAs have the same chromatin signatures. But more importantly, these differences can be explained to some extent using the modules and motifs.

CISDIVERSITY on EP300-bound regions in K562

Source We next ran CISDIVERSITY on four other datasets, which also report enhancer activity in K562. The first one is the EP300-bound sequences in K562 obtained from ChIP-seq experiment.

CISDIVERSITY output Here we once more obtain 9 motifs that are nearly identical to those in distal eRNAs but are present in different fractions of the sequences (Fig: 5.11 B). The bulk of the sequences in the EP300 set contain some motif, which is in contrast to the distal eRNAs set but similar to the other TF ChIP-seq datasets.

CISDIVERSITY on STARR-seq data in K562

Source After this, we examined the STARR-seq data. In a STARR-seq assay, random genomic fragments are inserted in the 3' UTR of a reporter gene with a minimal promoter and the resultant plasmids are then transfected into the target cells, which are K562 in our case. The 3' UTR regions of the reporter gene transcripts are then sequenced in order to determine the enhancer activity of these regions. This assay, unlike other approaches, takes into account regions outside their chromatin context, enabling it to detect regions that are inaccessible yet may have potential for enhancer activity [Liu et al., 2017]. Here we considered the peaks reported by the STARRPeaker method by Lee et al. (2020) [Lee et al., 2020] which produces just over 9000 sequences.

CISDIVERSITY output 7 motifs and 7 modules are discovered by CISDIVERSITY (Fig: 5.12 A).

Discussion The eRNAs or EP300 bound regions and the STARR-seq regions have relatively little sequence in common as shown in Fig: 5.13, but a lot of motifs are common among the three groups. Only YY1 is discovered additionally in the module 4 of STARR-seq regions, while NFYA/ NFYB and CTCF are absent. We examined whether the modules have accessibility profiles that were different from those of the first two datasets, since STARR-seq assays have the ability to consider regions outside the chromatin context. Indeed, it was observed that the DNase hypersensitive sites (DHS) content of modules 2 (which is devoid of any motif) and 4 (dominated by YY1) is significantly lower ($p < 10^{-5}$). This implies that these modules are suppressed by endogenous chromatin in K562. On the other hand, DHSs are significantly enriched in modules 1 (dominated by AP-1 motif) and 6 (GATA-dominated) which are also found in the eRNA and EP300 assays.

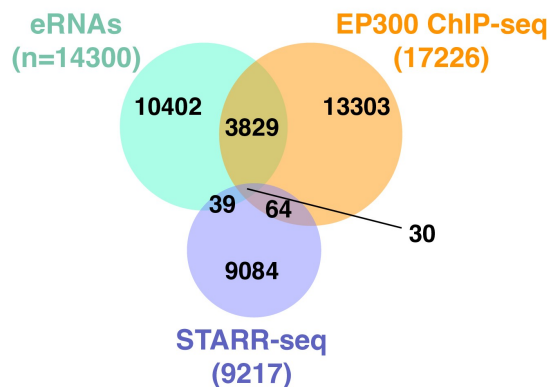


Figure 5.13: Sequences that are common across three assays that measure enhancer activity in K562.

CISDIVERSITY on proximal and distal enhancer like signatures

Source The final 2 datasets are from ENCODE phase III, where The ENCODE Project Consortium [Moore et al., 2020] published a registry of potential cis-regulatory elements based on the findings of numerous high-throughput experiments. They present two distinct groups that they claim contain both proximal

(pELs) and distal enhancer-like signatures (dELs). A high DNase and H3K27ac signals are required for a sequence to be included in any of these sets. While dELs are distant pELs are close (within 2000 bp) of an annotated TSS. In order to make sure that pELs are not active promoters, they also need to have a low relative H3K4me3 signal.

CISDIVERSITY results CISDIVERSITY finds 3 motifs and 4 modules in the pEL regions and 6 motifs and 5 modules in the dEL regions in K562.

Discussion From the CISDIVERSITY results on these 2 datasets, we observe that there are no shared motifs between these sets other than an SP1/SP2 motif (Fig: 5.12 B and C). CTCF is found in the dELs while ZNF143 is present in the pELs which is expected as the binding sites of these TFs are enriched in regions distal and proximal to TSSs respectively [Bailey et al., 2015, Kim et al., 2007]. In contrast to the first 3 datasets, CISDIVERSITY finds no motifs matching SPI1, CEBP, or NFYA/B. Also, no GATA motif is found in the pELs set, possibly because GATA proteins primarily binds to distal enhancers [Romano and Miccio, 2020]. In these enhancer-like regions from ENCODE III, fewer signatures that resemble TF motifs have been discovered altogether.

5.5.5 Modules in open regions have differing future fates

Single cell ATAC-seq on *Drosophila* embryos

Source We next looked at the data from two different technologies, viz. ATAC-seq and DNase-seq that assess the chromatin accessibility. *Drosophila* embryos at 3 different phases following egg laying were subjected to single cell ATAC-seq by Cusanovich et al. [Cusanovich et al., 2018]. Here we report the CISDIVERSITY results at the earliest time point: 2-4 hours after egg laying. We considered the areas that were open in at least 10% of the cells assayed at that stage. This resulted in 9963 distinct peaks.

CISDIVERSITY output CISDIVERSITY finds 28 motifs and 13 modules in the data (Figure in the Supplementary section of the paper). For the sake of clarity, only motifs that account for at least 25% of the sequences in a module are shown in Fig: 5.14 A.

Discussion Without any additional information CISDIVERSITY divides the data into modules that are significantly enriched with promoters (within 500 bp of an annotated gene) and those that are significantly deficient ($p < 10^{-4}$). The majority of the motifs contributing to the promoter-enriched modules are recognized core promoter motifs. In module 10, the 2 Pita motifs from the fly CTCF data (Fig: 5.6) are also recovered. Enhancers and/ or insulators are likely to be present in the modules that are far from the TSSs. Indeed, the enrichment of CTCF and Su(Hw) in module 11 implies the presence of insulators in this module. Similarly, dinucleotide repeats of GA (bound by GAGA binding factor Trl) and CA are recognized features of fly enhancers [Yáñez-Cuna et al., 2014]. One of the promoter modules contains tRNAs where CISDIVERSITY captures the highly conserved hair loop structure of tRNAs as a ‘motif’ (module 12).

There is little difference in the fraction of the cells open at the regions across the modules, with only modules 4 and 6 (distal regions with dinucleotide repeats) being significantly more open. However, we observe several modules significantly changing their accessibility when we contrast the fraction of cells that are open at these areas with the fraction open 4 hours later (6-8 hours after egg laying) shown in Fig: 5.14 B. Except for the tRNAs, almost all promoter modules are opening up. We found no evidence of a decrease in tRNA gene expression during these developmental stages. The tRNA genes, on the other hand, are known to have a role in chromatin remodelling and act as insulators or harbours origins of replication in other organisms [Sreekumar et al., 2021, Su et al., 2020]. The only other modules that are significantly closing are modules 1 and 9. Module 1 has very few predicted binding sites and no motif is present in even 10% of the sequences. This is consistent with recent findings suggesting that 2 classes of enhancers are active during early *Drosophila* embryogenesis, one of which has significantly lower TF occupancy [Arbel et al., 2019]. However, every sequence in module 9 contains a motif that is similar to the TF zelda (zld). zld is a pioneer TF that is essential for the transition from maternal to zygotic state, which happens during this stage [Hamm and Harrison, 2018]. CISDIVERSITY results indicate that these zld-enriched regions become less accessible in the subsequent stages.

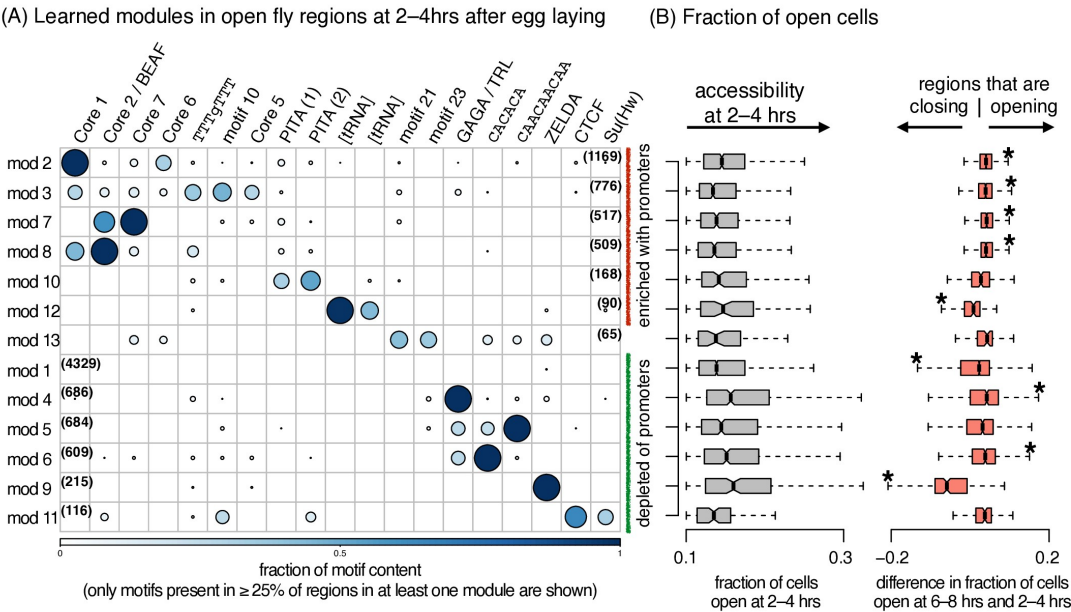


Figure 5.14: CISDIVERSITY run on open regions: (A) A total of 13 modules and 28 motifs were learned on the ATAC-seq regions which are open in at least 110% of the cell 2-4 hour after egg laying. Only the 19 motifs that contribute to at least a quarter of the sequences in some modules are shown here for clarity. Modules are reordered: the red and green modules are significantly enriched (hypergeometric $p < 10^{-4}$) with promoters and depleted of them, respectively. (B) Gray indicates that there are only a few differences in the fraction of cells that are open within each module. Orange indicates modules 2, 3, 7, 8, 4 and 6 are significantly more open in the cells 6-8 hours after egg laying, whereas modules 12, 1 and 9 are closing at that time point.

DNase-seq on H1 ESCs

Source CISDIVERSITY was next run on 47279 DNase hypersensitive regions of H1 ESCs with the entire results in the Supplementary section of the paper, and a subset shown in Fig: 5.15 A.

CISDIVERSITY output The original options of $r = 15$ resulted in 15 modules, so we re-ran CISDIVERSITY with $r = 20$ reporting a total of 18 modules and 21 motifs.

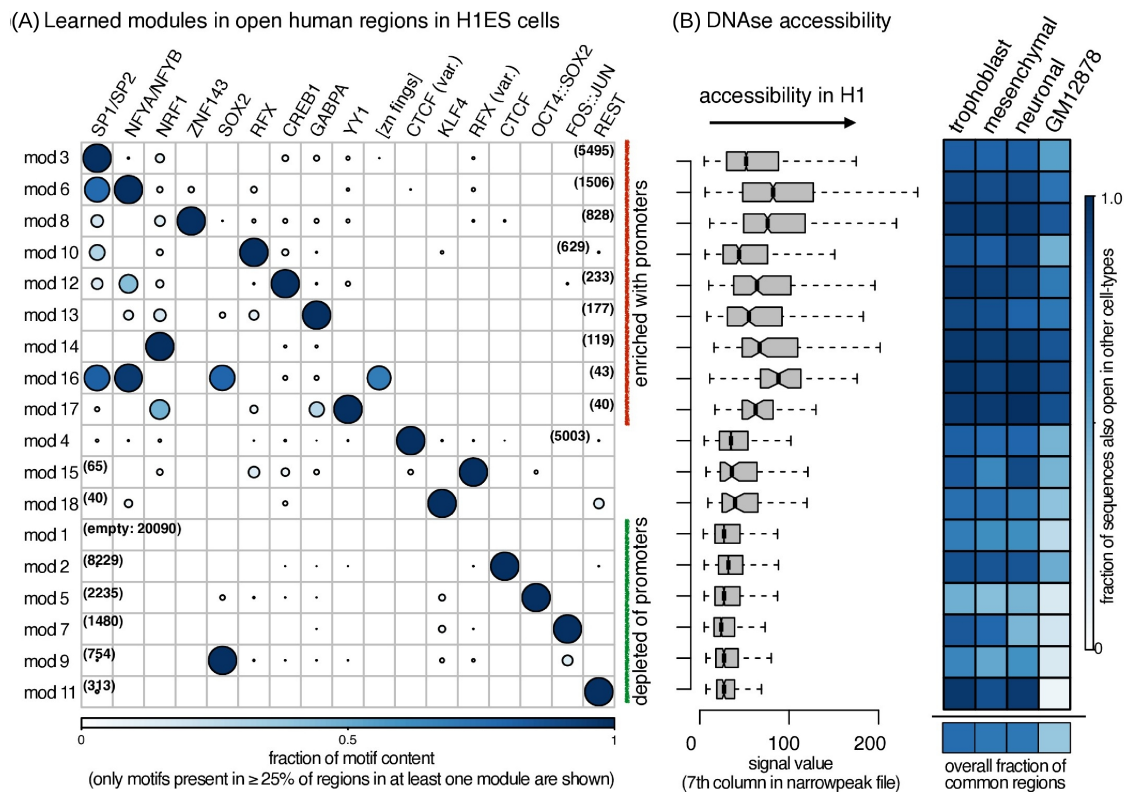


Figure 5.15: CISDIVERSITY run on open regions: (A) A total of 18 modules and 21 motifs were learned on the DNase-seq regions in H1 ESCs. Again, only the motifs appearing in a quarter of the sequences in some modules are shown here. The red and green modules are significantly enriched (hypergeometric $p < 10^{-4}$) with promoters and depleted of them, respectively. (B) Gray indicates promoter modules have a higher DNase signal in general, but there are variations among them. Blue indicates the fraction of each module (and total below) that is also open in trophoblast, mesenchymal and neuronal stem cells (all derived from H1 ESCs) and GM12878 and shows considerable variation in modules and cell types.

Discussion The modules are either considerably enriched with promoters or devoid of them, just like in fly data. Despite the fact that all promoter modules have strong DNase accessibility, there are nevertheless noticeable differences among them. For instance, module 6 exhibits a significantly greater ($p < 10^{-10}$) DNase hypersensitivity signal than module 3, despite the fact that both modules 3 and 6 are enriched with promoters and with the SP1/ SP2 motif and further with module

6 also containing an NFYA/NFYB motif. Module 16 also contain SP1/SP2 and NFYA/NFYB, but comprises primarily of core promoters of a KRAB family of zinc fingers. This similarity is captured as a distinct long motif, essentially creating a module distinct from module 6. Modules that are depleted of promoters are characterized with motifs of SOX2, POU5F1::SOX2, which are both TFs specific to H1 ESCs, as well as motifs of other pioneer factors AP1, REST and CTCF.

We examined the DNase accessibility at these locations in trophoblast, mesenchymal and neuronal stem cells produced from H1 ESCs that are available in ENCODE (Fig: 5.15 B). Despite the fact that these stem cells' overall DHS overlap fractions are comparable (0.76 for trophoblast, 0.71 for mesenchymal and neuronal), the overlap-fraction within modules varies from 0.42 to 1. Indeed, over 90% of the sequences in modules 8, 12, 14, 16 and 17 are accessible in these 2 stem cells. On the contrary, modules dominated by motifs of pioneering ES TFs(SOX2, POU5F1, KLF4) are less accessible, as would be expected. Module 11 containing the REST motif, has a high accessibility in all the stem cells. The canonical CTCF motif (module 2) is significantly more often accessible than the variant CTCF motif (module 4) in all other cells, although in H1 ESCs the variant has a higher accessibility signal ($p < 10^{-5}$). DHSs in the lymphoblastoid GM12878 cell line (derived from blood) are used as control. Although the overall overlap is lower across all modules, we still observe the same pattern of greater overlaps near promoters and smaller overlaps at distal modules.

5.5.6 Tissue specific modules discovered in open regions

Source We also analyzed the ATAC-seq data from 20 different mouse tissues [Liu et al., 2019]. In each tissue, the authors report regions with normalized read counts. In order to account for the fact that these regions were combined across experiments, we picked a longer 400 bp region around the center and considered all those regions with at least 10 normalized reads. Nonetheless, CISDIVERSITY was applied to each tissue independently. The authors compute an entropy value for each region based on the read counts obtained in each tissue. A lower entropy value indicates that the regulatory region is open in a tissue-specific manner, while a higher value suggests that the region is open in the majority of the tissues.

ATAC-seq data in murine brown fat tissue

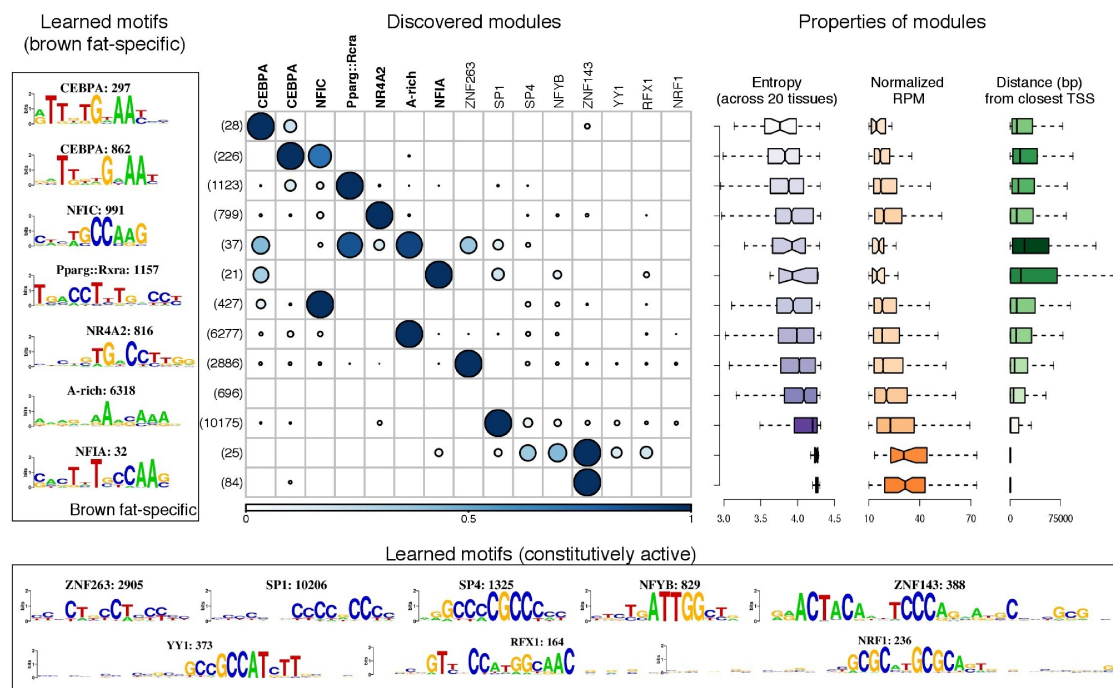


Figure 5.16: CISDIVERSITY run on accessible regions in mouse tissue: 13 modules are learned from the ATAC-seq regions from the brown fat tissue. They are sorted according to median entropy (most brown-fat specific on top). The 15 discovered motifs can be split into 2 sets: those that are enriched in tissue specific regions (with median entropy value < 4.0) and those that are constitutively active. Boxplots on the right indicate the relationship between tissue specificity, accessibility and proximity to TSS

CISDIVERSITY output Although CISDIVERSITY was run separately for datasets corresponding to each of the tissue types, here we will discuss output of CISDIVERSITY on the brown fat tissue. It discovers 13 modules and 15 motifs in brown fat tissue 22804 regions.

Discussion Fig: 5.16 shows the motifs and modules obtained in the brown fat tissue according to the median entropy value. The first 6 modules have a median entropy lower than 4.0 meaning that they are specifically reported in the brown fat tissue. These modules contain motifs that correspond to TFs known to be expressed in the fat tissue [Pradhan et al., 2017, Lee et al., 2019]. The fact

that these modules are tissue-specific and are less accessible than other modules is further supported by the low average read counts associated with them. They are also more likely to be distal components. In contrast, the modules distinguished by the constitutively active TFs, such as ZFP263, SP1 etc. are promoter-proximal and have a greater entropy and are more open than the other modules.

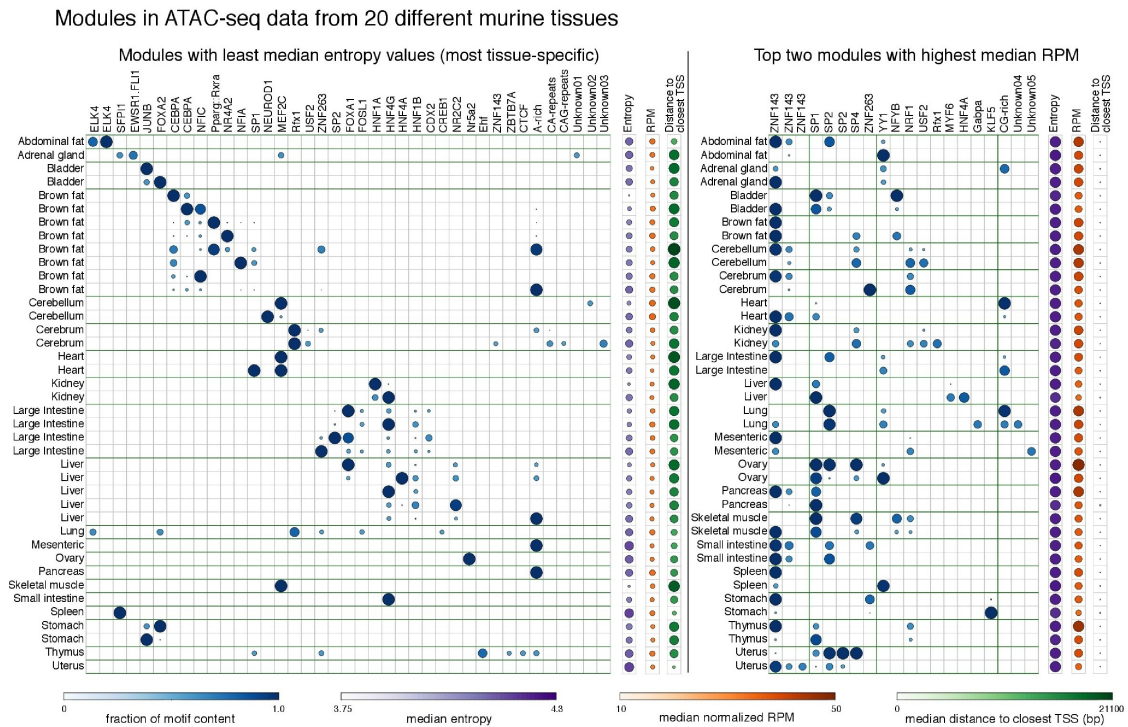


Figure 5.17: **Left panel:** shows modules with the lowest median entropy for each tissue. Many tissues have multiple modules with median values lower than 4.0: all such modules are shown. Motifs that are present in at least 10% of the sequences in any module are shown. **Right panel:** shows the top 2 modules based on highest median RPM (reads per million) values per tissue. Median entropy, RPM, and distance to closest TSS are displayed on the right for each module. TOMTOM matches were used to assign putative TF identities to motifs and combined across the CISDIVERSITY individual runs.

A similar trend is observed in the other 19 tissues shown in Fig: 5.17. Modules with significantly lower entropy value, i.e. those that are specifically reported in respective tissues, are characterized by motifs of corresponding tissue-specific TFs and also generally have reduced accessibility. They are also likely to be far from any

TSS. In contrast, modules having the highest accessibility are found to be proximal to promoters and are accessible across tissues. With a few notable exceptions, these modules are largely characterized by motifs of TFs that are constitutively active [Baldarelli et al., 2021]. As expected, these modules share a common set of TFs. We also note that MYF6 and HNF4a are present in certain modules in liver and GABPA in the lung. These are TFs known to be overexpressed in the respective tissues [Baldarelli et al., 2021]. We emphasize that CISDIVERSITY only detects the sequence signatures present in each module and not the identities of the TFs. Therefore, the matches to the TF motifs identified by TOMTOM form the basis of these results.

We note that ATAC-seq data is not typically examined in this way. Motif discovery is often targeted in regions pre-identified to have high tissue-specificity or be distal to any TSS, especially when multiple tissues are being analyzed, as in this case. Indeed, Liu et al. (2019) [Liu et al., 2019] use HOMER to find motifs in accessible areas relevant to different tissues. Despite not being given any knowledge about tissue-specificity CISDIVERSITY discovers these motifs and automatically partitions the data based on the sequence level information already available in the dataset.

5.6 Discussion

It is widely known that combinatorial binding of TFs play a crucial role in gene regulation [Reiter et al., 2017] and that the majority of high throughput experiments report regions with a variety of properties. There have been several attempts to find regulatory modules in high throughput sequencing data. CisModule [Zhou and Wong, 2004] was one of the earliest motif-module identification method. Its objective was to discover where a module was located in each sequence along with the motifs. The module however was of a singular kind. Genomic regions based on multiple ChIP-seq datasets were clustered using self-organizing maps [Xie et al., 2013], non-negative matrix factorization [Giannopoulou and Elemento, 2013], as well as topic models [Guo and Gifford, 2017], with the aim of discovering different modules. These methods have been utilized to predict TF co-localization and complexes developing along the chromatin. However, they specifically require regulatory information in the form of additional high throughput experiments for

each region, which serve as the features in clustering. CISDIVERSITY is the first attempt to cluster regions based on motifs which are themselves learned during the process of clustering necessitating neither more experimental data nor any prior TF-binding knowledge. The detected modules indeed correlate with experimental binding data of TFs with matching motifs, demonstrating the fact that these regions contain the requisite sequence information that can characterize functional modules. On simulated datasets, CISDIVERSITY perform better as a motif discovery method when compared to the standard approaches, especially in terms of precision or false positive prediction. This might be because of the model-based approach of CISDIVERSITY, where it aims to learn a model that explains the entire set of sequences and not learn motifs that individually explain a fraction of the sequences.

Regulatory modules are often studied in accessible areas or potential enhancers, assuming cooperativity between TFs in these regions. However, even with ChIP-seq datasets which document the binding sites of target TF, CISDIVERSITY could provide us with novel insights. For example, consider the extensively studied CTCF protein. In the original study, while many sequences did not have a strong match to the CTCF motif in the fly data, all subsequent locations and evolution analyses were nevertheless done considering all ChIP regions equally. However, in the CISDIVERSITY results, it clearly came up that some sequences are core promoters with no CTCF motif while some other sequences are enriched with different insulator-binding TF motifs. The evolutionary profile of the motifs is also diverse in the dataset. On the other hand, we observed that the human CTCF TF exhibits a different behavior and ZNF143 is the sole non-CTCF motif that was discovered in the data. Instead, a number of distinct CTCF motif variations are enriched in various modules, suggesting a possibility of differential usage of its zinc fingers while interacting with the DNA.

Putative enhancers have been discovered using different types of high throughput assays in the same cell type or context. These datasets generated by these experiments vary in cardinality, length distribution and evolutionary characteristics [Benton et al., 2019]. CISDIVERSITY illustrates these variations in different datasets in terms of motifs and modules discovered in them. For at least the cell type investigated here, CISDIVERSITY found no common motif across all the enhancer detection strategies. It is important to remember that each dataset utilized

in this study is derived from an experiment that assesses a unique biochemical activity or a combination of activities. CISDIVERSITY can be used to further iron out the differences between the sequences reported by these methods in terms of motifs and their combinations.

The advantage of CISDIVERSITY is that it can be applied to any collection of DNA sequences because it does not depend on any well known PWM or modules. The sole presumption it makes is that the sequences are from a single cellular context, implying that (i) a finite set of TFs is active during the experiment and (ii) binding sites for those TFs are the underlying cause for the regions to be reported. It is not necessary to provide it with a targeted set, such as distant regions or already-processed tissue-specific regions. Instead, it acknowledges that the data may be heterogeneous even when it comes from a single high throughput experiment. CISDIVERSITY uses the underlying motif information and automatically clusters accessible regions into promoter specific modules and TSS-distal regions into other distinct modules. It also captures a significant amount of sequence-level variability as well as tissue-specificity even within promoters and putative insulators/ enhancers. Although the modules in Fig: 5.14 A are rearranged according to their propensity to contain more or fewer promoters, no module is entirely made up of promoters or devoid of them. A promoter-enriched module's distal sequences should be further investigated for potential promoter activity and vice versa.

CISDIVERSITY does not explicitly model distance between or multiplicity of motifs. These features can be studied from the learned parameters. For instance, by learning multiple copies of the same motif, CISDIVERSITY can detect homotypic binding if it occurs in a significant fraction of sequences. In Figs: 5.6 A, and 5.14 A, we obtain 2 copies of the Pita motif in the fly data and multiple copies of CEBP motifs in mammals Fig: 5.14 and 5.10. The distance between the motifs as well as the relative orientation can also be measured from the CISDIVERSITY results. In case of the ChIP-seq investigations, the distance of each motif from the summit can provide information about the likelihood of direct binding of the profiled TF [Bailey and Machanick, 2012]. We can even use CISDIVERSITY to look for motifs only on the given strand, especially when dealing with datasets like 5' CAGE TSSs or UTRs which contain inherent directionality.

All the results reported here regard the highest scoring model as the best model, thereby treating this as a hard clustering strategy. We note that CISDIVERSITY

learns a probability distribution over the modules for each sequence and multiple module usage can also be explored [Guo and Gifford, 2017]. For a learned model, CISDIVERSITY does not yet report a significance value or a false discovery rate. It does however provide the posterior probability associated with the learned model. In datasets with random regions and no planted modules or motifs, these values are significantly lower (Fig: 5.18). Such distributions can be learned to report an associated empirical ‘p-value’. Run-time wise, CISDIVERSITY is similar to MEME, but significantly slower than HOMER. The model can theoretically be initialized with at least some user-defined motifs, which might hasten the model discovery process.

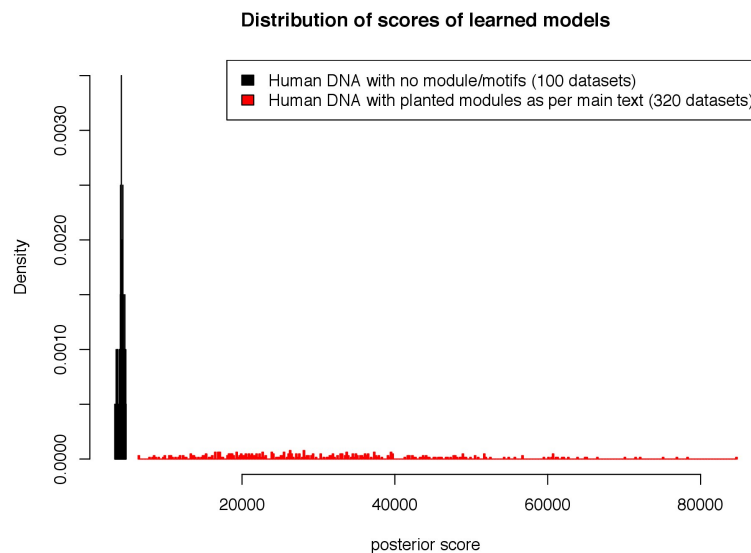


Figure 5.18: Distribution of posterior scores of models learned on the simulated datasets are shown in black (100 datasets containing non-repetitive human genomic regions with no planted motifs) and red (320 datasets with planted motifs and modules as per the Methods section)

In comparison to fly regions, modules in mammalian regions are characterized by far fewer TF binding sites across the datasets examined here. Cooperativity between TFs can arise from multiple mechanisms. The most straightforward scenario is when DNA mediates it by harboring the appropriate TF binding sites within

a brief linear area. This most likely explains the cases where different modules are defined by more than one motif. However, there is growing evidence that the interplay between the 3D genome architecture and the TF activity have an impact on TF cooperativity particularly in higher eukaryotes [Ma et al., 2018]. TF-driven loops and condensates are some mechanisms of cooperativity, which do not require the presence of multiple binding sites in a linear genomic neighbourhood [Kim and Shendure, 2019]. Additionally, it has been proposed that the presence of clusters of low-affinity binding sites, which may not be captured by CISDIVERSITY’s motif models, affects TF cooperativity [Malin et al., 2015]. It is currently unknown the extent to which these pathways influence gene expression. More experiments and analyses might be helpful in determining whether the sparse modules identified by CISDIVERSITY are indicative of these mechanisms.

Chapter 6

Software and datasets

Each of the statistical frameworks we discussed in the previous chapters have been developed from scratch using C, Python and in some of the cases R programming language. In this chapter, we will go over the software packages for each of the methods and the links to the GitHub repositories for them. We will then list out all the biological datasets that were used to obtain the results that have been discussed in this thesis previously.

6.1 Software for each of the methods

6.1.1 DIVERSITY

DIVERSITY is a mixture model-based Bayesian framework which is implemented as an unsupervised learning algorithm using Gibbs sampling to cluster a set of DNA regions obtained from a ChIP-based experiment. The goal is to identify different modes of protein-DNA binding based on the presence of a sequence component called motif. This motif is either the direct binding site of the target protein or a contact through intermediaries.

The main sampling algorithm of DIVERSITY is written in C. A Python wrapper is used on top of that to control multiprocessing, providing a parallel implementation of DIVERSITY. It learns models from different initializations and by varying the number of modes m in parallel. Finally, when all the models are learned, using Bayesian model selection the best model is identified.

DIVERSITY is available as web server <http://diversity.ncl.res.in/> (Fig: 6.1)

as well as a standalone software which can be downloaded from <https://github.com/NarlikarLab/DIVERSITY>.

The web-based version has the provision where a user can upload a FASTA file (which should be $< 2\text{MB}$) containing the DNA sequences or a BED file with chromosomal coordinates corresponding to the peak regions of a ChIP experiment. However, in case of a BED file, the reference genome has to be selected from a dropdown menu. The results of running DIVERSITY on a BED file also produces 3 additional images: aligned motifs based on their midpoint per sequence, phast-Cons scores at the sequences, and boxplots of distances between the modes and the closest transcription start site. DIVERSITY returns information about all the learned models in an HTML file which contains links to a table listing the mode identification of each sequence (I_i), position of the motif (Z_i) in each sequence as well as a text file listing all the model parameters (ϕ) as well as the sequence logos. It also produces a file `bestModel.html` (Fig: 6.2) which contains the motifs on the forward and reverse strands corresponding to the best model. DIVERSITY can be run with several additional options such as changing the range of the motif width, number of modes, restricting the motif detection to the given strand and more.

The standalone version can be downloaded and installed in any Linux or MacOS X machine. Once it is installed, it can be run using the command `diversity`. DIVERSITY returns the details of all models learned in an HTML file `all_model.html` where all models can be viewed at a glance. It contains the link to the best model with the optimal number of modes. The data of all the models learned are stored in separate directories of the form `modelWith_m_modes` with m denoting the number of modes in the model. Within each of these directories, DIVERSITY saves the results of all the trials corresponding to different Gibbs sampling runs with distinct initializations for the given mode. The model with the best posterior score out of the trial runs is considered as representative model for that mode.

The pre-requisite python packages for installation, instructions to download and install DIVERSITY, the optional parameters and example usage can be found at DIVERSITY documentation.

DIVERSITY

DIVERSITY uses a Bayesian approach to detect multiple modes of protein-DNA binding from high throughput ChIP data. It is available both as a [web based application](#) and as a [downloadable software](#).

• Home
• Downloads
• Documentation
• Help

DIVERSITY

Select the format of the INPUT file File Format

(Please refer to the help section for a sample input file)

Execution parameters

Input file (FASTA)* Browse... No file selected.

Minimum number of modes 3

Maximum number of modes 10

Mask repeats in sequences ☒ Yes ☐ No

Include reverse strand in training ☒ Yes ☐ No

Initial width of the motifs 12

Minimum width of the motifs 6

▶ User Information

▶ Advanced options

Reset
Submit

Figure 6.1: DIVERSITY server

6.1.2 Input and output of DIVERSITY

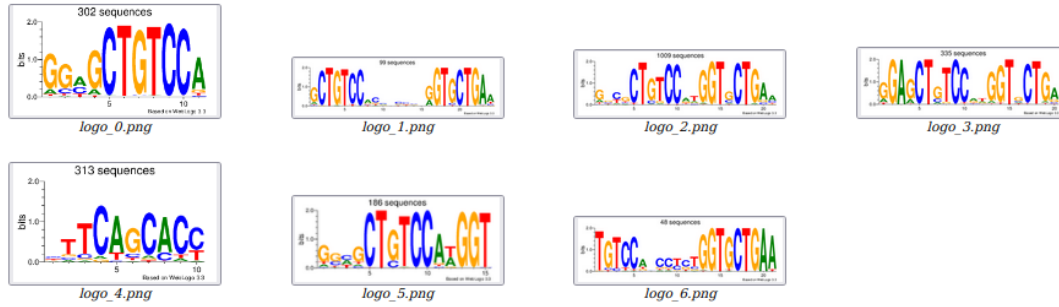
Input

DIVERSITY takes input as a fasta file corresponding to the ChIP-bound regions. The webserver allows the input to be in fasta as well as in bed format, but the

Model with 7 modes

For details about the sequences containing the different motifs please refer to [info.txt](#).

Logos on positive strand



Logos on negative strand

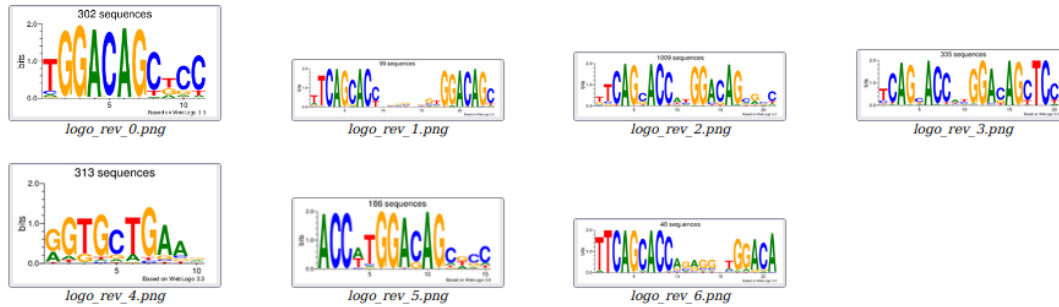


Figure 6.2: DIVERSITY bestModel.html file

total number of regions that can be given is limited to 5000 sequences. In case a larger number of sequences need to be examined, or simply if preferred, one can download the source code for DIVERSITY and compile it on a Linux or a Mac machine and run it on their own data sets. The default parameter values for DIVERSITY are as follows:

- -f filename
This is a compulsory parameter. File in fasta format in which motifs have to be identified.
- -o directory
Valid directory name. If it exists then a new one is created with the given name along with extension number. Default output directory is: diversityOut in the current working directory.

- -maskReps 0 or 1
Mask repeats by replacing them with 'N'. 0 implies do not mask. Default value: 1
- -r 0 or 1
Look for motifs in the reverse strand as well. 1 implies looking for motifs in both strands. Default value: 1
- -zoops $\langle \text{any real number} \in [0, 1] \rangle$
0 insists on all sequences to have a motif (OOPS model) and is the default value.
1 is to let the sequence length decide the prior on presence of motifs.
Values strictly between 0 and 1 is the probability of a sequence not having a motif.
- -minWidth $\langle \text{non negative integer} \rangle$
Minimum width of motifs while training models. Default value: 6
- -maxWidth $\langle \text{non negative integer} \rangle$
Maximum width of the motifs while training models. If this value is greater than the length of the sequence then maximum motif width is limited to the length of the sequence. Default value: no maximum width
- -initialWidth $\langle \text{non negative integer} \rangle$
Starting width of the motifs while training models. minWidth becomes initialWidth if minWidth value is greater than that of initialWidth. Default value: 12
- -minMode $\langle \text{non negative integer} \rangle$
Minimum number of modes possible for the given dataset. Default value: 3
- -maxMode $\langle \text{non negative integer} \rangle$
Maximum number of modes possible for the given dataset. Default value: 10
- -lcount $\langle \text{non negative integer} \rangle$
Number of models to be learned during training for each value of m in range $[\text{minMode}, \text{maxMode}]$. Only the best model is considered. Default value: 5

- -proc $\langle$ non negative integer $\rangle$
Maximum number of processors to be used for computation. Default value:
Number of processors that the system has.

These parameter values can be changed and DIVERSITY can be run for a wider range of modes or with a broader initial width of the motif as required.

The default values of the prior hyperparameters are as follows:

- The parameters in the Dirichlet prior α_{mode} for the modes have the non-informative value of 1.
- The parameters of the Dirichlet prior α_{motif} for the motifs have the sparsity promoting value of 0.5.

Output

DIVERSITY returns the details of all the models learned in an HTML file in which for each model it links to a table containing the identity of the mode in each sequence I_i and the position of the motif Z_i as well as a text file with all the mode parameters ϕ and corresponding sequence logos. It also calculates and reports the best model amongst the models learned.

6.1.3 ExoDiversity

ExoDiversity implements a mixture model using a Bayesian approach with Gibbs Sampling to identify diverse mechanisms of binding of a target TF from data taken from ChIP-exo like experiments. It learns a joint distribution using heterogeneous datasets, viz. sequence data along with read counts data, corresponding to the peak regions identified by the ChIP-exo experiment. While learning the model parameters it partitions the data into disjoint subsets where each partition is characterized by a sequence motif along with the read distributions around it on both the forward and the reverse strands. These binding modes can be the binding sites of TFs different from the target TF (in case of indirect binding) or they may involve small variations of the canonical binding site corresponding to a different TF-DNA structural configurations at those sites.

The entire algorithm of ExoDiversity is written in python and C while the heatmap plots in the end are plotted using R. The processing of the input files is

done in python after which with the help of python multiprocessing package the parallel runs for the Gibbs sampler are set up from multiple initial states of the model and for multiple values of m which denotes the number of modes the data is partitioned into. The Gibbs sampler is written using C and on completion it returns the parameters of the learned model. When all the models are created a python script writes the output to various files, the best model is selected using Bayesian Information Criterion (BIC) and a link file is created indicating the best model.

The code for ExoDiversity is available on <https://github.com/NarlikarLab/exoDIVERSITY>. It can be downloaded and installed on any Linux machine and can be run by using the command `exoDiversity`. For certain arguments ExoDiversity needs some additional tools to be installed in the system: bedtools to process the ChIP-exo read files and the UCSC tool twoBitToFa along with the necessary genome file in .2bit format to plot the heatmap of the aligned sequence motifs. However, if the read counts corresponding to the peak regions are available, then we do not require the presence of bedtools. Also, the plotting of the heatmap is optional, and thus it is not compulsory to have UCSC tool and the genome file available to run ExoDiversity. The pre-requisites, instructions to download and install the software, the possible values for the optional parameters, description of the output files and some examples are given in the README file in the GitHub repository.

6.1.4 CISDIVERSITY

CISDIVERSITY is a module discovery tool used for finding diverse sequence architecture based on only the information contained in the DNA sequences. It identifies the underlying sequence motifs and based on their different distributions in the sequences it partitions the data into multiple modules. Each of these modules represents a functionality that is defined by the presence of certain motifs in the sequence with a definite probability. CISDIVERSITY is a model-based framework whose parameter values describe the structure of the data. They are learned by using Gibbs sampling where each of them is sampled systematically and the maximum a posterior estimate for the model is calculated. The model with the highest posterior score is considered as the best one.

The entire algorithm for CISDIVERSITY is primarily written in C. It uses mul-

tithreading to run the sampler in parallel from different initial states. The parallel processing is implemented using OpenMP of gcc and by default it deploys the algorithm over 2 processor cores, although this value can be changed by the user as per convenience. The final results include the sequence logos of all the identified motifs which are plotted using seqlogo package in python along with a heatmap of the complete partitioning and a circle plot that captures the module architectures in the data. The last 2 images are plotted using R with the required R package already being provided with the CISDIVERSITY installation. CISDIVERSITY code is available at the GitHub repository: <https://github.com/NarlikarLab/cisDIVERSITY>. It can be downloaded and installed in any Linux or Mac machine and after installation one simply needs to execute the following command to learn multiple modules present in the data: `(path to CISDIVERSITY)/learnDiverseModules [options]`. The system requirements, instructions to download and install, and a detailed list of the possible values of the optional parameters are given in the README file in the GitHub repository. It also has some examples along with descriptions of the content of the output files generated by CISDIVERSITY.

6.2 Biological data sets

6.2.1 DIVERSITY

The datasets used to get the results of DIVERSITY are all publicly available. Their GEO accession numbers are listed below:

- Fly CTCF ChIP and RNA-seq data (GSE24449) is from Ni et al. [Ni et al., 2012].
- Su(Hw) data (GSE23537) is from the same white pre-pupa stage [Nègre et al., 2011].
- Pita data (GSE76997) is from 0–12h embryos [Zolotarev et al., 2016].
- Human REST ChIP data (GSE32465, GSE49570) and neuronal RNA-seq data (GSE46562) is as processed and compiled by Rockowitz et al. [Rockowitz et al., 2014].

- All other human TF data are narrowpeak files from ENCODE in K562 (ENCFF144DMD, ENCFF264QLP, ENCFF440KMN, ENCFF443TUR, ENCFF503LMD, ENCFF529CTW, ENCFF484BSF, ENCFF433PKW, ENCFF602YIK, ENCFF886EVL).

A 200 bp neighbourhood around the summit (or centers, if no summit was given) of the reported ChIP regions is used as the set of input sequences in all the cases. Based on repeatMasker, as per UCSC genome browser [Karolchik et al., 2014], regions with more than 150 bp of repetitive nucleotides were disregarded. With the exception of a greater number of sampling start points (10 as opposed to the default 5) and models with modes between 1 and 20, DIVERSITY was run with the default settings for the rest of the parameters. Sequence conservation was calculated using PhastCons scores, and gene analysis was performed using the UCSC genome browser’s `refGene.txt` file. When plots were created, the ChIP signal for each region as reported by the individual studies served as the ChIP score of the sequence. When available, JASPAR [Bryne et al., 2007] motifs were shown for comparison. Identities of the potential TFs interacting with the motifs were found using TOMTOM [Gupta et al., 2007]. These motifs come from well-known databases and are occasionally created using ChIP data, but never from the sets used in this study. The nucleosome occupancy data for GM12878 is taken from ENCODE [Sloan et al., 2016]. Weblogo [Crooks et al., 2004] has been widely used in plotting the PWM’s logos in DIVERSITY. The Wilcoxon test is used for any comparisons between sets that involve a p-value. The R `vioplot` package was used to create the violin plots.

6.2.2 ExoDiversity

In silico mixed data

The summits for each of the TFs, FoxA1 and CTCF, were obtained by running MultiGPS [Mahony et al., 2014] on the BAM files containing reads from the ChIP-exo experiments for FoxA1 and CTCF in the MCF-7 cell line (GSE110502). MultiGPS was run separately for both the TFs with the corresponding control file. We considered regions of length 240 bp about the summits

Full datasets

- ChIP-nexus data for the TFs Sox2 and Oct4 are taken from Avsec et al. (2021) [Avsec et al., 2021], where they had performed the experiments in mouse ESCs. The read counts in bigWig format and the peak files which they had obtained using the MACS2 peak finder [Feng et al., 2012] are taken from the GEO database (GSE137193).
- Data for the ChIP-nexus experiment for the TF Twist in *Drosophila* was taken from GSE55306.
- When the individual sets of FoxA1 and CTCF were analyzed by ExoDiversity, along with the individual read counts from the ChIP-exo experiments, MultiGPS peaks for those specific TFs were used as input without mixing.

For each of the datasets, 240 bp regions about the summit were taken, with the repeats masked and only regions with at least 100 bp non-repetitive DNA were used as input. ExoDiversity was run with its default parameters on all the datasets except in the case of CTCF where the initial motif width was set to 40 as we intended to capture the core motif along with the additional flanking motifs [Nakahashi et al., 2013]

Other data

Data for DNA shape is taken from GBShape website [Chiu et al., 2015]. The UCSC genome browser was used to get the placental phastCons scores, and gene annotations were taken from the refGene file [Karolchik et al., 2014].

6.2.3 CISDIVERSITY

All the datasets used in the results of CISDIVERSITY were processed using the following steps. A 200 bp (or 400 bp) neighbourhood of the summits (if they were given or else the midpoint) of the regions were selected. In case two regions overlapped with each other, only the one occurring first in the list was retained. This was done to make sure the model would not identify the same location in the genome more than once. Only those regions containing at least 100 bp non-repetitive bases were kept and given to CISDIVERSITY for module discovery. All

repeating bases were change to Ns and as such were skipped by CISDIVERSITY. Here we list out all the data sets with their accession numbers in the next table (Table: 6.1).

Assay	Cell type	Accession (Source)
Paired-End Analysis of TSSs	mixed stage embryos (fly)	Processed set (Ni et al., 2010 [Ni et al., 2010])
CTCF ChIP-seq	white pre-pupa (fly)	GSM602326 (GEO)
CTCF ChIP-seq	H1 ESCs (human)	ENCFF991GTC (ENCODE)
GR ChIP-seq	treated A459 (human)	ENCFF835HHK (ENCODE)
EP300 ChIP-seq	treated A459 (human)	ENCFF774PLX (ENCODE)
eRNAs from GRO-cap	K562 (human)	processed set of eRNAs (Azofeifa et al., 2018 [Azofeifa et al., 2018]) from SRR1552480 (SRA)
EP300 ChIP-seq	K562 (human)	ENCFF821GNB (ENCODE)
STARR-seq	K562 (human)	processed (Lee et al., 2020 [Lee et al., 2020]) from ENCFF717VJK & ENCFF394DBM (ENCODE)
dELSS (ENCODE III)	K562 (human)	dELS from ENCFF464BRU (ENCODE)
pELSS (ENCODE III)	K562 (human)	pELS from ENCFF464BRU (ENCODE)
scATAC-seq	2-4 hr after egg laying (fly)	Table S1 from Cusanovich et al. (2018) [Cusanovich et al., 2018]
DNase-seq	H1 ESCs (human)	ENCFF030XPN (ENCODE)
ATAC-seq	20 mouse tissues	processed set (Liu et al., 2019 [Liu et al., 2019])

Table 6.1: Datasets on which CISDIVERSITY was run in this study. Processed data or peaks were used from the original study in each case.

Chapter 7

Conclusions and future directions

In this chapter, we will first look at some of the advantages of using mixture models to understand high throughput experimental data. We will review all three of our methods and their contributions towards providing a comprehensive understanding of the TF-DNA interactions using the complete set of regions reported in an experiment. Next, we will discuss some of the limitations of our methods and propose some ways to ameliorate them. Lastly, we will discuss some of the other approaches that are being adopted to uncover the underlying TF motifs and their genomic arrangements in different cellular contexts.

With rapid advancements in sequencing technologies along with computational approaches, we now have a huge amount of experimental data that can be investigated using multiple complementary methods with the goal of understanding the architecture of the regulatory regions. Various experimental strategies have been developed to map different regulatory activities along the genome using a multitude of high-throughput sequencing assays. The large volume of data being generated from these studies have been used to build integrative approaches [Hawkins et al., 2010] to shed more light on the regulatory mechanisms at play during some of the critical biological processes such as development, homeostasis, response to stimuli, etc. In an effort to understand how proteins like TFs bind to the regulatory regions and interact amongst themselves in a spatial context, an extensive database called TFSyntax [Zhao, 2023] was recently made available that provide a global view of the arrangement of mammalian TF binding sites along with cooperative binding and co-localization of TFs. Although such a database provide us with a way to compare and explore the regulatory landscape, the question of how these TFs act

alone or with other TFs to affect regulation related activities still remains largely unanswered. Examining the data obtained from various high throughput experiments to determine the reasons behind the reporting of all the regions followed by analyzing them in context of chromatin structure, sequence conservation and genomic functionality can be helpful in answering multiple questions:

- Where do the TF binding sites lie within the regulatory regions? Knowing this will shed light on which variations/mutations are likely to cause misregulation and lead to disorders.
- In which cell types or under which conditions are these regulatory regions and binding sites active?
- Which genes do these regions regulate, and what role do they play in regulation (activating or repressing)?

Here we will go over the approach of modelling the data from the high throughput sequencing experiments as a mixture of different events based on the sequence composition alone or along with other nucleotide level information and how this helps to explain the complete set of reported regions.

7.1 Mixture model-based probabilistic frameworks

High throughput sequencing experiments generally report a set of genomic regions that share a common characteristic, which the experiment was designed to profile. The DNA sequence data corresponding to these regions contain the footprints of several TFs that endow these regions with a certain functionality, leading them to be reported in the experiment. We intended to identify these TF binding sites or motifs and decipher the reasons why each region was listed by the experiment. We considered the standard representation of the TF motifs as PWMs, where if w is the length of the motif, it is denoted by w independent categorical distributions over $\{A, C, G, T\}$. Although, there are approaches like inhomogeneous parsimonious Markov models [Eggeling et al., 2015] that can capture the higher order intra-motif dependencies between the positions of a TF binding site, we have assumed a simpler representation without any positional dependency within the motifs in this thesis.

7.1.1 DIVERSITY

Chromatin immunoprecipitation-based experiments are typically used to capture the interactions between a protein of interest (POI) and the DNA. They provide us with a set of genomic regions that are associated with our POI in some or the other way: they can be binding sites of our POI, or they can be binding sites of other TFs that interact with the POI forming protein complexes, or they can be the binding sites of the various configurations assumed by the POI. To elucidate these multiple modes of binding associated with our POI, our first objective was to model the set of genomic sequences as a mixture of regions, where each region might contain at most one binding site of a TF related to our POI. Our method DIVERSITY used a probabilistic partitioning scheme where it clusters the sequences based on the motif present in it. The positions of the motifs in the sequences and the mode identity for every sequence, widths of the motifs are all learned by using the collapsed Gibbs sampling algorithm while marginalizing over the other parameters. The optimal number of motifs present in a given dataset is not known apriori, thus, DIVERSITY learns multiple models by varying the number of modes in a given range and then uses Bayesian model selection to select the best model. By default, DIVERSITY uses the one occurrence per sequence (OOPS) scheme to look for motifs in the data, and thus the modes discovered by DIVERSITY account for all the sequences given in the input. The modes correspond to different features of sequence conservation, chromatin structure, as well as the distances from the nearest TSSs, indicating that they may be playing functionally important roles in the binding of the POI.

We considered the DNA sequences over a 200 bp window about the summits reported by ChIP-seq experiments and in case of peak regions, we took a 200 bp window about their mid-points. Broader regions have the possibility of having more than one TF binding site, which may or may not be associated with our POI. Thus, our assumption of having a single binding site per region will not be suitable in that case. Also, when considering these genomic regions, we ensure that they do not overlap with any other region in the dataset. This will prevent the same TF motif from being counted more than once in the algorithm, and will also not report the same chromosomal positions multiple times in the output. Parts of the sequences that contain repetitive nucleotides and are marked as repeat regions are masked by Ns and are not considered a part of any TF motif. Therefore, our

algorithm detects only contiguous subsequences without any Ns as TF motifs.

Another important aspect of motifs identified by DIVERSITY is that they can be of different widths and can even have gaps in them. This is possible because of the sampling of the motif widths which allows DIVERSITY to explore the regions on either sides of the motif, enabling it to capture motifs having different widths.

Now, although the premise for the mixture model framework of DIVERSITY considered at most a single TF motif per sequence, in case of homotypic binding or cooperative binding, there can be more than one motif present in a single region. For example, if we look at the ChIP-seq dataset for CTCF in fly [Ni et al., 2012], DIVERSITY detects a single occurrence of the Pita motif in every sequence of mode 5 (Fig: 3.2). In the same dataset, CISDIVERSITY detected a module (module 3 in Fig: 5.6) that contained 2 instances of the Pita motif. However, the two motifs in these sequences were not situated at a constant distance apart from each other, and thus even by taking a large "motif" width to encompass both the Pita motifs DIVERSITY would not have been able to detect them in those regions.

To identify motifs present in clusters or engaged in cooperative binding activity which are usually found in open chromatin regions, regulatory regions, we needed to consider each region as part of a distinct module, where a module would be represented by a probability distribution over the presence and absence of a set of motifs. Based on the proportion of the motifs present in a module, it will have a different functionality. To capture all these features, we developed a new method CISDIVERSITY that can partition the data into various modules based on the motifs present in the data while learning the motifs themselves along with their distributions over each module.

7.1.2 ExoDiversity

Exonuclease-based ChIP experiments like ChIP-exo and ChIP-nexus provide a highly improved near single base pair resolution of protein binding sites. This leads to an efficient downstream analysis, where sharp and precise peaks in the read distributions delineate the binding sites. These read distributions with their corresponding TF motifs can be used to study the diverse mechanisms and of binding of the TF as well as other TFs involved in cooperative binding, if any. Our second objective involved using the read distributions together with the sequence motifs to identify the modes of binding of a target TF from ChIP-exo or

ChIP-nexus data. We built upon our previous method DIVERSITY to include the reads component and learn a joint distribution over the DNA sequences and their nucleotide level read counts in our new method ExoDiversity. The set of sequences was considered a mixture of binding modes of a TF where each mode is characterized by a sequence motif along with distinct positive and negative strand read distributions. The sequence motif is defined as a PWM, and we consider PWMs over 0 and 1 for the short windows corresponding to the read counts on both the strands, presumably capturing the peaks of the read distributions about the motifs. The read counts are binarized based on the median of the dataset and are modelled using a Bernoulli distribution at each position within a read window. Thus, similar to the PWM matrices for the motifs, we now have two extra matrices of widths corresponding to that of the read windows on each strand. The probability of a mode is thereby calculated a product of categorical probabilities for the motif and Bernoulli probabilities for the two read windows. We used collapsed Gibbs sampling to determine the motif start positions in every sequence, the distance between the motif start positions and that of the read windows for each mode, the mode identity of each sequence along with the widths of the motifs while marginalizing over the other parameters. As the best number of modes that the data can be divided into is not known beforehand, ExoDiversity learns different models by varying the number of modes and then uses Bayesian Information Criterion to select the best model. Modes discovered by ExoDiversity were found to have distinct DNA shape features and varying conservation signals, providing insights on the structure and function of the complexes formed because of the TF-DNA interactions.

ExoDiversity was given 240 bp sequences about the summits obtained from the experiments along with the file containing the reads either in BAM or bedGraph format. The read counts for every position corresponding to the DNA sequences for the positive and the negative strands are extracted from the reads file using bedtools and given to the algorithm as input. We assumed that each sequence will contain a single binding mode. This ensures that each sequence in the data will be assigned to one of the modes detected by ExoDiversity, which can sometimes also be a non-informative mode that might be an artifact of the experiment. ExoDiversity partitions the data based on the variations in the sequence motifs and the read distributions, and can consequently identify modes with small nucleotide level

differences along with their corresponding differences in the read distributions.

As the exonuclease-based ChIP experiments have been shown to produce precise read counts peaks near the binding sites, we leveraged this fact and considered a small size for the read window (default width is 5) for the results shown in this thesis. The widths of the read windows for all the modes were kept constant for both the strands throughout the algorithm. This feature was later modified and sampling of the width for the reads was included. It was implemented similarly as the sampling of the motif width, where the width is allowed to increase or decrease or remain the same at both the left and the right side of the window. The width sampling for the read windows was done for both the strands and for all the modes every time the motif widths were sampled. The minimum width possible for the read windows was kept as 0 in which case there would essentially be no contribution of the reads from that window and the maximum width was considered as 20.

In a ChIP-exo experiment, when the 5' ends of the reads are mapped to the genome, they do not exactly coincide with the beginning of the protein binding site. They are usually located very near to but at a fixed distance away from the binding site [Rhee and Pugh, 2011]. To account for such a feature, we incorporated the parameter for the offsets or the relative distances between the starting position of the motifs and that of the read windows for every mode. The sampling for the offset was basically finding out the best position of the read window with respect to the motif. We assumed that although these offsets for the positive and negative strand read windows for a particular mode might be different from each other, they will remain the same across all the instances of TF binding in that mode. For the positive strand in a given mode, the starting position of the read window could be anywhere before the 3' end of the motif, and for the negative strand it was restricted to anywhere before the 5' end of the motif. This feature was also modified in the later version of ExoDiversity where the read windows in both the strands were allowed to vary along the entire length of the sequence and their optimal position was selected using sampling.

The original ExoDiversity algorithm considered the read counts as binary values and accordingly modelled them using independent Bernoulli distributions. Instead of binarizing the read counts, if we consider their actual values, we see that they are discrete count values and can be modelled using independent Poisson distributions

for each position in the read window. Since we already had the framework of ExoDiversity, we built an alternate Poisson-Gamma mixture model over the reads while keeping the motif part the same as before. We plan to examine the results in details before including this option in the ExoDiversity framework.

Proportion of CTCF and FoxA1 sequences	Recovered sequences	Percentage of unrecovered sequences
0% and 100%	Best model with 3 modes: 5000 of 5000 FoxA1	0% from FoxA1
20% and 80%	Best model is with 4 modes: 947 of 1000 CTCF 3171 of 4000 FoxA1	5.3% from CTCF 20.72% from FoxA1
50% and 50%	Best model is with 7 modes: 2171 of 2500 CTCF 2124 of 2500 FoxA1	13.16% from CTCF 15.04% from FoxA1
80% and 20%	Best model is with 5 modes: 3186 of 4000 CTCF 919 of 1000 FoxA1	20.35% from CTCF 8.1% from FoxA1
100% and 0%	Best model is with 6 modes: 4435 of 5000 CTCF	11.3% from CTCF

Table 7.1: ExoDiversity was run on the in silico mixed datasets by varying the proportion of sequences from both FoxA1 and CTCF ChIP-exo data.

The in silico mixed data which we had used to evaluate the performance of ExoDiversity contained sequences from ChIP-exo datasets for FoxA1 and CTCF in equal proportions (Fig: 4.3). To review the performance of ExoDiversity more thoroughly, we varied the proportion of sequences for both FoxA1 and CTCF and created 5 different datasets each containing a total of 5000 sequences. We

ran ExoDiversity on all of them and Table: 7.1 shows the results. We observe that decreasing the number of sequences for a certain TF does not increase the proportion of unrecovered sites for that TF.

7.1.3 CISDIVERSITY

We have seen that ChIP-based high throughput experiments, like ChIP-seq, profile all those regions in the genome that are associated with a target protein and modifications to the ChIP protocol (ChIP-exo/ ChIP-nexus) helped to locate these interactions with a higher precision. Many other gene regulation related biochemical activities are also commonly measured across the genome using different high throughput sequencing experiments. Some of the examples include detecting open chromatin regions, active transcription start sites, enhancer regions. These types of regions mostly harbor binding sites of multiple TFs that might be present nearby or are a part of the protein-DNA complex forming there. Our third objective was to study these regions in terms of the combinations of motifs present in their corresponding DNA sequences. We developed our method CISDIVERSITY with the goal of explaining the role of each of the reported regions in a given high throughput experiment, based on the sequence signatures present in them. CISDIVERSITY partitions the whole set of sequences into different modules, where each module represents a functionality that is defined by which motifs are present in it and in what proportions.

The model for CISDIVERSITY consists of three levels: the motifs present in the dataset, the distribution of the motifs in each module, the module to which each sequence belongs. The motifs are modelled as PWMs, with every position within the PWM denoting an independent categorical distribution over the 4 nucleotides. The modules are defined by the product of Bernoulli distributions over each of the motifs present in the dataset. The module identity of a sequence is established using the probability of the motifs it contains, taking into consideration that each motif can only appear once in each sequence and that the occurrences of the motifs cannot overlap. We use collapsed Gibbs sampling to identify the motif start positions for all the motifs present in each of the sequences, the module identity for each sequence and the width of each motif while marginalizing over the other parameters. The number of motifs as well as the optimal number of modules present in the data are both unknown. CISDIVERSITY assumes an upper

bound on both these counts and in the process of learning the parameters of the model, it detects the correct number of motifs and modules. In case this number is less than the upper bound on them, the remaining motifs and modules generally stay empty or contain very less number of sequences.

CISDIVERSITY was used on several datasets from different types of high throughput experiments. We considered sequences which were, in general, 200 bp regions about the summit, although in some of the cases we also ran CISDIVERSITY over 400 bp or 1000 bp regions about the summit as well. As we move farther away from the summits reported by the experiments, motifs of low complexity co-factor proteins become more frequent and their roles in the binding mechanisms of the target protein no longer remains clear and well-defined. Similar to previous methods, CISDIVERSITY does not provide the identity of the motifs discovered or the functionality of the modules detected by it. We need to perform other downstream analysis like examining the sequence conservation, distance from the closet TSS etc. at the regions in the modules to infer their roles in the cis-regulatory activities. Based on only the TF footprints present in the sequences, CISDIVERSITY could identify specific modules enriched in motifs of promoter TFs and other modules with a higher proportion of enhancer TFs. It could also capture the event where clusters of binding sites of the same TF were present. This demonstrated that CISDIVERSITY is able to cluster the sequences appropriately based on just their distinct underlying sequence architecture. In cases where co-operative binding between TFs takes place in the form of looping of the chromatin, then the binding sites of the TFs involved may not be present within a small linear neighborhood. In these situations, CISDIVERSITY clusters the sequences into modules, where it might be the case that only a single TF motif is enriched in each cluster.

The formulation of CISDIVERSITY is such that, given a new sequence and a pre-trained CISDIVERSITY model, we can predict the probability of the sequence belonging to each module in the given model as well as calculate the likelihood of the sequence given the parameters of the model. If we have two different models and a DNA sequence, we can subsequently calculate the likelihood of the sequence belonging to each of the models and thereby the log odds score. This measure can be used to classify genomic regions as likely to be generated from a certain type of experiment based on their sequence content and the models learned by CISDIVERSITY from the corresponding experimental datasets.

7.2 Limitations of sampling-based methods

In this thesis, we have developed different model-based Bayesian frameworks that models the regions reported in high throughput sequencing experiments as a mixture of diverse components. These components are presumed to follow certain distributions, and the goal is to learn the parameters of these distributions that will best describe our data. To estimate the values of the parameters, we use the Gibbs sampling algorithm to calculate the maximum a posteriori (MAP) estimate and the parameter values corresponding to that are considered to be optimal to describe the given dataset. Gibbs sampling allows us to estimate a complex joint posterior distribution by sampling the values of each of the components from their conditional distributions. These conditional distributions should be easily derivable, from which we can draw random samples economically. Although using Gibbs sampling allows us to deal with a variety of distributions, one major drawback is correlation between consecutive samples leads to a slow convergence. Under such circumstances, Gibbs sampler is often run for a long time so that the sampler can move on from the dependent samples. Also, the sampler can get stuck at a local maxima for a particularly unlucky initial condition, not reach close to the MAP estimate. To avoid such a situation, the sampler is generally run from multiple random initializations. The time for convergence of the sampler depends upon the structure of the parameter search space. For a dataset that contains distinctly separable modes or modules, the sampler will take lesser number of iterations to identify them, as compared to even a smaller dataset that contains closely related modes without any well-defined characteristics. All of these make sampling-based algorithms more time-consuming than the enrichment-based or optimization-based algorithms for motif discovery.

7.3 Deep learning to identify sequence signatures

With rapid advancements in the computational technologies, deep learning-based algorithms have now become economical and efficient to implement and have proven effective in modeling the sequence specificity of protein-DNA binding. DeepBind developed by Alipanahi et al. [Alipanahi et al., 2015] is one such method that employs deep convolutional neural networks to discover new patterns

within sequences without knowing their locations in the sequences. The convolutional stage scans for motif detectors that are $4 \times m$ matrices where m is the length of the motif, followed by the rectification stage that isolates the positions with a good pattern match. Finally, the values are fed into a nonlinear neural network to produce a score, which is then used in back propagation to improve the performance of the network. DeepBind can be applied to sequencing data with a large number of sequences and was shown to generalize well across several technologies without further tuning for technology-specific biases. Another methodology using deep learning and convolutional neural network is DeepLift (Deep Learning Important FeaTures) [Shrikumar et al., 2017] that has been applied to the motif finding problem. It decomposes the output prediction of a neural network on a specific input by using back propagation. DeepLift seeks to explain the difference in the output from the reference in terms of the difference in the inputs from the reference. Interpreting the sequence level information captured in the deep learning models require feature attribution over broad input DNA regions (500 bp to 1000 bp) and it typically does not adapt well across cell types. Nonetheless, it has been used with a higher degree of success to detect cis-regulatory motif syntax for high resolution TF occupancy (ChIP-nexus) data sets [Avsec et al., 2021].

Lately, unsupervised deep learning models to identify the TF binding landscape in open chromatin regions have gained prominence and have performed comparably well to current motif discovery methods. BindVAE, based on Dirichlet variational autoencoder (VAE) can capture cell type specific binding preferences of different TFs as well as cooperative binding patterns of TFs as latent variables of an autoencoder [Kshirsagar et al., 2022]. It considers the DNA sequences of the peak regions as a collection of k-mers that have specific distributions over different categories. These categories can be TF binding sites, low complexity regions, background genomic regions as well as bias signals corresponding to the DNA cleaving enzymes and each category is represented as a multinomial distribution over the k-mers. This method uses k-mers with wildcards with an inexact matching technique to capture degenerate TF binding motifs. The dimension of the latent space correspond to the number of topics it can learn, and the multiple dimensions are later used to detect the variations in the binding patterns of different TFs across cell types. Ultimately, BindVAE implements a probabilistic model that optimizes the likelihood of the data (decoder), while the latent variable is sampled from the

Dirichlet posterior distribution (encoder).

With increase in focus on the interpretability of the deep neural networks, together with their ability to identify complex sequence features, have made deep learning-based methods more popular in identifying regulatory elements from high throughput sequencing data. This opens up avenues to process experimental data not only from ChIP-based experiments, but also from other high throughput experiments that profile diverse biochemical activities along the genome. We can also have multimodal data which can be genomic annotations along with experimental data that can be used to provide correlations between cis-regulatory elements and their functionalities.

Bibliography

- Arttu Jolma, Jian Yan, Thomas Whittington, Jarkko Toivonen, Kazuhiro R Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, et al. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, 2013.
- Nikolay Zolotarev, Anna Fedotova, Olga Kyrchanova, Artem Bonchuk, Aleksey A Penin, Andrey S Lando, Irina A Eliseeva, Ivan V Kulakovskiy, Oksana Maksimenko, and Pavel Georgiev. Architectural proteins pita, zw5, and zipic contain homodimerization domain and support specific long-range interactions in drosophila. *Nucleic acids research*, 44(15):7228–7241, 2016.
- Shira Rockowitz, Wen-Hui Lien, Erika Pedrosa, Gang Wei, Mingyan Lin, Keji Zhao, Herbert M Lachman, Elaine Fuchs, and Deyou Zheng. Comparison of rest cistromes across human cell types reveals common and context-specific functions. *PLoS computational biology*, 10(6): e1003671, 2014.
- U Ohler, GC Liao, H Niemann, and GM Rubin. Computational analysis of core promoters in the, 2002.
- Amelia Casamassimi and Alfredo Ciccodicola. Transcriptional regulation: Molecules, involved mechanisms, and misregulation. *International Journal of Molecular Sciences*, 20(6):1281, March 2019. doi: 10.3390/ijms20061281. URL <https://doi.org/10.3390/ijms20061281>.
- Nils Weinhold, Anders Jacobsen, Nikolaus Schultz, Chris Sander, and William Lee. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, 46(11):1160–1165, 2014.
- Michael Levine and Robert Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945):147–151, 2003.
- Nathan C Sheffield and Terrence S Furey. Identifying and characterizing regulatory sequences in the human genome with chromatin accessibility assays. *Genes*, 3(4):651–670, 2012.
- Steven L Salzberg. Open questions: How many genes do we have? *BMC biology*, 16(1):1–3, 2018.
- Fátima Gebauer and Matthias W Hentze. Molecular mechanisms of translational control. *Nature reviews Molecular cell biology*, 5(10):827–835, 2004.

- Monica Tyagi, Nasir Imam, Kirtika Verma, and Ashok K. Patel. Chromatin remodelers: We are the drivers!! *Nucleus*, 7(4):388–404, July 2016. doi: 10.1080/19491034.2016.1211217. URL <https://doi.org/10.1080/19491034.2016.1211217>.
- Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7:29–59, 2006.
- Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin AM Semple, Martin S Taylor, Pär G Engström, Martin C Frith, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6):626–635, 2006.
- W Su, S Jackson, R Tjian, and H Echols. Dna looping between sites for transcriptional activation: self-association of DNA-bound Sp1. *Genes & development*, 5(5):820–826, 1991.
- ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *nature*, 447(7146):799, 2007.
- Albert J Courey and Songtao Jia. Transcriptional repression: the long and the short of it. *Genes & development*, 15(21):2786–2796, 2001.
- Qiliang Li, Kenneth R Peterson, Xiangdong Fang, and George Stamatoyannopoulos. Locus control regions. *Blood, The Journal of the American Society of Hematology*, 100(9):3077–3086, 2002.
- Annie Claringbould and Judith B Zaugg. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends in molecular medicine*, 27(11):1060–1073, 2021.
- David N Cooper. Regulatory mutations and human genetic disease. *Annals of medicine*, 24(6):427–437, 1992.
- Tong Ihn Lee and Richard A Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013.
- Laura A Lettice, Simon JH Heaney, Lorna A Purdie, Li Li, Philippe de Beer, Ben A Oostra, Debbie Goode, Greg Elgar, Robert E Hill, and Esther de Graaff. A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, 12(14):1725–1735, 2003.
- Xiao-Fen Chen, Yun-wu Zhang, Huaxi Xu, and Guojun Bu. Transcriptional regulation and its misregulation in alzheimer’s disease. *Molecular brain*, 6(1):1–9, 2013.
- Giovanna Ambrosini, Ilya Vorontsov, Dmitry Penzar, Romain Groux, Oriol Fornes, Daria D Nikolaeva, Benoit Ballester, Jan Grau, Ivo Grosse, Vsevolod Makeev, et al. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome biology*, 21(1):1–18, 2020.

- Valentina Boeva. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in genetics*, 7:24, 2016.
- Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin Van Der Lee, Adrien Bessy, Jeanne Cheneby, Shubhada R Kulkarni, Ge Tan, et al. Jaspas 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1):D260–D266, 2018.
- Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, et al. Hocomoco: towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic acids research*, 46(D1):D252–D259, 2018.
- Volker Matys, Olga V Kel-Margoulis, Ellen Fricke, Ines Liebich, Sigrid Land, A Barre-Dirrie, Ingmar Reuter, D Chekmenev, Mathias Krull, Klaus Hornischer, et al. Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl_1):D108–D110, 2006.
- Maxwell A Hume, Luis A Barrera, Stephen S Gisselbrecht, and Martha L Bulyk. Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein–dna interactions. *Nucleic acids research*, 43(D1):D117–D122, 2015.
- Thomas D Schneider and R Michael Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic acids research*, 18(20):6097–6100, 1990.
- Dominic Schmidt, Michael D Wilson, Christiana Spyrou, Gordon D Brown, James Hadfield, and Duncan T Odom. ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions. *Methods*, 48(3):240–248, 2009.
- Peter J Park. Chip-seq: advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10):669–680, 2009.
- Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):1–10, 2009.
- Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.
- Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoutte, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.
- Yuchun Guo, Shaun Mahony, and David K Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 2012.

- Dustin E Schones, Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898, 2008.
- Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897–903, 2008.
- Ho Sung Rhee and B Franklin Pugh. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current protocols in molecular biology*, 100(1):21–24, 2012.
- Aurelien A Serandour, Gordon D Brown, Joshua D Cohen, and Jason S Carroll. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome biology*, 14(12):R147, 2013.
- Qiye He, Jeff Johnston, and Julia Zeitlinger. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology*, 33(4):395–401, 2015.
- Matthew J Rossi, William KM Lai, and B Franklin Pugh. Simplified ChIP-exo assays. *Nature communications*, 9(1):1–13, 2018.
- Carl Wu. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase i. *Nature*, 286(5776):854–860, 1980.
- Lingyun Song and Gregory E Crawford. Dnase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, 2010(2):pdb-prot5384, 2010.
- Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1):21–29, 2015.
- Chain Monte Carlo. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, 581:540, 2004.
- George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
- Jun S Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.
- Sneha Mitra, Anushua Biswas, and Leelavati Narlikar. Diversity in binding, regulation, and evolution revealed from high-throughput chip. *PLoS computational biology*, 14(4):e1006090, 2018.

- Anushua Biswas and Leelavati Narlikar. Resolving diverse protein–dna footprints from exonuclease-based chip experiments. *Bioinformatics*, 37(Supplement_1):i367–i375, 2021a.
- Anushua Biswas and Leelavati Narlikar. A universal framework for detecting cis-regulatory diversity in dna regions. *Genome research*, 31(9):1646–1662, 2021b.
- Cricket A Sloan, Esther T Chan, Jean M Davidson, Venkat S Malladi, J Seth Strattan, Benjamin C Hitz, Idan Gabdank, Aditi K Narayanan, Marcus Ho, Brian T Lee, et al. Encode data at the encode portal. *Nucleic acids research*, 44(D1):D726–D732, 2016.
- Kevin Struhl. Mechanisms for diversity in gene expression patterns. *Neuron*, 7(2):177–181, 1991.
- Peggy J Farnham. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, 10(9):605–616, 2009.
- Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, 14(2):225–237, 2013.
- Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2:28–36, 1994.
- Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.
- Ralf Eggeling, Ivo Grosse, and Jan Grau. Inmode: tools for learning and visualizing intra-motif dependencies of dna binding sites. *Bioinformatics*, 33(4):580–582, 2017.
- Leelavati Narlikar. Mumod: a bayesian approach to detect multiple modes of protein–dna binding from genome-wide chip data. *Nucleic acids research*, 41(1):21–32, 2013.
- Leelavati Narlikar, Raluca Gordân, Uwe Ohler, and Alexander J Hartemink. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22(14):e384–e392, 2006.
- Jennifer E Phillips and Victor G Corces. Ctfc: master weaver of the genome. *Cell*, 137(7):1194–1211, 2009.
- Eilon Sharon, Shai Lubliner, and Eran Segal. A feature-based approach to modeling protein–dna interactions. *PLoS computational biology*, 4(8):e1000154, 2008.
- Xiaochun Ni, Yong E Zhang, Nicolas Negre, Sidi Chen, Manyuan Long, and Kevin P White. Adaptive evolution and the birth of ctfc binding sites in the drosophila genome. *PLoS biology*, 10(11):e1001420, 2012.

- Oksana Maksimenko, Marek Bartkuhn, Viacheslav Stakhov, Martin Herold, Nickolay Zolotarev, Theresa Jox, Melanie K Buxa, Ramona Kirsch, Artem Bonchuk, Anna Fedotova, et al. Two new insulator proteins, pita and zipic, target cp190 to chromatin. *Genome research*, 25(1): 89–99, 2015.
- Uwe Ohler and David A Wassarman. Promoting developmental transcription. *Development*, 137(1):15–26, 2010.
- Leelavati Narlikar. Multiple novel promoter-architectures revealed by decoding the hidden heterogeneity within the genome. *Nucleic acids research*, 42(20):12388–12403, 2014.
- Dhawal Jain, Sandro Baldi, Angelika Zabel, Tobias Straub, and Peter B Becker. Active promoters give rise to false positive ‘phantom peaks’ in chip-seq experiments. *Nucleic acids research*, 43(14):6959–6968, 2015.
- Jutta Vogelmann, Alessandro Valeri, Emmanuelle Guillou, Olivier Cuvier, and Marcello Nollmann. Roles of chromatin insulator proteins in higher-order chromatin organization and transcription regulation. *Nucleus*, 2(5):358–369, 2011.
- Nicolas Nègre, Christopher D Brown, Lijia Ma, Christopher Aaron Bristow, Steven W Miller, Ulrich Wagner, Pouya Kheradpour, Matthew L Eaton, Paul Loriaux, Rachel Sealfon, et al. A cis-regulatory map of the drosophila genome. *Nature*, 471(7339):527–531, 2011.
- Martha L Bulyk, Philip LF Johnson, and George M Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255–1261, 2002.
- Metewo Selase Enuameh, Yuna Asriyan, Adam Richards, Ryan G Christensen, Victoria L Hall, Majid Kazemian, Cong Zhu, Hannah Pham, Qiong Cheng, Charles Blatti, et al. Global analysis of drosophila cys2-his2 zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome research*, 23(6):928–940, 2013.
- Irfan A Qureshi, Solen Gokhan, and Mark F Mehler. Rest and corest are transcriptional and epigenetic regulators of seminal neural fate decisions. *Cell cycle*, 9(22):4477–4486, 2010.
- Deborah J Greenway, Miyoko Street, Aaron Jeffries, and Noel J Buckley. Re1 silencing transcription factor maintains a repressive chromatin environment in embryonic hippocampal neural stem cells. *Stem cells*, 25(2):354–363, 2007.
- Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2):W202–W208, 2009.
- Lezanne Ooi and Ian C Wood. Chromatin crosstalk in development and disease: lessons from rest. *Nature Reviews Genetics*, 8(7):544–554, 2007.

- Fabrice Bouilloux, Jérôme Thireau, Stéphanie Ventéo, Charlotte Farah, Sarah Karam, Yves Dauvilliers, Jean Valmier, Neal G Copeland, Nancy A Jenkins, Sylvain Richard, et al. Loss of the transcription factor *meis1* prevents sympathetic neurons target-field innervation and increases susceptibility to sudden cardiac death. *Elife*, 5:e11627, 2016.
- Nikolaj Dietrich, Mads Lerdrup, Eskild Landt, Shuchi Agrawal-Singh, Mads Bak, Niels Tommerup, Juri Rappsilber, Erik Södersten, and Klaus Hansen. Rest-mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS genetics*, 8(3):e1002494, 2012.
- Ye He and Patrizia Casaccia-Bonnel. The yin and yang of *yy1* in the nervous system. *Journal of neurochemistry*, 106(4):1493–1502, 2008.
- Kathryn J Brayer and David J Segal. Keep your fingers off my dna: protein–protein interactions mediated by c2h2 zinc finger domains. *Cell biochemistry and biophysics*, 50(3):111–131, 2008.
- Chenzhuo Feng, Jiejie Li, and Zhiyi Zuo. Expression of the transcription factor regulatory factor *x1* in the mouse brain. *Folia Histochemica et Cytobiologica*, 49(2):344–351, 2011.
- Maria Bergsland, Daniel Ramsköld, Cécile Zaouter, Susanne Klum, Rickard Sandberg, and Jonas Muhr. Sequentially acting sox transcription factors in neural lineage development. *Genes & development*, 25(23):2453–2464, 2011.
- Meenakshi B Kannan, Vera Solovieva, and Volker Blank. The small *maf* transcription factors *maff*, *mafg* and *mafk*: current knowledge and perspectives. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(10):1841–1846, 2012.
- Rory Johnson, Christina Hui-leng Teh, Galih Kunarso, Kee Yew Wong, Gopalan Srinivasan, Megan L Cooper, Manuela Volta, Sarah Su-ling Chan, Leonard Lipovich, Steven M Pollard, et al. Rest regulates distinct transcriptional networks in embryonic and neural stem cells. *PLoS biology*, 6(10):e256, 2008.
- Nicola Wiechens, Vijender Singh, Triantafyllos Gkikopoulos, Pieta Schofield, Sonia Rocha, and Tom Owen-Hughes. The chromatin remodelling enzymes *snf2h* and *snf2l* position nucleosomes adjacent to *ctcf* and other transcription factors. *PLoS genetics*, 12(3):e1005940, 2016.
- Deyou Zheng, Keji Zhao, and Mark F Mehler. Profiling *re1*/rest-mediated histone modifications in the human genome. *Genome biology*, 10(1):1–20, 2009.
- David C Bedford, Lawryn H Kasper, Tomofusa Fukuyama, and Paul K Brindle. Target gene context influences the transcriptional requirement for the *kat3* family of *cbp* and *p300* histone acetyltransferases. *Epigenetics*, 5(1):9–15, 2010.
- Axel Visel, Matthew J Blow, Zirong Li, Tao Zhang, Jennifer A Akiyama, Amy Holt, Ingrid Plajzer-Frick, Malak Shoukry, Crystal Wright, Feng Chen, et al. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, 2009.

- Raffaella Pippa, Ana Dominguez, Raquel Malumbres, Akinori Endo, Elena Arriazu, Nerea Marcotegui, Elizabeth Guruceaga, and María D Odero. Myc-dependent recruitment of runx1 and gata2 on the set oncogene promoter enhances pp2a inactivation in acute myeloid leukemia. *Oncotarget*, 8(33):53989, 2017.
- Douglas H Phanstiel, Kevin Van Bortle, Damek Spacek, Gaelen T Hess, Muhammad Saad Shamim, Ido Machol, Michael I Love, Erez Lieberman Aiden, Michael C Bassik, and Michael P Snyder. Static and dynamic dna loops form ap-1-bound activation hubs during macrophage development. *Molecular cell*, 67(6):1037–1048, 2017.
- John J Andreucci, Diane Grant, David M Cox, Lyn K Tomc, Ron Prywes, David J Goldhamer, Natalie Rodrigues, Pierre-André Bédard, and John C McDermott. Composition and function of ap-1 transcription complexes during muscle cell differentiation. *Journal of Biological Chemistry*, 277(19):16426–16432, 2002.
- Hideyuki Yanai, Hideo Negishi, and Tadatsugu Taniguchi. The IRF family of transcription factors. *OncoImmunology*, 1(8):1376–1386, November 2012. doi: 10.4161/onci.22475. URL <https://doi.org/10.4161/onci.22475>.
- Ronja Hollstein, Benedikt Reiz, Lucas Kötter, Alev Richter, Susen Schaaake, Katja Lohmann, and Frank J Kaiser. Dystonia-causing mutations in the transcription factor thap1 disrupt hcf1 cofactor recruitment and alter gene expression. *Human molecular genetics*, 26(15):2975–2983, 2017.
- Takashi Sekiya, Uma M Muthurajan, Karolin Luger, Alexei V Tulin, and Kenneth S Zaret. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor foxa. *Genes & development*, 23(7):804–809, 2009.
- Alexander W Bruce, Andrés J López-Contreras, Paul Flicek, Thomas A Down, Pawandeep Dhami, Shane C Dillon, Christoph M Koch, Cordelia F Langford, Ian Dunham, Robert M Andrews, et al. Functional diversity for rest (nr5f) is defined by in vivo binding affinity hierarchies at the dna sequence level. *Genome research*, 19(6):994–1005, 2009.
- Yongheng Chen, Darren L Bates, Raja Dey, Po-Han Chen, Ana Carolina Dantas Machado, Ite A Laird-Offringa, Remo Rohs, and Lin Chen. Dna binding by gata transcription factor suggests mechanisms of dna looping and long-range gene regulation. *Cell reports*, 2(5):1197–1206, 2012.
- Daechan Park, Yaelim Lee, Gurvani Bhupindersingh, and Vishwanath R Iyer. Widespread misinterpretable chip-seq bias in yeast. *PLoS one*, 8(12):e83506, 2013.
- Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science*, 339(6123):1074–1077, 2013.

- Marina Lizio, Jayson Harshbarger, Imad Abugessaisa, Shuei Noguchi, Atsushi Kondo, Jessica Severin, Chris Mungall, David Arenillas, Anthony Mathelier, Yulia A Medvedeva, et al. Update of the fantom web resource: high resolution transcriptome of diverse cell types in mammals. *Nucleic acids research*, page gkw995, 2016.
- Stephan R Starick, Jonas Ibn-Salem, Marcel Jurk, Céline Hernandez, Michael I Love, Ho-Ryun Chung, Martin Vingron, Morgane Thomas-Chollier, and Sebastiaan H Meijsing. Chip-exo signal associated with dna-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome research*, 25(6):825–835, 2015.
- Naomi Yamada, William KM Lai, Nina Farrell, B Franklin Pugh, and Shaun Mahony. Characterizing protein–dna binding event subtypes in chip-exo data. *Bioinformatics*, 35(6):903–913, 2019.
- Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.
- Shaun Mahony, Matthew D Edwards, Esteban O Mazzoni, Richard I Sherwood, Akshay Kaku-manu, Carolyn A Morrison, Hynek Wichterle, and David K Gifford. An integrated model of multiple-condition chip-seq data reveals predeterminants of cdx2 binding. *PLoS computational biology*, 10(3):e1003501, 2014.
- Ralf Eggeling, André Gohr, Jens Keilwagen, Michaela Mohr, Stefan Posch, Andrew D Smith, and Ivo Grosse. On the value of intra-motif dependencies of human insulator protein ctfc. *PLoS One*, 9(1):e85629, 2014.
- Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Froepf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021.
- Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, 10(7):e1003711, 2014.
- Irina M Bochkis, Jonathan Schug, Diana Z Ye, Svitlana Kurinna, Sabrina A Stratton, Michelle C Barton, and Klaus H Kaestner. Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators foxa1 and foxa2. *PLoS genetics*, 8(6):e1002770, 2012.
- Hiroataka Nakahashi, Kyong-Rim Kieffer Kwon, Wolfgang Resch, Laura Vian, Marei Dose, Diana Stavreva, Ofir Hakim, Nathanael Pruett, Steevenson Nelson, Arito Yamane, et al. A genome-wide map of ctfc multivalency redefines the ctfc code. *Cell reports*, 3(5):1678–1689, 2013.
- Raluca Gordân, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L Bulyk. Genomic regions flanking e-box binding sites influence dna binding specificity of bhlh transcription factors through dna shape. *Cell reports*, 3(4):1093–1104, 2013.

- Tsu-Pei Chiu, Lin Yang, Tianyin Zhou, Bradley J Main, Stephen CJ Parker, Sergey V Nuzhdin, Thomas D Tullius, and Remo Rohs. Gbshape: a genome browser database for dna shape annotations. *Nucleic acids research*, 43(D1):D103–D109, 2015.
- Andrew Brittain, Elizabeth Stroebele, and Albert Erives. Microsatellite repeat instability fuels evolution of embryonic enhancers in hawaiian drosophila. *PloS one*, 9(6):e101177, 2014.
- J Omar Yáñez-Cuna, Cosmas D Arnold, Gerald Stampfel, Łukasz M Boryń, Daniel Gerlach, Martina Rath, and Alexander Stark. Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome research*, 24(7):1147–1156, 2014.
- David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.
- Maartje J Vogel, Daniel Peric-Hupkes, and Bas Van Steensel. Detection of in vivo protein–dna interactions using damid in mammalian cells. *Nature protocols*, 2(6):1467–1478, 2007.
- Alan P Boyle, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322, 2008.
- Paul G Giresi, Jonghwan Kim, Ryan M McDaniell, Vishwanath R Iyer, and Jason D Lieb. Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6):877–885, 2007.
- Melissa J Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, et al. An oestrogen-receptor- α -bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.
- Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.
- Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, 2003.
- Ankit Agrawal, Snehal V Sambare, Leelavati Narlikar, and Rahul Siddharthan. Thicweed: fast, sensitive detection of sequence features by clustering big datasets. *Nucleic acids research*, 46(5):e29–e29, 2018.

- Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome biology*, 8(2):1–9, 2007.
- Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.
- Ting Ni, David L Corcoran, Elizabeth A Rach, Shen Song, Eric P Spana, Yuan Gao, Uwe Ohler, and Jun Zhu. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature methods*, 7(7):521–527, 2010.
- Nicholas E Matthews and Rob White. Chromatin architecture in the fly: living without ctcf/cohesin loop extrusion? alternating chromatin states provide a basis for domain architecture in drosophila. *Bioessays*, 41(9):1900048, 2019.
- Sheryl T Smith, Priyankara Wickramasinghe, Andrew Olson, Dmitri Loukinov, Lan Lin, Joy Deng, Yanping Xiong, John Rux, Ravi Sachidanandam, Hao Sun, et al. Genome wide chip-chip analyses reveal important roles for ctcf in drosophila genome organization. *Developmental biology*, 328(2):518–528, 2009.
- Sergi Cuartero, Ujué Fresán, Oscar Reina, Evarist Planet, and M Lluisa Espinàs. Ibf1 and i bf2 are novel cp 190-interacting proteins required for insulator function. *The EMBO journal*, 33(6):637–647, 2014.
- Nastaran Heidari, Douglas H Phanstiel, Chao He, Fabian Grubert, Fereshteh Jahanbani, Maya Kasowski, Michael Q Zhang, and Michael P Snyder. Genome-wide map of regulatory interactions in the human genome. *Genome research*, 24(12):1905–1917, 2014.
- Bingyu Ye, Ganggang Yang, Yuanmeng Li, Chunyan Zhang, Qiwen Wang, and Guoying Yu. Znf143 in chromatin looping and gene regulation. *Frontiers in Genetics*, 11:338, 2020.
- Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics*, 43(3):264–268, 2011.
- Simon C Biddie, Sam John, Pete J Sabo, Robert E Thurman, Thomas A Johnson, R Louis Schiltz, Tina B Miranda, Myong-Hee Sung, Saskia Trump, Stafford L Lightman, et al. Transcription factor ap1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Molecular cell*, 43(1):145–155, 2011.
- Lars Grøntved, Sam John, Songjoon Baek, Ying Liu, John R Buckley, Charles Vinson, Greti Aguilera, and Gordon L Hager. C/ebp maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *The EMBO journal*, 32(11):1568–1583, 2013.

- Ian C McDowell, Alejandro Barrera, Anthony M D'Ippolito, Christopher M Vockley, Linda K Hong, Sarah M Leichter, Luke C Bartelt, William H Majoros, Lingyun Song, Alexias Safi, et al. Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome research*, 28(9):1272–1284, 2018.
- Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014.
- Charles G Danko, Stephanie L Hyland, Leighton J Core, Andre L Martins, Colin T Waters, Hyung Won Lee, Vivian G Cheung, W Lee Kraus, John T Lis, and Adam Siepel. Identification of active transcriptional regulatory elements from gro-seq data. *Nature methods*, 12(5):433–438, 2015.
- Joseph G Azofeifa, Mary A Allen, Josephina R Hendrix, Jonathan D Rubin, and Robin D Dowell. Enhancer rna profiling predicts transcription factor activity. *Genome research*, 28(3):334–344, 2018.
- Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. Analysis of nascent rna identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics*, 46(12):1311–1320, 2014.
- Maximilian Haeussler, Ann S Zweig, Cath Tyner, Matthew L Speir, Kate R Rosenbloom, Brian J Raney, Christopher M Lee, Brian T Lee, Angie S Hinrichs, Jairo Navarro Gonzalez, et al. The ucsc genome browser database: 2019 update. *Nucleic acids research*, 47(D1):D853–D858, 2019.
- Fábio Madeira, Young Mi Park, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian RN Tivey, Simon C Potter, Robert D Finn, et al. The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, 47(W1):W636–W641, 2019.
- Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.
- Jemie Cheng, Roy Blum, Christopher Bowman, Deqing Hu, Ali Shilatifard, Steven Shen, and Brian D Dynlacht. A role for h3k4 monomethylation in gene repression and partitioning of chromatin readers. *Molecular cell*, 53(6):979–992, 2014.
- Yuwen Liu, Shan Yu, Vineet K Dhiman, Tonya Brunetti, Heather Eckart, and Kevin P White. Functional assessment of human enhancer activities using whole-genome starr-sequencing. *Genome biology*, 18(1):1–13, 2017.

- Donghoon Lee, Manman Shi, Jennifer Moran, Martha Wall, Jing Zhang, Jason Liu, Dominic Fitzgerald, Yasuhiro Kyono, Lijia Ma, Kevin P White, et al. Starrpeak: uniform processing and accurate identification of starr-seq active regions. *Genome biology*, 21(1):1–24, 2020.
- Jill E Moore, Michael J Purcaro, Henry E Pratt, Charles B Epstein, Noam Shores, Jessica Adrian, Trupti Kawli, Carrie A Davis, Alexander Dobin, Rajinder Kaul, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.
- Swneke D Bailey, Xiaoyang Zhang, Kinjal Desai, Malika Aid, Olivia Corradin, Batool Akhtar-Zaidi, Peter C Scacheri, Benjamin Haibe-Kains, Mathieu Lupien, et al. Znf143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature communications*, 6(1):1–10, 2015.
- Tae Hoon Kim, Ziedulla K Abdullaev, Andrew D Smith, Keith A Ching, Dmitri I Loukinov, Roland D Green, Michael Q Zhang, Victor V Lobanenko, and Bing Ren. Analysis of the vertebrate insulator protein ctf-binding sites in the human genome. *Cell*, 128(6):1231–1245, 2007.
- Oriana Romano and Annarita Miccio. Gata factor transcriptional activity: Insights from genome-wide binding profiles. *IUBMB life*, 72(1):10–26, 2020.
- Darren A Cusanovich, James P Reddington, David A Garfield, Riza M Daza, Delasa Aghamirzaie, Raquel Marco-Ferreres, Hannah A Pliner, Lena Christiansen, Xiaojie Qiu, Frank J Steemers, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, 555(7697):538–542, 2018.
- Lakshmi Sreekumar, Kiran Kumari, Krishnendu Guin, Asif Bakshi, Neha Varshney, Bhagya C Thimmappa, Leelavati Narlikar, Ranjith Padinhateeri, Rahul Siddharthan, and Kaustuv Sanyal. Orc4 spatiotemporally stabilizes centromeric chromatin. *Genome research*, 31(4):607–621, 2021.
- Zhangli Su, Briana Wilson, Pankaj Kumar, and Anindya Dutta. Non-canonical roles of trnas: trna fragments and beyond. *Annual review of genetics*, 54:47, 2020.
- Hamutal Arbel, Sumanta Basu, William W Fisher, Ann S Hammonds, Kenneth H Wan, Soo Park, Richard Weiszmam, Benjamin W Booth, Soile V Keranen, Clara Henriquez, et al. Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. *Proceedings of the National Academy of Sciences*, 116(3):900–908, 2019.
- Danielle C Hamm and Melissa M Harrison. Regulatory principles governing the maternal-to-zygotic transition: insights from drosophila melanogaster. *Royal Society Open Biology*, 8(12):180183, 2018.

- Chuanyu Liu, Mingyue Wang, Xiaoyu Wei, Liang Wu, Jiangshan Xu, Xi Dai, Jun Xia, Mengnan Cheng, Yue Yuan, Pengfan Zhang, et al. An atac-seq atlas of chromatin accessibility in mouse tissues. *Scientific data*, 6(1):1–10, 2019.
- Rachana N Pradhan, Johannes J Bues, Vincent Gardeux, Petra C Schwalie, Daniel Alpern, Wanze Chen, Julie Russeil, Sunil K Raghav, and Bart Deplancke. Dissecting the brown adipogenic regulatory network using integrative genomics. *Scientific reports*, 7(1):1–16, 2017.
- Ji-Eun Lee, Hannah Schmidt, Binbin Lai, and Kai Ge. Transcriptional and epigenomic regulation of adipogenesis. *Molecular and Cellular Biology*, 39(11):e00601–18, 2019.
- Richard M Baldarelli, Constance M Smith, Jacqueline H Finger, Terry F Hayamizu, Ingeborg J McCright, Jingxia Xu, David R Shaw, Jonathan S Beal, Olin Blodgett, Jeffrey Campbell, et al. The mouse gene expression database (gxd): 2021 update. *Nucleic Acids Research*, 49(D1):D924–D931, 2021.
- Franziska Reiter, Sebastian Wienerroither, and Alexander Stark. Combinatorial function of transcription factors and cofactors. *Current opinion in genetics & development*, 43:73–81, 2017.
- Qing Zhou and Wing H Wong. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the National Academy of Sciences*, 101(33):12114–12119, 2004.
- Dan Xie, Alan P Boyle, Linfeng Wu, Jie Zhai, Trupti Kawli, and Michael Snyder. Dynamic trans-acting factor colocalization in human cells. *Cell*, 155(3):713–724, 2013.
- Eugenia G Giannopoulou and Olivier Elemento. Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome research*, 23(8):1295–1306, 2013.
- Yuchun Guo and David K Gifford. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC genomics*, 18(1):1–16, 2017.
- Mary Lauren Benton, Sai Charan Talipineni, Dennis Kostka, and John A Capra. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *Bmc Genomics*, 20(1):1–22, 2019.
- Timothy L Bailey and Philip Machanick. Inferring direct dna binding from chip-seq. *Nucleic acids research*, 40(17):e128–e128, 2012.
- Xiaoyan Ma, Daphne Ezer, Boris Adryan, and Tim J Stevens. Canonical and single-cell hi-c reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome biology*, 19(1):1–23, 2018.
- Seungsoo Kim and Jay Shendure. Mechanisms of interplay between transcription factors and the 3d genome. *Molecular cell*, 76(2):306–319, 2019.

- Justin Malin, Daphne Ezer, Xiaoyan Ma, Steve Mount, Hiren Karathia, Seung Gu Park, Boris Adryan, and Sridhar Hannenhalli. Crowdsourcing: spatial clustering of low-affinity binding sites amplifies in vivo transcription factor occupancy. *bioRxiv*, page 024398, 2015.
- Donna Karolchik, Galt P Barber, Jonathan Casper, Hiram Clawson, Melissa S Cline, Mark Diekhans, Timothy R Dreszer, Pauline A Fujita, Luvina Guruvadoo, Maximilian Haeussler, et al. The ucsc genome browser database: 2014 update. *Nucleic acids research*, 42(D1): D764–D770, 2014.
- Jan Christian Bryne, Eivind Valen, Man-Hung Eric Tang, Troels Marstrand, Ole Winther, Isabelle da Piedade, Anders Krogh, Boris Lenhard, and Albin Sandelin. Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic acids research*, 36(suppl_1):D102–D106, 2007.
- Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome research*, 14(6):1188–1190, 2004.
- Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740, 2012.
- R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476–486, 2010.
- Yongbing Zhao. Tfsyntax: a database of transcription factors binding syntax in mammalian genomes. *Nucleic Acids Research*, 51(D1):D306–D314, 2023.
- Ralf Eggeling, Teemu Roos, Petri Myllymäki, and Ivo Grosse. Inferring intra-motif dependencies of dna binding sites from chip-seq data. *BMC bioinformatics*, 16:1–15, 2015.
- Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.
- Babak Alipanahi, Andrew DeLong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.
- Meghana Kshirsagar, Han Yuan, Juan Lavista Ferres, and Christina Leslie. Bindvae: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin. *Genome Biology*, 23(1):174, 2022.