Development of bioinformatics methods for the analysis of large collections of transcription factor binding motifs: positional motif enrichment and motif clustering

Jaime Abraham CASTRO-MONDRAGON

2017-06-28

Contents

Chapter 1

Acknowledgments

I would like to remark that this work is a collection of pieces (ideas) coming from different people from different places of the world. During the last four years that I have lived in France, I have visited many places and met many people than ever, many of them have changed me as a person. Although the next lines could not be understood at all for many of you, let me tell you that your contribution is reflected on (the strength to finish) this work.

I want to express my appreciation to Professor Jacques van Helden. Thanks for let me work freely and develop my own ideas. Usually the intellectual freedom is achieved after several years of work, but in my case, and thanks to you, this was possible at an early age. Thanks for the confidence to manage several projects. Thanks as well for the biscuits, coffee and unhealthy food. I feel so lucky to have worked with you during this time.

A very special thanks to Morgane Thomas-Chollier, for the supervision and guidance during my PhD, writing the manuscript of *matrix-clustering*, discussion of technical issues, to choose the right colors and more. Thanks to Denis Thieffry for hosting me at your lab, for the contributions, corrections and ideas. Thanks to Samuel Collombet for the ideas and for test my programs. To Alejandra Medina-Rivera for the guidance, confidence and the ideas in our projects and Bruno Contreras-Moreria for visit us in Marseille and bring ideas for the projects.

Thanks to all members of our team: Lucie, Claire, Moustafa and Najla; to Lan for take care of my plants during my absence, the *Petit Dej team* (Alberto (Albertronco), Alejandro (el buen hombre), Ana, Diogo (El Meco), Elena, Guillaume, Michael (el gran danes), Lamia, Lolita), Aldo (el pariente), the *Cafe des Langues* members (Tom Grainer *et al*), Ariel Galindo (y toda su perrada) for prepare *mole*, David Santiago for the nice and funny conversations and print many sheets of paper when I asked; Aitor Gonzales for bring useful information and discussions; and Salvatore Spicuglia for let me collaborate with your team. To all TAGC members for their contributions and beer sessions. Thanks to Julio Collado-Vides and his team (RegulonDB) for their hospitality and invitation to the parties; and Jose Alquicira for your knowledge in R.

Thanks to *los morros* team (Amhed, Blanca, Violena, Yuvia) for hosting me in Paris, Jhonas Ibn-Salem for do the same in Mainz. To my friends that have followed closely this thesis: specially to Carmen Sandoval (Csandova), Gustavo Ruiz (monroik), Lucia Pannier, Carlos Salgado, Fidel Velez, Gabriel Klimek, Adriana Rubio, Daniel Parra, Lito y Ramiro (el pollo) for the nice conversations, laughs and beers. Special thanks to the Peterca Family for welcomed me in Austria (and for the nice time we have spent together), to my parents Jaime and Lucrecia for their support and for trust in my ideas, to my brother Emanuel for the funny moments and conversations, to my sister Abril for let me be Thiago's godparent and to the (big) family Mondragon (we are more than 100, sorry for not list all of you).

I want to thank Ariana (la doctora) Peterca, for your love and support, thanks for stay with me during this time and for your patience, specially during the last months. Thanks for be part of my life outside the scientific world and for show me a different way to see the world.

Chapter 2

Resume

Les facteurs transcriptionnels (TF) sont des protéines qui contrôlent l'expression des gènes en activant ou en réprimant la transcription. Leurs motifs de liaison (TFBM, également appelés «motifs») sont généralement représentés sous forme de matrices de scores spécifiques de positions (PSSM). L'analyse de motifs est utilisée en routine afin de découvrir des facteurs «candidats» pour la régulation d'un jeu de séquences d'intérêt (par exemple les promoteurs d'un groupe de gènes co-exprimés). L'avénement des méthodes à haut débit a permis de détecter des centaines de motifs, qui sont disponibles dans des bases de données.

Durant ma thèse, j'ai développé deux nouvelles méthodes et implémenté des outils logiciels pour le traitement de collections massives de motifs, afin d'extraire une information interprétable à partir de données à haut débit: matrix-clustering regroupe les motifs par similarité; position-scan détecte les motifs présentant des préférences de position relativement à une coordonnée de référence (par exemple les sommets de pics de ChIP-seq).

Actuellement, les bases de données de motifs sont couramment utilisées pour l'annotation. Cependant elles ont tendance à croître rapidement, et leur contenu devient redondant. Une autre source de redondance est la découverte de motifs sur base d'approches multiples, qui s'avère utile pour évaluer la robustesse des motifs, mais présente un coût en termes de redondance. La découverte et l'annotation de motifs sont deux tâches communes pour les études à échelle génomique. Cependant, à mesure qu'on découvre et annote plus de motifs dans les bases de données, cette redondance rend les analyses ultérieures plus complexes et coûteuses en ressources. Afin de faciliter l'analsye de motifs avec des collections de motifs étendues, j'ai développé matrix-clustering, un outil qui réduit la redondance des motifs et permet de visualiser les groupes de motifs alignés pour montrer leur similarité. Le résultat du clustering est représenté de différentes façons (alignements de logos, arbres, carte de couleurs), et des collections multiples peuvent être analysées simultanément. En profitant de cette fonctionnalité, j'ai effectué un clustering de collections taxonomiques de motifs afin de créer des collections non-redondantes pour les insectes, les plantes et les vertébrés.

L'utilisation de collections non-redondantes de motifs présente un avantage pour certaines méthodes, par exemple la détection de motifs enrichis. Ces méthodes se basent sur une collection de motifs connus, et chaque analyse requiert donc d'analyser de grandes collections de motifs, ce qui complète la découverte de motifs. Pour certains jeux de données, tels que les pics de ChIP-seq, on s'attend à observer certains motifs au centre des pics, l'enrichissement est donc relatif à une position de référence. Puisque les méthodes d'enrichissement positionnel existantes sont spécialisée pour le ChIP-seq, elles prennent uniquement en compte les sites à haute affinté. Cependant, les sites à affinité plus modérée peuvent s'avérer pertinents pour moduler la transcription. J'ai donc développé une méthode qui détecte les motifs soit enrichis soit appauvris localement, sans se limiter aux sites de haute affinité. Cette méthode a été utilisée pour édtecter des motifs enrichis dans un jeu de promoteurs humains capables d'activer l'expression de gènes à distance («Epromoters»).

Les méthodes que j'ai développées ont été évaluées sur base de cas d'études, et utilisées pour extraire de l'information interprétable à partir de différents jeux de données de Drosophila melanogaster et Homo sapiens.

Les résultats démontrent la pertinence de ces méthodes pour l'analyse de données à haut débit, et l'intérêt de les intégrer dans des pipelines d'analyse de motifs.

Chapter 3

Abstract

Transcription Factors (TFs) are DNA-binding proteins that control gene expression by activating or repressing transcription. TF binding motifs (TFBMs, more simply called "motifs") are usually represented as Position Specific Scoring Matrices (PSSMs), which can be visualized as sequence logos.

Motif analysis is routinely used to discover candidate TFs regulating a set of sequences of interest (e.g., a set of promoters of co-expressed genes) and the results are key to infer regulatory networks between TFs and genes. The advent of high-throughput methods has allowed the detection of thousands of motifs which are usually stored in databases.

In this work I developed two novel methods and implemented software tools to handle large collection of motifs in order to extract interpretable information from high-throughput data: (i) matrix-clustering regroups motifs by similarity and offers a dynamic interface to visualize them; (2) position-scan detects TFBMs with positional preferences relative to a given reference location (e.g. ChIP-seq peaks, transcription start sites, …).

Currently, motif databases are highly used for motif annotation, however they grow up rapidly and their content becomes redundant. Another source of redundancy is the discovery of motifs using distinct approaches, which is useful to obtain robust results, but has a cost in terms of motif redundancy. Both motif discovery and motif annotations are common tasks in genome-wide studies. However as more motifs are discovered and annotated using the databases, the redundancy makes the analysis more complex and time consuming, and obfuscate the interpretation of the results.

In order to ease the motif analysis with large collection of motifs I developed matrix-clustering, a tool to reduce motif redundancy and visualize groups of similar motifs, aligned to highlight their similarities. The clustering is represented in different ways (logo alignments, trees, heatmap) and many input collections can be clustered in a single run. Taking advantage of this latter capability, I clustered taxon-wise motif collections creating thus non-redundant motif collections for insects, plants and vertebrates.

The use of non-redundant motif collections can be an advantage for some methods, for example those detecting the enrichment of TFBSs relative to a reference position, (e.g., to Transcription Start Sites of to the center of ChIP-seq peaks). The motif enrichment methods use as input a collection of known motifs and therefore a large motif set can be analyzed in a single run, complementing the results of motif discovery. Since the current positional motif enrichment tools are specialized in ChIP-seq, they only consider the TFBSs with the higher affinity enriched at the center of the peaks. However weaker affinity binding sites can be relevant to modulate transcription. For these reasons, I developed a method that detect positionally biased motifs either enriched and depleted and the analysis is not limited to the strongest sites. This method was used to detect TFBMs enriched in a particular set of human promoters involved in long-rage interactions with other promoters (Epromoters).

The methods I developed have been evaluated based on control cases, and applied to extract meaningful information from different data sets from Drosophila melanogaster and Homo sapiens. The results show that

these methods enable to analyse motifs in high-throughput data sets, and can be integrated in motif analysis workflows.

Chapter 4

Abreviations

- ChIP-seq: Chromatin Immunoprecipitation followed by hight-troughput sequencing
- CRM: Cis-Regulatory Module
- DNA: Deoxyribonucleic Acid
- DBD: DNA-Binding Domain
- FBP: Familial Binding Profile
- GTF: General Transcription Factor
- IC: Information Content
- Nucleosome-Depleted Region
- NGS: Next-Generation Sequencing
- PSSM: Position Specific Scoring Matrix
- PWM: Position Weight Matrix
- RNA: Ribonucleic Acid
- RNAP: RNA-polymerase
- mRNA: Messenger Ribonucleic Acid
- miRNA: Micro Messenger Ribonucleic Acid
- PBM: Protein-Binding Microarray
- RSAT: Regulatory Sequences Analysis Tools
- SELEX: Systematic Evolution of Ligands by Exponential Enrichment
- TF: Transcription Factor
- TFBM: Transcription Factor Binding Motif
- TFBS: Transcription Factor Binding Site
- TSS: Transcription Start Site

Chapter 5

Transcriptional Regulation

5.1 The importance of transcriptional regulation

One of the most intriguing and studied questions in biology is that one related to transference of information starting from DNA on multicellular organism: how is possible that a single cell with a genome has the potential to develop different tissues, organs and systems?

Now we know how this information flows through different biochemical processes, starting from DNA to DNA (replication), from DNA to RNA (transcription) and from RNA to proteins (translation), see Figure 5.1. The first on reveal aspects of gene regulation were Jacob and Monod, they demonstrated that the synthesis of proteins (i.e., the final product of most genes after being expressed is a protein), starting from DNA transcription, was mediated by a special class of proteins that they called *repressors* and these could regulate the gene activity bound in specific short sequences of DNA (they called them *operators*) located near the genes (Jacob and Monod, 1961). This discovery opened a research field on molecular biology (the transcriptional regulation) and shed light in the understanding of gene regulation showing (at that time) that gene expression was mediated by a group of proteins (*repressors*) and short DNA sequences (*operators*).

Many years after, it was discovered that there was another kind of regulatory proteins, the *activators*, that in contrast to *repressors*, they can regulate positively the gene expression. Nowadays both are known as Transcription Factors (TFs). Today we have a better understanding of these processes and we know that the DNA transcription is the first step of gene expression, hereafter I will refer gene expression as the process where a gene is transcribed to produce messenger RNA (mRNA), that generally leads to synthesis of proteins. Initially, as most experiments were done in bacteria, the researches discovered that the DNA regions where the TF bind were located upstream near the gene, a region known as *promoter*, this concept is valid for bacteria but we will see that metazoan transcriptional regulation is driven by more regulatory elements.

However, although the information flux looks very simple, we must take into account that metazoan genomes develop complex cellular states that give rise to different tissues with specialized functions, this make us set

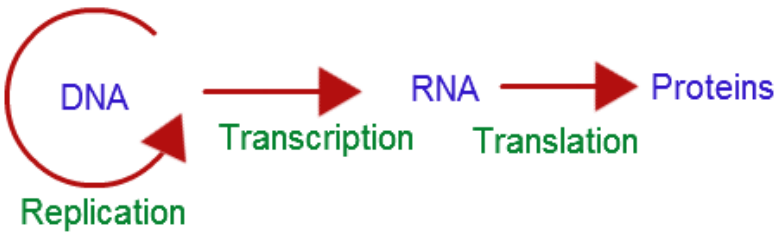


Figure 5.1: Flux of biological information

another question: how is coordinated the cell differentiation, for example to produce neuron or a kidney cell? A simple answer could be: the gene repertoire (that is partially true); but once again this make us set more questions: how is regulated the gene expression?; do all genes are active at the same time? Nowadays we know that the differential gene expression through time (development, ageing) is what give rise to different tissues, cell lines, organs and homeostasis; and the dysregulation of gene expression could be associated with diseases (Mathelier et al., 2015). Some examples are the *Drosophila melanogaster* body segmentation (Tautz and Pfeifle, 1989), T-cell differentiation (Zhu et al., 2010), cell reprogramming (Takahashi and Yamanaka, 2006) or cancer (Ell and Kang, 2013).

The answers of these questions are far from trivial, but a very naive answer could be the next one: a coordinated combination of several factors (including metabolites, proteins, RNAs or the DNA itself) are responsible of the gene expression (from DNA to RNA to proteins). Each step is highly regulated by several components and even at different cell compartments. One of this steps of gene regulation, called epi-genomics, involve chemical changes on DNA or DNA-associated proteins (e.g., histones) that does not alter the DNA sequence itself, but may alter the gene expression or DNA accessibility. Some examples are DNA methylation, modification of histone-tails, and 3D structure of the chromatin. The term epi-genomics used in this work is not the same term by Waddington (epi-genetics), which refers to the connection between the genotype and phenotype related to cell differentiation (Waddington, 1942).

For the scope of this thesis I will focus on the transcriptional regulation (Figure 5.2) specially in the TFs and the cis-regulatory elements, but the readers must not forget that other layers of regulation could also modulate the gene expression (e.g., mRNA translation and post-translational regulation).

To summarize, cells sense the internal and external stimuli which consequence are changes on gene expression to adapt to these changes. Molecularly gene gene expression is driven by different elements, two of them, the TF and cis-regulatory sequences drive the transcriptional regulation of the genes. The changes expression goes from activation or inactivation of a gene, to a complex processes as cell differentiation or organogenesis. This awesome phenomena is of my interest and this is my motivation to study and contribute to this field.

5.2 Generalities of DNA transcription

Transcription is defined as the process in which the RNA is synthesized from a DNA template, the produced mRNA will be further processed to produce a protein. In any living organism (from bacteria to metazoa), the transcription is performed by a protein complex called RNA-polymerase (RNAP) which has affinity for short sequences located in the promoter (e.g. TATA-box, BRE-elements), but usually this affinity is not sufficient to start transcription and the RNAP requires help from other elements.

These short sequences at the promoters are capable of recruit the RNAP. Those promoters capable to start the transcription of its downstream gene by themselves (i.e., recruiting the RNAP without help of other elements) are considered *strong* promoters. By contrast, the co called *weak* promoters require the help from TFs (or other proteins) to stabilize the RNAP or recruit some of its sub-units (Qin et al., 2010) (Figure 5.3).

Although the first studied cis-regulatory elements were the promoters, another class of regulatory elements was discovered, that are the so-called enhancers, which are regions that can activate distally the gene expression (Banerji et al., 1981), by contrast to the promoters that do it locally. This discovery revealed other cis-regulatory regions acting distally in the silencing of gene expression (silencers) or other sequences capable of avoid the enhancer activity (insulators). However, we must not forget that the interaction of these elements depends on the accessibility for the regulatory proteins to these cis-regulatory sequences, see Figure 5.3 and table 3.1 for a summary of the regulatory elements participating in the transcriptional initiation (Lenhard et al., 2012).

A detailed description of TFs and cis-regulatory regions will be showed in the next sections with a brief review of the other transcriptional regulatory elements.

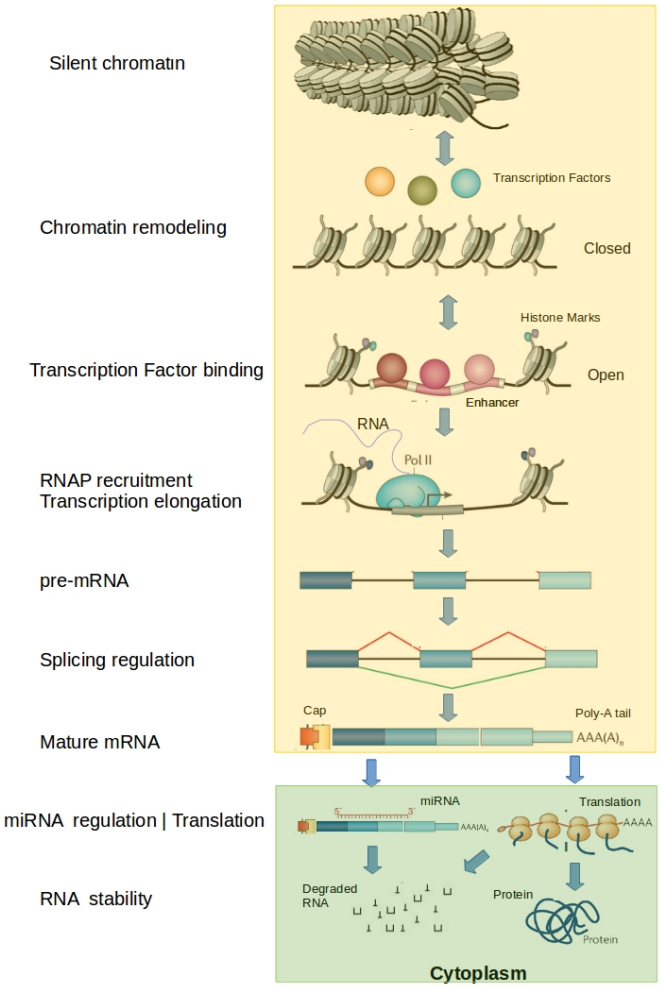


Figure 5.2: Different layers of transcriptional regulation. The chromatin conformation and histone marks determine the regions where regulatory proteins can bind DNA, an open region allow the acces of TFs which can help to recroute RNAP, the transcript elongation is mediated by the RNAP resulting in a mRNA that is further processed by the splicing machinery, the mRNA stability is determined by the degradation rate of the mRNA or by miRNA. Figure adapted from Komili(2008), Cole (2008), Bentley (2014), and Shlyueva (2014)

Table 5.1: classification of transcriptional regulatory elements.

Element	Description	Class
RNA-polymerase (RNAP)	Complex of proteins that transcribe the DNA to RNA	Protein
Transcription Factor (TF)	DNA-binding proteins regulators of gene expression	Protein
Transcription Start Site (TSS)	First nucleotide transcribed by the RNAP	DNA
Cis-Regulatory Module (CRM)	A DNA region whit a high concentration of distinct TFs	DNA
Transcription Factor Binding Site (TFBS)	A short DNA sequences bound by a TF	DNA
Enhancer	Distal regulatory (activation) region	DNA
Silencer	Distal regulatory (repression) region	DNA
Insulator	boundary between hetero- and eu-chromatin.	DNA

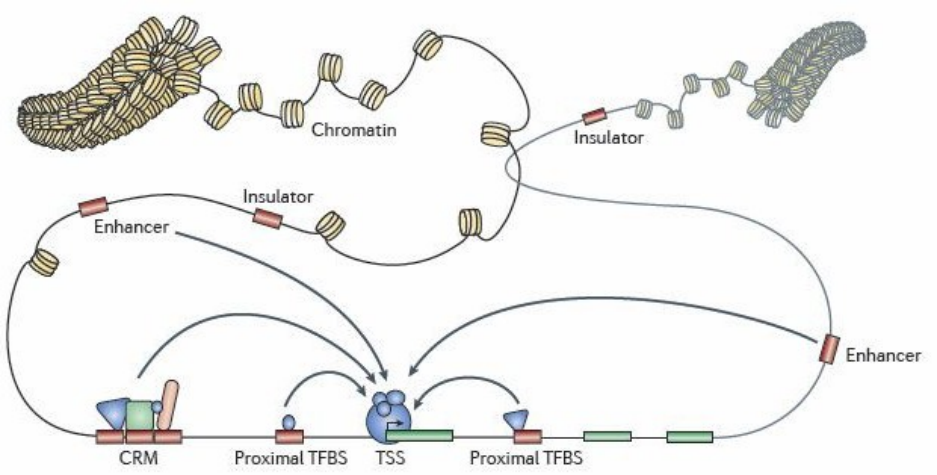


Figure 5.3: Transcriptional regulatory elements in metazoa. CRM: Cis-Regulatory module; TSS : Transcription Start Site. Figure adapted from Lenhard (2012).

5.3 Chromatin structure and histone modifications

All the genomes should be efficiently packed into a small volume to fit into the nucleus of a cell. In eukaryotes, the large scale 3D organization of the genome consist in the so-called chromosome territories, the specific region of the nucleus occupied by a chromosome. At intermediate scale, the DNA is folded and can form Topological-associated domains (0.5-1Mb) and within them, the DNA can from smaller loops (hundreds of kilobases) (Figure 5.4) (Bonev and Cavalli, 2016; Rao et al., 2014; Stevens et al., 2017). At small scale, the DNA is wrapped (147bp) in structures called nucleosomes which consist on a octamer of proteins called histones (two units of each H3, H2A, H2B and H4). The histones contain domains which particular aminoacid residues can be reversely and covalently modified, by adding or removing compounds (e.g., methyl or phosphate groups) by specialized enzymes as methylases/demethylases or kinases/phosphatases(Tsankova et al., 2007). Many residues can be modified in the same tail, these modifications are denoted as histone marks and have a particular notation. For example if the lysine at position 27 of the histone 3 is acetylated, this is represented as H3K27ac.

The histone marks may be recognized by chromatin remodeler proteins, which in turn modify the local structure of the DNA, for example the nucleosome compaction, as consequence the DNA can be open and accessible to regulatory proteins as TFs or RNAP, or the DNA can be closed, silencing the local gene expression (Plass et al., 2013). The histone marks are commonly associated to transcriptional states (active or inactive genes) or cis-regulatory regions (Figure 5.5) (Lawrence et al., 2016). For example, H3K27ac is associated with active promoters and distal regulatory elements, H3K4me3 and H3K36me3 are both associated with transcribed chromatin, H3K36me3 is found along gene body of transcribed genes. By contrast to these active marks, H3K9me3, H3K27me3 and H4K20me3 are generally related to gene repression (Barski et al., 2007). It is important to note that these marks are associated to these regions or activities, but it should not be considered as the cause of these phenomena.

5.4 DNA methylation

DNA methylation is a reversible process where a methyl group (CH3) is added to the DNA (specifically to the 5th carbon in a cytosine nucleotide ring; 5mC). This covalent modification alters the nucleotide but not the DNA sequence. This modification, in mammals, occurs mainly at the CpG sites (a cytosine followed by a guanine). The methylation is driven by proteins belonging to the DNA methyltransferase (DMT) family but some histone marks can block *de novo* methylation (Figure 5.6) (Jones, 2012).

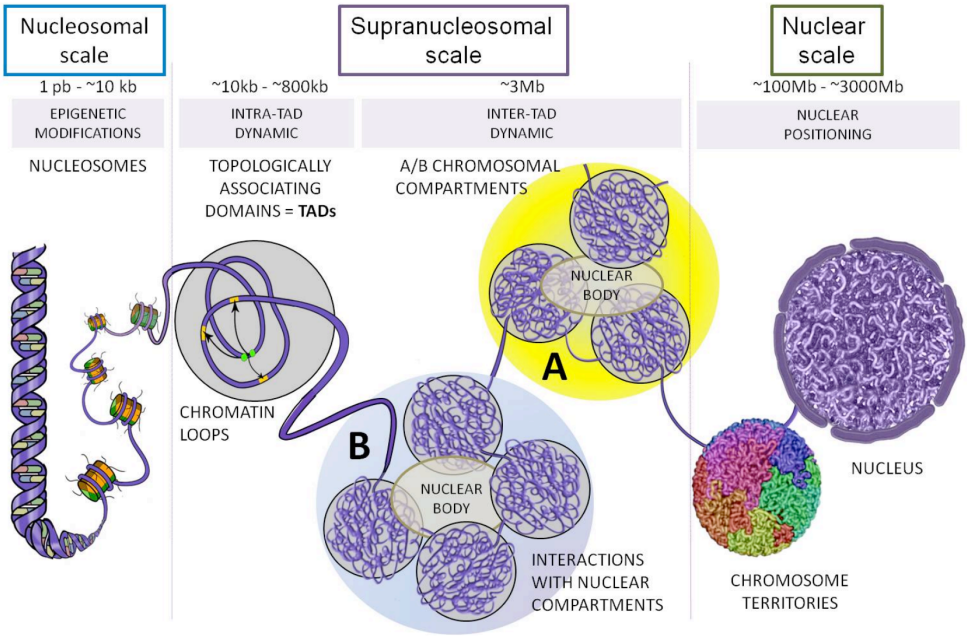


Figure 5.4: 3D organization of the eukaryote genome. Figure from Ea (2015).

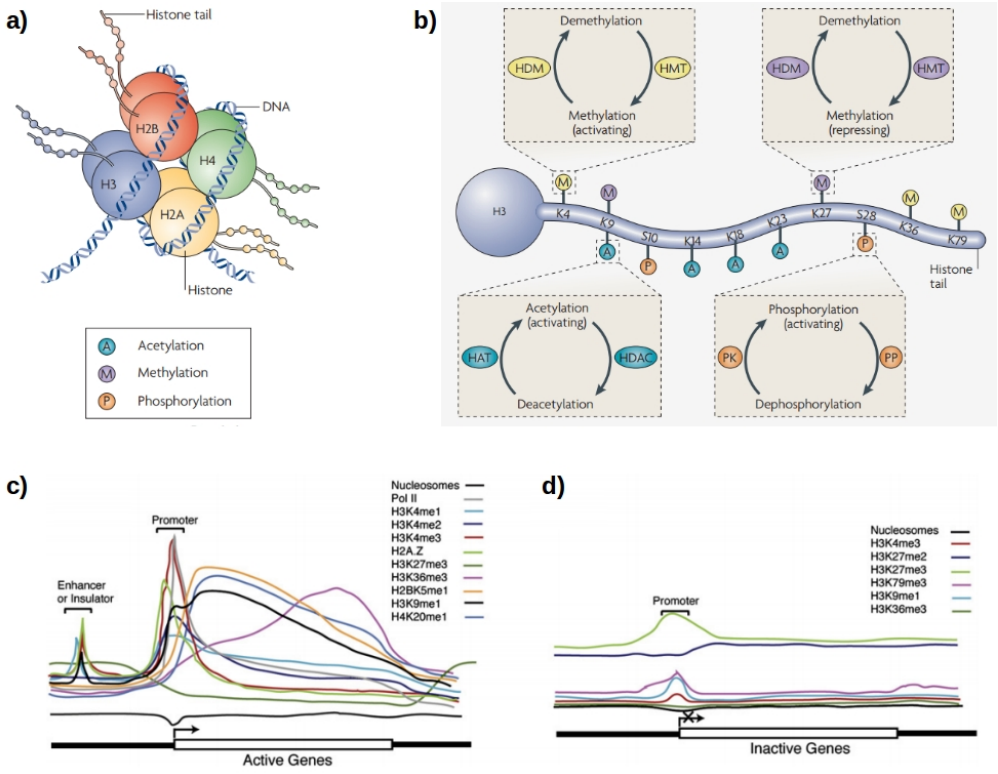


Figure 5.5: Nucleosome and histone modifications. a) A nucleosome and its components. b) Histone tail and examples of modifications. Figure adapted from Tsankova (2007) and Barski (2007).

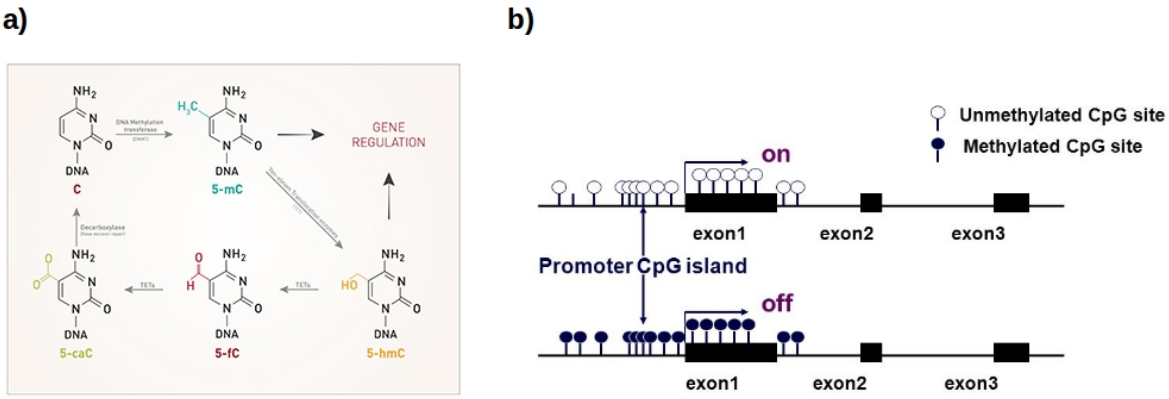


Figure 5.6: DNA methylation. a) DNA methylation pathway. b) Example of gene activity when the CpG islands are either methylated or unmethylated.

Genomic regions with a high concentration (>60%) of CpG dinucleotides are known as CpG islands. Paradoxically, most of the CpG islands located at promoters (80-90%) are not methylated, by contrast those CpG islands located at transposons are constitutively methylated, therefore contributing to repress the transposon activity.

Usually the DNA methylation is associated to gene silencing by the following reasons (Figure 5.7a):

- Methylated cytosines may alter the binding specificities for many TFs (Hu et al., 2013; Lercher et al., 2014).
- DNA methylation directly increases affinity of certain sequences for histone octamer, therefore increasing nucleosome occupancy and compaction (Collings et al., 2013).
- 5mC is a marker for methyl-cytosine binding domain proteins, which may recruit chromatin remodelers that induce chromatin compaction (Lande-Diner et al., 2007).

The 5mC can be converted into 5'-hydroxymethyl-cytosine (5hmC), as part of the methylation/demethylation pathway of the cytosine, this process achieved by the enzymes belonging to the ten-eleven translocation (TET) family. Conversely to the 5mC, the 5hmC is correlated to gene expression, since it has been observed on promoters, enhancer and active genes (Figure 5.7b) and this could be explained because the TET enzyme may block the activity of DMT, maintaining the promoters in a unmethylated state (Sérandour et al., 2016; Branco et al., 2011).

The methylation/demethylation of the cytosines is cell-type and it also depends on the stage of the cell cycle.

Altogether, the gene regulation driven by chromatin structure, the histone marks and DNA methylation is known as epigenomic regulation. The regulation of transcription driven by these factors is a result of the DNA local structure and the DNA covalent modifications that are key players that limit or facilitate the recruitment of regulatory proteins.

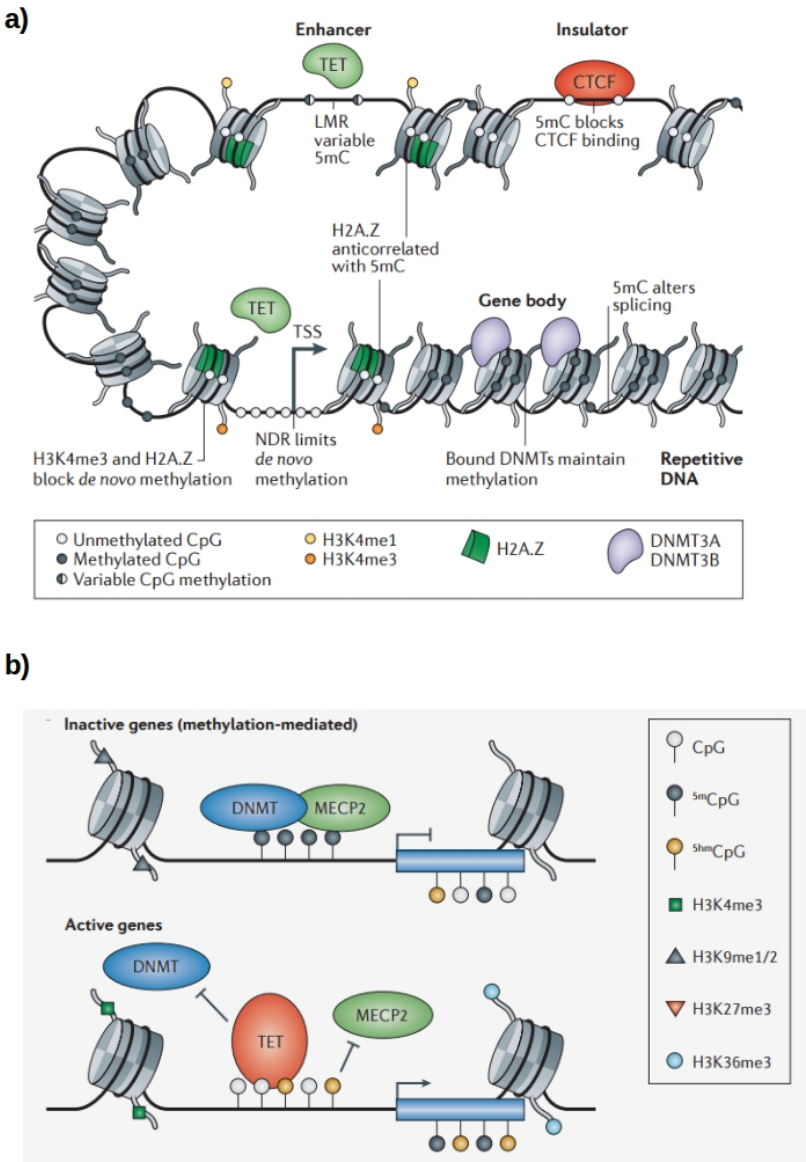


Figure 5.7: Examples of gene regulation mediated by (a) 5-methyl-cytosine (5mC) and (b) 5-hidroxy-methyl-cytosine (5hmC). figure adapted from () and (),

Table 5.2: Summary of TF regulatory mechanisms and their effects on gene regulation.

Mechanism	Effect	Description
Activation	Direct	Recruitment/stabilization of RNAP subunits
Activation	Indirect	DNA conformation change
Activation	Indirect	Co-binding with other factors (synergy)
Activation	Indirect	Recruitment of chromatin remodeller (pioneer TF)
Activation	Indirect	Preventing nucleosome repositioning
Repression	Direct	Blocking RNAP binding
Repression	Indirect	Modulating a TF activator
Activation/Repression	Indirect	Looping

5.5 Transcription Factors

The Transcription Factors (TF) are regulatory proteins that can activate or repress the gene transcription. Their main particularity is that they recognize very short DNA sequences, known as Transcription Factor Binding Sites (TFBS), varying in length from 6-20 base pairs (bp). The TF binding on regulatory sequences (either promoters or enhancers) is a crucial step in the transcription initiation and modulating the transcriptional rates. In general, TFs are classified as activators or repressors, some TFs however, can act as both, depending on the condition (Lee et al., 2012). In addition, the mechanism the TFs help to activate or repress transcription is widely variable but in general can be classified as direct or indirect ways (see table 3.2 for a summary), reviewed in (Spitz and Furlong, 2012) and (Browning and Busby, 2016).

The feature that distinguish a TF from other regulatory elements, it is the ability to bind DNA via a DNA-binding domain (DBD) (see Figure 5.8), other regulatory proteins lacking the DBD are considered co-factors. The DBDs can read the DNA minor or major groove and create short-term weak interactions between the amino acids of the DBD and the nucleotides of the TFBSs. Usually the TFBSs of a particular TF use to be similar at many positions, but not identical. However using computational methods we can infer the consensus sequence for binding (i.e., a representation of the collection of sequences bound by the query TF) for a huge number of TFs (Wasserman and Sandelin, 2004), and as more binding sites are available for a TF, we would have more accurate consensuses (see Figure 5.9), but this is not always feasible since the number of TFBSs of each TF varies, some TFs (e.g., HipB on *Escherichia coli K12*) has a handle of experimentally validated TFBSs whilst other TFs (e.g., cMyc on humans has thousands of reported TFBSs).

The genome-wide identification of TFBSs is a complex task either at computational and experimental level, and at the same time it is not completely understood how the TFs recognize their binding sites: looking for short sequences (6-20bp) in a whole genome (thousands of nucleotides). Several models have proposed that TF spent a lot of time on DNA searching for their binding sites and by four different modes of motion: (i) 3D diffusion (i.e., the TF moves freely in the nucleus), (ii) 1D sliding (i.e., the TF moves through short regions of DNA), (iii) intersegmental transfer (i.e., the TF moves from one DNA segment to another that are not linearly close) and (iv) hopping (i.e., the TF make short 'jumps' away the DNA) (Schmidt et al., 2014; Metzler, 2009). Through these movements TF can scan hundreds or thousands of nucleotides in a short period time (see Figure 5.10) and recent studies suggest that TF use to bind *bona fide* binding sites that are located in regions with a similar GC-content to the consensus binding sequence (Slattery et al., 2014; Dror et al., 2016) (Figure 5.11), reducing thus the universe of TF-scannable sequences.

It is already known that within the TFBSs not all the nucleotides contribute with the same strength for the TF binding, usually the strongest nucleotide contributors are the most conserved positions in the consensus (Figure 5.8), however many recent studies showed the importance of the flanking regions as determinants of strong or weak TF binding specificity (Jurk et al., 2016; Gord??n et al., 2013).

In addition to all the mechanisms listed before, another important feature to list about TFs is their mechanisms to activate or express genes. In bacteria, where the transcriptional regulation use to be simpler,

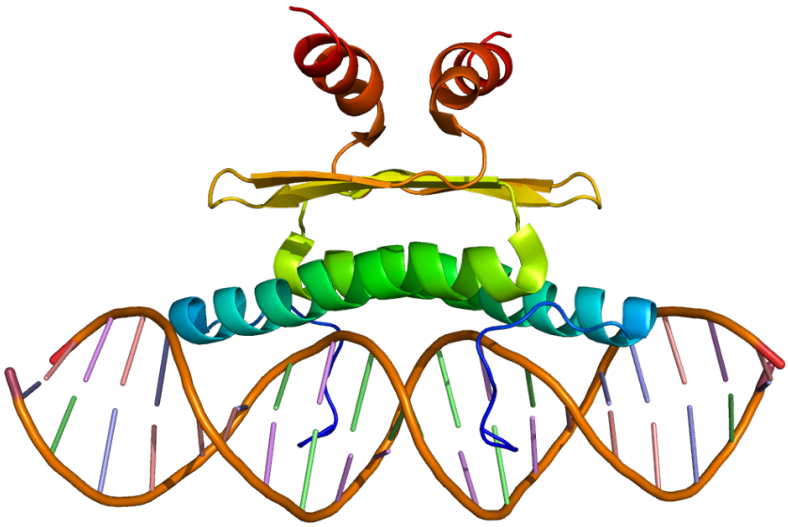
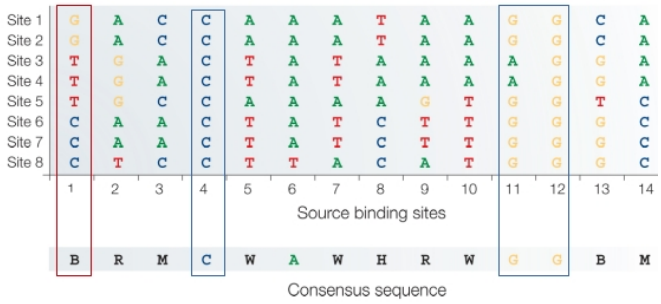


Figure 5.8: 3D structure of TF MEF2C bound to DNA.



IUPAC DNA alphabet

code	type	bases
A	adenine	A
C	cytosine	C
G	guanine	G
T	thymine	T
R	purine	A or G
Y	pyrimidine	T or C
W	weak H-bond	A or T
S	strong H-bond	C or G
M	amino	A or C
K	keto	G or T
H	not G	A, T, or C
B	not A	C, G, or T
V	not T	A, C, or G
D	not C	A, G, or T
N	any	A,C,G, or T

Figure 5.9: Example of binding sites for human MEF2 and table of IUPAC for DNA sequences. A collection of TF binding sites can be represented as IUPAC consensus sequence. Note that whilst some positions of the alignment are highly conserved in all the binding sites (blue rectangle), others are highly variable (redrectangle). The IUPAC table shows the alphabet used to represent all the possibles nucleotides at each columns of the binding site alignment. Figure adapted from Wasserman (2004).

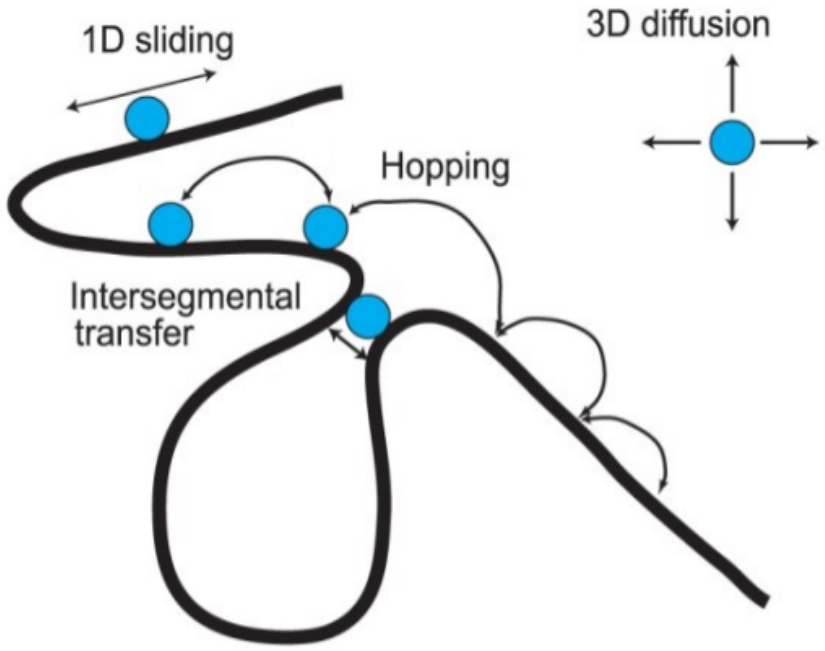


Figure 5.10: TF searching modes. Figure from Schmidt (2014).

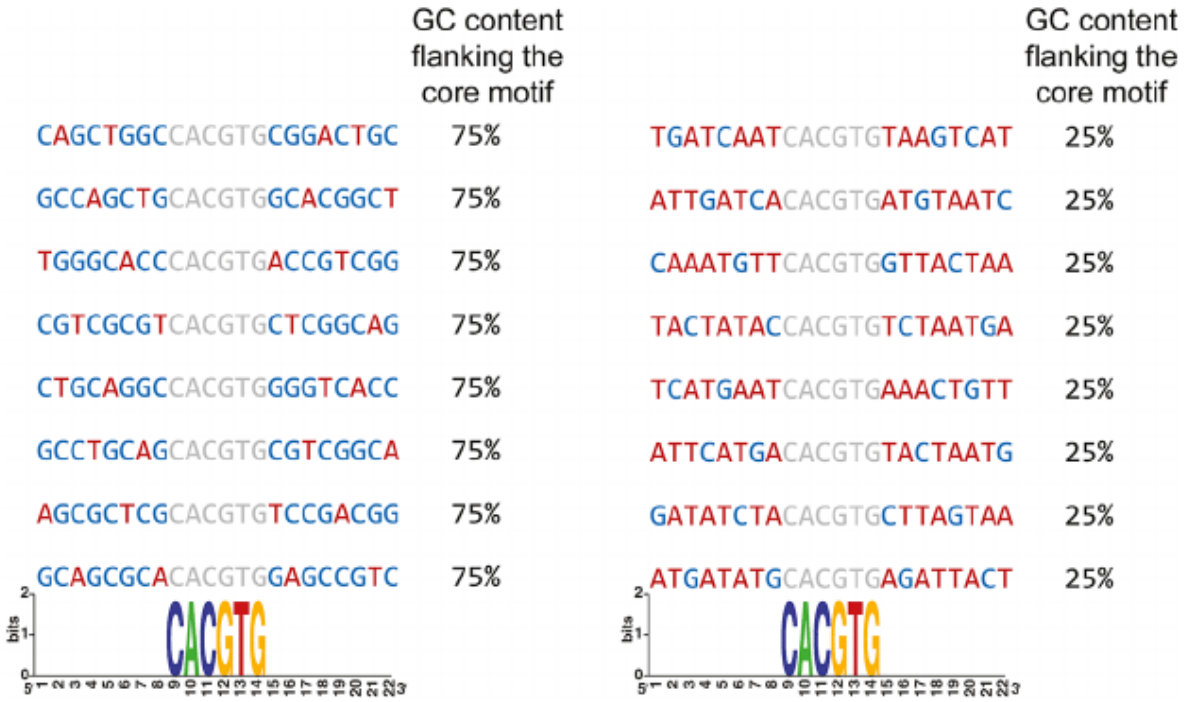


Figure 5.11: Motif environment. TFs tend to bind to regions with highly similar GC content relative to their consensus sequence. Figure from Dror (2016).

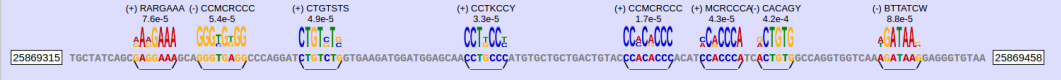


Figure 5.12: Example of Cis-Regulatory Module. Figure from the meme-suite.org, mcast sample.

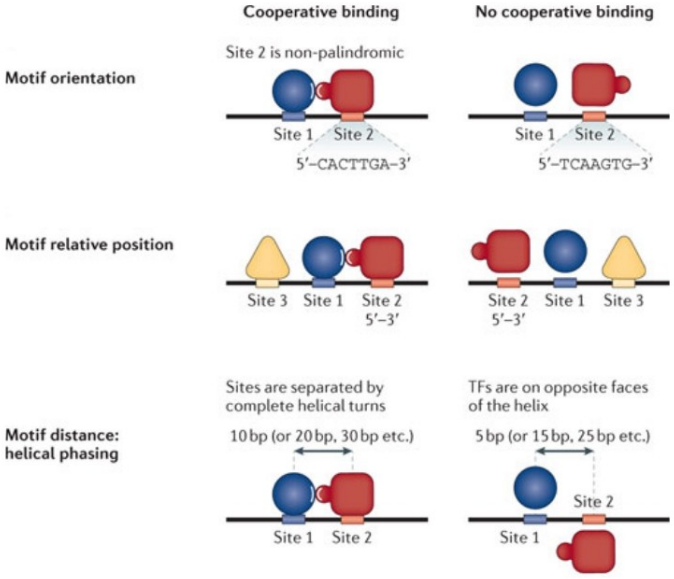


Figure 5.13: Transcription Factor Binding Motif grammar. Several examples of motif grammar considering the orientation, strand and relative position. Figure from Spitz (2012).

relative to metazoa, usually the presence of a particular TF and the position where it is bound relative to the Transcription Start Site (TSS) is enough to infer if such TF would be an activator or repressor. In bacteria, usually the TF activators interact directly with some RNAP sub-units, helping to stabilize the complex before starting transcription, by contrast the repressors usually block the RNAP binding sites at the promoters; the most complex cases of transcriptional repression involve looping of DNA mediated by TFs or histone-like proteins known as Nucleoid-Associated Proteins (Browning and Busby, 2016, Grainger and Busby (2008)).

In metazoan, TFs use to be concentrated at the cis-regulatory regions (e.g., promoters and enhancers), those regions with a high density (i.e., a cluster) of TFs in close proximity are known as Cis-Regulatory Modules (CRMs) (Figure 5.12), and although it is understood that the transcription activation can be tuned up by the cooperative binding among several TFs, there is no (at this time) one way to know the individual contribution of each TF in the regulation of a gene (Hardison and Taylor, 2012), and even now it is not fully understood whether the combinatory and motif positioning (grammar) of TFs matters within a CRM. One challenge in this research area is that the presence/absence of TFBSs is not solely required to study this complexity, other features must be considered (e.g., inter-motif distances, relative orientation, order of motifs, presence of co-factors) (Spitz and Furlong, 2012) (Figure 5.13). In addition, the studied conditions may affect the behavior of a TF, for example RNX1 (a TF involved in blood cell differentiation) can act as either activator in certain conditions and as repressor in others (Whitfield et al., 2012), this behavior could be explained by the presence of co-activators or co-repressors (Zabidi and Stark, 2016; Reiter et al., 2017; Stampfel et al., 2015) or biochemical modifications of the bound sequences, for example methylation of the cytosines.

In brief, although co-factors are regulatory proteins, they do not bind directly the DNA, but they can regulate gene expression (i) interacting with the TFs and change the TF conformation, therefore the T's DNA-affinity, or (ii) making a direct contact with RNAP (using the TF as a scaffold to reach the RNAP) or (iii) through

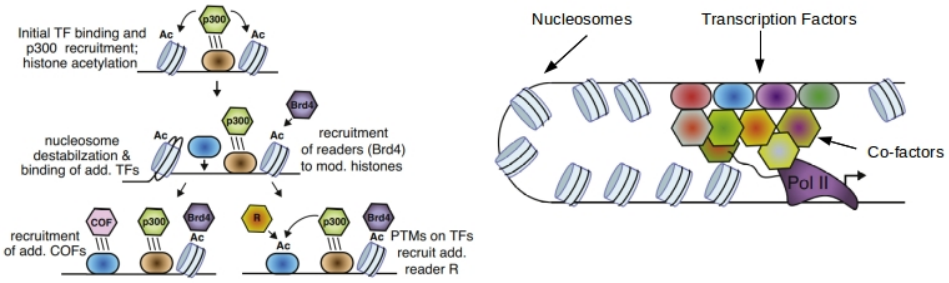


Figure 5.14: Example of co-factors, their functions and how they interact with TFs and RNAP. Figure from Reiter (2017).

an indirect way, for example, recruiting chromatin remodelers, histone modification enzymes or nucleosome destabilizators (Reiter et al., 2017) (Figure 5.14).

Another important property of TFs is their ability to interact among them, through protein-protein interactions, to form TF complexes. When two molecules of the same TF interact, the TF-TF complex is known as homo-dimers (e.g., TF from STAT family), when the TFs are different, it is called hetero-dimer (e.g., Sox2-Oct2). Although dimers are commonly observed, other complexes can be formed, as tetramers or octamers. These TF-TF interactions are frequently observed in some cases of transcriptional repression, for example on *E coli K12*, when a tetramer made of AraC or the lambda-repressor forms DNA-loop that avoid the RNAP binding in the promoter. In other cases a TF alone cannot act unless there is another TF (e.g., CytR is a repressor that only inhibits the activity of bound CRP in *E coli K12*). In metazoan TFs, two recent studies detected TF dimers not observed before and a detailed analysis suggested that the TFBSs recognized by a TF dimer could be different from that recognized by one of its own components (monomers) (Jolma et al., 2015; Isakova et al., 2017), increasing thus the possible combinations of TF binding (lexicon).

The regulation of transcription initiation is also modulated by a set of proteins called General Transcription Factors (GTFs), they are six protein complexes (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH) that bind specifically in the promoters and altogether form the Transcription Pre-Initiation Complex, which interacts with RNAP and helps it to bind to and open the DNA at promoters (Sainsbury et al., 2015). However, although some of these protein sub-units interact directly with DNA (e.g., the TATA-binding protein (TBP) from the TFIID complex), altogether with the RNAP are considered the classical transcriptional apparatus required for transcription initiation in almost all promoters, for this reason they will not be considered for the further chapters and results analyses.

5.5.1 Transcription Factor Families

Proteins are usually made of hundreds of aminoacids that can form structures called domains. The protein domains has a particular function (e.g., bind a metabolite or bind DNA) and a single protein can have multiple domains. We can find several domains within the TF structure, as I mentioned early all the TFs have a DNA-binding Domain (DBD), which is the responsible for the recognition of a particular sequence on DNA, however other domain commonly found in TFs (mainly in bacteria) is the domain that senses environment stimuli (e.g., binding a metabolite) and thus the TF becomes active/inactive.

The TFs can be classified according the similarity of their DBD aminoacid sequences, the resulting groups are called TF families. Usually the TFs from the same family use to have at least 25% of aminoacid sequence similarity among them, this means that the domains should fold similarly and therefor the TFs would recognize similar DNA sequences (Pérez-Rueda et al., 2015). An important feature of DBDs is that they can recognize not only contiguous sequences (monad), but also sequences separated by a spacer (dyads), some times with a fixed or a variable length (Figure 5.15). In many cases, the TFBSs of members belonging to the same TF Family are almost identical (e.g., STAT1 and STAT2, from the family STAT), this is due because

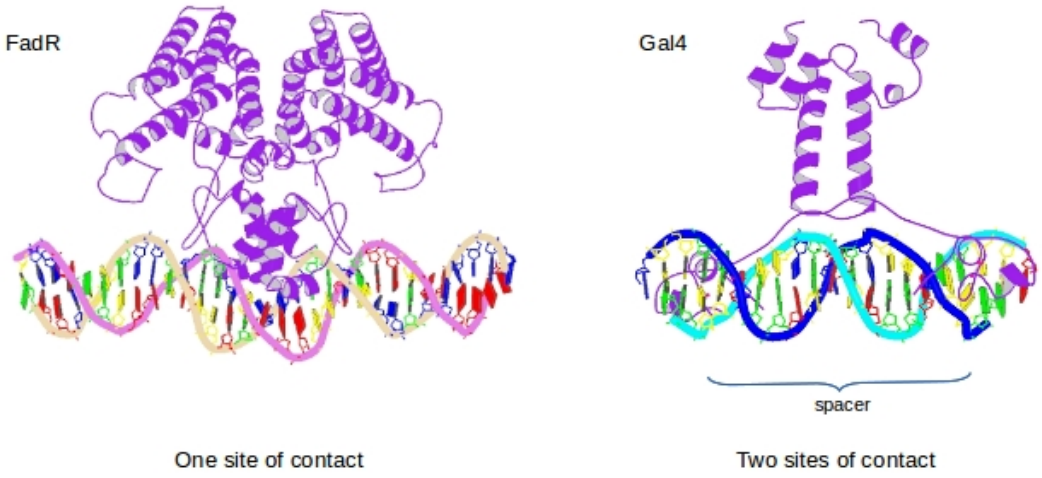


Figure 5.15: Examples of 3D structures of TFs interacting with the DNA in one and two points. Figure adapted from Protein DataBank; ID: 4p9u.

Table 5.3: Annotation of TF Families.

Name	Organisms	PMID	Website
TFClass	Human and mouse orthologs	23180794	tfclass.bioinf.med.uni-goettingen.de/tfclass
-	E coli K12	26094112	-
TEC	E coli K12	26843427	www.shigen.nig.ac.jp/ecoli/tec/
DBD	Multi	20675356	www.transcriptionfactor.org/
PAZAR	Multi	18971253	www.pazar.info/cgi-bin/index.pl

these TFs are actually paralogs, without sufficient time of divergence. But in other cases, although the DBD is similar, it does not recognize similar sequences (e.g., zinc fingers because of the variable spacer between the nucleotide recognized by the DBD).

The number of TF families varies on each organism, there are some TF families associated to a particular taxon for example Zinc clusters on fungi or zinc fingers in vertebrates. Within each family, the number of members differ, from one to hundreds of members. Similarly, not all the member of a family are activators or repressors (as I mentioned in the previous section, this capability depends not only in the TF itself), but one particular family could be associated to a particular function or process (e.g., Hox TF Family involved in development).

Currently, exist different studies and resources that have classified the TFs from different species (e.g., human, mouse, bacteria), and constitute key resources in the identification and annotations of novel TFs (Table 3.2).

5.6 Cis-Regulatory Sequences

As I mentioned in the previous section, the TFs search and bound specific short-sequences in open DNA regions where they can regulate the transcription. If these regions are located near the TSS of a gene are called promoters; if they activate gene expression at distance (relative to TSS of the target gene) and independently of their orientation are called enhancers; if they repress gene expression distally are called silencers; if they delimit the euchromatin from heterochromatin and isolate the enhancer activity are called

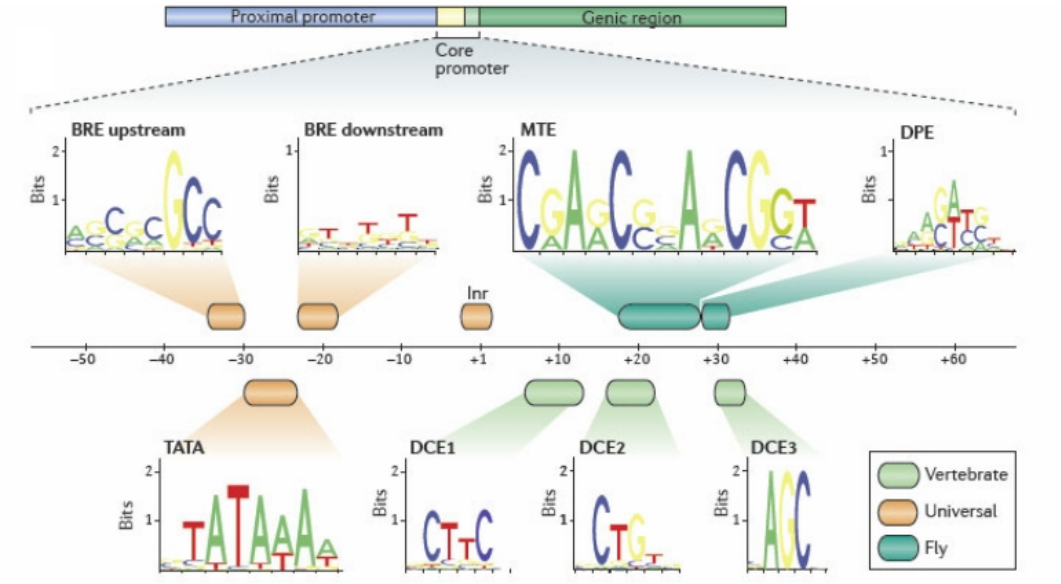


Figure 5.16: The metazoa core-promoter and its elements. BRE: B recognition element; DCE: downstream core element; DRE: DNA recognition element; MTE: motif ten element. Figure adapted from Lenhard (2012).

insulators (Kolovos et al., 2012), of note that in bacteria the promoter is the classical cis-regulatory element and only a few cases of distal regulation have been reported (Beck et al., 2007). In this section I will explain with more details the genomic and epigenomic features of these elements and how they interact (e.g. physically) to activate or repress a target gene.

5.6.1 Promoters

The transcription of a gene starts when the RNAP complex is stable. The region near the TSS (+/- 50 bp) where the RNAP is loaded is called core promoter (Kadonaga, 2012). Within the core promoters there are short regions that are recognized by the GTFs (e.g., TATA-box, Initiator (Inr), upstream and downstream TFIIB Recognition Elements (BRE)) or RNAP sub-units, some of them are taxon-specific, and normally a promoter has only a few of these elements (the Inr is the most common), rarely all of them (Lenhard et al., 2012). As consequence, a core promoter *per se* can rarely activate the transcription of a gene, and it requires the help of other proteins (e.g., TFs, GTFs) which are usually bound immediately in the region upstream the core promoter. This region, known as the proximal promoter, is indispensable for those promoters lacking the TATA-box and usually the GTFs are recruited in this region (Sainsbury et al., 2015). Hereafter the term promoter will be used to refer the region including either the core and proximal promoters (Figure 5.16).

Although this work is focused in the regulatory sequences, is important to mention the contribution of the epigenomic context in the regulation of gene expression. It must be taken in consideration that an active promoter is always located at an open DNA region, a nucleosome-depleted region (NDR) flanked by two nucleosomes (called promoter-associated nucleosomes), these open regions allows the access and assembly of regulatory elements (e.g., TFs, GTFs, RNAP). In addition, many histone modification have been associated with the promoter-associated nucleosome (e.g., H3.3/H2A.Z) and specific histone marks (e.g., H3Kme3 and H3K27ac) have been associated with active promoters (Figure 5.17), but other marks, for example H3K27me3, are associated with repressed promoters (Lawrence et al., 2016).

Another property of some promoters is their ability of bidirectional transcription, usually the transcription is unidirectional, however, recent studies have revealed that bidirectional transcription is a common phenomena across metazoa promoters and it is not completely understood (and is still under debate) whether the

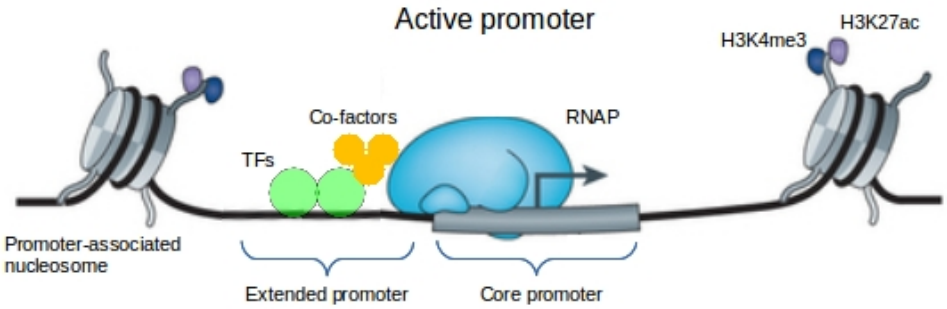


Figure 5.17: An active promotre with its epigenomic features. Figure adapted from Shlyueva (2014).

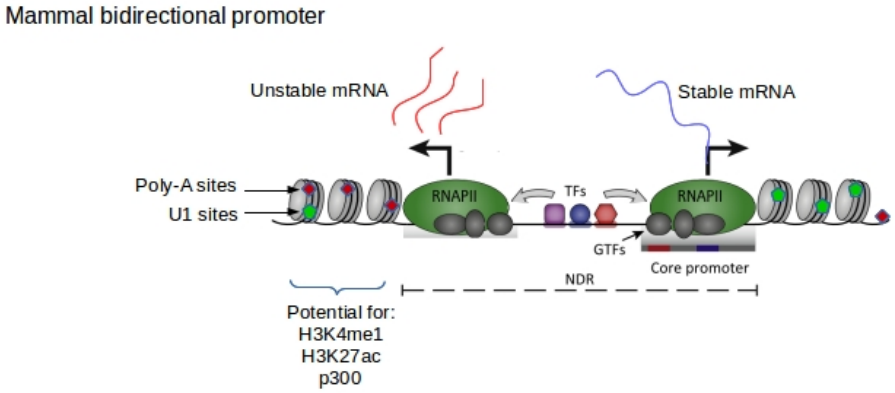


Figure 5.18: Bidirectional promoters in bacteria and human. Figure adapted from Andersson (2015).

bidirectional transcription is due to two close promoters or as a consequence of the RNAP recruited in an open DNA region with a high concentration of TFs (other cis-regulatory regions (e.g., enhancers) show bidirectional transcription as well) (Andersson, 2015; Scruggs et al., 2015; Bagchi and Iyer, 2016). Only one of the transcripts, that one transcribed from downstream gene, will produce an stable mRNA, the other transcript (antisense) normally is rapidly degraded (because of the lack of splice sites for U1 and a higher concentration of poly-A sites). It has been discovered many features associated to promoter bidirectionality, for example the presence of promoter elements (TATA-box), over-representation of TFs including NF-Y, Nrf-1, YY1, GABP, MYC, E2F1, and E2F4, or special histone-tail modifications (H3K4me2/3) or histone variants (H2A.Z and H3.3). In addition, the NDP of the bidirectional promoters use to be longer that those for the mono-directional, allowing thus the binding of more TFs (Bagchi and Iyer, 2016) (Figure 5.18).

5.6.2 Enhancers

In addition of the TSS-proximal regulatory regions (promoters), exist other class of regulatory elements capable of activate transcription from a TSS-distal position, these group of cis-regulatory regions are known as enhancers. Although the enhancers were discovered since many years ago, it is until recent times that they became widely studied, and at least in the human genome, a huge number of regions with potential enhancer activity has been reported by the ENCODE project (Encode Consortium, 2012).

Usually, the enhancers are defined as a cis-regulatory regions that active genes distally, their size vary in length from 100-1000bp and contain a large number of TFBSs for a multitude of TFs, which in turn recruit other regulatory proteins (e.g., co-factors, chromatin remodelers) (Figure 5.20). Another characteristic of the enhancers is their capability to interact physically with other cis-regulatory regions (e.g., promoters) that can be distant in the genome (usually the enhancers are located on intergenic regions, however they

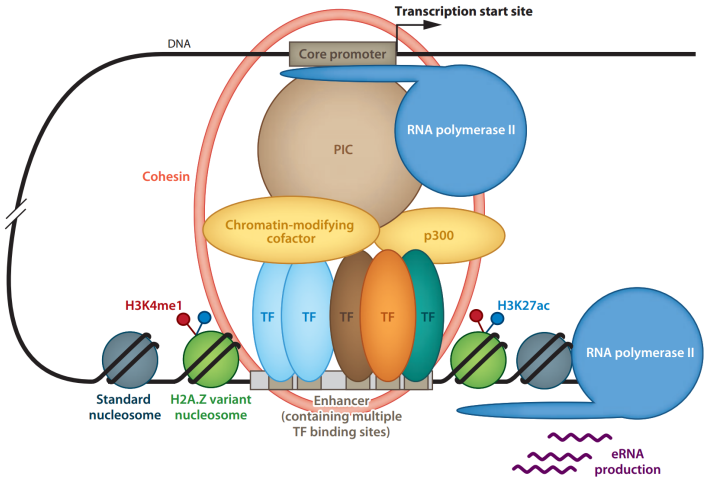


Figure 5.19: An enhancer and its genomic and epigenomic features. Figure taken from Maston (2015).

have been found in introns as well), and they can activate gene expression independently of their orientation. These interactions are, in part, orchestrate the transcriptional regulation (Pennacchio et al., 2013; Shlyueva et al., 2014).

Although once an enhancer is identified, it is usually associated (naively) to interact with the closest TSS, or using complex methods it is associated to a single distal TSS. Recent studies showed that an enhancer can interact with several promoters and one promoter could be associated to several enhancers (distinct enhancers for distinct conditions). Given that the enhancers can recruit a high concentration of TFs (and other regulatory proteins), first, their chromatin environment should be open. As the chromatin regulation is cell-type dependent, it is important to note that not all the enhancers are active at the same time (Andersson et al., 2014).

Similarly to promoters, many epigenomic features (i.e., histone marks) have been associated to active enhancers. Starting from the chromatin environment, enhancers are regions with low nucleosome occupancy and high DNaseI hypersensitivity, as an open DNA region, the flanking nucleosome use to have particular histone variants (e.g., H3.3 and H2A.Z) that use to be unstable in order to a rapid chromatin remodeling. Other feature is the enrichment and depletion of H3K4me1-H3Kme2 and H3Kme3, respectively, relative to promoters, and the histone mark H3K27ac have been associated specifically to active enhancers (Engel et al., 2016; Maston et al., 2012). Although many features have been associated to enhancer, it is difficult to find enhancers using simply the epigenomic features, many studies report that the variability in epigenomic features in enhancers is higher than those observed in other regions as promoters (Heintzman et al., 2007). Although the epigenetic features can indicate the DNA accessibility, other features (e.g., presence of TFs) could complement the identification of *bona fide* active enhancers. Some examples are the two co-factors p300 or CBP (CREB-binding protein), both are acetyl-transferases that interact with TFs usually bound at enhancers.

In addition to the epigenomic and genomic features already described, another property of the enhancers it that, similarly to promoters, given the high occupancy of TFs and RNAP, some enhancers can present bidirectional transcription, resulting in the production of a class of RNA known as eRNA (enhancer RNA), and currently are considered a key feature for a large part of active enhancers (Andersson et al., 2014) and this phenomena occurs at genome-wide scale (Figure 5.19). But similarly to some anti-sense transcripts produced at promoters, usually the eRNAs are rapidly degraded, because the lack of U1 splice-sites and enrichment of poly-A sites (Andersson, 2015; Nguyen et al., 2016). Although the function of eRNAs is not completely understood and it is currently debated (Rahman et al., 2017), it has been associated with the modulation of enhancer-promoter looping stability and recruitment of co-factors (Hsieh et al., 2014).

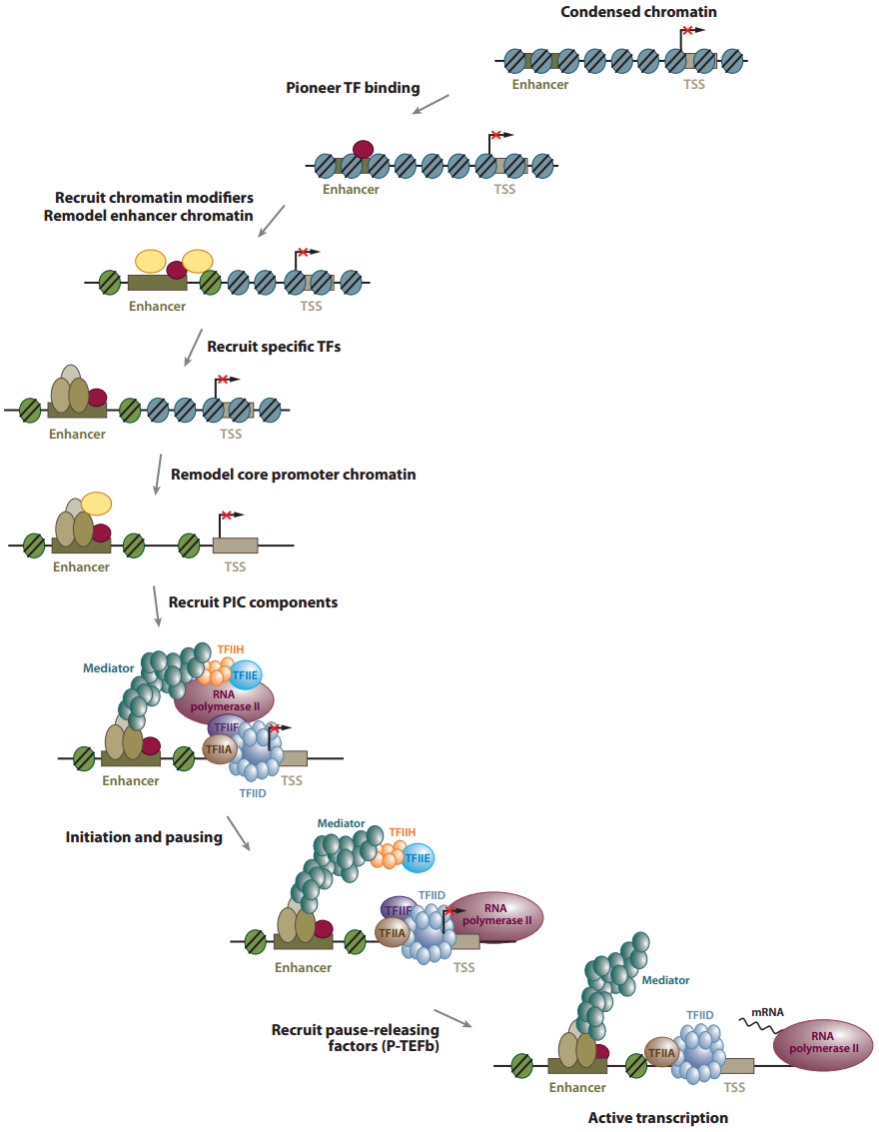


Figure 5.20: Scheme with the steps of enhancer activation and recruitment of regulatory elements. Figure from Maston (2015).

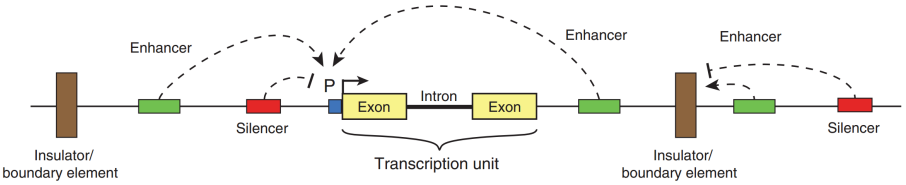


Figure 5.21: Scheme with the steps of enhancer activation and recruitment of regulatory elements. Figure taken from Maston (2015).

5.6.3 Silencers and insulators

In addition to promoters and enhancers, that altogether control the activation of genes, there are other cis-regulatory regions that control the repression. One of these regions are called silencers, which similarly o enhancers act TSS-distally, and independently of their orientation with the difference that their activity is always repressive (conversely to enhancer activity that is always gene activation).

Silencers recruit combinations of TFs or co-factors that repress the gene activity and their activity can be direct, interacting physically with a promoter, indirect by modulating an activator element (i.e., interacting with an enhancer and annihilating its activity). It is difficult to assign whether the TF recruited at the silencers are strictly repressors. One possibility is that TFs bound or the co-factors recruited on silencers are exclusively repressors or co-repressors, respectively, but there are many TFs that can act as activators and repressors (e.g., RUNX1), according to different conditions. Currently, a short number of silencers have been identified in mammal genomes and hence their genomic and epigenomic features are not well described (Liu et al., 2006; Hao et al., 2015).

Another cis-regulatory regions, with repressive activity are the so-called insulators, they are called so because they prevent the activation of a gen by an enhancer (i.e., they isolate the enhancer activity) (Figure 5.21). In addition, insulators limit the heterochromatin boundaries. Similarly to enhancers and silencers, insulators can act distally and independently of their orientation relative to their targets. Although the regulating mechanism are not fully understood, the function of insulators is associated with CTCF (that is also considered a TF) (Ong and Corces, 2014); and two models of regulation have been proposed (Herold et al., 2012): (i) looping model, where two or more insulators interact physically (using cohesin) and this loop alter the 3D genome conformation, affecting thus the promoter-enhancer interactions; (ii) decoy model: the insulator interacts directly with other cis-regulatory elements and inhibits thus their activity.

5.6.4 Similarity between enhancers and promoters

Enhancers and promoters are classically considered as independent regulatory elements, based on their relative location to TSSs and their histone modifications. Recent studies, however, have higlighted the functional, genomic and epigenomic similarities between enhancers and promoters (Andersson, 2015; Kim and Shiekhattar, 2015; van Arensbergen et al., 2016; Arnold et al., 2016; Dao et al., 2017) listed below (Figure 5.22):

- Epigenomic features:
- Enhancers and promoters are located at open chromatin regions.
- Their flanking nucleosomes contains the histone variants H3.3 and H2A.Z (Barski et al., 2007).
- Although some studies define enhancers and promoters based on histone marks, it has been shown that both share similar histone marks although the amount of these marks differs. For example, the ratio of H3K4me1/H3K4me3 is low at promoters and high at enhancers, whilst the ratio of H3K4me3/H3K27ac is high at promoters and low at enhancers (Barski et al., 2007; **?**; **?**). The changes on these ratios occurs

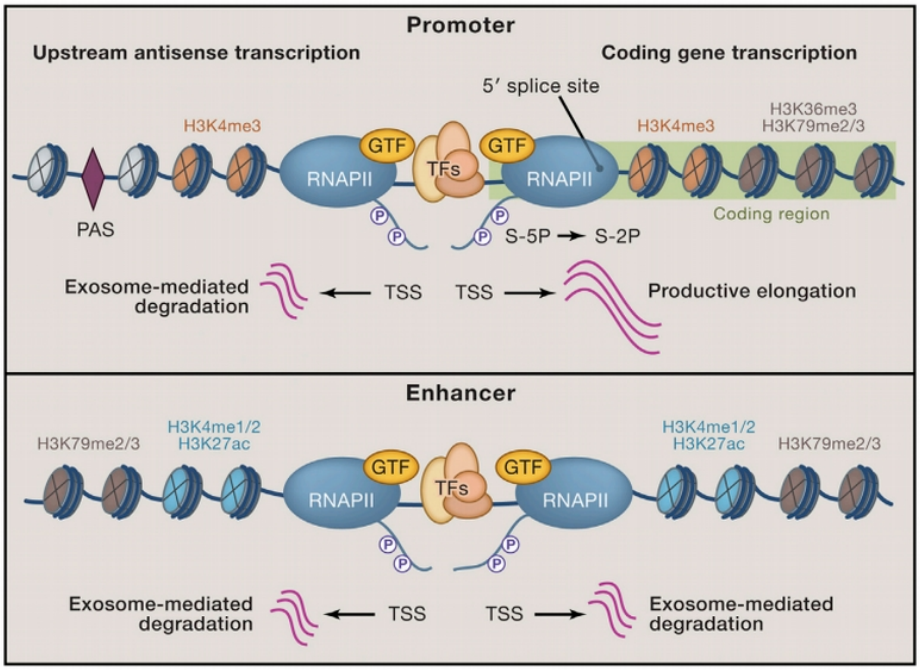


Figure 5.22: Similarities between enhancers and promoters. Figure taken from Kim (2015).

during the cell differentiation and development, which reflects the dynamic changes on transcription activity, rather than represent functional regulatory elements.

- Genomic features:
- Enhancers and promoters contain typical core promoter sites (TATA box, INR motifs, GTFs binding) which enable the RNAPII recruitment and therefore the transcription initiation (**?**).
- Recruitment of RNAPII, although at enhancers occurs at a lower rate (**??**).
- Contain binding sites for different TFs.
- Bidirectional transcription.
- Evidences:
- In a recent study, it was shown that a genomic region may display promoter-related histone modification in a cell line, and enhancer-related histone marks in another cell line (Leung et al., 2015).
- An enhancer located at an intronic regions may function as an alternative promoter, producing thus a protein isoform in a particular condition (**?**).
- The study by Li et. al (**?**) showes that promoters more frequently interact with other promoters than enhancers. In addition, they also showed that some promoters interacting with other promoters displayes enhancer-related epigenomic features.
- Using a genome wide enhancer-assay, STARR-seq, Zabidi et al. (**?**) found that some *Drosophila melanogaster* randomly fragmented regions with enhancer activity overlapped or where proximal to TSSs. They suggested that these regions may act as *bona fide* enhancers.
- Other study compared the activity of hundreds of promoters and enhancers using Massively Parallel Reporter Assays. The authors showed that some sequences displayed both enhancer and promoter activities. Interestingly, although promoters displayed more frequently promoter activity, they found that both enhancers and promoters displayed similar enhancer activity (Nguyen et al., 2016).

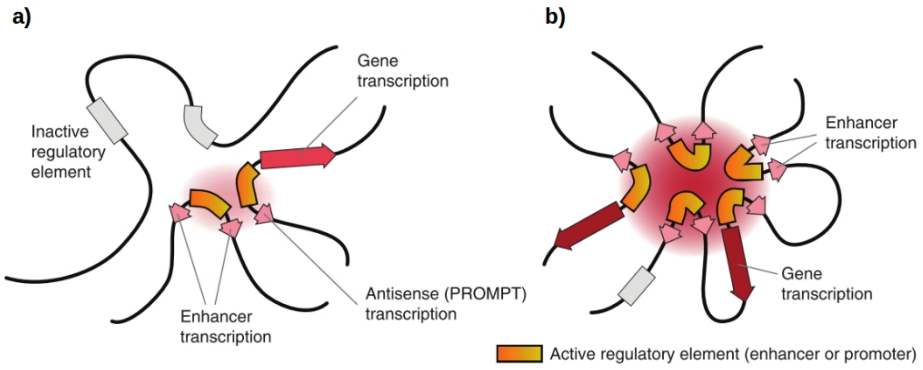


Figure 5.23: Transcriptional output increases with interaction between regulatory elements. a) With a low number of regulatory elements, the gene activity is low. b) With more elements interacting, the gene activity increases. Figure taken from Andersson 2015 (2015).

- It was shown that in human, ~2-3% of promoters display enhancer activity, by activiating near promoters. These class of promoters are called Epromoters and are mainly related to cell stress induced upon viral or bacterial infection. In the same study, the authors showed that some promoters gain the enhancer activity upon an stimulus, suggesting a complex dynamic interaction between promoters (Dao et al., 2017), see chapter 13.

Even when enhancer and promoters have a large list of similarities, there are a few differences that should be noted:

- CpG islands, they are common at promoters but enhancers are poor on CpG islands, suggesting that the TF binsing site composition may differ among these cis-regulatory regions.
- Although enhancers and promoters have bidirectional transcription, the functions of their produced RNAs may differ. The eRNAs are short, unstable and rapidly degraded. Conversely, the mRNAs produced are long, stable and exported to the cytoplasm for translation.

Altogether, the evidences described above showed that enhancers and promoters are similar, call for a revision of established distintion between these elements, specially in the fact that they may be considered as the same class of regulatory elements, independently if they are near or far from TSSs or the epigenomic features associated. According to the model proposed by Andersson (Andersson, 2015) the cis-regulatory sequences often interact in close physical proximity in RNAPII foci, and depending on the context, the elements can act as either promoter or enhancer, therefore influenting on the activiy of the other physically close elements in a synergistic way (Figure 5.23).

Chapter 6

Experimental detection of TF binding events

The mapping of the TF binding sites is key for understand regulatory interactions between TFs and their target genes, this information can be further used in order to infer transcriptional regulatory networks, to detect regulons (i.e., a set of genes regulated by a TF), and recently to detect regulatory biding variants that affect the binding of a TF with consequences on the expression of the regulated genes. The evolution of the methods to detect TF binding events has progressing since the detection of a handful of TFBSs in a single experiment, for example using low-throughput methods as EMSA (Cann, 1998) and DNAse footprint (Galas and Schmitz, 1978) to recently developed high-throughput methods.

Before revising every of the low and high-throughput methods to detect TF binding events, it is important to note that there are two main differences on these methods:

- Detection of binding events: whilst the low-throughput methods can detect the exact position of the TF binding site at a single nucleotide resolution, the high-throughput methods do not detect the precise location of TFBS, but a region of variable size where the TF is bound. In order to generalize these difference in binding sites and binding regions, I will use the term binding events.
- Throughput: even when the low-throughput methods have a high resolution, they are limited to detect a low number of TF binding events (~10). By contrast the high-throughput methods can detect from hundreds to thousands TF binding events in a genome in a particular experimental condition.
- Analysis of results: all the TF binding sites for a particular TF detected with low-throughput methods can be collected in order to create a TF binding model. By contrast the TF binding regions detected by high-throughput methods must be processed in order to find the putative TFBSs.

6.1 Low-troughput TFBS detection methods

6.1.1 Electrophoretic Mobility Shift Assay

The Electrophoretic Mobility Shift Assay (EMSA), also known as gel shift assay, is a technique to study DNA-RNA or DNA-protein interactions. It can be used to determine if a protein (or a set of proteins), for example TFs, are bound to a DNA sequence of interest.

The logic behind this method is that the interest sequence (with no proteins of RNA bound) has a molecular weight that can be visualized as a band detected by eletrophoresis on a polyacrylamide or agarose gel, this band is used as a control. The speed at which the molecules migrates trough the gel depends on their molecular weight and charge. In case when the tested sequence is bound to another molecule, the molecular

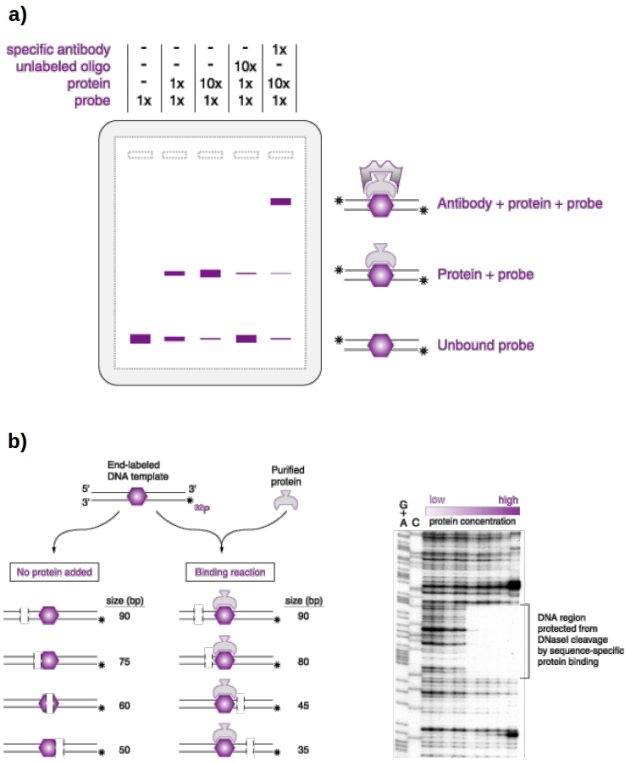


Figure 6.1: Graphical representation of low throughput method for TFBS detection. (a) EMSA and (b) DNAse footprint assays.

weight of this complex (e.g., DNA-TF) will be higher and hence the molecules are less mobile than the control. In the gel this is visualized as a band shifted relative to the control (Cann, 1998) (Figure 6.1a).

With this approach it is possible to evaluate several control sequences (one per lane) in a single run. These sequences might be *bona fide* TFBSs and EMSA assay can be used to detect the strongest site (i.e., the site with the highest affinity). Once a set of of TF binding sequences have been identified, they can be represented as a TF binding motif.

This method is one of the most used to study individual TFBSs, it can be used to detect *bona fide* TFBS from a set of candidate sequences, and can be used as well to classify strong or weak TFBSs for a particular TF. Their limitations are that a few sites can be evaluated in a single run and that requires *a priori* knowledge of the evaluated sequence.

6.1.2 DNAse footprint

Whilst the EMSA can be used to determine the presence/absence of bound proteins in a given sequence, it does not detect the precise location of the TFBS. To do so, another method can be used, that is the DNAse footprint assay which take advantage of the molecular properties of the deoxyribonuclease (DNAse), an enzyme that degrades DNA. Similarly to the EMSA assay, the DNAse footprint is visualized on a gel.

The logic behind this method is that DNA alone (control) will be degraded by DNAse. However, if the DNA is bound by a protein (e.g., a TF), this union will protect the DNA from DNAse cleavage at the binding site but the surrounding DNA will be degraded, revealing a pattern (or footprint) on the gel where the DNAse could not degrade the DNA, in other words, revealing the exact location where a protein interacts with DNA (e.g., TFBS). The remaining (protected) fragments, can be further isolated and amplified in order to detect the exact sequence (Galas and Schmitz, 1978) (Figure 6.1b).

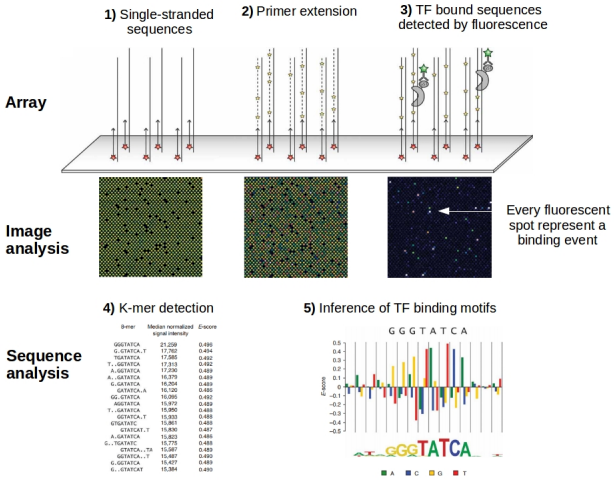


Figure 6.2: TF binding analysis with Protein Binding Arrays. The single-stranded sequences containing the k-mers are double-stranded by primer extension. The TFs marked with an epitope are added to the PBM, next an antibody tagged with a fluorophore is added, producing a fluorescent signal in the spot where the TF is bound to the DNA. This fluorescent signal that is proportional of the binding strength. The k-mers with highest fluorescence are ranked and can be assembled as a TF binding motif.

In addition to identify TFBSs, DNASe footprint method can be used to detect the minimum amount of protein (by increasing its concentration) required to observe the DNAse footprint pattern, i.e., the minimum concentration of a given protein required to be bound to DNA. The limitation of this method is that only few sites can be evaluated in a single run.

6.2 High-troughput TFBS detection methods

6.2.1 Protein binding Microarrays

The Protein binding Micro arrays (PBM) are the first high-throughput method able to detect *in vitro* TF binding events at genome-wide scale and can be used to measure the TF binding affinities independently of the genome (Berger and Bulyk, 2009; Mukherjee et al., 2004).

The method works as follows (Figure 6.2):

1. All the oligonucleotides of a given size k (e.g., 8-mer for k = 8) are first fixed in an array as single-stranded. Every k-mer is appears at 16 times in the array and their flanking sequences are different on each instance.
2. The k-mers are doubled-stranded by primer extension.
3. A TF tagged with an epitope is bound to the DNA on the array.
4. The array is washed to discard non-specific binding events.
5. The TF molecules bound to the array are labeled with a fluorescent antibody.
6. The bound k-mers are detected the fluorescence signal intensity.
7. The signal intensities are used as score to rank the k-mers.
8. The list of selected k-mers can be further converted to a TF binding motif.

In addition to include the 10-mers, the PBM also evaluate all the combination of gapped 8-mer with at most 4 gaps, this allow the detection of spaced motifs, a common TF conformation observed on bacteria and yeast.

The PBM has the advantage of detect and measure the strength of TF binding events in any genome, independently of the genome annotation or if the genome of interest has been already sequenced. All the analysis can be achieved in a short period of time (two days according to the authors).

The simultaneous evaluation of all k-mers allow to detect strong and weak binding events, in addition the k-mer analysis allow the detection of nucleotide interdependencies, a feature that could not be detected by motif discovery methods at the time when the PBM were released, although for visualize the results, the most representative k-mers are summarized as a Position-Specific Scoring Matrices (PSSMs), that simplifies the representation of these k-mers.

The main limitations of PBMs are the following:

- It does not detect the exact location of TFBSs in the genome.
- Some k-mers with high *in vitro* affinity could not be relevant *in vivo* because hetero-dimers of TFs or interaction of TF with co-factors are not evaluated by PBMs.
- The method used to infer the TF binding motifs should be carefully selected, since distinct methods could detect different motifs and lead to misleading conclusion about motif heterogeneity (Zhao, 2013; Badis et al., 2009).

6.2.2 Chromatin Immunoprecipitation (ChIP-x) methods to detect TF binding events

Chromatin Immunoprecipitation (ChIP) is a method commonly used to study *in vivo* interactions between DNA and proteins, the most studied proteins by this method are TFs and histones (Orlando, 2000). ChIP is the base of recently developed techniques that allow to detect all the TF binding events at a genome-wide scale (cistrome), this particularity represents and advantage relative to the PBMs, however these methods are prone to detect a high number of false positives due to the cross-link of transient proteins to the DNA, and depend on the quality and specificity of the antibody chosen.

The conventional ChIP method is as follows (Figure 6.3):

1. Cross-link: the DNA and its associated proteins (e.g., TFs) are cross-linked (covalently bound) using formaldehyde or another molecule. This step assures that the DNA-binding proteins remain fixed to the DNA. This step occurs in living cells.
2. DNA fragmentation: the DNA is randomly fragmented by sonication or DNAse digestion, producing DNA fragments of ~500bp.
3. Immunoprecipitation: the DNA fragments cross-linked with a protein (the fragments of interest) are immuno-precipitated (pulled-down) using a protein-specific antibody.
4. Purification: the pulled down DNA fragments are reversely cross-linked in order to release the bound protein.
5. Enrichment: the DNA fragments that are recurrently pulled down in a significant proportion relative to a control, in other words, that are enriched, represent those regions of the genome where the protein of interest is bound *in vivo*.

This steps are done in two population of cell with the difference that in one population, the protein of interest (e.g., a TF) is immunoprecipitated but not in the second population, whose results will be used as a negative control as a reference to detect the enrichment.

It is important to note that ChIP methods are not limited to detect TF binding events, the positioning of other DNA-binding proteins as histones or RNAPII can also be studied with these methods (Schones et al., 2011).

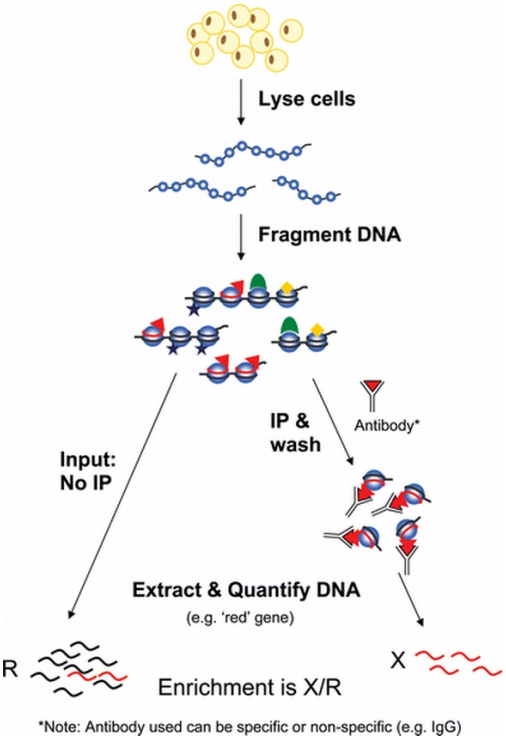


Figure 6.3: Steps for Chromatin immunoprecipitation. Figure from Wu (2009).

The results of these ChIP-x methods have been used to create catalogs of TF binding regions in a genome (e.g., REMAP (Griffon et al., 2015)). The integration of such results is useful for the detection of cis-regulatory modules and cis-regulatory sequences at genome-wide scale, therefore contributing to the annotation of the analyzed genomes.

In order to detect the TF (or other proteins) binding events, the detection of the enriched regions can be done using different methods, as sequencing of microarrays.

6.2.3 ChIP-chip

The ChIP-chip technique combines the ChIP method and the detection of the enriched regions is done with microarrays (chip) (Jothi et al., 2008). This technique follows the first four steps of the ChIP, but the enrichment is detected as following (Figure 6.4):

6. The enriched DNA regions are denaturalyzed (to a single-stranded (ss)DNA).
7. The ssDNA is hybridized to a ssDNA microarray containing a selected set of sequences (e.g., all the promoter of yeast, all the intergenic sequences of *Escherichia coli* K12).
8. Mapping probes to a reference genome in order to identify TF (or other protein) binding regions with a resolution of ~200bp.
9. The identification of TFBSs should be done *a posteriori* with motif discovery tools.

Although two of the main issues of this method are that does not bring a precise location of the TFBSs, since it reports large DNA regions that should be further analyzed by motif discovery tools and the high signal to noise ratio, two of its main advantages (and is the same for the genome-wide ChIP-based methods) are that eliminate the bias reported on the PBM, it is possible to infer gene regulatory networks given that the binding regions can be mapped in a genome, that actually is not possible using the PBM data *per se*, and

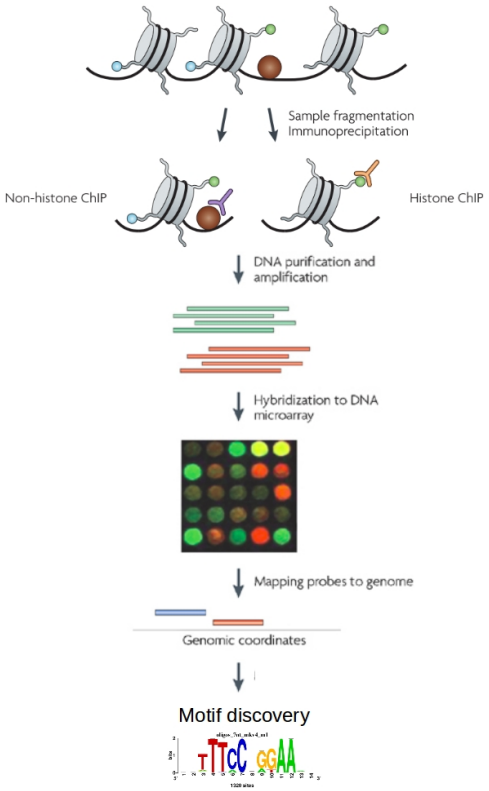


Figure 6.4: Steps for ChIP-on-chip technology. Figure adapted from Park (2009) and Zhao (2008).

that it permits to discover TF binding events on regions that were unanticipated (e.g., at introns) (Gilchrist et al., 2009).

6.3 ChIP-seq

Since 2007, the advent of high-throughput sequencing methods marked a milestone in the genome-wide analysis, with great benefits for the genome-wide detection of TF binding regions through ChIP-seq. The main methodological difference relative to ChIP-on-chip is that the immunoprecipitated sequences are sequences and can be directly mapped on a reference genome, without an microarray hybridization step, as consequence all the binding events can be detected at any position of the genome and are not limited to a selected set of sequences (Park, 2009; Furey, 2012).

This technique has considerable advantages over its predecessor, the ChIP-on-chip: (i) requires a lower quantity of input DNA, (ii) the resolution of the results is higher, allowing to detect binding events in regions of ~150bp, (iii) it is less noisy and (iv) has a higher coverage. See (Park, 2009; Gilchrist et al., 2009) for a detailed comparison between ChIP-seq and ChIP-on-chip technologies (Figure 6.5).

The ChIP-seq technology follows the same steps for the ChIP but the enrichment is detected as following (Figure 6.6):

6. The enriched DNA regions (and the control sequences) are sequenced with a short-read sequencer.
7. Every short-read is mapped to a reference genome.
8. The detection of the read-enriched regions (commonly named *peaks*) are detected using specialized software known as *peak-callers* (Steinhauser et al., 2016), that compare the immunoprecipitated reads

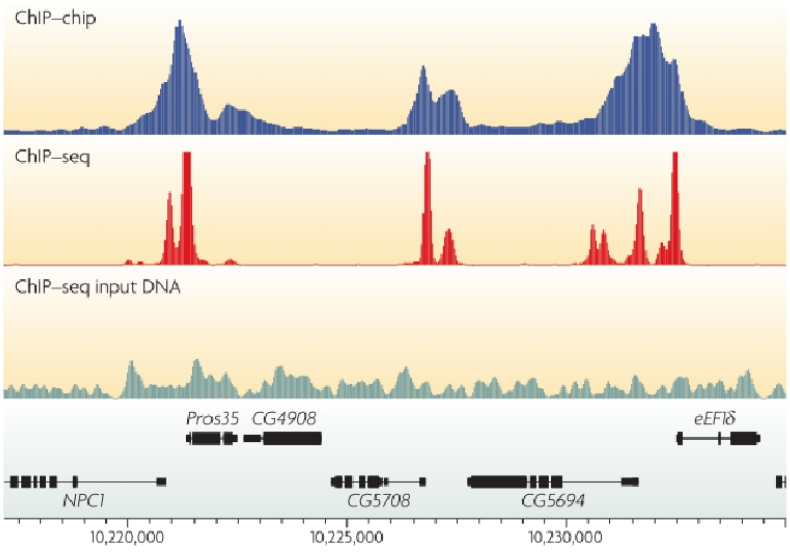


Figure 6.5: Diferences of binding regions length detected by ChIP-on-chip and ChIP-seq. Figure adapted from Park (2009).

at a particular region relative to the read concentration observed in the control. The selection of the peak-caller is crucial since some of them are focused on histone and others on TF peaks.

9. The identification of TFBSs should be done *a posteriori* with motif discovery tools.

ChIP-seq has been used to detect either TF binding events, as well as histone marks, nucleosomes and RNAPII, the original technique has been slightly modified to detect particular proteins or chromatin modifications (Landt et al., 2012; Schones et al., 2011). Depending on the analyzed, the peak length can vary, for example whilst the TF peak use to be narrow (~150bp) the histone peaks encompass hundreds or even thousand of nucleotides (Figure 6.7).

Nowadays, ChIP-seq is the most popular method for genome-wide detection of TF binding regions, and thousands of experiments are publicly available at Gene Expression Omnibus (GEO) website. And, as every method has its own limitations: (i) the fragmentation of the sequences is not equal in the samples, as consequence, the peak-callers can detect false positive given the uneven distribution of reads; (ii) the repetitive sequences may be detected as enriched regions; (iii) dependency on several bioinformatic methods.

Every step of the ChIP-seq after the ChIP, from the read alignment, mapping, peak calling and motif discovery can be analyzed using several tools, with different parameters making the results difficult to reproduce, although some guidelines (Landt et al., 2012; Bailey et al., 2013) have been proposed, there is no a standard method to analyse ChIP-seq data.

6.4 ChIP-exo and ChIP-nexus

The ChIP-exo and nexus techniques are extensions of the ChIP-seq method that add one extra step of exonuclease digestion that degrades the DNA in the 5'-3' direction except the protected regions (bound by a TF). As consequence, the detected peaks are shorter than those detected by ChIP-seq, almost reaching the single nucleotide resolution (Rhee and Pugh, 2011; He et al., 2015). The detection of the binding events are detected by high-throughput sequencing.

It is important to note that these methods are the state of the art for the *in vivo* TF binding event detection, and their resolution is so high that a peaks almost correspond to a TF binding site. This enable to detect

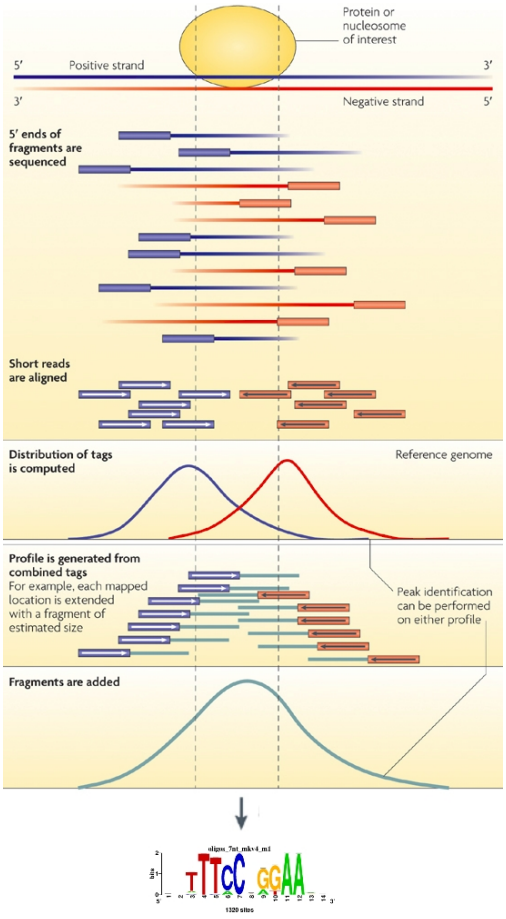


Figure 6.6: Detection of ChIP-seq peaks after immunoprecipitation steps. Figure adapted from Park (2009) and Zhao (2008).

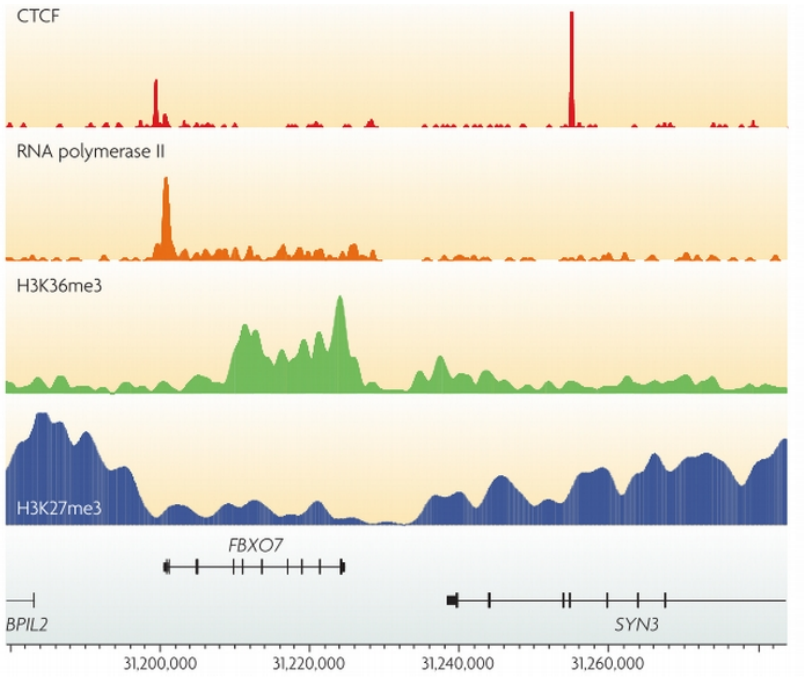


Figure 6.7: Differences of length and loclization of ChIP-seq peaks for TFs, RNAP and histone marks. TF peaks are narrow and can be located near the gene promoters, enhancer or introns. The RNAP peaks are long (through the gene body) with a high peak at the gene promoter. The histone peaks are broad, and are located depending on the modifications (e.g., associated to active genes or associated to promoters). Figure adapted from Park (2009).

complex binding events, for example dimers, tetramers or closely spaced binding events (Mahony and Pugh, 2015) and even the organization of histones (Rhee et al., 2014), i.e., it was shown that the distribution of histone marks within the nucleosome may be assymetrical.

Two of the main differences between ChIP-sea and -exo/-nexus are the following: (i) in ChIP-exo, there is no a background (control), since the free DNA is degraded by the exonuclease, as consequence, the ChIP-seq peak-callers do not fit the ChIP-exo results because the they usually require the background to calculate the enrichment of reads in order to detect the peaks. Therefore, another approach is required to detect the peaks; (ii) Whilst in ChIP-seq the sequences reads corresponds to sites where the protein is not bound, on ChIP-exo, the sequences reads are those where the TF is bound (Hartonen et al., 2016). See (Mahony and Pugh, 2015) for detailed revision of the differences between these methods.

The steps for ChIP-exo are the first 4 steps of ChIP whit additional steps as follows:

5. The immunoprecipitated fragments are treated with an exonuclease that degrades the free DNA.
6. Reversal cross-link of the remaining DNA fragments.
7. Sequencing of the fragments with high-throughput methods.
8. Detection of peaks (peak-calling).
9. the discovery of motifs is done *a posteriori*.

The steps for ChIP-nexus are the first 6 steps of ChIP-exo whit additional steps as follows:

5. The immunoprecipitated fragments are treated with an exonuclease that degrades the free DNA.
6. Reversal cross-link of the remaining DNA fragments.
7. Auto-circularization and amplification of the fragments.

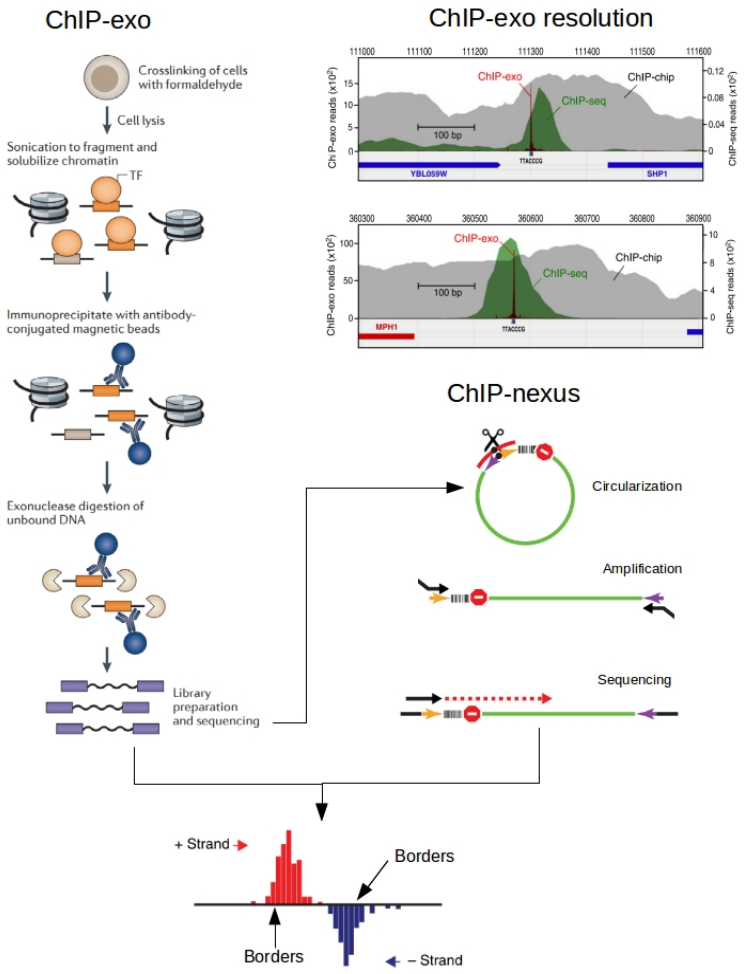


Figure 6.8: ChIP-exo and ChIP-nexus methodologies. Inset: two examples of the ChiP-exo resolution compared with ChIP-seq and ChIP-on-ChIp data for the same genomic coordinates. Figure adapted from Rhee (2011), He (2014) and Zentner (2014).

8. Sequencing of the fragments with high-throughput methods.
9. Detection of peaks (peak-calling).
10. the discovery of motifs is done *a posteriori*.

The additional circularization and amplification of the detected fragments with ChIP-nexus provides a better coverage of the sequences fragments than ChIP-exo (He et al., 2015) (Figure 6.6).

Regarding the peak-calling algorithm, this is not detecting enrichment relative to a control. Taking advantage of the exonuclease footprints, these algorithms search for the 'borders' (i.e., the limits of a bound TF at the positions where the exonuclease stopped) at the sense and antisense strands (Hartonen et al., 2016; Starick et al., 2015; Wang et al., 2014).

Altogether ChIP-exo and nexus allow to detect allele-specific binding events or regulatory variants with a higher specificity than ChIP-seq, because the read can be mapped directly on the TF binding site and this allow to discriminate mutations overlapping the core TF binding motif and those on the flanking positions. In addition, the exonuclease treatment allow to detect peaks that are either bound by the Tf of interest and at the same time bound at open chromatin regions. This is not possible with ChIP-seq alone, it should require for example add additional information of open chromatin (e.g., DNase-seq or ATAC-seq).

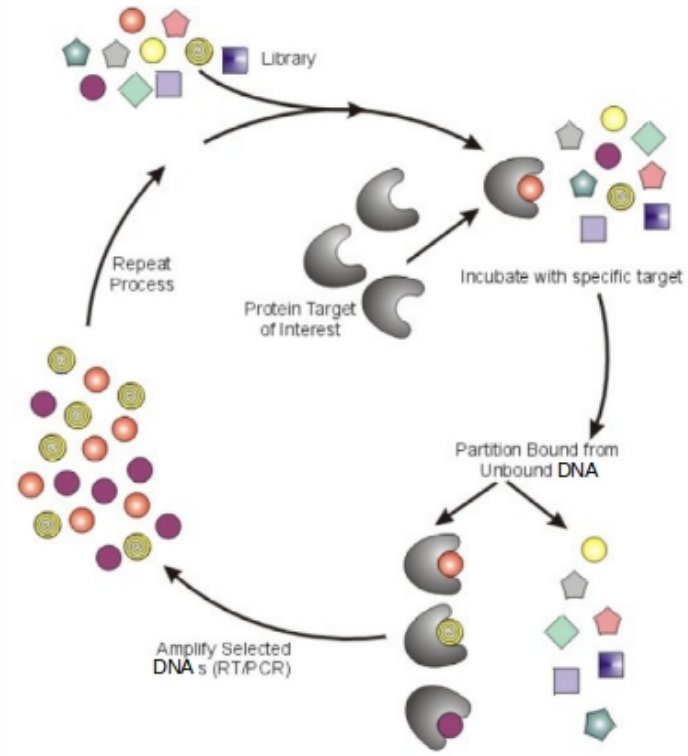


Figure 6.9: Schematic representation of SELEX procedure. Figure adapted from Ray (2010).

ChIP-exo and ChIP-nexus have considerable advantage over ChIP-seq, specially the resolution and the peak size, this advantage can be useful for detection of allele specific binding, but given that these methods are more expensive and the amount of work for process the samples is higher than ChIP-seq, their use is limited. In addition, the additional washes and digestion steps in the ChIP-exo/nexus result in less complex DNA libraries compared with ChIP-seq.

6.5 Systematic evolution of ligands by exponential enrichment (SELEX)

SELEX is an *in vitro* technique that detects oligonucleotides that specifically bind an interest protein or RNA (Tuerk and Gold, 1990).

The steps of SELEX (Figure 6.8) are the following:

1. Synthesis of oligonucleotide libraries. The oligos are randomly generated and must have the same length, in addition they are flanked by specific 5' and 3' ends that will be used as primers.
2. Exposition: The molecule of interest (e.g. a TF) is exposed to the oligos.
3. Wash: The unbound oligos are removed.
4. Amplification. The bound oligos are amplified by PCR.
5. The previous steps are repeated in several rounds. Every round the strongest oligos bound to TFs are conserved.

The main limitation of this method is since the oligomer are randomly generated, they might not be exist

in a given genome and they cannot be mapped. In addition, the iterative selection of strong sites does not take into account the weak binding sites that also have regulatory roles. Beside these limitations, SELEX has becoming popular since it allows to test sequences of any size (PBM are limited to sequences of 10nt) (Jolma et al., 2010).

Initially the SELEX method was used to study a single TF at time, however, Taipale group have recently developed different high-throughput versions of SELEX as follows:

- High-Throughput (HT)-SELEX: the oligos detected by SELEX are sequenced by high-throughput sequencers, this approach allow to analyze hundreds of proteins of interest in a single analysis (Jolma et al., 2010, 2013). This method has been used to determine the TF binding affinities for hundreds of human TFs.
- Consecutive Affinity-Purification (CAP)-SELEX: similar to HT-SELEX, but selects those oligos interacting with two different TFs at the same time (Jolma et al., 2015).
- Methyl-SELEX: similar to HT-SELEX, with two versions of the oligos, one methylated at the Cytosines and the other with no methylation. This method is able to detect TF which binding affinity is affected by the presence of methyl groups at the nucleotides (Yin et al., 2017).

6.6 Detection of open chromatin regions

TFs and other DNA-binding proteins are usually bound at open chromatin regions, the following techniques were designed to detect all the genomic positions where the chromatin is open. For the scope of this manuscript I will describe only two methods: DNase and ATAC-seq, since they allow to study TF binding sites. These techniques are not specific for a given TF (conversely to ChIP-x methods), but they can detect all TF binding events in a single run, see (Meyer and Liu, 2014) for a review of these methods and their limitations.

6.6.1 DNAseI-seq

This method could be considered as the high-throughput and *in vivo* version of the DNAse footprint assays, since it follows the same principle (accesible DNA sites are sensitive to the DNAseI activity) but it is coupled with high-throughput sequencing to detect the *footprints* (Neph et al., 2012; Hesselberth et al., 2009; Boyle et al., 2011), this method is also refered as digital genomic footprinting (Figure 6.9).

This method has been used in eukaryote genomes, and in a single run is able to detect all the unprotected (free) regions that could be protected by a bound DNA-binding protein (e.g., a TF), allowing the creating of DNA accesibility maps for tens of cell lines or genomes for distinct species (Hesselberth et al., 2009). the detected regions can be selected by their size, in order to focus the analysis in the detection of TF binding sites.

Taking advantage of the high-throughput sequencing methods, the detected sensitive regions can be sequenced with a high depth revealing the signatures of the TF binding at nucletide resolution (Neph et al., 2012). Within this signatures, the nucleotide conservation has been studied, revealing that the nucleotides making contact with the TFs are the most conserved in the footprint, and the flaking residues are less conserved. In addition, this analysis have revealed novel motifs for distinct TFs that were not detected before by the conventional methods as ChIP-seq (Figure 6.10).

It has been shown that a large amount of the ChIP-seq peaks for a given TF overlaps with the genomic location of the sensitive DNAseI sites, as consequence, it is highly recommended to have information of the TF binding and chromatin accesibility. This accesibility information could be useful to discard those false positive peaks.

The advantages of this method is that all the TF binding events in a given condition can be detected with a single experiment, revealing novel motifs for certain TFs which suggest unknown regulatory mechanism or

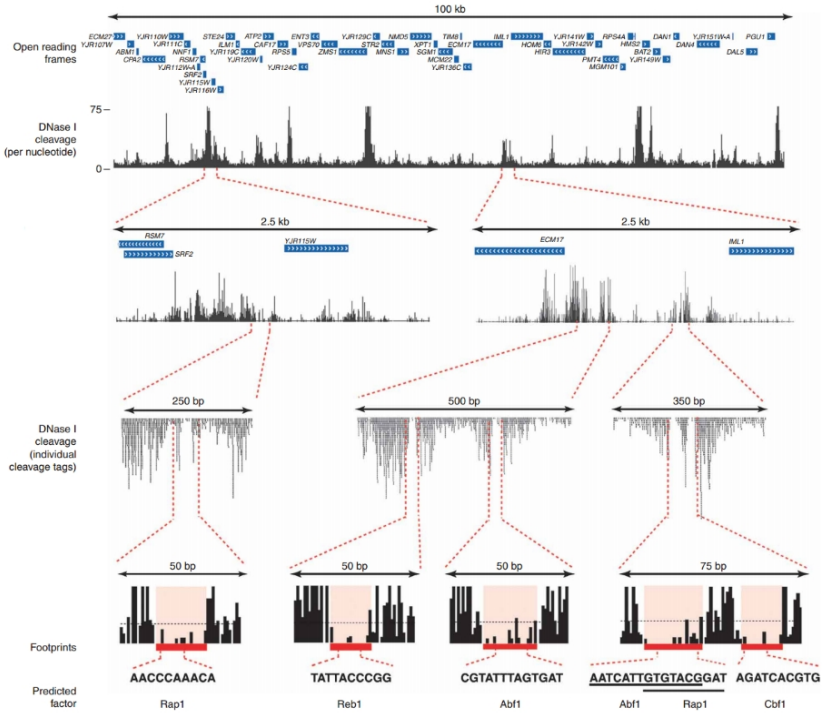


Figure 6.10: Digital DNase I analysis from chromosomal to nucleotide resolution. Figure adapted from Hesselberth (2009).

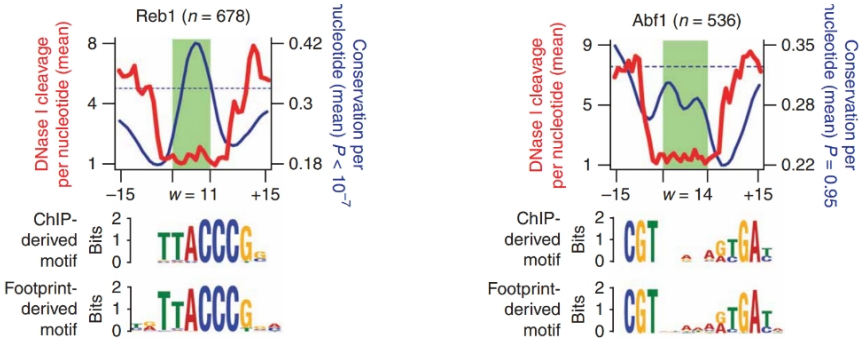


Figure 6.11: Mean cleavage score is anti-correlated with sequence conservation score. Figure adapted from Hesselberth (2009).

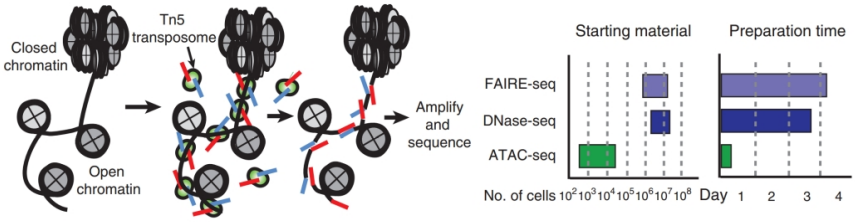


Figure 6.12: Schematic representation of ATAC-seq and comparison with alternate methods to detect chromatin accesibility. Figure adapted from Buenrostro (2013).

unknwon TF-TF interactions (e.g., dimers). Some issues of the method is that its efficiency depends highly in the sequence depth, which in turn depends on the number of cells and could be expensive. In addition, the footprint quality is TF dependent and there is an intrinsic cleavage bias, i.e., some regions are more easily cleaved than others, this bias should be considered to avoid missinterpretation of the found motifs (He et al., 2013).

6.6.2 ATAC-seq

Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq) is a method to detect chromatin accesibility, alternative to the DNAseI-seq method. This method uses a mutated hyperactive version of the transposase Tn5, which allows to insert specific sequences, called adapters, at the open chromatin regions. The adapter-containing regions are isolated, amplified and sequenced (Buenrostro et al., 2013).

Two advantages of ATAC-seq relative to other methods to detect chromatin accesibility (e.g., DNAseI-seq, MNase-seqm FAIRE-seq) are the following: (i) low input material, ATAC-seq can be performed with 50,000 cells, whilst DNAseI-seq requires at least 1 million cells; and (ii) the ATAC-seq protocol can be ran in ~3 hours, the alternate methods require days (Figure 6.11).

Similarly to DNaseI-seq, the results from ATAC-seq allow to detect TF binding sites at high resolution and nucleosome positioning as well (Figure 6.11). ATAC-seq also has .

One of the main limitations of ATAC-seq is the bias inserting the adapters to certain sequences (similar to the enzymatically induced cleavage shown in the DNaseI-seq), although this bias is not fully understood, it is suggested to use naked DNA as control in order to detect the regions with potential bias (Meyer and Liu, 2014). Another issue is related with the input number of cells, the addition of too many cells produce 'under-transposition' therefore producing large fragments, conversely, with too few cells occurs the so-called 'over-transposition' producing short fragments. In addition, to have optimal results, the input number of cells may vary depending on the analyzed genome or the cell line (Buenrostro et al., 2016).

6.7 Other methods

The methods described are (or have been) the most used for detection of TF binding events, but of course there are not all the existing ones. More methods less popular are under development, see (Levo and Segal, 2014) for a revision of novel methods to detect TF-DNA interactions. The modified version of some methods are developed to adapt such methods to deal with genomic limitations of certain genomes (e.g., ATAC-seq for plant genomes (Lu et al., 2017)).

For the moment the ChIP-exo/nexus are the most precise *in vivo* methods to detect TF binding events, but ChIP-seq is the most used method. This is due because ChIP-exo/nexus require greater amount of work compared to ChIP-seq and also because is a growing methodology.

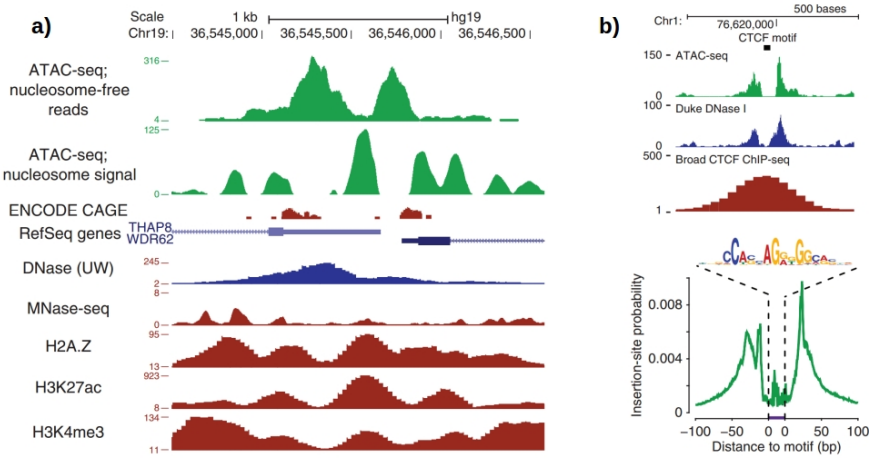


Figure 6.13: ATAC seq can be used to study (a) nucleosome dynamics and (b) TF binding sites with high resolution. Figure adapted from Buenrostro (2013).

Chapter 7

Bioinformatics methods to study transcription factor binding sites

Since the discovery that TFs recognize specific sequences in the DNA, two subsequent questions were the next: (i) how to locate all these sites for a given TF? (ii) Is there a way to model the binding affinities of a given TF? These questions gave rise to a generation of computational methods that allowed the representation, discovery and detection of the TFBSs.

As the TFs recognize short sequences (varying from 5-20bp), that are similar but not identical, there was a necessity to create models to represent a set of sequences bound by the same TF. These models should be good enough to either detect the already known and unknown binding sites. These models representing the sequences recognized by a TF, are known as Transcription Factor Binding Motifs (TFBMs), or simply called *motifs*.

The key problems to face to answer the previous questions are: (i) *de novo* motif discovery; the detection of signal (e.g., short sequences over-represented) that could be associated to a particular TF in a set of related sequences (e.g., set of regulatory regions of differentially repressed genes in a particular condition), and (ii) pattern matching; the detection of instances of the TF motif in a whole genome or in a selected set of sequences.

The first methods to analyze motifs were developed around twenty years ago, at that time the data availability was scarce and distinct methods to detect the TFBSs were developed, however, as the technologies to discover the TFBSs evolved and more data was available, for example using Next-Generation Sequencing (NGS) technologies, many of these computational methods were adapted to handle high-throughput results and others became obsoletes.

The aim of this section is to describe the evolution of the methods specialized in the detection and analysis of TFs, this includes the methods to discover the motifs, detection of individual binding sites and comparison of motifs.

7.1 Representation

The first experimental methods to detect TFBSs were the EMSA and DNAse footprint assay, combining their results it is possible to detect the exact location of TFBSs (or others DNA-binding proteins) on a small sets of sequences. The TFBSs identified could be collected and aligned to further generate a model (TFBM) representing the shared features between these sequences and summarizing an alignment of TFBSs (e.g., the most conserved positions in the alignment) in either a text string or as a numeric matrix (Figure 7.1).

The first and simplest model to locate TFBSs is searching for exact matches of an already known TFBS, an other identical TFBSs could be easily found in a huge sequence, however given that TFs recognize similar (not always identical) sequences, this method is limited by the available repertoire of TFBSs. However, this method can be easily extended to represent the TFBM either as a regular expression or using the International Union of Pure and Applied Chemistry (IUPAC) nomenclature (Comnish-bowden, 1985) (consensus) in order to use a single letter to represent all the nucleotides found at each column of the TFBS alignment (e.g., R and Y represents purines (A or G) and pyrimidines (C or T), respectively) (Figure 7.2). Although these models are simple to understand and can be applied easily to search TFBSs, they are focused in the presence/absence of nucleotides at each column without considering the importance of nucleotide frequencies, therefore many functional motifs could not be detected if the consensus is strict.

In order to consider the nucleotide frequencies at each column of the TFBSs alignment, a more refined model called Position Specific Scoring Matrices (PSSMs) was generated by Stormo (Stormo, 2000; **?**). The PSSMs contain the frequency of each nucleotide at each column (mono-nucleotide model) of the binding site alignment and therefore every possible sequence with the same size of the PSSM (i.e., the TFBS alignment length) can be scored (i.e., weighted).

Since this representation could be hard to interpret by human eyes, Schneider developed a method to visualize the PSSM using their information content (IC), showing the relative importance of each nucleotide at each position of the alignment (Schneider and Stephens, 1990), this visual representation of the PSSM is called a sequence logo (Figure 7.1). The IC is based on the Shannon's uncertainty (Schneider et al., 1986; Shannon, 1948) and reflects the capability of the PSSM to make the distinction between a binding site (represented by the PSSM) and the background model (Aerts et al., 2007) (the prior nucleotide probabilities, for example, the estimated nucleotide frequencies in the upstream region of all human TSSs), see Figure 7.3.

The PSSM model is the most used representation of TFBM. This model however, assumes the independence of each position (i.e., the frequency of a nucleotide on a particular position does not depend on the frequency of the previous nucleotides) that is true for a large set of TFs, although this observations was meda even before the creation of the PSSMs []. With current amount of data it is possible to study nucleotide inter-dependencies for certain TF Families and the mono-nucleotide PSSM model has been extended (e.g., the di-PSSMs proposed by Kulakovskiy (KULAKOVSKIY et al., 2013), Transcription Factor Flexible Models (TFFMs) proposed by Mathelier (Mathelier and Wasserman, 2013) and the extension of N-nucleotides proposed by Siebert (Siebert and Johannes, 2016)), the logo representation has been also modified, although for the moment, there is no a standard format to represent it (Figure 7.4).

The option to use a mono- or di-nucleotide model depends on the TF biology and additional information (e.g., structural) can help to choose the correct model to use. For example the detection of TFBSs for human TFs belonging to E2F, MADS and Zinc-Fingers were improved using di-nucleotide or higher models (Mathelier et al., 2016; Siebert and Johannes, 2016; Jolma et al., 2013), although for a large set of TFs, they can be modelled by simpler (mono-nucleotide) models (Zhao, 2013). An important condition of di-nucleotide models is that they require a large set of TFBSs in order to be trained, for some TFs with a lot of data available this should not be a limitation, but in other cases, when a TF has a low number of TFBSs (e.g., the HipB TF on *E coli K12* with only four known TFBSs) a mono-nucleotide model could be used.

7.2 Pattern Matching

The prediction of TFBSs is the main use of TFBMs, here the key problem relies in identify *bona fide* TFBS in a set of sequences (varying from a bunch of regulatory sequences to a whole genome) and requires prior knowledge of the motif (i.e., the PSSM or consensus string) that describes the binding specificity of a TF.

Depending on the method used, the pattern matching methods are classified as *string-based* and *matrix-based*. In the *string-based* approaches, a single string summarizes a collection of binding sites (Figure 7.1), if this string is represented with the four DNA letters only (Figures 7.1 and 7.2), it is denoted as strict consensus, otherwise, if the IUPAC alphabet is used (15 letters), this representation is known as degenerate consensus. The advantage of use this method is that is easy to program in a computer, as a regular expression to look

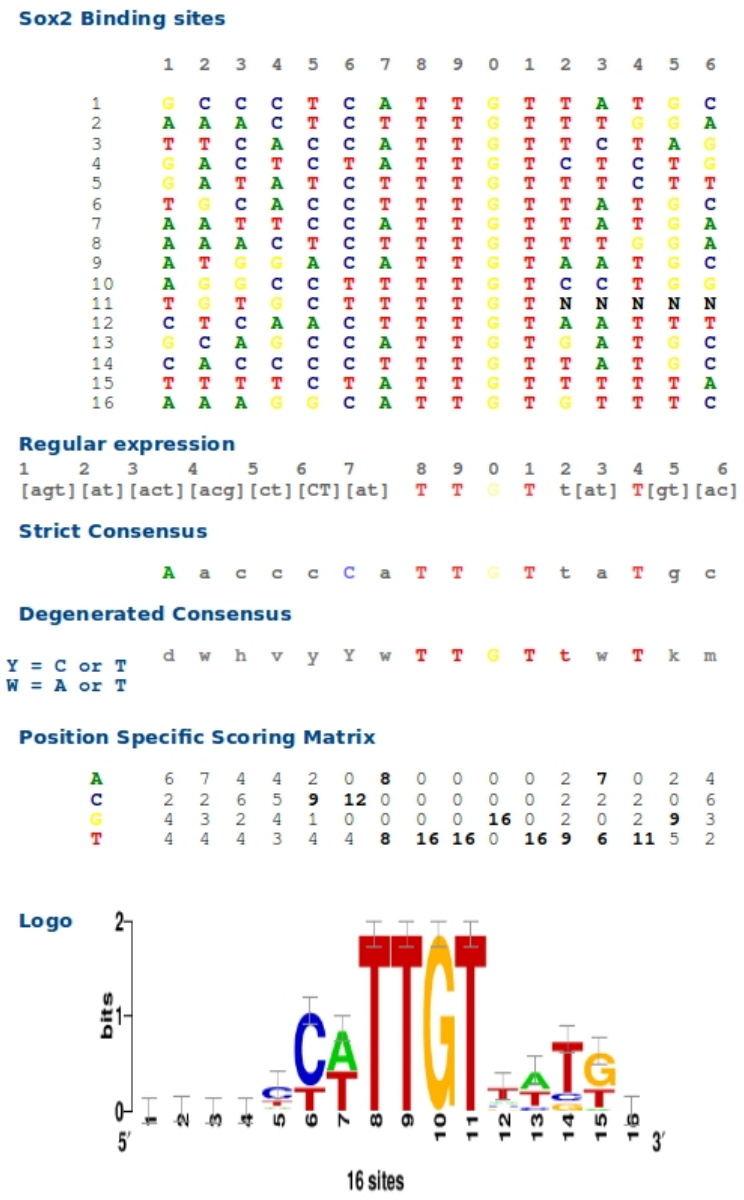


Figure 7.1: A collection of Sox2 binding sites and different motif representations.

IUPAC ambiguous nucleotide code		
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
R	A or G	puRine
Y	C or T	pYrimidine
W	A or T	Weak hydrogen bonding
S	G or C	Strong hydrogen bonding
M	A or C	aMino group at common position
K	G or T	Keto group at common position
H	A, C or T	not G
B	G, C or T	not A
V	G, A, C	not T
D	G, A or T	not C
N	G, A, C or T	aNy

Figure 7.2: IUPAC alphabet used to represent DNA sequences.

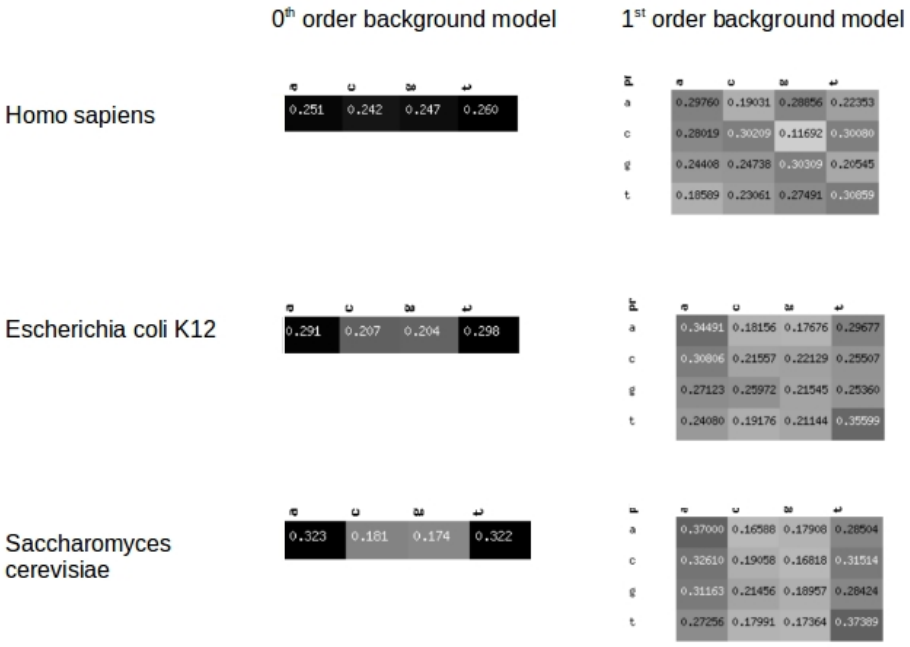


Figure 7.3: The 0-order and 1-order background models of promoters of humans, E coli K12 and budding yeast. Note the low frecuency of CG (1st order) in the human promoters.

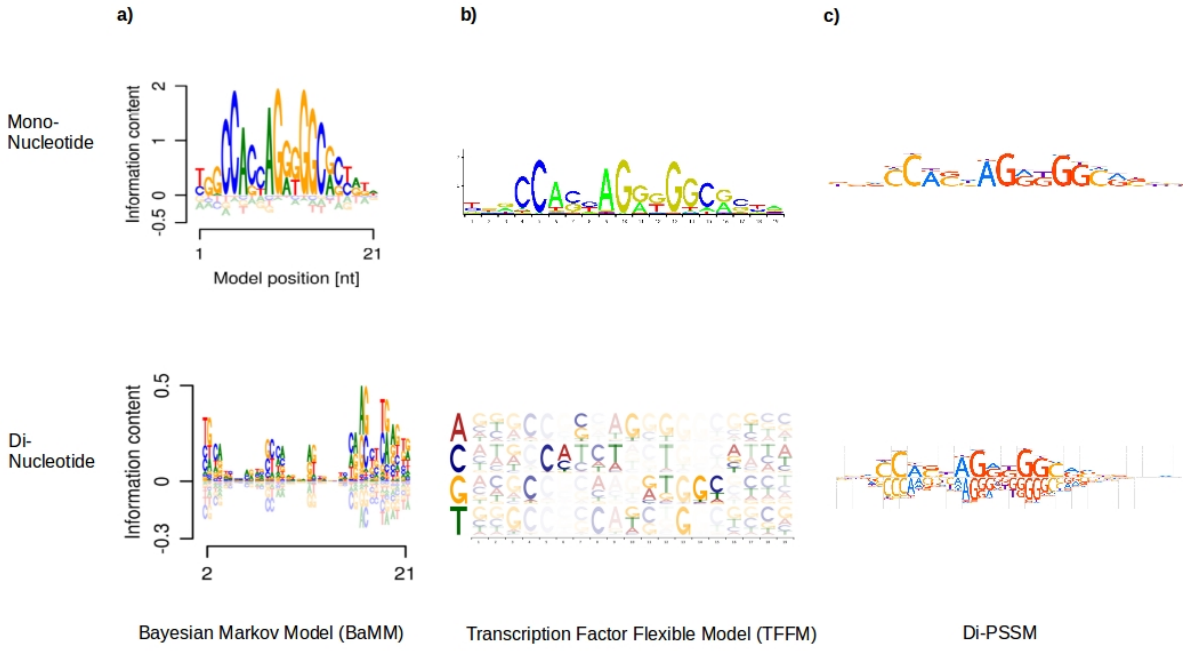


Figure 7.4: Examples of mononucleotides and dinucleotides TFBMs for CTCF represented with three different methods. (a) Bayesian Markov Models. (b) Transcription Factor flexible Models (TFFMs) from JASPAR. (c) diPSSMs from HOCOMOCO.

for strings matching the pattern, and no requires special software. The weaknesses of this approach is that does not consider the frequencies of nucleotides at each position and discovered TFBSs with slight variations that were not considered by the consensus will not be detected.

The *matrix-based* methods rely on a more complex model, the PSSMs (or its extended versions), this model considers the relative frequencies of nucleotide at each position of the motif and considers the nucleotide frequencies of a set of background sequences at the moment to search TFBSs. At each evaluated position, the following probabilities are calculated: (i) the probability of a TFBS (S) according to the nucleotides frequencies stored in the PSSM (M):

$$P\left(S|M\right) = \sum_{i} f'_{i,j} = \frac{n_{i,j} + p^{i}k}{\sum_{r=1}^{A} n_{i,j} + k}$$

and the probability of the same sequence expected from the background model (B):

$$P\left(S|B\right) = \sum_{i} p_i$$

where:

$$A = Alphabet\,size = 4$$

$$n_{i,j} = occurrences\,of\,residue\,i\,at\,position\,j\,of\,the\,matrix$$

$$w = matrix\,width$$

$$p_i = prior\,residue\,probability\,for\,residue\,i$$

$$f_{i,j} = relative\,frequency\,of\,residue\,i\,at\,position\,j\,of\,the\,matrix$$

$$k = pseudo\,weight\,(arbitrary)$$

$$f_{i,j} = corrected\,frequency\,of\,residue\,i\,at\,position\,j$$

$$W_{i,j} = weight\,of\,residue\,i\,at\,position\,j$$

$$I_{i,j} = information\,of\,residue\,i\,at\,position\,j$$

The log-ratio of these two probabilities is also called *weight* score, a positive weight score indicates that a sequence segment is more likely to be an instance of the motif than an instance of the background (the genomic context), and could be considered a TFBS.

$$W_{i,j} = ln\left(\frac{P\left(S|M\right)}{P\left(S|B\right)}\right)$$

The background model is the expected nucleotide frequencies in a particular set of sequences (e.g., all human promoters). A simple model might be assume that the expected nucleotide frequencies are equiprobable (0.25), however, this assumption is rarely true, because the nucleotide frequency on each organism may vary according their living environment (e.g., some organisms living in extreme environments (e.g., hot springs) use to have a high frequency of both cytosine and guanine). The background models can be represented using Markov chains, where the frequency of a nucleotide depends on the M preceding nucleotides (M is the Markov chain order). A Markov models of order m determine the frequencies of words of length k = m + 1. A Markov order of 0 is also known as Bernoulli model (frequency of mono-nucleotides), and it assumes dependency on the nucleotides, however, higher orders do not make this assumption and in addition may reveal interesting properties of the sequences. For example in mammals, the CpG di-nucleotide is usually lower than the frequencies of the others di-nucleotides, this dependency can be observed in the first-order model but not with the zero order (see Figure 7.3 with some examples of background models in distinct organisms).

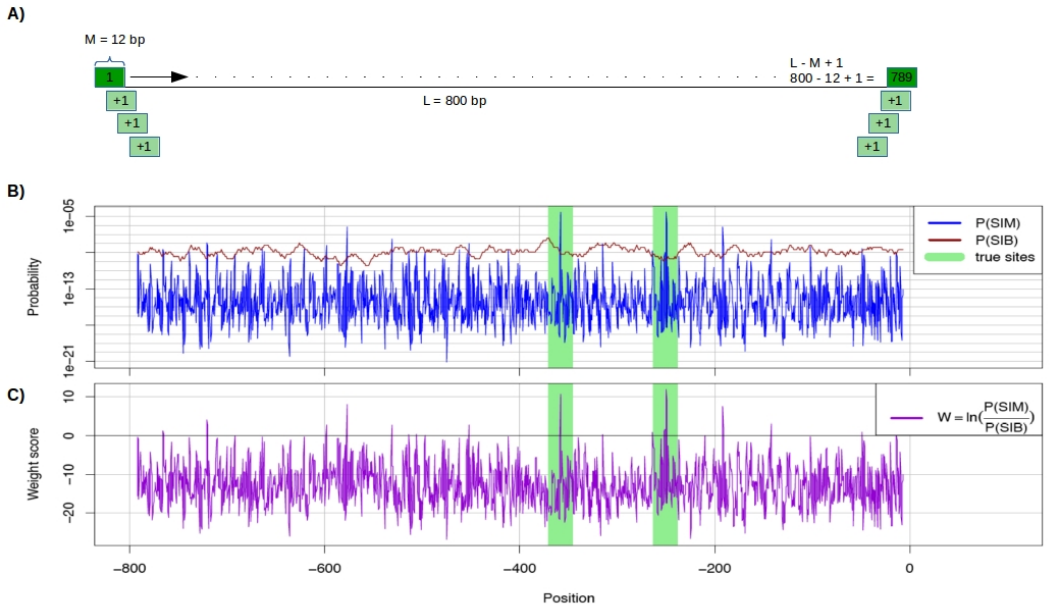


Figure 7.5: Motif scanning and weight calculation. (A) A motif of width W = 12 is scanned in sequence of length L = 800. The motif is shifted one position (+1) each time, therefore a total of 789 scanned positions were analized. (B) At each scanned position, two probabilities are calculated: (i) the probability of the sequence given the PSSM (P|M; red line) and the probability of the sequence given the background (P|B; blue line). (C) The weight is calculated as the log-ratio between the two previously calculated probabilities. The higher the weight, the higher likelihood a sequence is an instance of the motif scanned, see true binding sites highlighted in green have the highest weights. Adapted from Sand (2008)

The choice of an appropriate background model affects the TFBS prediction, this could be observed in the scanning results when sometimes the predicted binding sites are consequence of a bias in the sequence compositions of the analyzed regions. To avoid this bias, at least two tools, HOMER2 (Heinz et al., 2010) and BiasAway (Worsley Hunt et al., 2014) normalize the GC content of the sequences to generate background models.

In order to detect TFBSs, the PSSM is aligned with the sequence of interest at the left, and a weight score is calculated at this position, then the PSSM is moved one position to the right, and a new weight score is calculated. This procedure, known as *motif scanning*, is repeated until the right side of the PSSM is aligned with the right side of the sequence (Figure 7.4). The total number scanned positions T in a sequence of length L with a motif of width W is defined by the next formula:

$$T = L - W + 1$$

Once all the position have been scanned, the next step is to find the true TFBSs. Ideally a weight score greater than zero (positive) should be an indicative of true TFBSs, however in the practice this is not true (see Figure 7.5 where many positions have a positive weight score, but only two of them are known TFBSs). The simplest solution is to set a threshold based on the weight (for example, in the Figure 7.5C a threshold of a weight score >= 10 should be enough to discriminate the true TFBSs from the noise).

Weight-based thresholds are not however the optimal solution: (i) the range of weight scores produced by a PSSM depends on the PSSM width (i.e., higher weights are observed in larger motifs) (Figure 7.5) and (ii) usually a set of PSSMs with distinct widths are used to scan the sequences. Setting a single weight-based threshold (e.g., 5) should be stringent for some PSSMs but loose for others.

The use of p-values is one way to set a threshold based on the risk associated to each predicted TFBS. Here, the p-value of a given weight score is the probability that the background model achieves a weight score

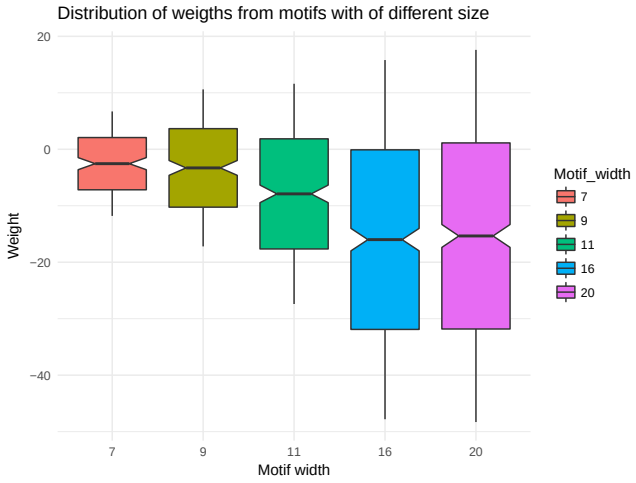


Figure 7.6: The range of weight scores depends on the PSSM width.

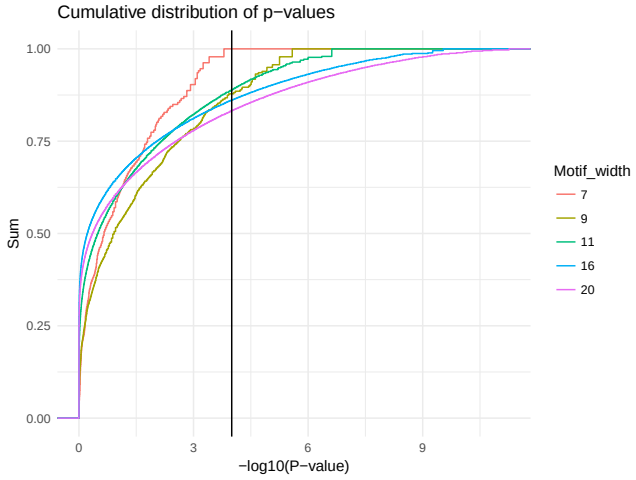


Figure 7.7: Cumulative distribution of p-values from motifs with different lengths. Setting the threshold in a p-value (1e-4, black dotted line) ease the interpretation of TFBS predictions for multiple PSSM in a single analysis.

greater or equal than the observed value (e.g., given the nucleotide frequencies in the PSSM) (Touzet and Varré, 2007). This approach associates the weight scores to p-values assuming a Markov order (e.g., 0, 1, 2) for the background model (Staden, 1989; Turatsinze et al., 2008; Grant et al., 2011). The logic behind this approach is that some weight scores can be generated by several sequence segments, hence they are more frequent than other weight scores. The expected frequencies of each possible weight score are calculated and a distribution of expected weight scores can be obtained. The p-value threshold can be applied in analysis with tens or hundreds of PSSMs, independently of PSSMs' width.

Similarly to the weight-based threshold, the p-value based threshold is also arbitrary, but the user, however, can tune it according the total length of scanned sequences and the accepted level of risk. For example, setting as threshold a p-value 0.0001, one false positive prediction is expected every 10 Kb (Figure 7.6). Ideally, a PSSM can be used to scan a whole genome, but giving the genome size (e.g., 1Gb) and the TFBS size (e.g., 10bp), it is expected by chance that all the 10-mer appears several times in the genome, therefore even setting stringent thresholds will produce a lot of false positives. In addition, the sequence composition through the genome changes, for example, the nucleotide frequencies at promoters are not the same as in the coding or intragenic regions, this justify the selection of an appropriate background model.

In order to reduce the number of false positives, which actually is one of the main issues in the computational prediction of TFBSs, the single-TFBS prediction via scanning, which is the most commonly used approach can be complemented with other sources of information (see (Aerts, 2012) for a detailed revision of methods and tools to detect TFBSs) :

- **Single TFBSs prediction supported by sequence conservation**: the predicted TFBSs are selected whether they are in a conserved region across multiple species (i.e., with a high conservation (phasCons) score) or if the position of the TFBSs is strictly conserved (ortholog position) in two close species (e.g., -40nt relative to TSS).
- **Identification of clusters of TFBSs**: this approach is focused on searching a high density of TFBSs (cluster) from the same (homotypic) or distinct (heterotypic) TFs relative to background sequences. In metazoa, it has been observed that TFs use to be grouped in high concentration in the cis-regulatory sequences (e.g., CRM at enhancers).
- **Identification of clusters of TFBSs supported by sequence conservation**: this approach search for predicted TFBS clusters that are located in conserved regions across multiple genomes. As every region of a genome, the cis-regulatory mutate but they are under pressure to conserve their function, so it is expected to conserve the TFBSs for a group of TFs (e.g., the CRM components) although the TFBSs might not have the same exact location (ortholog position) (Ballester et al., 2014; Schmidt et al., 2010; Villar et al., 2015).
- **Using chromatin activity data to identify TFBSs** : in this approach the idea is to reduce the number of scanned sequences and focus on those sequences with signal of open chromatin as DNAseI hypersensitive sites, ATAC-seq, co-activators (e.g., p300 for enhancers), or a particular histone modification related to gene expression or enhancer activity (e.g., H3K27Ac).
- **Using gene expression data** : the regulatory regions (e.g., promoters) of a set of differentially expressed genes are scanned to search TFBSs on CRM. This method works well with bacteria, yeast and drosophila data, however in other metazoa the cis-regulatory regions goes beyond the promoters (e.g., enhancers and insulators).
- **Using DNA shape information data to identify TFBSs** : recent studies have shown that the observed interdependence between successive nucleotides at the TFBSs are given by physical interactions of the nucleotides (Zhou et al., 2013; Yang et al., 2014; Zhou et al., 2015) and proposed four DNA structural features: minor groove width (MGW), roll, propeller twist and helix twist. The DNA structural features of a collection of known TFBSs (i.e., curated from literature or obtained from motif databases) can be learned and then evaluate whether each individual predicted TFBS (for example, using PSSMs) is shows the *a priori* known structural features (Yang et al., 2014) (Figure 7.7).
- **Using the motif environment to identify TFBSs** : A recent study by Dror and colleagues (Dror et al., 2015) showed that true TFBSs, for different human TF Families, use to be surrounded by sequences with similar GC-content. An approach to measure the motif environment of every TFBS might be developed in order to improve the TFBS (and CRM) predictions.

Although each of these methods contributes in a different way to reduce the number of false positives, a combination of them (e.g., chromatin activity data + DNA shape information + PSSM scan) might result in an additive contribution to find true TFBSs (Mathelier et al., 2016; Lu et al., 2017). The methods to detect TFBSs and some guidelines of how to improve the predictions are reviewed on the next references (Aerts, 2012; Jayaram et al., 2016; Boeva, 2016).

7.3 Motif Discovery

If we known a set of genes that are co-expressed, we might expect that some of them share a common regulator (e.g., the same TF bound on their promoters). If this regulator is a TF and there is a PSSM available for this TF, a simple scanning would reveal the presence or absence of TFBSs; but what happens

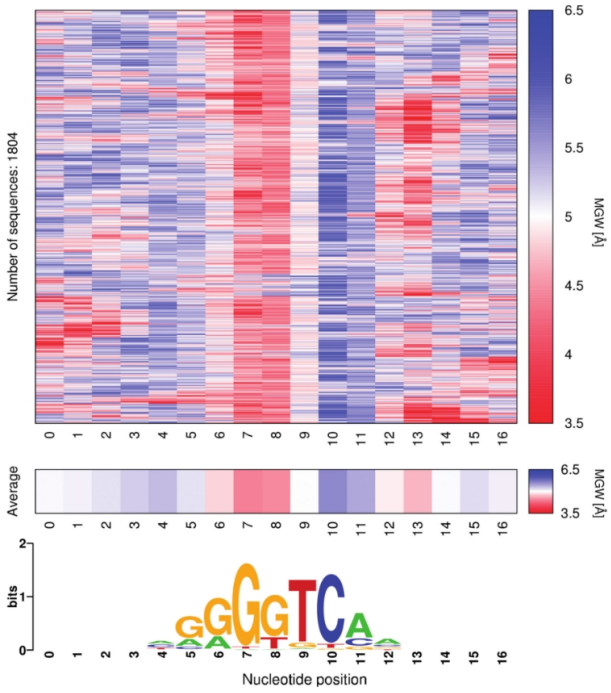


Figure 7.8: DNA structural features in TFBSs. Predicted MGW profiles for a large collection of curated Hnf4a TFBSs (top) can be represented as a heatmap (middle) showing the narrowness (red) or widerness (blue) of minor groove at the specific positions of the binding sites. (Bottom) PSSM generated with the curated sequences and aligned with the structural features heatmap. Adapted from Yang (2013)

when there is no such PSSM ? In these cases, when we guess that a set of sequences might share a signal for a given TF or other DNA-binding molecules, we can use an approach known as *motif-discovery*.

The logic behind motif discovery methods is to discover short sequences (also known as oligomers) that are over-represented in a set of sequences that could be obtained from different methods (**?**): by PBM, by ChIP-seq, by ChIP-exo, by SELEX, or by a set of co-regulated promoters. The over-represented oligomers can be further represented as a PSSM.

The development of novel PSSM representation (e.g., di-PSSM or TFFM) is related with the development of the motif-discovery methods. Initially, given the low-throughput of the experimental methods to detect TFBSs (e.g., 10 sequences obtained from EMSA assays) the first motif-discovery methods were developed to work with a small set of short sequences and therefore the PSSM did not consider the nucleotide inter-dependencies. Nowadays, with the advent of high-throughput methods, the pattern-discovery methods were adapted to deal with a big number of large sequences (e.g., ChIP-seq peaks), reviewed in (Boeva, 2016). In general, pattern-discovery methods are divided in two categories: string-based and matrix-based methods (see (Ma et al., 2012), (Tompa et al., 2005), (Tran and Huang, 2014), and (Weirauch et al., 2013) for an evaluation of distinct motif-discovery algorithms).

String-based methods are simple and rely in the count of oligomers of an arbitrary size k (k-mers) (van Helden et al., 1998; Bailey, 2011). In others words, with k = 7 these methods will search for all the possible 7-mers. Once the k-mers have been counted, their expected frequencies are estimated given a background model (Figure 7.9). The choice of background model is crucial, since the most observed k-mers may reflect the nucleotide composition of the genome. A p-value is calculated for every k-mer and the most significant ones can be further assembled in order to produce a PSSM. Some examples of programs using these approaches are RSAT *oligo-analysis* (van Helden et al., 1998) and DREME (Bailey, 2011).

Some advantages of pattern-based methods are: fast and simple, they use to run fast and their complexity

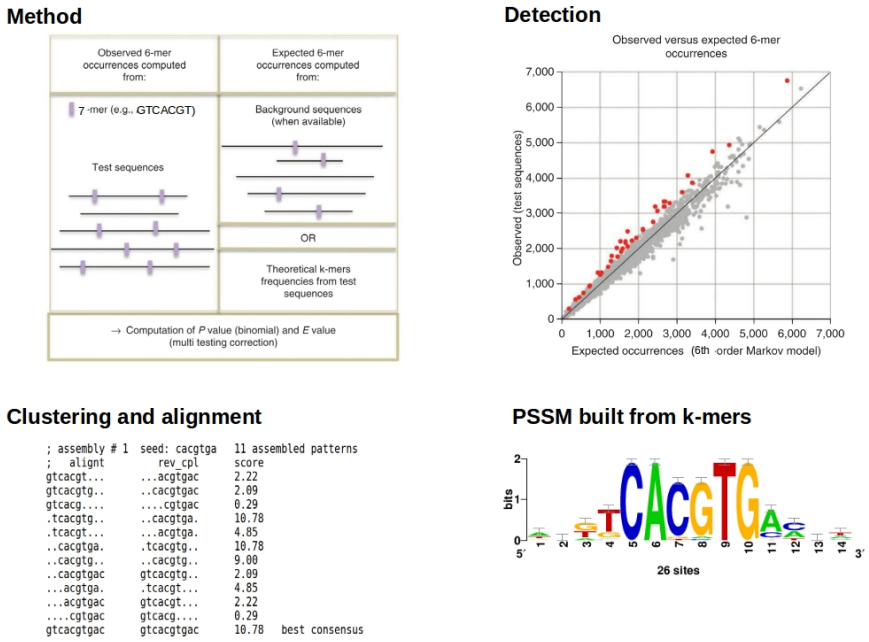


Figure 7.9: Detection of over-represented k-mers. A p-value is calculated for every k-mer comparing the observed (tested) vs the expected (background) frequencies. the over-represented k-mers are clustered and aligned in order to build a PSSM. Adapted from Thomas-Chollier (2012).

is linear (the running time is proportional to the input data size); exhaustive in the way that all possible k-mers are evaluated; detection of under-represented k-mers (e.g., enzyme restriction sites); possibility to report negative results (when no significant k-mer was found); and assembly of several significant k-mer produce PSSMs larger than the k-mer length. By contrast one weakness of these methods are: k-mers does not consider the nucleotide degeneracy (variable nucleotide in a specific position of the TFBS), but *a posteriori* analysis can be done to study degeneracy.

Matrix-based methods are probabilistic models that maximize the IC or log-likelihood ratio (LLR) of the resulting PSSM. The logic behind these methods is that if the selected sites to built the PSSM are randomly chosen, the PSSM produced should not be informative (e.g., a low IC), but if the PSSM is built using the 'correct' sites, we expect to observe a high IC. The optimal solution should test all possible PSSMs that can built from the sequences by aligning x sequence fragments of length w. In this exhaustive solution, however, the possible number of PSSMs is untraceable, so heuristic techniques (starting by a random seed pattern) or the expectation maximization algorithm (**?**) allowing the rapid exploration of the sequences in order to find PSSMs that can be further iteratively refined until reach a maximum of IC or LLR (Figure 7.10). Some examples of programs using these approaches are XXmotifs (Luehr et al., 2012), ChIPmunk (Kulakovskiy et al., 2010), Dimont (Grau et al., 2013) and MEME (Bailey and Elkan, 1994) based on expectation maximization and currently is the most popular motif discovery algorithm.

Some advantages of matrix-based methods are: better description of motif degeneracy than string-based pattern methods and optimization of the IC or LLR, which is generally a relevant criterion for estimating the relevance of a PSSM. By contrast, some weakens of these methods are: non-linear running time (e.g., MEME has a quadratic complexity); stochasticity makes that not always return exactly the same PSSMs; at least one motif should be returned, i.e., in cases of sequences with no significant signature a noisy motif could be found (e.g., low complexity motifs); and possibility to be trapped at a local maximum.

Additional information can be integrated in some motif discovery programs in order to improve the motif discovery results. This information can be related to the technique used to detect the TFBS or the TF

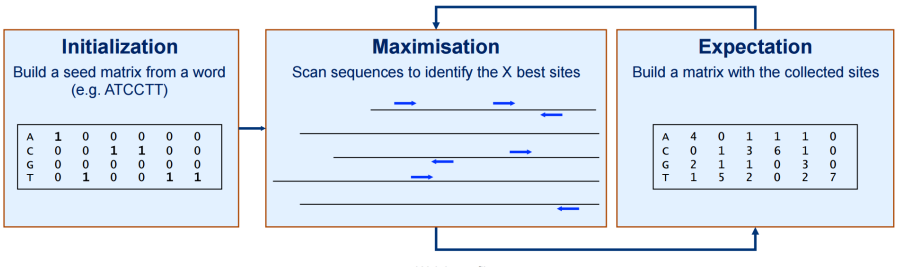


Figure 7.10: Workflow of a motif-discovery approach based on Expectation Maximization (MEME). MEME receives a set of sequences, the width of the motifs to be discovered and the expected number of motif ocurrences per sequence (0 or more, 1 or more, many). A seed is randomly generated and their sites are identified, with these new sites a novel matrix is built trying to maximize likelihood of the nucleotides described by the matrix. The process is repeated until there is a little o no improvement in the likelihood.

biology.

- Positional prior information: the regions whit a high density of reads where TFs use to bind (bigwig files) detected with a high-throughput experiment, for example ChIP-seq, can be used in order to focus the search of motifs in a particular region (e.g., at the center or summit of ChIP-seq peaks), this information improves the quality of the detected motifs relative to the motifs detected without the priors (Ma et al., 2012; Kulakovskiy et al., 2010; Bailey et al., 2010).
- Spaced motifs: the members of some TF families bind DNA as homo-dimers, where each dimer recognize a short sequence and there is a spacer between the monomers (dyads), for example, the Helix-turn-helix family in bacteria or some zinc fingers in mammals. Given that the nucleotides at the spacer are not always conserved in the TFBS, the motif discovery of these motifs is a challenge and for these reason at least two motif-discovery programs are dedicated exclusively to found spaced motifs: RSAT *dyad-analysis* (van Helden et al., 2000) and *GLAM2* (Frith et al., 2008).
- Positionally constrained motifs: the binding of some TFs can be constrained at certain regions (e.g., at the center of ChIP-seq peaks (Thomas-Chollier et al., 2012b), upstream the TSSs, downstream the Transcription Terminal Sites (TTS) (Helden, 2000), or relative to replication origins (Cayrou et al., 2015)). In these cases, the use of positional priors improves the motif discovery, but in some cases the prior information might be not available.

Two pattern-based methods that detect positionally constrained motifs without requiring positional prior information are RSAT *position-analysis* (Helden, 2000) and *local-words-analysis*. The *position-analysis* algorithm aligns the sequences relative to one reference position (start, center or end), divide these sequences in bins (non-overlapping windows) of a given size, counts the k-mers, and assumes that the distribution of every k-mer is homogeneously through the bins. Those k-mers which distribution deviates from the homogeneous distribution will have a high significance, that is calculated using a chi-squared test (Figure 7.10). The detected k-mers are used as seed to further built PSSMs.

Some advantages of this method are: (i) a predefined background model is not required since the total count of every k-mer is distributed homogeneously across the bins and this count becomes the expected k-mer frequencies and (ii) the input sequences can have variable length. The weakens is that the significance of the k-mers is affected by the bin size and this method is sensitive to small number of sequences.

More applications of the motif-discovery are the so called phylogenetic footprinting and the differential motif-discovery.

- Phylogenetic footprinting: the use of ortholog sequences in order to discover PSSMs (Wang and Stormo, 2003). It is useful for certain organisms, for example Bacterias, in cases when a TFs has a low number of reported TFBSs (either the TF has a low number of targets or the TFBSs are unknown). In this cases, the information is not enough to produce a PSSM with good quality, however, the information

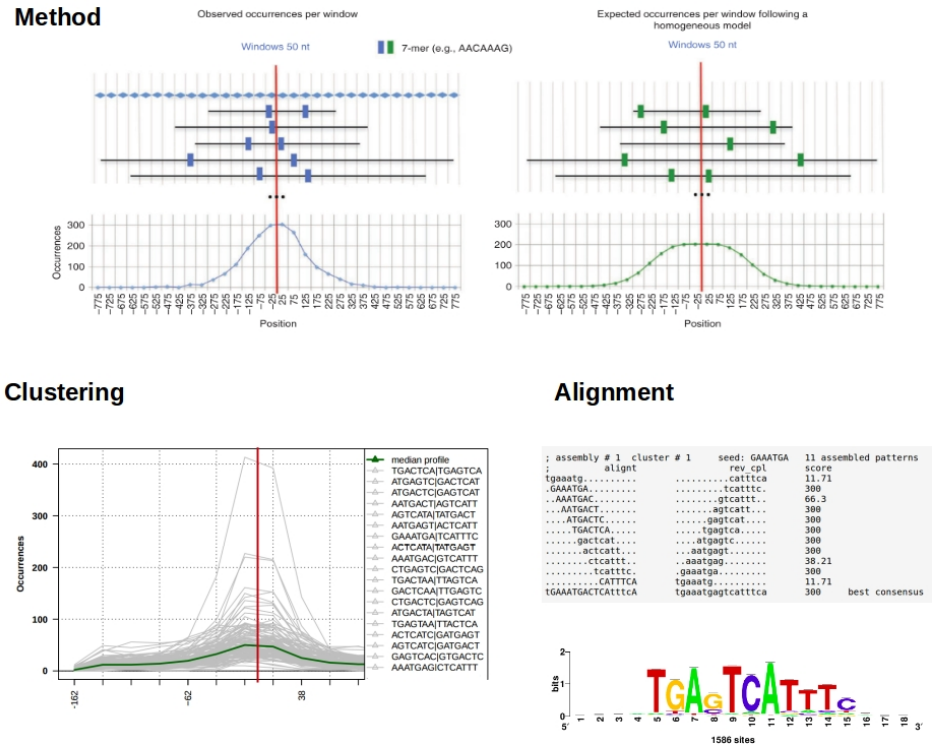


Figure 7.11: Discovery of positionally constrained motifs with RSAT position-analysis. Method: the sequences are aligned relative to their center and are divided in bins of the same width. Every k-mer (blue box) occurence is counted in the sequences and the total count is distributed homogeneously (green box) through the same sequences. Given that the sequences have variable size, the count decreases relative to the center of the sequences. Clustering: The k-mers following the same distribution are grouped. Alignment: the clustered k-mers are aligned to further build a PSSM. Adapted from Thomas-Chollier (2012).

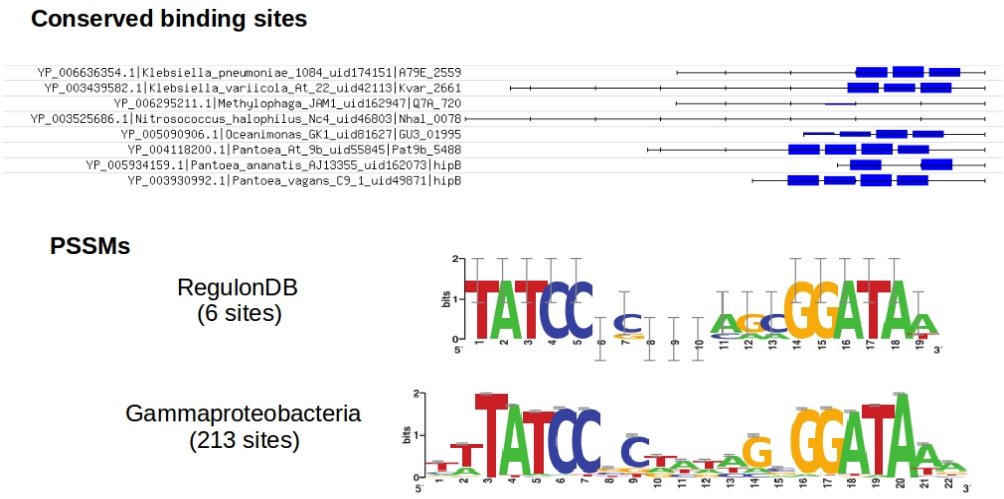


Figure 7.12: A set of orthologous regulatory sequences are collected and a motif-discovery program is applied in order to find PSSMs with the information of multiple genomes. Comparison between a PSSM obtained from RegulonDB and the PSSM obtained from the phylogenetic footprint. Note the difference in the error bars between them. Adapted from Medina-Rivera (2010).

of other close related organism (e.g., all the Enterobacteriales or Gammaproteobacteria) can be used to have a large number of sequences and therefore to discover (multigenome) PSSMs, that in many cases are better predictors of TFBSs than those PSSM generated using a single organism information (Figure 7.11). Some programs as RSAT *footprint-discovery* can do automatically this type of analysis (Janky and van Helden, 2008). In addition, this approach can be used to infer gene regulatory networks (Brohée et al., 2011).

Once again, the background model choice is crucial, and given that phylogenetic footprinting uses sequences from many genomes (that might have different nucleotide frequencies), a solution is to create a taxon-wise background model from all the regulatory sequences of all the organisms belonging to a given taxon. It is important to note that given the genomic architecture of bacteria and fungi, where the promoters are the main cis-regulatory regions and there are few cases of distal regulation, the phylogenetic footprint is easy to apply on these genomes. In more complex genomes, for example vertebrates, the collection of conserved sequences can be done using conservation scores to infer conserved sequences (Schmidt et al., 2010).

- Differential motif-discovery: this approach takes two sets of sequences (query and control) and tries to identify the regulatory elements that are specifically enriched in on set relative to the other. This method is extremely useful to avoid motifs over-represented given the genome composition and also avoids the so called low-complexity motifs (artifact motifs usually with repetitive elements). For example, using a single sequence set for the motif-discovery algorithms might find ubiquitous regulators, using two sequence set, however, may reveal the specific regulators of the query sequences (Thomas-Chollier et al., 2012b).

7.4 Motif comparison

After finding a motif (PSSM), we want to know whether the found motif resembles to some previously identified PSSM. This task is common in the studies including motif analysis, where we want to know which is the most similar motif (the 'best hit') from a collection of motifs provided by the user or public motif databases as JASPAR (Mathelier et al., 2015) or HOCOMOCO (Kulakovskiy et al., 2016) (Figure 7.13). The best hit motif could be considered as a candidate TF for the unknown motifs and is necessary for the further motif annotation. This area of the motif analysis is called motif comparison.

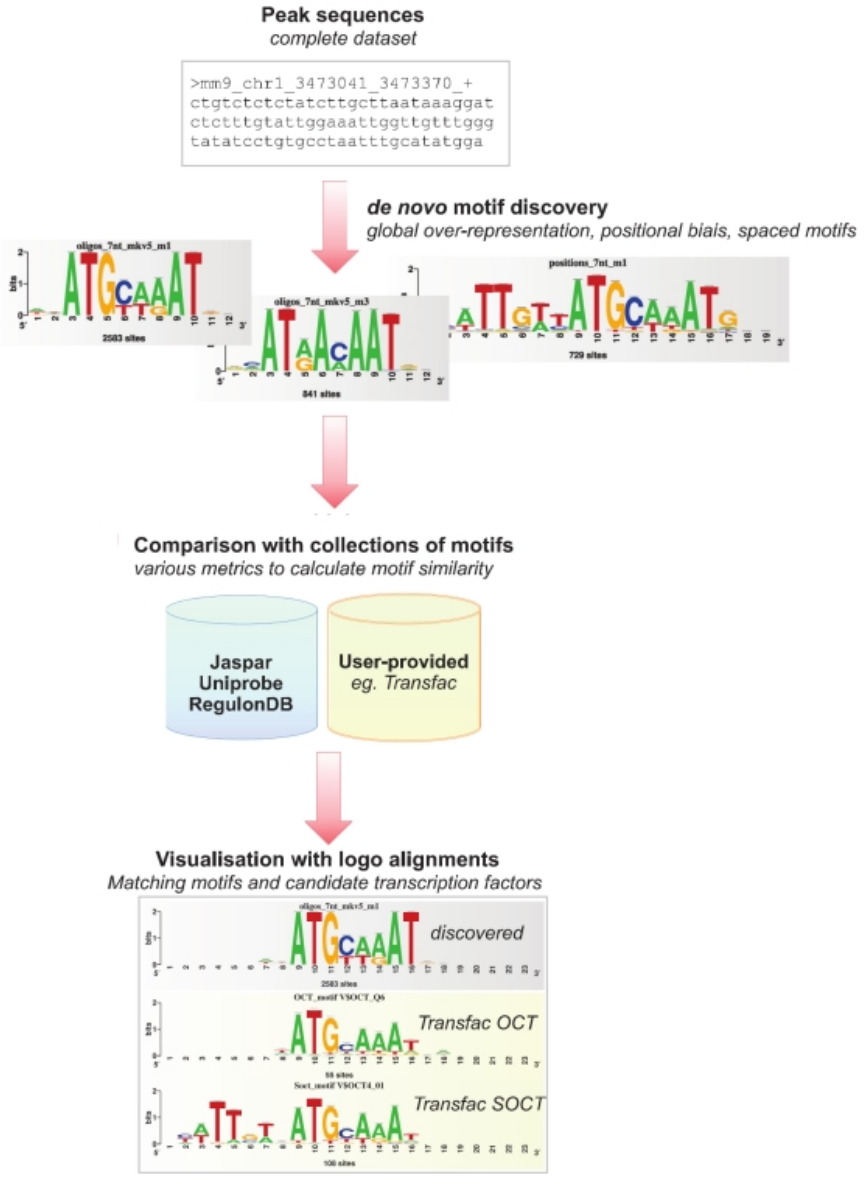


Figure 7.13: The motifs discovered using different algorithms are compared with a collection of motifs provided by the user or taken from motif databases. The unknown motifs are compared and aligned in order to highlight the similar positions and identify the candidates TFs for the discovered motifs. Adapted from Thomas-Chollier (2012).

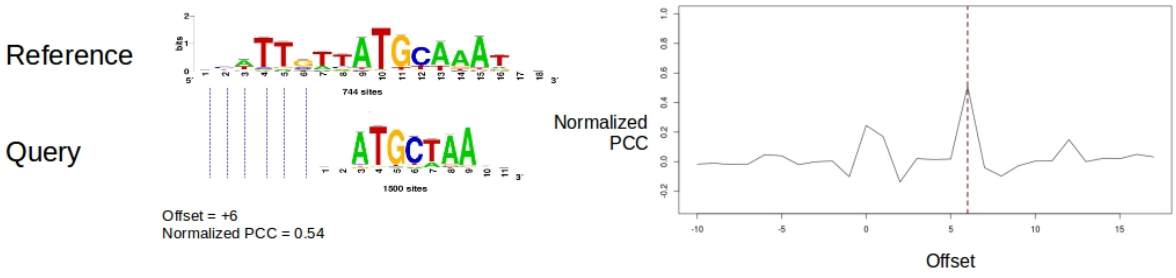


Figure 7.14: Comparison of two PSSMs. (Right) Example of PSSM alignment, note the differences in number of sites, motif length and IC content. The best alignment corresponds to an offset of +4 positions, with a high Pearson Correlation Coeficient (PCC). (Left) Each offset produces a PCC score, the offset with the highest score is used to align the motifs.

In most of the cases, the motif comparison algorithms assume that the nucleotides at each column of the TF binding sites alignment are independent and the simplest way to compare two motifs is through their columns (column-based comparison) until find a position where the columns are the most similar possible. The column comparison, in addition, allows to align the PSSM in order to visualize the common positions between the PSSMS. This type of comparison must take into account the next considerations:

- Evaluate all the possible offsets: an score is produced at every possible alignment between the two compared motifs (even when they are aligned at a single column).
- The forward and reverse complement of the motifs are compared: the PSSMs represent only one strand of DNA, however, all the possible offset in the reverse complement must be evaluated as well.
- Motifs with different length: since every possible offset is evaluated, the comparison is not limited to motifs with the same length. Sometimes two motif discovery algorithms could find the same motif, but the flanks in one motif could be larger than the flanks of the other motif, or a short motif (e.g., monomer) could be part of a larger one (e.g., hetero-dimer). Some motif comparison metrics, however, consider the fraction of aligned columns between two motifs, this information can be useful to match *de novo* PSSMs with known PSSMs with a similar length and avoid alignments between short and large motifs.
- Motifs with different number of sites: PSSM may be built from a small (tens) or large (thousands) sets of TFBSs. This difference is not a limitation in the motif comparison since the proportion of every nucleotides considered. However, in the PSSMs built with a small number of TFBSs, the fraction of nucleotides at each position could be over or underestimated.
- Motifs with different IC: usually the flanks of the PSSM use to have low IC whilst the central positions (e.g., the core motif) have high IC. The IC can be used as information to compare and align the motif correctly.

At every offset in both orientations (forward or reverse), the columns are compared and a score is produced, the offset that maximizes the score is used to align the motifs relative to the most similar positions (Figure 7.14).

In order to compare the PSSM columns, many metrics have been proposed so far (Table 7.1), for example some groups suggest to use metrics commonly used on bioinformatics as the Pearson Correlation Coefficient (cor) (Hughes et al., 2000), the Euclidean distance (ED) (Choi et al., 2004), the chi-squared (Schones et al., 2005), the Jaccard index (Vorontsov et al., 2013), and the Kullback-Leiber divergence (Aerts et al., 2003), whilst other metrics have been developed specifically for the motif comparison, for example the Average Log-Likelihood Ratio (ALLR) (Wang and Stormo, 2003), the Sandelin-Wasserman score (Sandelin and Wasserman, 2004), the Similarity with Information Content (SPIC) (Zhang et al., 2013), the Information Coverage (Stegmaier et al., 2013), the Bayesian Likelihood 2-Component (BLiC) score (Habib et al., 2008), the Asymptotic Covariance (Mosta) (Pape et al., 2008), and the k-mer frequency vectors (KFV) (Xu and Su,

2010). The last to methods do not compare the columns of the PSSMs, and they could be used to compare the novel PSSM models (e.g., di-nucleotide PSSMs).

Table 7.1: formulae for motif comparison metrics.

Metric	Formula
Euclidian Distance (ED)	$d_{Eucl} = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{w}(n_A(i,j) - n_B(i,j))^2}$
Pearson Correlation (cor)	$cor_{A,B} = \frac{\frac{1}{r \cdot w}\sum_{i=1}^{r}\sum_{j=1}^{w}(n_A(i,j) - \bar{n}_A) \cdot (n_B(i,j) - \bar{n}_B)}{\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{w}(n_A(i,j) - \bar{n}_A)^2 \cdot \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{w}(n_B(i,j) - \bar{n}_B)^2}$
Chi-squared	$\chi^2(A,B) = \sum_{K=A,B}\sum_{b=A}^{T}\frac{(n_k(b) - n_k^e)^2}{n_k^e(b)}$
Jaccard Index	$J(X,Y) = \frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Kullback-Leiber divergence	$KL(X,Y) = 10 - \frac{\sum_{b=A}^{T} f_x(b) \cdot log\frac{f_x(b)}{f_y(b)} + \sum_{b=A}^{T} f_y(b) \cdot log\frac{f_y(b)}{f_x(b)}}{2}$
Average Log-Likelihood Ratio (ALLR)	$ALLR(X,Y) = \frac{\sum_{b+A}^{T} n_x(b) \cdot log\frac{f_y(b)}{p_{ref}(b)} + \sum_{b+A}^{T} n_y(b) \cdot log\frac{f_x(b)}{p_{ref}(b)}}{\sum_{b=A}^{T}(n_x(b) + n_y(b))}$
Sandelin-Wasserman score	$SW = 2 - \sum_{b \in \{A,C,G,T\}}(X_b - Y_b)^2$
Similarity with Information Content (SPIC)	$Sim(M_1(X), M_2(Y)) = min\left\{1, \frac{max\{S(P_1(X), F_2Y), S(P_2(Y), F_1X)\}}{max\{S(P_1(X), F_1X), S(P_2(Y), F_2Y)\}}\right\}$
Information Coverage	$ICov = \frac{\sum_{i=s_x..sx+w-1} \cdot I(p_x^i)}{\sum_{k=1..L_x \cdot I(p_x^k)}}$

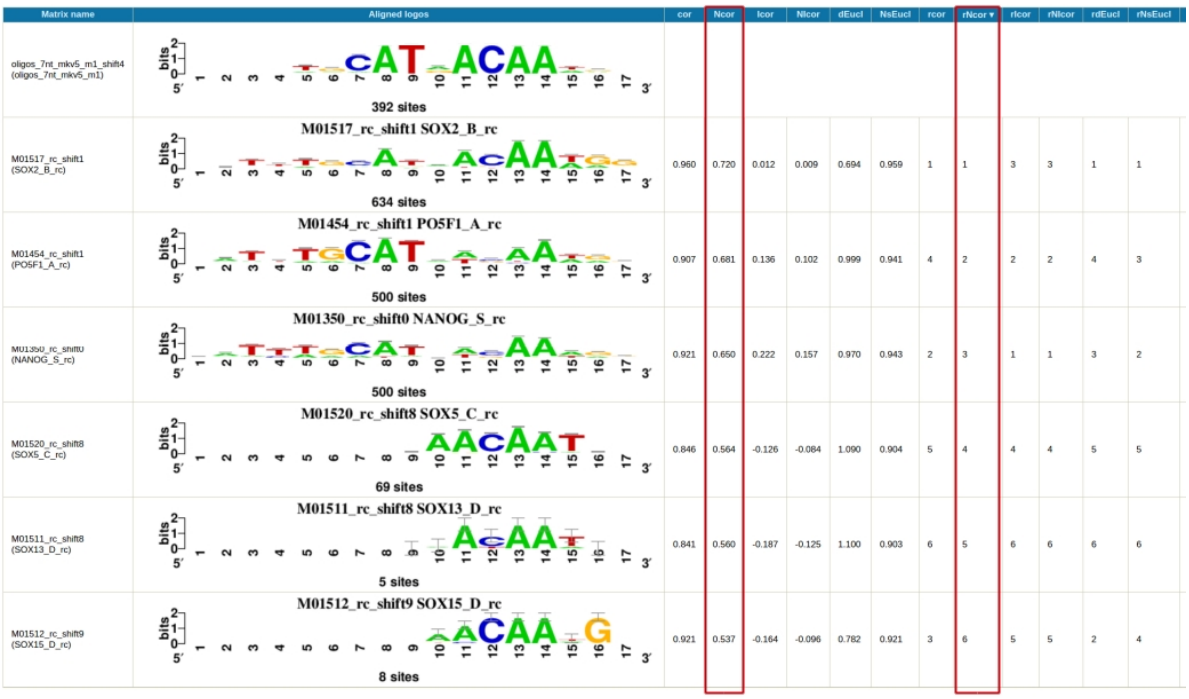


Figure 7.15: Example of motif comparison computing several metrics. A motif discovered was compared with the HOCOMOCO human motifs using RSAT compare-matrices. The results are ranked by width-normalized pearson correlation (Ncor), red rectanlges. Note that the best score of Ncor does not correspond to the best score of the others metrics (e.g., SW). cor: pearson correlation; Icor: information content correlation; dEucl: euclidian distance; SW: Sandelin-Wasserman score; and the width-normalized version (Ncor, NIcor). The last columns of the table (starting with r, for example rNcor) show the rank of the result for the given metric from highest to lowest.

Metric	Formula
Bayesian Likelihood 2-Component (BLiC)	$BLIC = log\frac{P\left(n^1, n^2 \vert \hat{p}^{1,2}\right)}{P(n^1 \vert \hat{p}^1) \cdot P(n^2 \vert \hat{p}^2)} + log\frac{P\left(n^1, n^2 \vert \hat{p}^{1,2}\right)}{P(n^1, n^2 \vert p^{bg})}$
Asymptotic Covariance (Mosta)	$AC(A,B) = \lim_{m \to \infty} m^{-1} cov(N_A(m) + N_{A'}(m), N_B(m) + N_{B'}(m))$

The methods to compare motifs have evolved rapidly, however, currently there is no a standard metric to measure the similarity, and there is a debate of which is the best metric (Habib et al., 2008; Tanaka et al., 2011), since every metric presents its own drawbacks and these metrics are not directly comparable, some of them measure correlations (ranging from -1 to +1) and other distances (e.g., ED goes from 0 to values >= 1). One solution has been to implement software supporting many of these metrics, for example STAMP (Mahony et al., 2007), TomTom (Gupta et al., 2007), m2match (Stegmaier et al., 2013) and RSAT *compare-matrices* (Thomas-Chollier et al., 2011a), using rank-statistics combining the results of several metrics computed in a run (*compare-matrices*) or calculating a p-value (TomTom) (Figure 7.15).

One drawback of the column-based comparison methods is that some non-informative columns could be highly correlated, for example, the flanks with low IC, therefore producing the so-called spurious alignments.

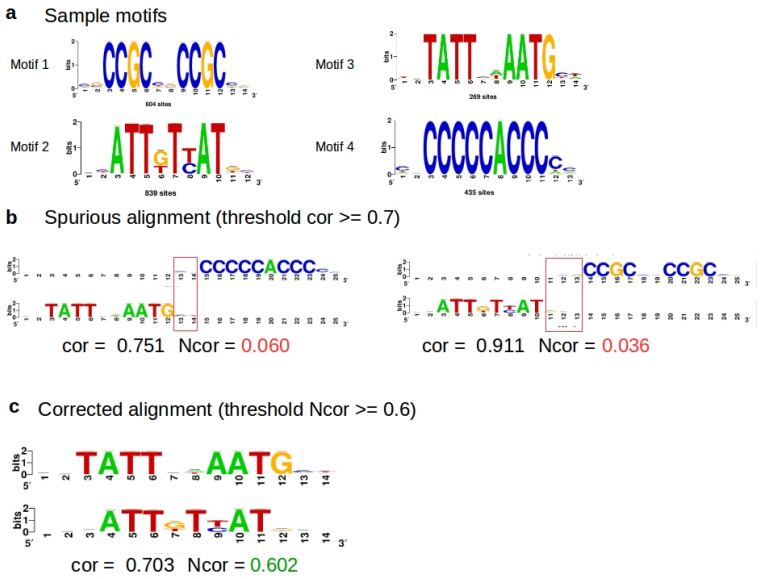


Figure 7.16: Spurious alignments and how to avoid them. (A) Four sample motifs to be aligned. (B) Setting a stringent threshold based on a single metric (cor: Pearson Correlation), the motifs could be wrongly aligned relative to non-informative positions (red rectangle). (C) Using as threshold a metric that considers the fraction of aligned columns (Ncor: Width-normalized Pearson Correlation) and the cor the spurious alignments are not observed anymore.

A solution to these drawbacks is consider the fraction of the columns that are aligned, thus a low fraction might indicate a spurious-alignment and a high fraction might indicate a good alignment, some programs as TomTom, *compare-matrices*, and m2match have modified some metrics (e.g., ED, cor) to consider the fraction of the alignment, avoiding the output of spurious alignments. The results of the comparison could be normalized by the fraction of aligned columns. Another solution is to calculate several comparison metrics in a single row, and set a threshold on two or more metrics (Figure 7.15).

Regarding the alignment of the motifs, is different from the alignment of DNA sequences, because it is focused on short sequences (~20 nucleotides) and because most algorithms for PSSM comparison/alignment do not consider internal gaps (except STAMP (Mahony et al., 2007), that allows gapped or ungapped motif alignments). The biological reason for most programs to do not allow internal gaps is because the TF recognize particular short sequences, a internal gap should change the TF specificity, however this could be useful to visualize regulatory variants as insertions or deletions (e.g., mutations affecting the TF binding) (Shi et al., 2016).

Other methods to compare motifs are not based on the columns. In one of them, Mosta, the motif comparison is based on the putative TFBSs that each PSSM is able to detect. If a pair of PSSM detect the same TFBSs, they should be similar (Pape et al., 2008), this approach, however, does not allow the alignment of motifs.

7.5 Motif clustering

A common conclusion of the studies related to motif comparison, independently of the metric used, is the usefulness of the motif comparison results to cluster similar motifs. This task has become more used since the current amount of PSSM is growing exponentially with the results obtained with high-throughput experiments.

The clustering of motifs can be used to identify TFs belonging from the same family (Mahony and Benos,

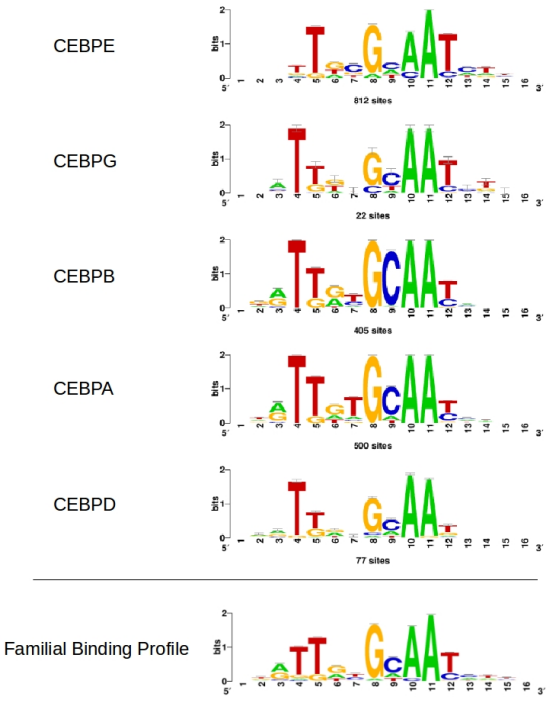


Figure 7.17: A cluster of motifs belonging to the CEBP family can be summarized as a Familial Binind Profile.

2007) and further represent the cluster as a Familial Binding Profile (FBP), i.e., a motif that summarizes all the motifs in the cluster. This representation is possible since the TF families are classified according to the TF DBDs (Wingender et al., 2013), and combined with the alignment of the motif logos is useful to visualize the similarities or differences in a motif cluster (Figure 7.16). In order to build the FBP, the clustered motifs should be aligned and merged (the counts of nucleotides at each column can be summed or averaged) (Habib et al., 2008), the flanks could be trimmed in order to obtain a FBP with the most relevant positions (Mahony et al., 2007).

The idea to represent a group of PSSMs as a FBP arose before the high-throughput methods became popular and that time, the total number of PSSM available was small (around ~100 PSSM for human TFs). However, nowadays, the number of available PSSM has increased (~600 PSSMs for human TFs) and the studies focused on TF binding (e.g., using ChIP-seq, ChIP-exo, SELEX-seq) have became popular, making available from hundreds to thousands of PSSMs in a publication (Forrest et al., 2014; Kheradpour and Kellis, 2013; Jolma et al., 2013; Weirauch et al., 2014), even more PSSMs than those stored in the databases.

The problem of having thousand of motifs is the redundancy, that could be a consequence of using multiple tools to discover the motifs, this is recommended since some tools could find a motif that the others do not, however a particular motif could be found by several tools, returning thus several redundant PSSMs for the same TF. Every discovered motif should be further compared with hundreds or thousands of known motifs stored in databases as JASPAR (Mathelier et al., 2015), HOCOMOCO (Kulakovskiy et al., 2016), FootprintDB (Sebastian and Contreras-Moreira, 2014) or Cis-BP (Weirauch et al., 2014), therefore the comparison becomes unnecessary for the redundant motifs. The advantage of a motif clustering step before the comparison should reduce the analysis time, in other words, with the clustering we could identify a FBP for a set of PSSMs, and only this FBP should be compared with those motifs in the databases (Figure 7.17), another application that might accelerate the searches on databases is searching motifs by pre-built FBPs.

The redundancy of motifs, however, is also present on motif databases. The biological explanation is that TFs from the same family share their DBD, hence they recognize similar TFBSs, even when the TFs belong to distal species (Weirauch et al., 2014), and this is reflected in the motif logo (Figure 7.17), although in some

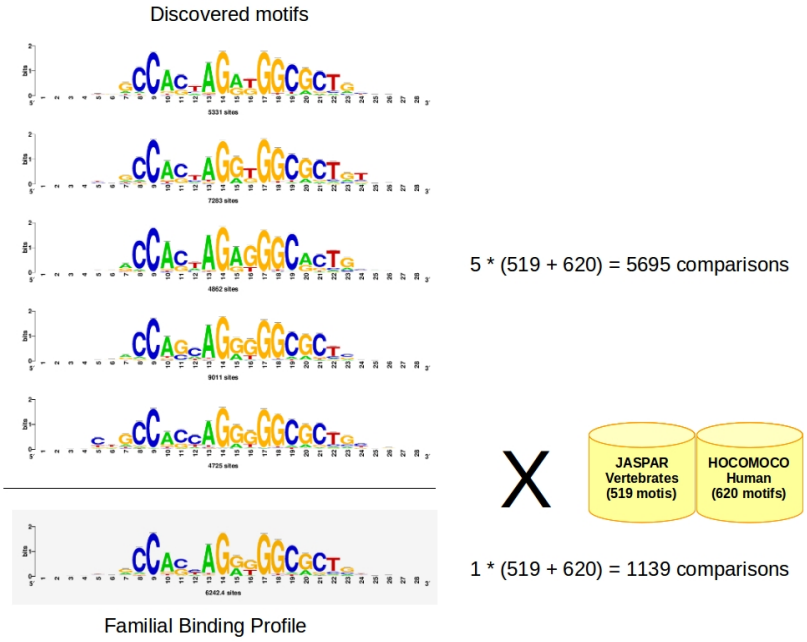


Figure 7.18: A single Familial Binding Profile obtained after the clustering of redundant discovered motifs is compared only once with the motif databases, reducing the analysis time and the number of comparisons.

cases, motifs with similar logos do not belong to the same family and in other cases TFs from the same family (e.g., zinc fingers) have completely unrelated logos (Figure 7.18). However, in databases, motif redundancy occurs when the motifs are discovered from different experimental sources (e.g., ChIP-seq, SELEX or PBM) or with data from close related organism following the phylogenetic footprinting approach (Figure 7.19). The information of similar DBDs (with a high similarity in their aminoacids) has nede possible to infer infer thousands of TF binding motifs in tens of eukaryotes genomes starting from a relativelly small sets of characterized DBDs. The results are available in a database called Cis-BP (Weirauch et al., 2014), that is at this time the most comprehensive motif database.

Nowadays, the importance of the clustering of motifs is demonstrated on large scale projects as FANTOM5 (Forrest et al., 2014) or ENCODE (Kheradpour and Kellis, 2013), where hundreds of datasets (e.g., ChIP-seq) are analyzed with several motif discovery tools that produce thousands of motifs. In most of the cases, in order to cluster the motifs, the authors have developed their own pipelines, starting from the results of motif comparison, however, currently there are at least seven tools specialized on motif clustering: Matlign (Kankainen and Löytynoja, 2007), STAMP (Mahony et al., 2007), m2match (Stegmaier et al., 2013), DMINDA (Ma et al., 2014a), motifstack (Ou and Zhu, 2012), GMACS (Broin et al., 2015) and RSAT *matrix-clustering* (Castro-Mondragon et al., 2017).

The motif clustering tools rely on the same metrics to compare motifs, therefore, the clustering tools based on column comparisons (Matlign, STAMP, m2match, motifstack,and *matrix-clustering*) use similar clustering approaches (e.g., hierarchical clustering) and the motifs can easily be aligned, allowing the visualization of the clusters. Only one tool, GMACS, use a genetic algorithm to cluster the motifs and returns the list of motifs of every detected group without visualization of the motifs.

Given the large amount of motifs analyzed in a single study and the resulting redundancy of discovered motifs given the use of several tools, there is a increasing need for efficient tools to cluster the motifs and ease the further motif analysis. Some of the existing motif clustering tools have some limitations, for example, the number of input collections of motifs to be clustered and the visual representation of the clusters. For these reasons, I developed a motif clustering algorithm, RSAT *matrix-clustering* that present some advantages over the other existing tools, see Results chapter (Castro-Mondragon et al., 2017).

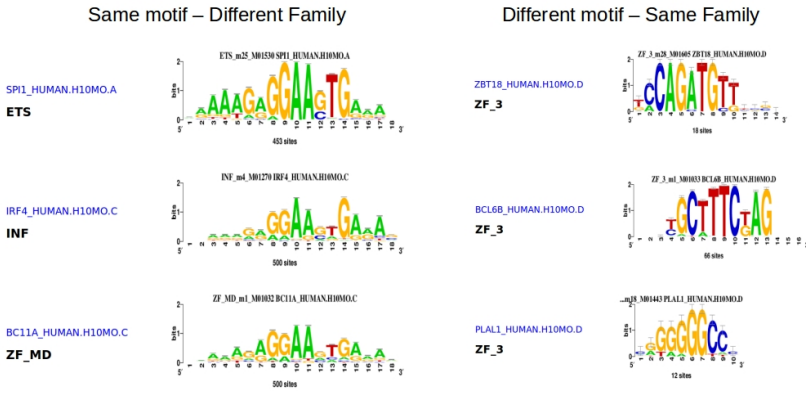


Figure 7.19: Example of logos for a) different TFs from different families sharing similar motifs and b) unrelated motifs for TFs from the same family. Motifs and family information taken from HOCOMOCO. ETS: Ets-related factors, INF: interferon-regulatory factors, ZF MD: multiple dispersed zinc fingers, ZF 3: 3 adjacent zinc finger factors.

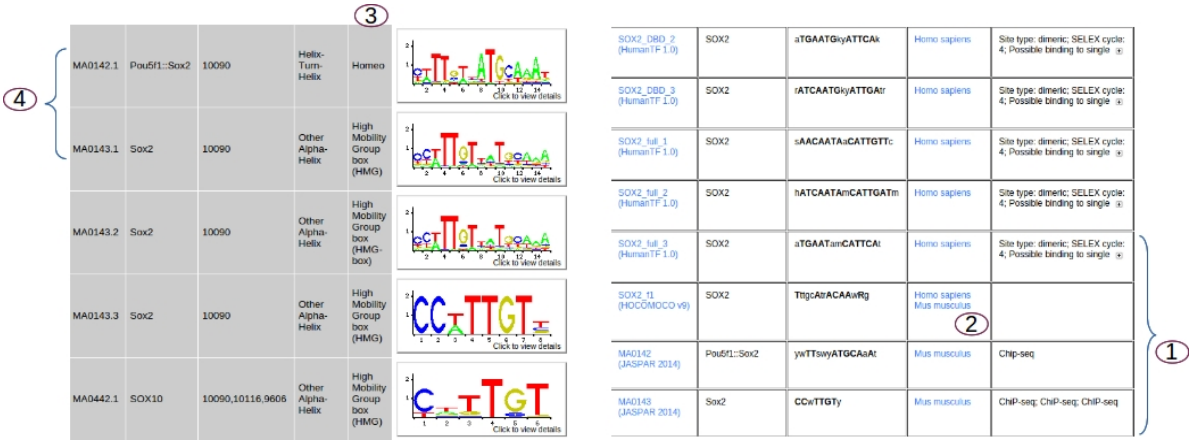


Figure 7.20: Examples of sources of redundancy on motif databases, Jaspar and FootprintDB, using as query "Sox2". (1) Distinct experimental methods to detect TF binding. (2) Motifs can come from data of several species. (3) TFs from the same Family use to have conserved structure of DBD. (4) Difficulties in annotation.

The clustered motifs (FBP) are useful to represent a group of motifs, for example a family of TFs, or can be used to reduce the computing time of algorithms affected by motif redundancy (e.g., motif comparison or motif enrichment), however, I recommend not to use them to scan sequences, since the IC for some positions could be biased in cases when motifs of different sizes are aligned, one solution could be trim the flanks with low IC, however, these flanks could be informative for some TF models (Jurk et al., 2016; Gord??n et al., 2013).

It must be noted that the methods for comparison and clustering of motifs, at this moment are limited to the mono-nucleotides PSSMs. Only one comparison metric has been upgraded to deal with dinucleotides PSSMs, that is the Jaccard index method (Vorontsov et al., 2013) implemented by the authors of HOCOMOCO (Kulakovskiy et al., 2016), as these models become more used, the comparison metrics should be upgraded in order to deal with them.

7.6 Motif Enrichment

Another common question regarding the motif analysis is when we have a set of sequences and we want to known if they are enriched by certain TFs (e.g., if a set of sequences thought to be related with interferon response are enriched with IRF TFs). In this approach, we start from a set of already known PSSMs (e.g., a complete motif database as JASPAR). The term enrichment refers to observe a higher number of TFBSs relative to a control (e.g., random expectation or relative to another set of sequences).

The advantage of using motif enrichment are that (i) a small set of TFs could be analyzed, in a large set of sequences (e.g., ChIP-seq peaks) without motif discovery or (ii) all the TFs of a complete databases can be analyzed in a single run, therefore some motifs that are not detected by the motif discovery approaches could be detected by measuring the enrichment.

In general, the motif enrichment methods can be divided in the following two groups (Figure 7.20):

- Global enrichment: the approach measures the TFBS enrichment in a set of sequences independently of the location of the TFBSs, specially helpful for analyzing promoters of co-regulated genes. Some existing tools are PASTAA (Roider et al., 2009), CLOVER (Frith et al., 2004), *cisTargetX* (Aerts et al., 2010), AME (McLeay and Bailey, 2010), and RSAT *matrix-quality* (Medina-Rivera et al., 2011) and *matrix-enrichment* (unpublished).
- Positional enrichment: this approach is used to detect enrichment of TFBSs located relative to a reference (e.g., upstream TSS, center of ChIP-seq peaks) (Bucher and Bryan, 1984). Some existing tools are CentriMo (Bailey and Machanick, 2012), TFBSLandscapes (Worsley Hunt et al., 2014), *ChIP-Seq tools* (Ambrosini et al., 2016) and RSAT *position-scan* (unpublished).

It must be noted that the enrichment of motifs relies on the detection of individual TFBSs (pattern matching), and the results depend on the background model and on the threshold to detect TFBS (e.g., PASTAA (Roider et al., 2009) and CLOVER (Frith et al., 2004)). Others methods do not require a threshold and the enrichment can be measured for low and high-scored TFBSs (McLeay and Bailey, 2010; Thomas-Chollier et al., 2011b; Medina-Rivera et al., 2011), in addition different statistical methods can be used or even additional data as transcriptome information can be also used (Aerts et al., 2010) for a more precise detection of the enriched motifs.

In the threshold-free methods, the expected number of TFBSs can be estimated with the distribution of weight scores of a PSSM given a background model, for example, if the TFBSs with strong weight score for a given PSSM are observed more than expected under the background (Medina-Rivera et al., 2011), then the PSSM can be considered as an enriched motif. Usually a p-value can be calculated for every PSSM (McLeay and Bailey, 2010; Thomas-Chollier et al., 2011b) returning a list of top enriched motifs, or the enrichment could be calculated for every weight score (Medina-Rivera et al., 2011). In these cases, if the analyzed collection of motifs is redundant (e.g., measuring the enrichment for all the motif databases of insects), similar TFs could be detected as enriched and the higher the number of analyzed motifs, the higher correction factor for the multitesting correction used in these programs.

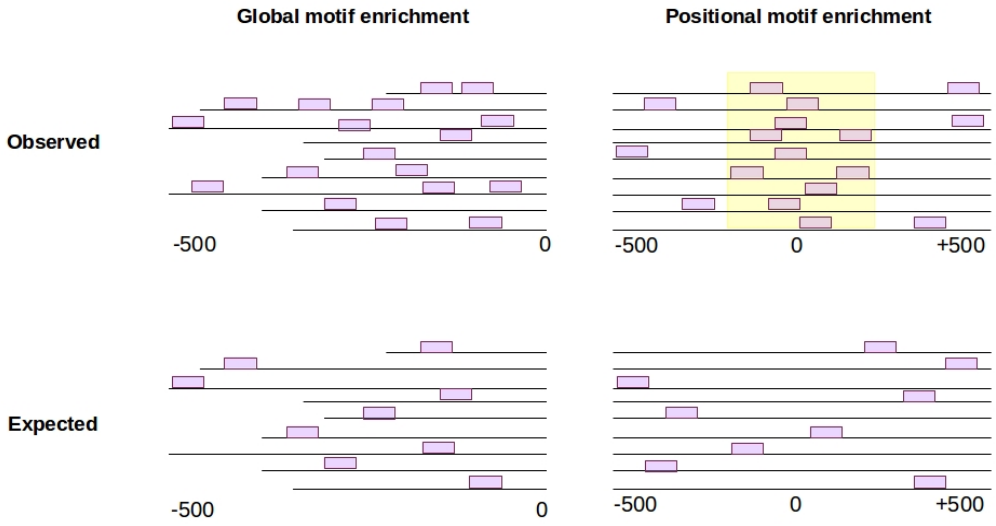


Figure 7.21: Global and positional motif enrichment. The global enrichment is estimated relative to expected number of TFBSs, it is specially useful for sequences with variable lengths. The positional enrichment detect the TFBSs concentrated at certain location of the sequences (e.g., at the center of ChIP-seq peaks of the same size).

The current global motif enrichment methods are limited to use only one set of sequences, and they have no visualization of the enriched motifs. To face these limitations, I extended, in collaboration with Alejandra Medina-Rivera the algorithm RSAT *matrix-quality* (Medina-Rivera et al., 2011) in order measure enrichment in several sequences and I created a dynamic visualization interface for the results. This new algorithm RSAT *matrix-enrichment* (unpublished), allows to easily detect TF enriched in a particular set of sequences (Figure 7.21).

With the advent of high-throughput experiments, specifically with ChIP-seq, the motif enrichment methods have been adapted in order to find TFBSs concentrated at certain locations of the sequences (e.g., the center of the peaks) where it is expected that the strongest TFBS are located. At least three tools have been developed to achieve this task: CentriMo (Bailey and Machanick, 2012), TFBSLandscapes (Worsley Hunt et al., 2014) and RSAT *position-scan* (unpublished). In the positional enrichment, at least for ChIP-seq data, every sequences is scanned with a set of motifs (discovered or from databases) and only the strongest TFBS per sequence is considered, therefore the further calculation of the enrichment is based with these sites (Figure 7.22). This approach, conversely to some global motif enrichment algorithms, requires a threshold to detect the putative TFBSs.

Given that CentriMo and TFBSLandscapes are focused on finding enriched motifs, I recently developed *position-scan*, following the principle of RSAT *position-analysis* (Helden, 2000), where the input sequences are divided in bins, but rather than using the k-mer counts, this method counts the predicted TFBSs, and the expected frequencies per bin (null hypothesis) are estimated by distributing the total homogeneously over the bins. The main difference between *position-scan* and the other three existing methods is the capability to find locally depleted motifs (e.g., under-represented), for example a TF that could alter gene regulation and therefore their binding has been counter-selected in these sequences (Telorac et al., 2016) or probably the local depletion of a motif results from the nucleotidic composition of the sequences (Worsley Hunt et al., 2014). See chapter 10 for a detailed explanation of *position-scan* algorithm.

Although the current motif enrichment tools work with mono-nucleotide PSSMs only, since the calculation of the methods is a count problem (and not related directly with the PSSM model), these tools could be easily extended to use the TFBSs predicted by di-PSSMs.

Motif enrichment methods allow to execute exhaustive search of motifs (e.g., all the vertebrate motif

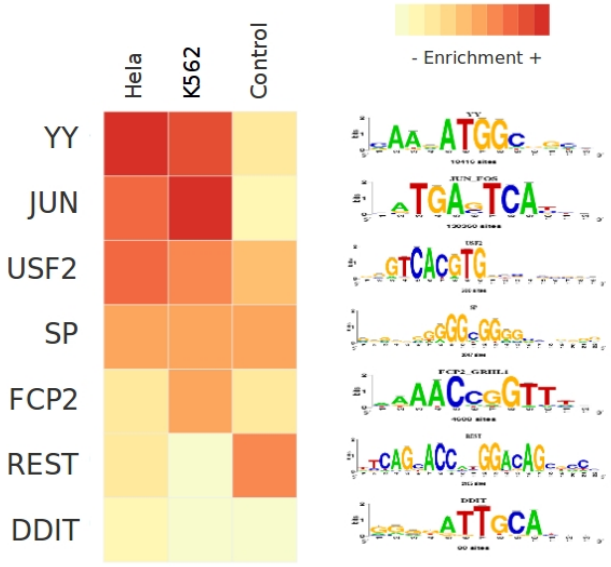


Figure 7.22: Motif enrichment on two set of test sequences (HeLa and K562) and one set of control sequences. The visualization of the enrichment shows two TFs (YY and Jun) are enriched on the test sequences only, whilst other TFs (UFS2 and SP) are enriched in the three sequence sets

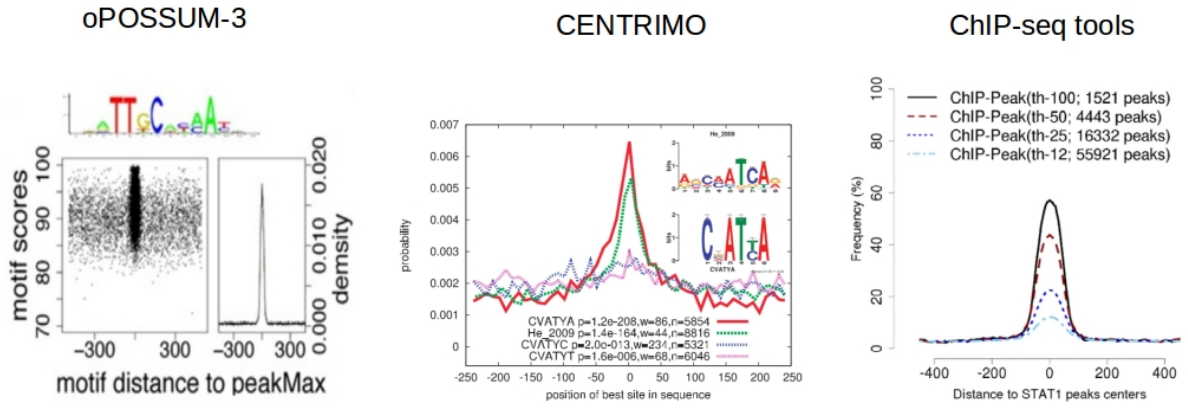


Figure 7.23: Examples of positional enriched motifs by TFBSLandscapes, CentriMo and ChIP-seq tools. In these algorithms, the enrichment is calculated based on the best hit per sequence. Adapted from Worsley-Hunt (2014), Bailey (2014) and Ambrosini (2016).

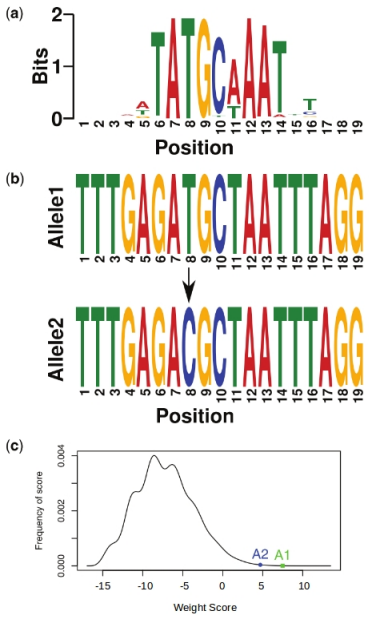


Figure 7.24: Detection of regulatory variants using PSSMs. (A) Logo for OCT1. (B) The two alleles analyzed, see position 8. (C) Distribution of all the weight scores with the OCT1 PSSM, note the score difference between the alleles. Adapted from Macintyre (2010).

databases), this is specially useful when the motif discovery approaches have not the statistical power to detect motifs (Bailey and Machanick, 2012), although it should me mentioned that enrichment methods are affected by the motif redundancy.

7.7 Identification of TF binding variants

The most recent application for the PSSMs is in the discovery of regulatory variants (i.e., mutations within a regulatory region as enhancer, promoters or TFBSs) that might affect the downstream regulation. Although some of these variants affect others aspects of gene regulation as looping, others directly affect the TF binding and these can be detected using PSSMs. See (Mathelier et al., 2015) for a revision about the regulatory variants, their effects in health, and the experimental and computational methods to detect them.

The logic behind this approach is that the variant could be reflected in the weight score produced by a PSSM. A difference in the weight score between two alleles could indicate a true regulatory variant (Figure 7.23). One way to detect is calculating the complete distribution of weight scores (or p-values) for a single PSSM and calculate the difference in weight scores between the alleles A (query) and B (control), this method is used by the tools is-rSNP (Macintyre et al., 2010), Regsnps (Teng et al., 2012), that are specialized on regulatory Single Nucleotide Polymorphisms (rSNPs). Others tools as RSAT *variation-scan* (Medina-Rivera et al., 2015) and sTRAP (Thomas-Chollier et al., 2011b) can detect SNPs and insertions or deletions at the TFBSs. Although using PSSM only is the most used method, other methods however have been proposed, for example combining TFBSs prediction and sequence conservation (Andersen et al., 2008), TFBS prediction (Shi et al., 2016) and k-mer analysis (Fletez-Brant et al., 2013; Lee et al., 2015) combined with supervised classification (training a model with a set of regulatory sequences detected with DNaseI-seq versus a set of random genomic sequences).

It should be noted that the identification of the variant gives only partial information, to complement it should be also analyzed (when data is available) wheter the variant is located in a locus whose genotype can be quantified, that is a quantitative Trait Loci (QTL) (Kumasaka et al., 2015). If these variant show

differences on gene expression (measured as mRNA production) between two alleles, it is considered an expression QTL (eQTL). So, in addition to find a variant with a significant difference in weight score, this information should also be supported by the fact that there is a difference on the expression of the gene associated to such variant.

Depending on where the QTLs are located, they can be classified as local (i.e., if they affect the gene expression by changing the TF binding in a promoter) or distal (located at enhancers, introns or far away from their associated gene) (Kumasaka et al., 2015). In addition, depending on the phenotype affected, others QTLs may be related to local chromatin access (caQTL), identified with DNase-seq, DNase-sensitive QTLs (dsQTLs) (Degner et al., 2012) or measured with ATAC-seq (Kumasaka et al., 2015); others may regulate histone marks (hQTLs) (Grubert et al., 2015); others may be CpG sites in which changes in DNA methylation are associated with genetic variation (Taudt et al., 2016). Of note that either dsQTLs, caQTLs, meQTLs and hQTLs does not always are eQTLs.

Since the methods to detect regulatory variants mentioned in this section rely on the scanning of PSSMs, the high rate of false positives is a current issue that should be carefully treated. Threshold selection is crucial to distinguish real rSNPs from noise. A previous step selecting the threshold for every PSSM and a selected set of PSSMs related to the SNPs may improve the analysis and reduce the number of false positives, see (Andersen et al., 2008) for detailed guidelines of *in silico* detection of regulatory variants.

The detection of regulatory variants is widely studied, specially now that we have data for entire populations (Qu et al., 2015, Kasowski et al. (2010)) and it is possible to study differences in TF binding from one individual to another, for example the allele specific binding or groups of individuals from the same population, the results, however, require further investigation in order to know if a mutation is or not the cause of some phenotype and it should be noted that association of a regulatory variants with a QTL should not be considered as a causation of the observed phenotype.

7.8 Resources

The bioinformatic methods to study TFs have been developed since the late 80's, starting from the study of a few sequences related in a particular condition to the study of all the TF binding events in a whole genome. Given this change in the amount of the data analyzed, many of the early developed methods have been adapted to deal with large data amount, other methods have been specially designed to work with large data amount but others became obsolete.

Whilst a lot of algorithms (either for motif discovery, motif scan, comparison, etc) have been published and can be used independently, only a few research projects have been focused specially in the development of software to study cis-regulatory sequences and currently they are the most highly used in almost all publications involving motif analysis. These projects (*suites*) have in common that support a collection of tools that allows to analyze the motif following a pipeline (i.e., motif discovery starting from input sequences followed by motif comparison and further identification of TFBSs), allowing thus the creation of workflows and reproductibility of results (Table 4.1). At the moment only one of these projects, AUTOSOME, has expanded their methods to work with di-nucleotide PSSMs.

Since every one of these suites contain different tools for motif discovery (or other tasks), some tools in one suite could be better for an specific task than those tools in the other suites, therefore in order to have complementary results, their output can be combined (i.e., a workflow including tools from different suites). The users are not limited to use a single suite for all their analyses (Shi et al., 2016; Kuttippurathu et al., 2011).

7.8.1 Motif databases

In addition to those projects focused on the development of tools for TF motif analysis, others no less important projects are the databases for TF binding motifs and their binding sites. The development of

Table 7.2: Description of some motif analysis resources.

Name	Year	Website	PMID	Description
RSAT	2003	rsat.eu	25904632	Tools to analyze TF binding motifs, download sequences and
TOUCAN	2005	goo.gl/lH7kR1	15980497	Motif analysis tools, specialized in finding cis-regulatory mo
MEME	2006	http://meme-suite.org/	25953851	DNA, RNA, and protein motif analysis tools.
HOMER	2010	homer.ucsd.edu/homer/	20513432	Tools for analyze high-throughput results and TF binding m
AUTOSOME	2010	http://autosome.ru/	NA	Tools for analyze TF binding motifs, including dinucleotide

these databases (e.g., collecting the TFBSs) started even before the development of many tools for motif discovery, at that time, the search of TFBSs was based on manual curation of the literature, giving rise to popular databases as TRANSFAC (Wingender et al., 1996, Kaplun et al. (2016)) and RegulonDB (Huerta et al., 1998; Gama-Castro et al., 2015), for eukaryotes and bacteria, respectively. At the beginning, they stored TFBSs only, however nowadays they have been extended to include PSSMs and other regulatory data, however, still from literature curation.

Currently there are tens of motif databases either public or private, some of them specialized in a single or few organisms, and others containing information from different taxa. The methods to obtain the TFBSs and PSSMs vary on each databases. See Table 7.2 for a summary of the most representative motif databases.

In the case of RegulonDB, at least for *E coli K12*, given that most of its TF use to recognize a small number of TFBSs, the manual curation is possible, however, for metazoas to gather data of high quality (i.e. with references to the literature for individual binding sites might be time-consuming and all the databases (except TRANSFAC) obtain the data from the results of high-throughput methods (e.g., ChIP-seq, PBM, SELEX-seq). The advantage of literature curation is that the TFBSs are validated experimentally (i.e., *bona fide* TFBSs) but it is time consuming, by contrast, the data obtained from high-throughput experiments contains a higher number of TFBSs but with a lot of false positives (Weiss et al., 2013).

Many databases as JASPAR, RegulonDB, Cis-BP or HOCOMOCO are constantly updated, and the increase in the amount of information can be observed through their releases (for example JASPAR 2016 (Mathelier et al., 2015) doubled its size relative to the 2014 version (Mathelier et al., 2014)).

This is expected, since during the time each version is released more TFs are studied or those TF with available data are studied in new conditions (e.g., cell lines) or using distinct experimental methods (e.g., binding regions detected by ChIP-seq or ChIP-exo), and therefore PSSMs with low quality can be improved with more data. However the cost of these updates is the redundancy in the motif databases. Similarly, some recent studies have produced as many data (e.g., discovered motifs) as the data stored in the motif databases (Jolma et al., 2013, 2015; Whitaker et al., 2015), however these studies are not focused in produce and maintain a database, although the discovered motifs are freely available.

As the databases become bigger and other studies produce large number of PSSMs, it is difficult to choose a single database to use it in a study, the simple solution should be create a meta-database (database of databases) containing all the published motifs, at least two meta-databases are available: Cis-BP (Weirauch et al., 2014) and FootprintDB (Sebastian and Contreras-Moreira, 2014), although they contain different information. Cis-BP integrates information from several motif databases and contains its own motifs, some of them obtained directly from PBM data and other inferred motifs with similar PSSM and high amino-acid conservation of their DBD in several species, currently is the biggest motif database. FootprintDB contains 16 motif databases and for most motifs include the structural information of their DBD, the binding sites used to build the PSSMs and can be considered as a public alternative for the commercial database TRANSFAC (Matys, 2003).

Although the meta-databases store thousands of motifs, the content is specially redundant. One motif (e.g., Sox2) can appear several times, since any merged collection could have its own version of this motif. A partial solution is to show the motifs grouped by similarity, as part of my thesis, I developed an algorithm RSAT *matrix-clustering* [ref] that presents a dynamic visualization of similar motifs, and might be integrated

Table 7.3: Description of representative motif databases for vertebrates, insects, plants and bacteria. Hs: Homo sapiens, Mm: Mus musculus, Dm: Drosophila melanogaster, At: Arabidopsis thaliana, Ec K12: Escherichia coli K12, Multi: databases with data from multiple species, V: vertebrates, P: plants, I: insects, B: bacteria.

Taxon	Database	Size	Sp	PMID	Description
V	Cis-BP	734	HS	928674	Inferred TFBSs from multiple (>300) species. Incorporates da
V	ENCODE	2065	HS	22955990	Discovered TFBMs for the ENCODE TF ChIP-seq datasets
V	epigram	589	Hs	25240437	TFBMs identified in ChIP-Seq data of six histone modification
V	Fantom5 novel	169	Multi	24670764	Motifs discovered in clustered CAGE TSS (not matching know
V	Hocomoco	641	Hs	26586801	Hand-curated Human TFBS models constructed by integratior
V	Hocomoco	427	Mm	26586801	Hand-curated Mouse TFBS models constructed by integration
Vs	homer	332	Hs	20513432	Human TFBMs discovered in public ChIP-seq and promoter d
V	hPDI	437	Hs	19900953	Human TFBMs discovered from PBM (until 2009)
V	JASPAR	519	Multi	26531826	Curated TFBMs derived from published collections of experim
V	HumanTF	818	Hs	23332764	Human TFBMs obtained by high-throughput SELEX and ChI
V	HumanTF_dimers	664	Hs	26550823	TFBMs of human TF pairs that bind cooperatively to DNA ol
V	Uniprobe	386	Ms	25378322	TFBMs generated by universal protein binding microarray (PI
P	ArabidopsisPBM	108	At	24477691	Arabidopsis TFBMs discovered from PBM
P	Athamap	84	At	18842622	Genome-wide map of potential TFBSs in Arabidopsis
P	Cistrome	862	At	27203113	Arabidopsis TFBMs discovered from DAP-seq data
P	Cis-BP	309	At	928674	Inferred TFBSs from multiple (>300) species. Incorporates da
P	JASPAR	227	Multi	26531826	Curated TFBMs derived from published collections of experim
I	dmmpmm	41	Dm	19605419	Drosophila TFBMS discovered from DNAse-seq data
I	DrosphilaTF	61	Dm	17238282	Drosophila promoter motifs discovered using NestedMICA
I	Cis-BP	361	Dm	928674	Inferred TFBSs from multiple (>300) species. Incorporates da
I	FlyFactorSurvey	652	Dm	21097781	Drosophila TFBMS discovered from the bacterial one-hybrid s
I	idmmpmm	39	Dm	19605419	Drosophila TFBMS discovered integrating data from different
I	JASPAR	133	Multi	26531826	Curated TFBMs derived from published collections of experim
I	OntheFly	608	Dm	24271386	Drosophila TFBMS discovered from DNAse-seq data, SELEX-
B	RegulonDb	93	Ec K12	26527724	TFBMs obtained from literature curation

in databases in order to browse the motifs, see the Results for the publication and a summary of this algorithm. This is relevant specially for those methods affected by redundancy, as motif comparison and motif enrichment, where the algorithm already include the motif databases to ease the analysis (Mahony et al., 2007; Gupta et al., 2007) (i.e., the users have not to download the motif database since it is part of the program).

7.8.2 All in one: motif analysis for high-troughput data

At this point I have presented separately the different algorithms for motif analysis. Nowadays, however, several workflows have been developed to automate the application of successive analysis steps on large of collections of sequences with large collection of sequences (e.g., ChIP-seq peaks). These programs, are modularly designed to achieve the next steps (Figure 7.24):

- Nucleotide composition: the mono or di-nucleotide composition of the input sequences brings information about the expected discovered motifs.
- Motif discovery: several algorithms are ran in the same sequences in order to discover motifs in an exhaustive way. This is recommended, but the cost is redundancy in the results.
- Motif comparison: the discovered motif are compared with motif databases (e.g., JASPAR, HOCOMOCO) in order to find the most similar motif(s) and annotate the discovered motifs. This step, in addition, reflect if the experiment were well performed, for example, if the motif though to be expected is not found could indicate an error in the experimental procedure.
- Motif scan: the discovered motifs are scanned in order to find their putative TFBSs.
- Motif enrichment: the enrichment of motifs (global or positional) from one or more databases can be measured in order to reveal motifs that were not detected in the discovery step.
- Motif clustering: the motifs are grouped by similarity in order to ease the analysis and detect the redundancy.
- Visualization: the TFBSs can be exported as tracks to be visualized in a genome browser.

At least four of these workflows have been used for several studies: MEME-CHIP (Machanick and Bailey, 2011; Ma et al., 2014b), RSAT *peak-motifs* (Thomas-Chollier et al., 2012b,a), XXmotifs (Luehr et al., 2012) and Dimont (Grau et al., 2013), see (Lihu and Holban, 2015; Tran and Huang, 2014) for a revision and comparison of several ChIP-seq motif analysis workflows. The most popular of these tools is MEME-ChIP that is a workflow that combines results from MEME (Bailey and Elkan, 1994), DREME (Bailey, 2011) and CentriMo (Bailey and Machanick, 2012). However given the algorithm complexity of its core program MEME, the motif discovery is limited to 600 sequences randomly selected from the input, in addition the input sequences should have the same length which is a limitation for some studies; XXmotifs has limitation in the amount of input data but not on sequence length, by contrast *peak-motifs* have no limitation on amount neither on sequence length, although it lacks of clustering and enrichment of motifs from databases.

The development and maintenance of these workflows allows the reproductibility of motif analysis results and the annotation of motifs in an automatic way. A general limitation of these workflows is that they start from the sequences and not from the raw data (e.g., short or long reads from sequencers), although this topic is outside of the scope of this thesis, it is important to note that the results of the peak-caller algorithms affect the downstream motif analysis. Another limitation, although now some solutions started to arise, is the annotation of the regulatory regions (i.e., the association of the regulatory region with its target gene. At this time there is only one motif analysis workflow under development involving the complete process from peak calling to peak annotation, that is CRUNCH (Berger et al., 2016), although for the moment is only accessible via a web interface.

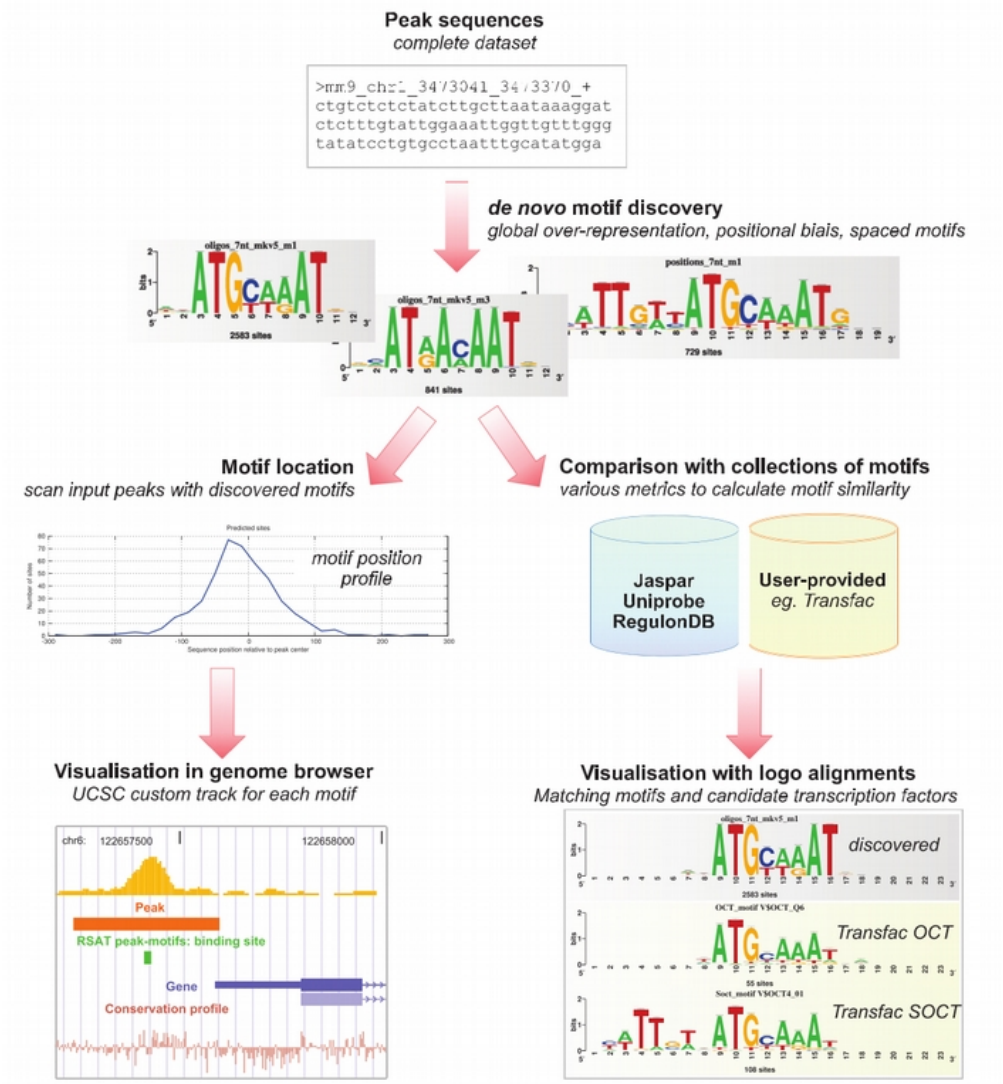


Figure 7.25: Schematic representation of a workflow of motif analysis in large sequence datasets. Adapted from Thomas-Chollier (2012).

Chapter 8

RSAT 2015: Regulatory Sequences Analysis Tools

8.1 Motivation and state of the art

RSAT (Regulatory Sequences Analysis Tools) is a modular suite of software tools dedicated to the analysis of cis-regulatory sequences and TF binding motifs. This project is led by Jacques van Helden since 1998 (van Helden et al., 1998) and it has been constantly updated with new programs at every release (Van Helden et al., 2000; van Helden, 2003; Thomas-Chollier et al., 2008, 2011a; Medina-Rivera et al., 2015). The added programs are adapted to the current necessities of the scientific community working with motif analysis. The programs from the RSAT suite can be used online (http://www.rsat.eu/), as stand-alone via command-line, remotely via SOAP/WSDL Web services, as a VirtualBox virtual machine or instantiated on the Institut Français de Bioinformatique (IFB) cloud.

The first version included two programs for motif discovery (Van Helden et al., 2000) and three years later new programs were added allowing users to run complete motif analysis, including the visualization of the results and the option to retrieve sequences (i.e., regulatory regions) for a large number of organism, specially bacterial genomes (van Helden, 2003). The 2008 version had 30 programs including *matrix-scan* one of the first programs to scan sequences with PSSMs using the p-values as threshold rather than the weight scores. In addition this version included software to create negative controls and phylogenetic footprinting. Between the 2008 and the next version in 2011, started the era of high-throughput experiments producing large amount of data to analyze, for this reason the 2011 RSAT version included several programs focused on the high-throughput experiment results as *peak-motifs* that is a workflow for analyze large sequence sets, other programs to download sequences from the UCSC server, and software to compare and evaluate the quality of motifs (Thomas-Chollier et al., 2011a). The latest version was released in 2015 and contains 52 tools, among the added tools there is a program to detect and download regulatory variants and a program to cluster TF binding motifs. In this version, part of the work was focused on ensuring the reproductibility of the results, for example using virtual machines with a particular version of RSAT. In addition, since the number of genomes growth considerably since the last RSAT version, the RSAT servers were separated by taxon-specific servers containing the genomes and other information for their different organisms (Medina-Rivera et al., 2015).

The main applications of RSAT are the following:

- Motif discovery, appropriate to high-throughput data sets like ChIP-seq, PBM or groups of promoters of co-expressed genes.
- TF binding motif analysis (quality assessment, comparisons and clustering).
- Motif scanning either to detect single TFBSs or cis-regulatory modules.
- Comparative genomics for a large set of organisms in different taxa.

- Analysis of regulatory variations.
- Creation of background models and negative controls for motif analysis.

RSAT is one of the most comprehensive suites of programs for analysis of cis-regulatory sequences, it goes beyond the motif analysis and also includes programs for comparative genomics and detection of regulatory variants. In addition of this versatility of programs, RSAT is compatible with the output of external tools (thanks to inter-conversions between file formats). The accessibility of RSAT made itself a choice for biologists with little or no experience in programming skills, in addition, the modularity of the programs allows to create complex workflows for motif analysis, including external tools.

8.2 Contribution

Since 2014 I have contributed to the development of new tools and maintenance of RSAT. As part of this PhD thesis I developed three new tools for motif analysis: *matrix-clustering* (Castro-Mondragon et al., 2017), *position-scan*, and *matrix-enrichment*. The output of these tools is highly interactive and are the first programs in RSAT with interactive visualization features, using the D3 javascript library (Bostock et al., 2011), which ease the interpretation of the results.

These three novel tools can be used for the analysis of high-throughput data: *matrix-clustering* (Castro-Mondragon et al., 2017) (Chapter 11) can group similar and redundant motifs returned by motif discovery tools, the other tools, *position-scan* (Chapter 11) and *matrix-enrichment* detect the positional and global motif enrichment, respectively. The last two programs are new developments posterior to this publication and will be published elsewhere.

As part of the maintenance of RSAT, I have contributed to this project by collecting and updating the set of motif databases that is used by several tools (e.g., *compare-matrices peak-motifs*).

8.3 Conclusion

In this thesis, most of my results were obtained using programs from RSAT, some of these programs were developed by myself and others were already available. In my opinion, RSAT is the most comprehensive suite for tools focused on cis-regulatory sequences, my reasons to think this and choose RSAT for my projects are the following:

- Modularity of the RSAT programs: the programs can be called individually or integrated either in automated workflows for motif analysis (e.g., *peak-motifs*) or on in-home developed workflows using workflow managers such as snakemake (Köster and Rahmann, 2012) or script languages (e.g., bash, make)“]. for example, the tool *matrix-clustering* was partially made by parts of existing tools (e.g., *compare-matrices*).
- Multiple genomes supported: the RSAT programs are not limited to a handful of model organisms, during my thesis I worked with data from humans, flies, mouse, bacteria and plants. The organism information can also be downloaded via RSAT programs which save time for the further analysis.
- Inter-compatibility: the RSAT suite can be used with input from other tools (e.g., MEME or HOMER), this allow to analyze data combining distinct tools.

The development of software dedicated to motif analysis and specially the distribution and maintenance of this software in projects as RSAT, is specially useful for the scientific community that is not familiar with programming, the interaction between biologists and bioinformaticians is key for the development of user-friendly, simple and accessible interfaces. By the side of the software developers the main task is to ensure the reproductibility of their results.

Chapter 9

RSAT *matrix-clustering* : dynamic exploration and redundancy reduction of transcription factor binding motif collections

9.1 Motivation and state of the art

Transcription Factor Binding Motifs (TFBMs), simply called motifs, are models describing the binding specificity of a transcription factor (TF). Such motifs are generally obtained by aligning the sequences of several binding sites, and summarizing the nucleotide frequencies per position (Stormo, 2000). Motifs are commonly represented as position-specific scoring matrices (PSSMs) and visualized as sequence logos (Schneider and Stephens, 1990). Although the adequacy of PSSMs has been questioned for some particular TF classes (Weirauch et al., 2013; Jolma et al., 2013; Mathelier and Wasserman, 2013; Keilwagen and Grau, 2015), e.g. in cases of dependencies between adjacent nucleotides, they are still the most widely used method to represent the binding specificity of a TF.

Thousands of motifs are available in motif databases which constitute key resources to interpret functional genomic resources. A well known issue of these databases is the motif redundancy resulting from various sources (Mathelier et al., 2014):

- For a given TF, multiple PSSMs can be built from different collections of sites characterized with alternative methods (i.e. DNase-Seq, SELEX, PBMs, ChIP-seq).
- The binding specificity is often conserved between TFs of the same family.
- Some databases contain PSSMs obtained from orthologous TFs in different organisms.
- Some unrelated TFs recognize similar DNA motifs.

Another example of motif redundancy is observed in the results of motif discovery tools, where is recommended to use several motif discovery approaches in order to have robust results (Thomas-Chollier et al., 2012a; Ma et al., 2014b). While some motifs could be discovered exclusively by a given tool, most will be found independently by different tools, hence producing redundant motifs with small variations in length and/or nucleotide frequencies at some positions.

In addition, it must be noted that current large projects as ENCODE (Kheradpour and Kellis, 2013) or FANTOM5 (Forrest et al., 2014) can produce thousands of motifs, even more motifs than those stored in a single databases (with the difference that these produced motifs are not curated as in the motif databases).

This constant increase in the number of motifs and redundant collections represents a real challenge for the community. Which collection to use? How important is the overlap between the different collections?

Many efforts have been done in order to reduce motif redundancy using the results obtained from motif comparison. The most similar motifs can be grouped in clusters of motifs. therefore a cluster can be represented as a single motif (also known as Familial Binding Profile (FBP) (@ Mahony and Benos, 2007)). Currently a handful of tools are specialized in motif clustering: STAMP (Mahony et al., 2007), m2match (Stegmaier et al., 2013), MATLIGN (Kankainen and Löytynoja, 2007), GMACS (Broin et al., 2015), DMINDA (Ma et al., 2014a) and motIV (Bioconductor package). However, each of these tools presents some limitations: analysis based on a single metric, restricted number of input motifs, static visualization interfaces.

During this thesis I developed a tool called *matrix-clustering* as part of RSAT motivated by the need of a tool to cluster similar motifs with a strong visualization component. In addition to the clustering step, the motifs are aligned and the clusters are represented in different ways (alignment of consensuses and motif logos, heatmaps) and many collections can be used as input, if this is the case, *matrix-clustering* will compare as well the input collections measuring the similarity among them. The final result is an interactive website where the users can browse the motifs including a file with the non-redundant motifs.

There is a large list of metrics to measure motif similarity, usually the tools are limited to use the results from a single metric, in *matrix-clustering* many comparison metrics can be used in order to group the motifs, and we demonstrated that a threshold to separate the motifs based on two similarity metrics makes a better separation of the motifs than thresholds based on a single metric.

In the publication (Castro-Mondragon et al., 2017), we show four applications of *matrix-clustering*:

- Integration of results from multiple motif discovery tools (RSAT *peak-motifs* (Thomas-Chollier et al., 2012b,a), MEME-CHIP (Machanick and Bailey, 2011; Ma et al., 2014b), and HOMER (Heinz et al., 2010)) in a single analysis. The alignment of motifs allowed us to detect many binding variants (homo and hetero-dimers) for the TF OCT4 (Tantin et al., 2008).

- Integration of results from motifs discovered in 12 ChIP-seq peaksets. The results integrate the information in a visual way to detect motifs found exclusively in one peaksets or motif found on peaks of functionally related TFs.

- Identification of motifs belonging to the same TF Family. We used the motifs from HOCOMOCO (Kulakovskiy et al., 2016) because they include the TF Family information taken from TFClass (Wingender et al., 2013). We show that a complete collection of motifs can be reduced to a set of non-redundant motifs, where each motif represent a FBP. We also noted in these results that some families as Zinc Fingers, have one motif for each of its members, according to the reported wide variability of binding motifs (Najafabadi et al., 2015).

- Creation of taxon-wise motif databases. We collected several motif collections for insects, plants and vertebrates. For each taxon, the total number of collected motifs was reduced to a non-redundant collection with ~20% of the original size. We measured the inter-similarity between the databases and we noted that although many of them have similar content, few databases has unique content. As part of the results, we made available the non-redundant motif collections.

9.2 Contribution

I developed the algorithm integrating the results of *compare-matrices* (Thomas-Chollier et al., 2011a) with the further clustering step and the partition of the resulting tree. I also developed the interactive website with advice from Morgane Thomas-Chollier and Denis Thieffry. I wrote the first version of the manuscript and participated in the correction of the revised versions. I collected the motif databases used in the publication and made them available through the RSAT website (Medina-Rivera et al., 2015).

9.3 Conclusion

I think that as long as large collections of motifs are produced or handled, the clustering will become a necessary step for the analysis, it is specially useful for approaches affected by redundancy as motif comparison and enrichment. At the moment, there is only one tool that shows the motifs grouped by similarity, that is MEME-ChIP (Ma et al., 2014b) but the results of others tools as RSAT *peak-motifs* (Thomas-Chollier et al., 2012a) can be improved showing the motifs by clusters.

Regarding the motif databases, currently there is only one database, that is RegulonDB (Gama-Castro et al., 2015) for *Escherichia coli K12* transcriptional regulation, where the TFs are shown by similarities or by TF Families, however for the remaining databases a clustering of motifs should be useful to highlight the similarities of motifs. I believe that the interactive visualization of *matrix-clustering* can be integrated in maintained databases as JASPAR (Mathelier et al., 2015), HOCOMOCO (Kulakovskiy et al., 2016) or FootprintDB (Sebastian and Contreras-Moreira, 2014) in order to improve the searching of motifs.

For the moment, *matrix-clustering* does not support the di-PSSM, only one tool MACRO-APE (Vorontsov et al., 2013) is able to compare di-PSSM. As long as novel models become more used (e.g., di-PSSMs (KULAKOVSKIY et al., 2013; Mathelier and Wasserman, 2013; Siebert and Johannes, 2016) or PSSMs for methylated sequences (Viner et al., 2016; Ngo and Wang, 2016)) I plan to extend the algorithm to analyze these new PSSMs.

Chapter 10

RSAT::Plants: Strategies for Motif Discovery in Plant Genomes

10.1 Motivation and state of the art

During this thesis I participated in the publication of two protocols for motif discovery in plant genomes, one for ChIP-seq peaks and other for promoters of co-expressed genes. Plants are well known for the high percent of repetitive elements (RE) distributed along their genomes (Biscotti et al., 2015), this particularity makes a challenge for the motif discovery on these genomes, therefore more strategies should be considered than only use *de novo* motif discovery tools.

For the first protocol, dedicated to motif analysis in ChIP-seq peaks (Castro-mondragon et al., 2016) we highlight the importance of the following points:

- Use distinct motif discovery approaches (e.g., over-represented and positionally biased k-mers) in the same analysis, although both approaches can find redundant motifs, each one brings different information about the found motifs.
- Use clustering of motifs approaches *a posteriori*. Once the motifs have been found the redundant motifs can be discarded in order to create a non-redundant set of motifs.
- Use of negative control for motif discovery. In the case of plant genomes, REs are randomly distributed in the genome, including the cis-regulatory regions (e.g., promoters or enhancers). In these cases, the negative controls could help to discard those motifs found as consequence of the REs but that may not bind any TF.

Those motifs found either on the sequences of interest (query) and in the negative control sequences, may result from the repetitive sequences but also for over-represented k-mers in the whole genome. Therefore we could consider them as non-specific of the query sequences and then discarded, they are discovered simply because the genome composition and might not correspond to binding motifs.

For the second protocol, dedicated to motif analysis on promoters of co-expressed genes (Contreras-moreira et al., 2016) we highlight the importance of the following points:

- Use distinct motif discovery approaches (e.g., over-represented k-mers and spaced k-mers) in the same analysis.
- Use of negative control(s) for motif discovery (e.g., several replicates of motif discovery in groups of randomly selected genes of the same size as the input co-expression clusters).
- Use the motif comparison metrics as criteria to select relevant motifs. For example, the distribution of comparison scores of the discovered motifs against motif databases has higher values than the

distribution of scores of the motifs found at random sets of genes).

- Use the distribution of the discovered motif scores (e-values) either in promoters and controls as criterion to select significant motifs.

Regarding the negative control used, for the ChIP-seq analysis, we randomly selected fragments from a reference genome with the same same length as in the sequence query set. Given that these sequences contain a mixture of different types of genomic regions (i.e., including coding, intergenic, centromeric, or cis-regulatory regions) we expect to discover (or better if we do not) different motifs relative to those motifs found on the query sequences. For the analysis of promoters of co-expressed gene sets we used promoters of randomly selected genes, therefore, although they are still regulatory regions we do not expect to find over-represented motifs, since these promoters are not functionally related.

The use of negative control should it is an empirical validation of the statistical model, and can serve as evaluation of the specificity of the motifs. When we use a set of sequences as negative control, some the motifs could be discovered in both sequence sets, a further clustering of motifs will allow to visualize the common motifs between both sequence sets.

The motif comparison is normally used to annotate the discovered motifs using already known motifs stored in databases as FootprintDB (Sebastian and Contreras-Moreira, 2014), JASPAR (Mathelier et al., 2015) and Cis-BP (Weirauch et al., 2014), in one of these protocols we showed that the distribution of comparison scores can be used, in combination with the significance of the discovered motifs as criterion to detect relevant motifs and discard those that could be artifacts (e.g., low complexity motifs with repetitions of a single nucleotide).

The distribution of motif scores might suggest if a motif is an artifact, since we expect highly significant scores for motifs in functionally related promoters, and low significant motifs obtained from random promoters. If we find more significant motifs in random promoters than in the query promoters, this could indicate such artifact motifs.

These protocols shows complementary strategies not applied in the previous protocol of RSAT *peak-motifs* (Thomas-Chollier et al., 2012a) or RSAT *oligo-analysis* (Defrance et al., 2008), these strategies can be considered for users interested in plant or other genomes with a high percentage of REs.

10.2 Contribution

For these protocols I developed a workflow for motif analysis integrating several tools from RSAT, in collaboration with Claire Rioualen and Bruno Contreras-Moreira. We used the software *make* to make it reproducible by the users on command-line. I participated on the writing of the two protocols and generated some of the figures.

10.3 Conclusion

The analysis of motifs goes beyond running a program and find the motifs, depending on the genome analyzed the users could be confronted with some additional challenges (e.g., ER), that can be intrinsic to the genome studied, therefore a single motif analysis tool cannot overcome these challenges by itself. However, using the results from separated tools can be helpful to complement and give robustness to the results. In addition, including negative controls and clustering of motifs improves the quality of the analysis.

At this time ChIP-seq analysis workflows as RSAT *peak-motifs* (Thomas-Chollier et al., 2012b), MEME-CHIP (Machanick and Bailey, 2011) or XXmotif (Luehr et al., 2012) can be expanded to include these strategies: (i) the clustering of motifs (only MEME-CHIP does this task) or (ii) the generation of negative controls to improve the analysis. The importance of integrate these strategies on automated workflows ease the analysis for scientist not familiar with programming.

These protocols showed that the comparison of motifs can be used not only to compare and annotate the motifs, but also to reduce the motif redundancy and help to identify motifs found on two conditions (e.g., ChIP-seq peaks and negative control regions), the integration of this strategy in RSAT *peak-motifs* will be helpful for the users.

Chapter 11

RSAT *position-scan* : identification of transcription factor binding sites with positional bias

11.1 Motivation

During this thesis, I have collaborated with two groups of biologists interested in the regulatory mechanism of human promoters with enhancer activity (Spicuglia's team) and detection of regulatory motifs on ChIP-seq peaks for Polycomb Repressive Elements and for some TFs (Cavalli's team). In both collaborations, it was required to detect motifs that were locally enriched (near promoters for the first project and at the center of the peaks in the second one). Since the current available tools are specialized on ChIP-seq peaks, and only consider the strongest TFBSs per sequence for their statistical analysis I developed an algorithm called *position-scan* that detect TFBS heterogeneously distributed in the sequences, therefore detecting either enriched and depleted motifs in a single run.

Given that TFBSs with low or median affinity may contribute to the regulation, this algorithm considers all the TFBSs (not only the strongest per sequence). In addition, several representation of the positional enrichment are shown, and the motifs are clustered based on their positional distribution. The results are displayed in a dynamic interface that allow users to select motifs based on their positional profile, significance or enrichment/depletion of the TFBSs. This algorithms was used to detect motif enriched in human promoters with enhancer activity (see chapter 13).

The following text correspond to a draft that I wrote and I plan to publish it after my PhD defense.

11.2 Introduction

Transcription Factors (TFs) are DNA-binding proteins that control gene expression, they bind short sequences called TF binding sites (TFBSs) that are usually located at cis-regulatory regions. The TFBSs can be detected *in silico* using Position Specific Scoring Matrices (PSSMs, simply called motifs), that are models to represent the TF binding affinities.

The advent of high-throughput technologies allows to detect experimentally thousands of sequences bound by a TF at genome-wide scale, for example using ChIP-seq (Jothi et al., 2008) or ChIP-exo (Rhee and Pugh, 2011). In these cases the TFBSs can be analyzed relative to a reference position, e.g. peak center, peak summit. Approaches known as positional motif enrichment focus on the detection of over of TFBSs at certain positions of the analyzed sequences. Although currently is a popular approach, the idea is not recent

and was originally used to study TF and RNAP binding motifs near promoters (Bucher and Bryan, 1984), i.e., relative to Transcription Start Site (TSS).

For example in ChIP-seq and ChIP-exo, it is expected that most TFBSs for the ChIP-ped TF are found (positionally biased) around the peak summit, i.e. the position with of maximal read coverage (Bailey and Machanick, 2012; Worsley Hunt et al., 2014; Thomas-Chollier et al., 2012b). Similar examples of positionally constrained motifs are observed around TSSs at gene promoters (Whitaker et al., 2015), at the upstream region of introns (Yeo et al., 2007) and replication origins (Cayrou et al., 2015). Although the enrichment of motifs is commonly studied, recent studies have noted that the depletion (under-representation) of motifs at certain positions has consequence in gene regulation (Whitaker et al., 2015; Telorac et al., 2016).

The measure of enrichment/depletion is based on the detection of TFBSs from collections of already known motifs, for example JASPAR (Mathelier et al., 2015) or HOCOMOCO (Kulakovskiy et al., 2016), making the analysis exhaustive for thousands of motifs, which is a big difference relative to the matrix-based motif discovery methods.

Currently, specialized software has been developed to detect motif positionally constrained at ChIP-seq peaks based on a library of known motifs, that are CentriMo (Bailey and Machanick, 2012) and TFBSLandscape (Worsley Hunt et al., 2014). These tools are focused on the positional motif enrichment, however they do not consider motifs that are positionally depleted. To the best of our knowledge, there is only one motif discovery method that detects under-represented positionally constrained motifs (and over-represented as well), that is *position-analysis* (van Helden et al., 2000; Thomas-Chollier et al., 2012a).

Given the recent observation of positionally depleted TFBSs and the lack of tools to detect them, we developed *position-scan*, a method to detect either enriched and depleted positionally constrained motifs in large set of sequences. It can be used to detect positionally constrained motifs in a single set of sequences or differentially in two sets (query and control). This program can be used to analyse *de novo* motifs detected in high-throughput sequences (e.g., ChIP-seq) or a complete motif database. *position-scan* was developed at the end of this thesis work and it has not been published yet, however it is already publicly available is part of the RSAT suite (Medina-Rivera et al., 2015) and can be used via website or as stand-alone tool (command-line) for its integration in workflows of motif analysis.

11.3 Material and methods

11.3.1 Input formats

position-scan supports different motif formats: TRANSFAC (default), MEME, HOMER, JASPAR, etc. The users can easily inter-convert the motifs to different formats using the tool RSAT *convert-matrix* (Thomas-Chollier et al., 2011a). Given that users would scan many motif databases (e.g., JASPAR, HOCOMOCO) in a single analysis, one or more motifs collections (in separate files) can be provided, each one with a given collection name to ease the identification of the motifs in the results.

11.3.2 Motif scan

The sequences are scanned using the program RSAT *matrix-scan* (Defrance et al., 2008; Turatsinze et al., 2008) with a user-specified threshold on p-value (default: 10^{-3}) to detect the TFBSs. By default *position-scan* generates background models from the input sequences using a Markov chain of order 1, whose the transition frequencies are estimated from the input sequences. This capability to use a Markov background model is important to account for interdependencies between adjacent nucleotides (e.g. CpG depletion in vertebrate genomes, poly-A or poly-T enrichment in non-coding sequences). Users are also allowed to load custom background models.

position-scan can be operated in two modes: (1) consider all the detected instance of each motif; (2) similarly to CentriMo (Bailey and Machanick, 2012) and TFBSLandscapes (Worsley Hunt et al., 2014), it only consider

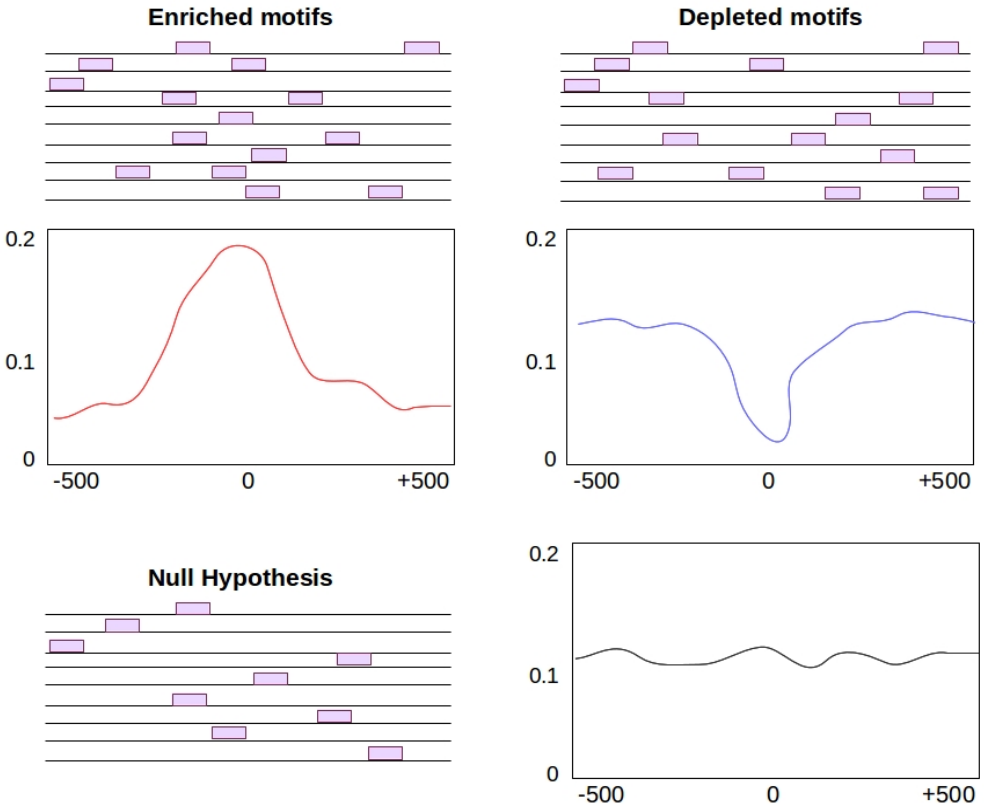


Figure 11.1: Graphical representation of distribution of TFBSs: enriched and depleted at the center of the sequences and the null hypothesis where the TFBSs are distributed homogeneously.

the best hit per sequence. In addition, the user can select to scan both or a single strand, in cases when the orientation of the sequences must be considered.

11.3.3 Chi-square test

The sequences are divided in bins (non-overlapping windows) of a given size (default: 25nt). For each motif, the number of instances per bin is counted. The sum of TFBSs is then divided by the number of bins in order to obtain an estimate of the expected counts per bin under null hypothesis (i.e., assuming that the sites are distributed homogeneously throughout the sequences).

$$E_k = \frac{\sum TFBS}{k}$$

where E_k corresponds to the expected sequences at the bin k.

Those counts deviating the null hypothesis correspond to positionally constrained motifs (enriched and depleted) (Figure 11.1). In case the expected number of TFBSs does not satisfy the chi-square applicability (the expected number at each class should be at least five), this warning is indicated in the results. The p-values produced by the chi-square test are corrected by the number of analyzed motifs, producing thus an e-value.

The chi-squared formula is defined as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O and E correspond to the observed and expected counts (i.e., TFBSs), respectively, at each bin k.

When two sequences are given as input, the frequencies of one of these sequences are used as expected frequencies for the chi-square conformity, and are calculated as follows:

$$E_k = \frac{\sum_{i=1}^{k} TFBS_k}{\sum k}$$

where E_k corresponds to the frequency of TFBSs at the bin k.

11.3.4 Motifs and sequences for the case studies

For study case one I used the complete JASPAR (Mathelier et al., 2015) vertebrate collection (519 motifs). I downloaded the IRF1 and STAT1 ChIP-seq peaks from ReMap (Griffon et al., 2015), in bed format. The peak summits were extended 300bp to each side of peak centers and the sequences were retrieved from the Human genome version hg19 using RSAT *fetch-sequences*.

11.3.5 Implementation

position-scan is implemented in Perl and R. The dynamic profiles are implemented in HTML5 using the JavaScript library C3 (http://c3js.org/). Motif logos are produced using weblogo (Crooks et al., 2004).

11.4 Results

11.4.1 *position-scan* overview

position-scan takes as input a collection of motifs and one (single-set analysis) or two (test versus control) sets of sequences of identical lengths. The sequences are first scanned with the motifs and then divided in bins of a given size.

In single-set mode, the program runs a chi-squared homogeneity test, in order to detect motifs whose TFBSs are distributed heterogeneously in the sequences relative to some reference position (start, center, end), assuming a homogeneous distribution of the TFBSs. This test enables to detect not only local enrichment but also local depletion or more complex heterogeneous profiles (e.g. two enriched peaks separated by a local depletion).

An alternative modality is to specify a control set of sequences in addition to the query sequence file, in which case the positional profiles of TFBSs of the control sequences are used as expected frequencies for a chi-squared conformity test.

The visual interface is dynamic and enables users to select a subset of motifs based on their significance, the shape of their profiles, select clusters of motifs with similar profiles, or perform a selection of motifs of interest based on their individual profiles (Figure 11.1).

11.4.2 Visualization

position-scan produces three different ways to visualize the positionally constrained motifs: TF binding profile plot, qualitative distribution of TFBSs and heatmap with positional profiles.

11.4.2.1 TF binding profiles

For each motif, the number of TFBSs per bin is summed and the result is normalized by the number of sequences with at least one TFBSs. All the profiles are shown in an interactive plot where the users can choose which profile to display (Figures 10.2 and 10.3a).

11.4.2.2 Qualitative distribution of TFBSs

For each motif, all the predicted TFBS are shown in a scatter-plot where the abscissa corresponds to the position relative to a reference (i.e., the center) and the ordinate to the significance ($-log10(pval)$). The TFBSs are coloured according their significance: a dark color for the most significant a a lighter color for the less significant (Figures 10.2 and 10.3b).

11.4.2.3 Heatmap of positional profiles

All the positional profiles are clustered and represented as a heatmap (Figures 10.2 and 10.3d). The profiles are clustered based on the Pearson Correlation Coefficient and the ward method is used as linkage rule for the hierarchical tree.

11.4.3 Case study 1: positionally constrained motifs in STAT1 ChIP-seq peaks

STAT1 is a TF involved in interferon pathway that can form homo- and hetero-dimers with members of its TF Family (STAT). I chose to study since the dimers may have differences that can be reflected in the motif analysis of positionally enriched motifs (Ehret et al., 2001).

I ran *position-scan* with a merged collection of ChIP-seq peaks made from several experiments taken from ReMap (Griffon et al., 2015) and the 519 matrices from the complete JASPAR vertebrate motif collection (Mathelier et al., 2015). The results sorted by chi-squared p-value ranked Stat4, Stat3 and Stat1 as the motifs deviating most significantly from the homogeneous distribution, followed by motifs from the JUN/FOS family and CTCF.

The distribution of TFBSs coloured by significance (Figure 11.4) highlights those motifs whose strongest TFBSs are concentrated at certain location, for example, at the center of the sequences. Regarding the Stat motifs, in the cases of Stat1 and Stat4 which have similar motifs, the strongest TFBSs are concentrated at the center of the peaks (Figures 10.4a and 10.4c), similarly for the Stat1::Stat2 dimer, although it has a different motif (Figure 11.4b). The strongest TFBSs for Stat6, however, are not concentrated at the peaks center (Figure 11.4d). In order to visualize the differences between the Stat1, Stat4 and Stat6 motifs I used *matrix-clustering* (Castro-Mondragon et al., 2017), the alignment shows that Stat1 and Stat4 have a 2bp spacer whilst Stat3 has a 3b spacer. This difference is reflected in the distribution of TFBSs (Figure 11.4e).

11.5 Discussion

I present RSAT *position-scan* a method to detect motifs with positional bias in large set of sequences aligned on some reference position. This method can be typically used for motif analysis of ChIP-seq or promoters where a high concentration of TFBSs are expected to be found at around peak centers or near

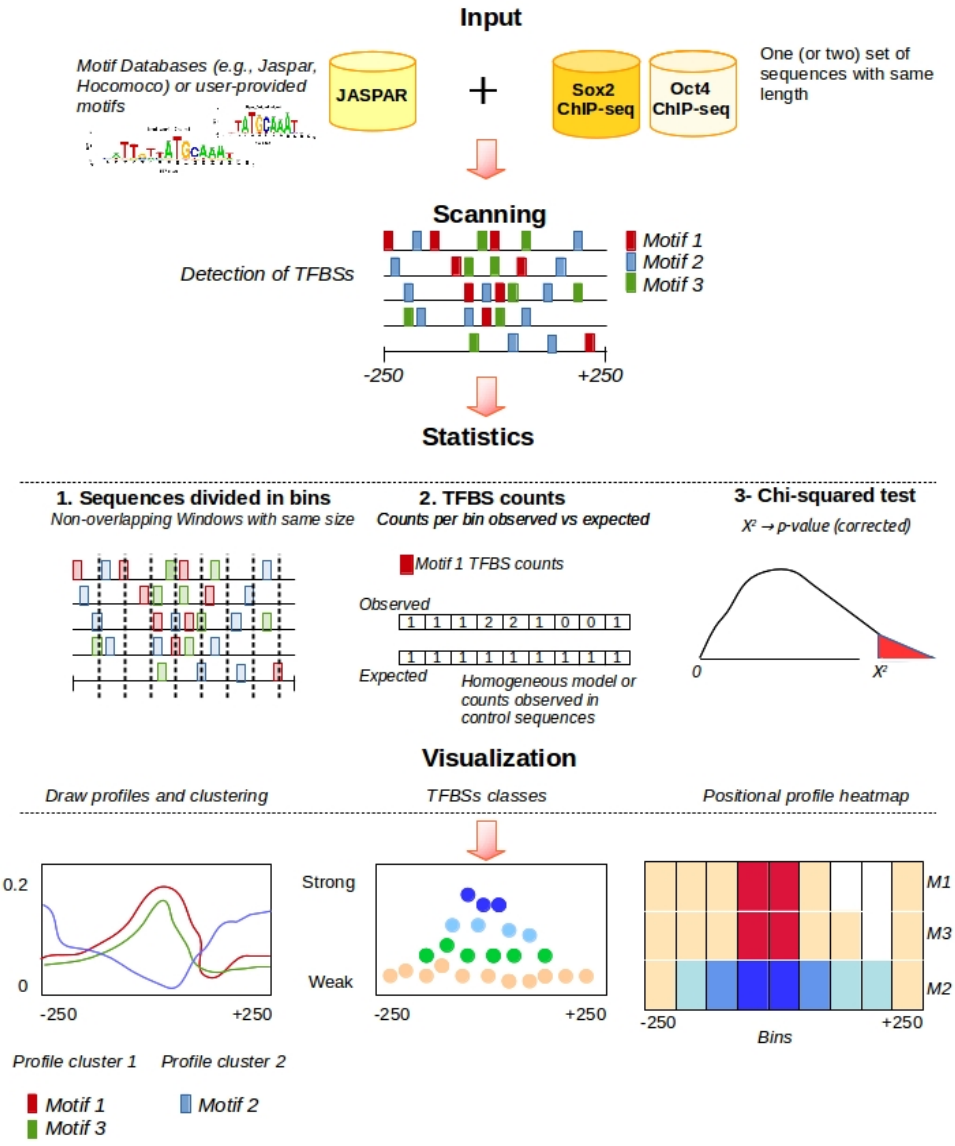


Figure 11.2: Schematic flow chart of the position-scan algorithm. The program takes as input one (or several) collection(s) of motifs, and one or two sequence sets with the same length. Each motif is scanned and the sequences are divided in bins of the same lenght. For each motif the program sums the TFBS counts per bin and the total is distributed homogeneously (null hypothesis) and both distributions (observed and expected) are compared using a chi-square test. The distribution of TFBSs is visualized in three ways: (i) the binding profile showing the frequency of TFBSs per bin, (ii) the heatmap and clustering of the binding for all the motifs, and (iii) the distribution of TFBSs separated by classes (from weak to strong TFBSs) of every motif.

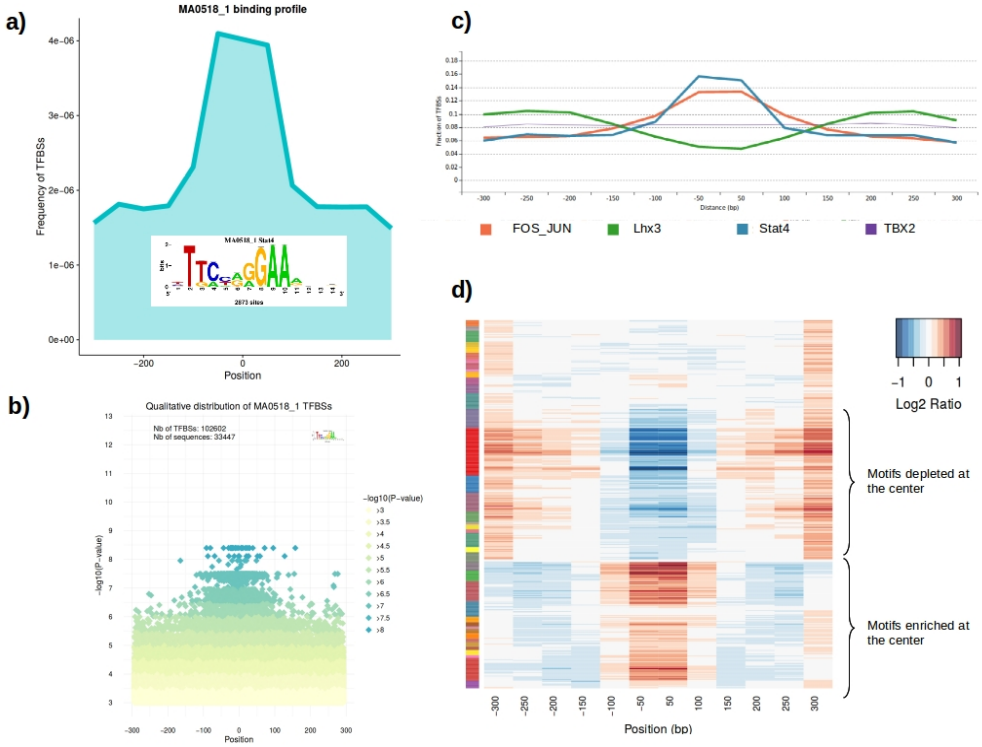


Figure 11.3: Examples of position-scan results for STAT1 ChIP-seq peaks. (a) Example of TFBS binding profile for the motif STAT4. (b) Qualitative distribution of TFBSs of STAT4 motif in the sequences. Every class of TFBSs is coloured from yellow (weak sites) to blue (strong sites). (c) Examples of binding profiles for four different motifs: JUN FOS and STAT4 enriched at the center; Lhx3 depleted at the center; TBX2 follows and homogenoeus distribution. (d) Heatmap with clustering of binding profiles showing motifs enriched and depleted at the center of the peaks. The colours at left indicate the clusters.

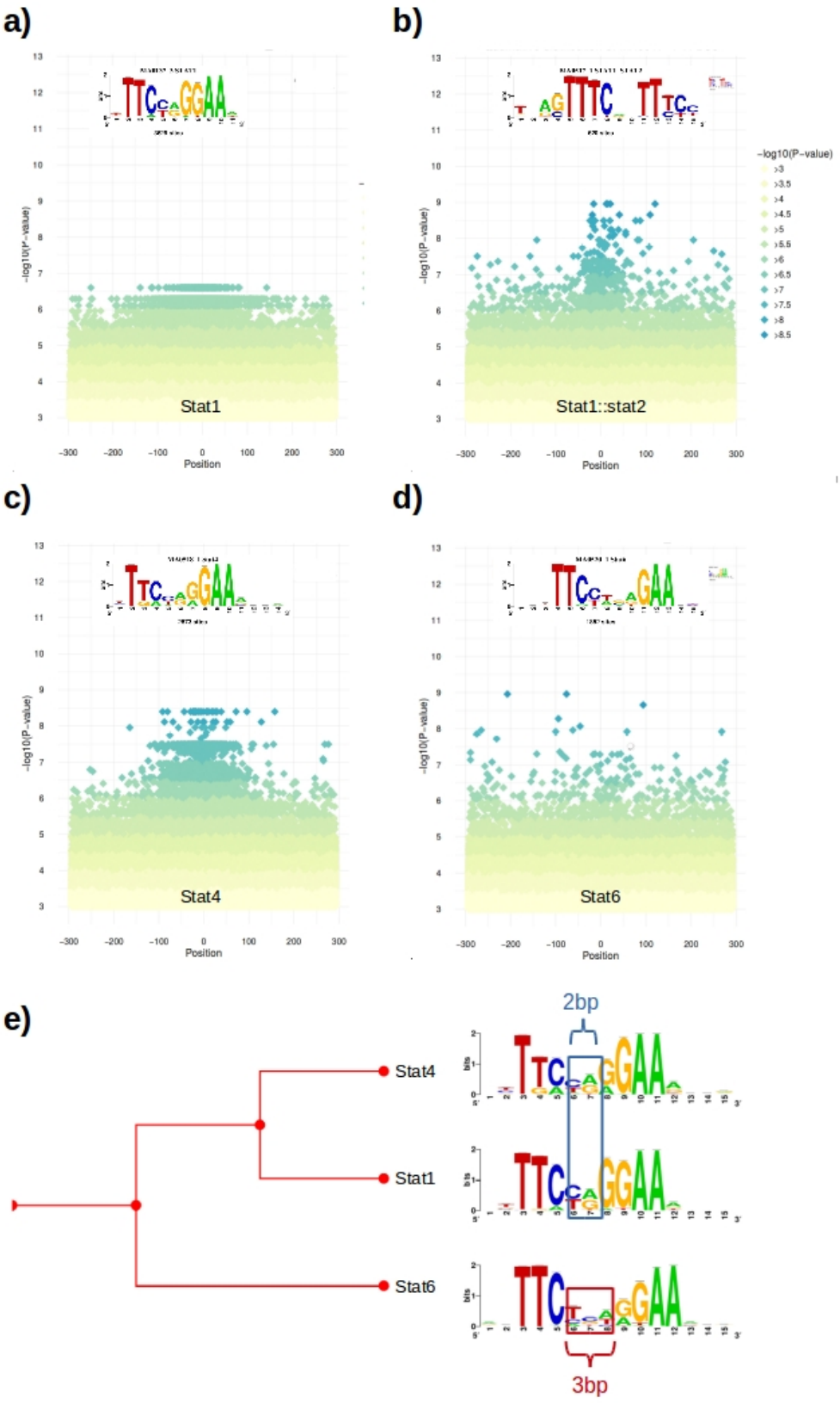


Figure 11.4: Distribution of TFBSs for a) Stat1, b) Stat1::Stat2, c) Stat4, and d) Stat6. The color scale indicates the -log10(pvalue) of the predicted TFBSs. e) Alignment of Stat motifs. Note the 2 and 3bp spacer and the difference in the distribution of strongest TFBSs.

TSSs, respectively. *position-scan* follows the similar principle than *position-analysis* (van Helden et al., 2000; Thomas-Chollier et al., 2012a), both programs detect positionally constrained motifs, however the later is a motif discovery algorithm whilst *position-scan* use the counts of TFBSs from known motifs and can be used as a topological motif enrichment tool. For such large data sets (e.g. 10,000 peaks of 600bp each), the usual multiple testing corrections would impose a stringent threshold on TFBS p-values. However, the rationale of the tool is not to assess individual binding sites, but to scan sequences with a relatively lenient threshold (e.g. 10^{-3}), and to test, for each motif, the positional distribution as a whole.

Currently three others methods already exist (CentriMo (Bailey and Machanick, 2012), TFBSLandscape (Worsley Hunt et al., 2014) and ChIP-seq tools (Ambrosini et al., 2016)), however they focus on enriched motifs and they only consider the top-scoring TFBSs per sequence for their enrichment statistics. In this method, the analysis is not limited to the best match per sequence, but it considers all the TFBSs. In addition, given that *position-scan* uses the homogeneous distribution of TFBSs as the null hypothesis, the method can detect not only local enrichment, but also local depletion or more complex patterns (e.g. alternances of enriched and depleted windows).

The visualization of the results allows to identify those TFs following similar binding profiles for their TFBSs in the analyzed sequences, which often result from similarities between the motifs themselves, but can also reflect interactions between TFs recognizing distinct motifs acting in the same condition (e.g., cooperative binding or co-occurrence of TFs). In some cases, the observation of both enriched and depleted motifs might suggest the mutual exclusion of two factors as was observed by Telorac and co-workers (Telorac et al., 2016), where a set of sequences (although not bound by TFs) avoid the binding of the glucocorticoid receptor. Note that this vision is a bit too mechanistic. Depleted motifs may be of poor-complexity, and their depletion reflects a compositional bias of the regions bound by a factor (a context-dependent composition) rather than the specific exclusion of a particular factor.

In the current version, *position-scan* returns motifs based on the global shape of their position profile. In future releases I will update the program to also report the particular windows of enrichment/depletion.

There is no obvious way to choose the optimal bin size, which depends on the sequence lengths, their numbers, and the p-value threshold, and should thus be chosen on a case-per-case basis.

Since this tool can scan thousands of motifs from several databases (e.g., JASPAR, HOCOMOCO), this tool could be affected by motif redundancy by making the running time longer, however, the chi-squared p-values can be sufficiently small to afford for multiple testing corrections.

11.6 Conclusion

The detection of positionally constrained motifs is of special interest for the motif analysis of large sequences data sets as ChIP-seq (or similar technologies) and motifs locally enriched/depleted relative to a reference position (e.g., promoters, replication origins). This program can be used validate discovered motifs or to identify already known motifs that were not find by motif discovery algorithms.

Chapter 12

RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond

12.1 Motivation and state of the art

RegulonDB (http://regulondb.ccg.unam.mx) is one of the most useful and important resources on bacterial gene regulation, as it integrates the scattered scientific knowledge of the best-characterized bacteria, *Escherichia coli K12* in a database that organize large amounts of data since 1998 (Huerta et al., 1998).

One particularity of this database is that the information is curated from the literature manually and semi-automatically, and it integrates genomic annotations for transcriptional regulation, as transcription start and termination sites, TF binding sites and miRNAs, in a strong visualization of these elements in the genome.

Initially, RegulonDB was thought as a catalog of TFBSs for *Escherichia coli K12* TFs but later other layers of information were integrated and nowadays RegulonDB also includes metabolic and functional annotation of the genes.

Regarding the TFBSs stored in RegulonDB they are taken from literature and from low-throughput methods as EMSA or gel retard assays, that bring the exact location of the TFBSs on the bacterial genome, in other words they are made from *bona fide*, yet validated TFBSs which can be used to built PSSMs. It is important to note that RegulonDB does not include PSSM discovered with motif discovery tools, that is the normal procedure for some databases as JASPAR (Mathelier et al., 2015) or HOCOMOCO (Kulakovskiy et al., 2016).

In every version of RegulonDB, as more TFBSs are collected, the PSSM are improved and novel PSSM are made available. The current version (9.0) contains 93 PSSMs. In order to built a PSSM, the minimum number of TFBS must be four, giving to each nucleotide the chance to appear at least once on each position of the PSSM (Medina-Rivera et al., 2011).

Since 1997 there is a strong collaboration between the RSAT and RegulonDB teams, many RSAT programs have been developed using RegulonDB information, e.g., *oligo-analysis* (van Helden et al., 1998), *dyad-analysis* (van Helden et al., 2000), and *matrix-quality* (Medina-Rivera et al., 2011).

Given that the number of PSSM stored in RegulonDB is growing, the latest version of RegulonDB includes many tools to visualize information, among them there are two motif browsers based on the TF Families

(Pérez-Rueda et al., 2015) and from motif similarity using *matrix-clustering*. This visualization highlight the similarities (and differences as well) for TF belonging to the same family.

12.2 Contribution

I ran *matrix-clustering* with the novel RegulonDB PSSMs and the output was integrated in RegulonDB version 9.0. In addition, I also implemented a PSSM tree browser where the logos of each PSSM are depicted in the leaves. clicking on the logos the users can go directly to the PSSM information website. I contributed with the writing of the manuscript on the part related with the clustering of motifs and the motif browser.

12.3 Conclusion

For the analysis of bacteria genomes, in my opinion the high-throughput experiments have not the same success as they have for metazoa genomes, of course that many genome-wide experiments have been done, but they are scarce, and there is no one database collecting all the information, but currently RegulonDB team is working on the integration of these results in the database.

The scarce of high-throughput methods is mainly due to the biology of bacterial TFs, some of them recognize a handful of sequences in the genome, and a genome-wide experiment could not be required for such TFs.

Although currently RegulonDB contains 93 PSSMs, if the results from high-throughput experiments are integrated we will observe an increase in the number of PSSMs (including redundant motifs discovered from different algorithms and experiments). The integration of a browser based on motif similarity or TF Families will allow to easily integrate novel discovered PSSMs and will ease the search of TFs.

The integration of *matrix-clustering* in RegulonDB shows that other motif databases, even those containing larger number of PSSM can take advantage of the interactive output in order to browse the motifs by similarity.

Chapter 13

Genome-wide characterization of mammalian promoters with distal enhancer functions

13.1 Motivation and state of the art

In mammals, transcriptional regulation is driven by cis-regulatory sequences and regulatory proteins (TFs, RNAP, GTFs) and RNAs. Regarding the cis-regulatory sequences that positively regulate gene expression, they have been classified according their distance relative to the TSS of their associated genes: the promoters and enhancers regulate genes proximally and distally, respectively (Kim and Shiekhattar, 2015).

This basic definition of promoters and enhancers has been challenged by recent studies showing similarities among them, for example both enhancers and promoters can recruit TFs and RNAP, produce bidirectional transcripts, and are associated with open-chromatin histone marks (Andersson et al., 2014; Pennacchio et al., 2013; Andersson et al., 2015). Altogether these evidences suggest that promoters might play enhancer functions (i.e., regulate distal genes) and it has been demonstrated in isolated cases, however it is unknown what fraction of promoters may have enhancer activity involved in distal regulation.

The massive detection of enhancers have evolved from detection of enhacer activity on synthetic sequences (Kheradpour and Kellis, 2013) to novel technologies capable of measure the activity in genomic conditions. One of these novel methods, STARR-seq (Arnold et al., 2013; Muerdter et al., 2015), allows to detect sequences with enhancer activity based on the sequence itself (i.e., by function) and not by epigenomic features or location criteria, however this method was developed and tested in *Drosophila* genomes or in human cell lines using BACS (Arnold et al., 2013), it has not been adapted to study larger genomes.

In order to study genome-wide enhancer activity in mammals, Spicuglia's team adapted the STARR-seq with a capture step that allows to enrich the interest regions before measure the enhancer activity, the new method is called capStarr-seq and was validated in mice cell lines (Vanhille et al., 2015).

Now, for the current project we used the capStarr-seq method in order to measure the enhancer activity for all the humans promoters (~20,000) defined by RefSeq in two cell lines (HeLa and K562). We found that 2-3% of the promoters display enhancer activity in the analyzed cell lines. These TSS-overlapping enhancers (hereafter called Epromoters) display specific genomic and epigenomic features that differs from either enhancers and promoters, in addition these Epromoters were associated with TF and genes related to stress response. Even a small set of Epromoters was identified after stimulation of interferon. By using CRISPR/Cas9 deletions we demonstrated that Epromoters are involved in cis-regulation of distal gene expression in their natural context, therefore functioning as *bona fide* enhancers (Dao et al., 2017).

We suggest that regulatory elements playing a dual role as transcriptional promoters and enhancers might ensure rapid and coordinate regulation of gene expression upon stress response.

13.2 Contribution

This work is a collaboration between experimental biologists and bioinformaticians, many of the results obtained from bioinformatic analysis were further experimentally validated. My contribution was at the bioinformatc side, here I developed a workflow to analyze the Epromoter sequences combining several tools from RSAT (Medina-Rivera et al., 2015), this workflow was implemented in snakemake (Köster and Rahmann, 2012) and allow to reproduce the results showed in the publication (github.com/arielgalindoalbarran/Epromoters).

The motif analysis was separated in the following steps:

- Motif discovery: The motifs were discovered with RSAT *peak-motifs* (Thomas-Chollier et al., 2012b,a), which run several motif discovery algorithms in the same analysis.

We discovered many already known motifs (e.g., Jun, YY), however the analysis was not comprehensive since it did not include the information about the already known motifs stored in databases, for this reason we change the analysis to a motif enrichment approach, which enable us to study all the known TF motifs in a single run, rather than limit the analysis to the motif discovery approach.

- Motif clustering: We used the motifs stored in JASPAR vertebrates (Mathelier et al., 2015) (519 motifs) and HOCOMOCO human (Kulakovskiy et al., 2016) (622) motifs because their motifs are curated and built from high-throughput data, however as we merged these databases we have redundant motifs. In order to avoid such redundancy we run *matrix-clustering* and obtained 489 non-redundant motifs that were used for further analysis in this project (e.g., detection of regulatory variants).
- Motif enrichment: We looked for motifs enriched at the Epromoter region (-250 to +50 relative to the TSS), therefore a positional motif enrichment approach should be the solution. However when this project started, the tools for positional enrichment (CENTRIMO (Bailey and Machanick, 2012) and TFBSLanscapes (Worsley Hunt et al., 2014)) were specialized on ChIP-seq peaks, considering only the strongest TFBS per sequence. As we wanted to consider all the TFBS detected, I developed a script that later became the tool *position-analysis* (see chapter 8) that can detect over and under representation of TFBSs positionally constrained.
- Negative controls: We ran a negative control on every step of the workflow.

The combination of these programs showed a set of enriched TFs specifically on the Epromoters (and not in the promoters with no enhancer activity). In addition, I also participated in the writing of the methods of the manuscript.

13.3 Conclusion

The main contribution of this study to the current knowledge of the enhancers and promoters is that it reveals a proportion of promoters displaying either local and distal regulation, by contrast to previous studies showing isolated cases of promoters regulating distal genes.

The fact that many Epromoters are associated with stress response genes is in agreement with previously validated enhancers associated to rapidly induced genes (e.g., viral immediate early genes, heat shock genes and the anti-viral interferon genes) that are located near their associated TSSs (Schaffner, 2015). In addition, we also observed that the enhancer activity is not correlated with gene expression. However, both observations supports the transcription factory model were genes are located closely to each other in order to increase the concentration of regulatory factors or RNAP (Feuerborn and Cook, 2015), although for the moment the possible contribution of Epromoters to transcription factories remains unknown.

The fact that many Epromoters are related with interferon response, and even some of them displayed the enhancer activity after interferon stimulation, suggests that the enhancer activity is condition-dependent and the number of detected Epromoters might be underestimated. Future studies focused on specific stimulation should be done in order to reveal novel Epromoters and regulatory mechanisms.

The results suggest that could be two classes of Epromoters: Epromoters displaying both enhancer and promoter activity (e.g., regulating their associated and close genes) and Epromoters acting independently as enhancer or as promoters.

The (cap)Starr-seq is a method useful to measure the enhancer activity based on the genomic properties of the analyzed sequences (e.g., having binding sites for particular TFs)and not based on associations with epigenomic features, as histone marks. The reported enhancers by this method are in agreement with the original functional definition of enhancers, that is a cis-regulatory sequence capable of drive distally the activation of a gene, independently of the orientation (Banerji et al., 1981). However is important to note that many current studies consider as enhancers those regions associated to H3K27ac (Chatterjee and Ahituv, 2017; Heintzman et al., 2009), although it has been observed that when many of these regions are tested using enhancer assays, they do not display the enhancer activity, even a same region can display enhancer epigenomic features in one cell line and promoter features in other (Leung et al., 2015). In my opinion, the functional definition should be respected for future analysis .

The (cap)Starr-seq method could be adapted in order to study silencers which are less studied and understood in comparison to enhancers, and it could be used also to experimentally validate those enhancers defined by histone marks and not functionally.

The large projects like this one require collaboration of several groups, for experiments and bioinformatics. In regards with this thesis, here is an example of integration of results from several motif analysis tools through a workflow that assures the reproductibility of the results. Furthermore, the development of *position-analysis* is an example that bioinformatic tools should be developed side by side in collaboration with biologists and bioinformaticians.

Chapter 14

General discussions and prospects

14.1 The cis-regulatory code

Since decades, the identification of TF binding motifs and the identification of their targets genes through the detection of TF binding sites, either *in silico* or experimentally, has represented a challenging step for the understanding of regulatory networks. The *in silico* analysis of TF binding motifs is a field of bioinformatics that has been developed since the 1980s, but is currently in fast development, as consequence of the large amount of sequence data produced and the new insights about TF binding (e.g., DNA shape, interactions between TFs and co-factors, genomic context).

In bacteria, the transcriptional regulatory networks have been built using the information of TFs and their target genes (regulons), however, in more complex organisms, as metazoa, these information is not sufficient to create such networks: in metazoa the transcriptional regulation is driven either from close and distal (relative to TSSs) cis-regulatory regions, clusters of TFs bound, usually known as CRMs and the epigenetic modifications. The combination of these elements orchestrate the transcriptional programs that give rise to distinct cell types, developmental stages and distinct responses to environmental changes.

It has been demonstrated that cis-regulatory regions can still recruiting TFs and activating genes even outside of their endogenous genomic context (van Arensbergen et al., 2016; Arnold et al., 2016; Dao et al., 2017), this suggests that the information to drive the transcriptional programs is encoded in the DNA sequences (Meireles-Filho and Stark, 2009), which accessibility for the TFs or other regulatory elements depends on epigenomic mechanisms. Given these facts, it has been proposed the existence of a so-called *cis-regulatory code* stating that the regulatory information follows a set of defined rules, regarding sequence composition (e.g., GC content, enrichment of TF binding sites) and organization (e.g., the additive enhancer activity) (Istrail and Davidson, 2005; Yáñez-Cuna et al., 2013).

The cis-regulatory code also includes the grammar (i.e., presence of certain TFs and their arrangements) in the TF binding sites located at the cis-regulatory sequences. For example, although it has been observed that certain enhancers require a set of specific TFs to activate genes and that functionally related enhancers use to bind similar TFs (Erives and Levine, 2004; Lecellier et al., 2016), it has also been observed that generally the order in which these TF are bound does not have an effect on the enhancer activity (Zinzen et al., 2009; Igg et al., 2007), which is partially explained by motif re-arrangements during evolution (Schmidt et al., 2010). The current view is that the combined input of TFs is more important than the binding site location and orientation.

One of building blocks for transcriptional regulation, according to the *cis-regulatory code* are the TFBSs (Figure 14.1a). Although usually the strongest ones are considered, it is demonstrated that low affinity binding sites at cis-regulatory sequences may be key for gene regulation (Parker et al., 2011). In addition to the affinity, the simple presence/absence of a given TF is not always and indicative of enhancer activity, it has been shown that certain TFs use to work as pairs, for example forming heterodimers or acting as partner

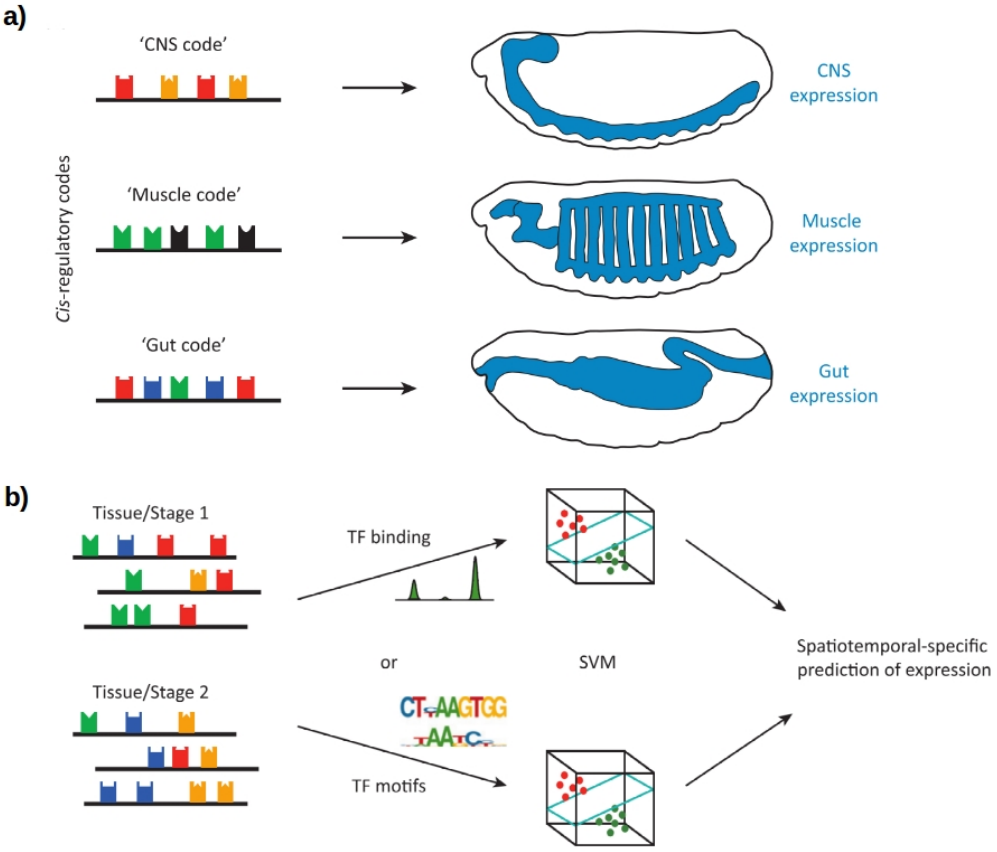


Figure 14.1: Cis-regulatory code: (a) the cis-regulatory sequences have the information to bind TFs that will drive the transcriptional regulation of specific cell lines; (b) such TF combinations (grammar) may be inferred using machine learning methods. Figure taken from Yanez-Cuna (2013),

factors (Heinz et al., 2010; Jolma et al., 2015), such conformation are of special interest to study particular conditions, e.g., cell line or tissue specificity.

Given that the *cis-regulatory code* states that most of the transcriptional information is encoded on the DNA, pattern recognition techniques based on machine learning (Tarca et al., 2007) combined with *in silico* detection of TFBSs have been done in order to identify rules that may drive transcriptional regulation (Figure 14.1b), (Mathelier et al., 2016; Kelley et al., 2016; Whitaker et al., 2015), although given the complexity of the transcriptional regulation, it is important to note that the TF binding depends on additional features such chromatin accessibility, DNA nucleosome occupancy, presence of co-factors or DNA methylation. Considering all of these features makes the study of such code, very hard to understand (Slattery et al., 2014).

14.2 Experimental and computational TFBS detection

The experimental methods to detect the binding preferences of TFs have rapidly evolved from low-throughput (but precise) methods where the exact binding site is revealed, to high-throughput methods (e.g., ChIP-seq) detecting genomic regions where the TF might be directly or indirectly bound and necessarily require further computational analysis to detect the exact location of TFBSs.

In two studies, it was shown that ~60% of the ChIP-seq peaks published by ENCODE (Encode Consortium, 2012) do not contain a binding site for the immunoprecipitated TFs (the sites are detected *in silico*) (Worsley

Hunt et al., 2014; Worsley Hunt and Wasserman, 2014). the remaining sequences are enriched by TFs related to CTCF, ETS, JUN and THAP1. This observation suggests that the detection of these false positive peaks might be a consequence of DNA open region (Yan et al., 2013; Teytelman et al., 2013) (e.g., peaks located near cohesin-bound segments) rather than a *bona fide* binding event of the analyzed TF, other explanations might be the indirect TF binding via protein-protein interactions or because the computational methods to detect these peaks (peak-callers) are far from perfect (Castro-mondragon et al., 2016).

Regarding the *in silico* methods, the main problem is the trade-off between sensitivity and specificity. Even here, it is not obvious that a medium-scoring site should be considered as a false positive. The TF might be bound with lower affinity but still have some preferential binding for this site, and the site might become relevant or not for regulation depending on interactions with other TFs within enhancers (Parker et al., 2011).

The improvement of methods or the development of novel strategies to detect TFBSs is necessary not only to reduce the large amount of false positive, but also because the detection of TFBSs is key for other methods (e.g., motif enrichment) and now is becoming popular for the *in silico* detection of regulatory variants. Future efforts may complement the detection of regulatory variants with the DNA shape and the tools should not be limited to detect only regulatory SNPs, but also indels.

14.3 TF binding motifs representation

The representations of TF binding motifs have evolved from simple character strings (regular expression or IUPAC consensus), PSSM for mono-nucleotides, models based on hidden markov models to the novel di-PSSMs that model nucleotide interdependencies. Every generation of TF binding motifs incorporates more information and therefore the models become more complex. The most recent models, the di-PSSMs are available since 2013 but they are not widely used as the simpler mono-PSSMs. This could be due to the fact that di-PSSMs require large TFBS sets to be built (whilst a mono-PSSM can be built from a handful of and this requirement is not always feasible for some TFs (those with a low number of TFBSs reported). Another possibility is that the the community does not dispose of sufficient elements to prove that dinucleotide dependencies are crucial for many TFs, rather than for a very few selected cases, and that there is thus no strong incentive to rewrite all the existing algorithms. Another reason is that the software incorporating DNA features as DNAshape (Yang et al., 2014) and TFBSshape (Zhou et al., 2013) are simpler than the di-PSSMs, since they require less features to model the TFBSs and the nucleotide interdependencies are already considered, as consequence they became more popular and they are started to be used in recent studies, (i.e., by combining motif scan predictions with DNA shape features), a recent study propose an unifying model representing as logos either the nucleotide preferences and the DNA shape (Yang et al., 2017) (Figure 14.2).

At this time there are two ongoing projects developing a novel representation of TF binding motifs, by including a modified version of the IUPAC alphabet for DNA that includes the epigenetic modifications of cytosine (Figure 14.3), see the following preprints: (Viner et al., 2016; Ngo and Wang, 2016). These modifications may alter the TF binding on the DNA (Hu et al., 2013; Lercher et al., 2014). These new PSSM models are represented as mono-PSSMs with new symbols for the modified cytosines. However in a recent study using a novel method named Methyl-Selex, the authors have demonstrated that the binding preferences (i.e., motifs) for hundreds of TFs may be slightly different when the motifs are built from sequences with and without epigenomic modifications (e.g., 5mC, 5hmC) (Yin et al., 2017). Interestingly, for this publication the authors use the mono-PSSM model, with no extended IUPAC alphabet for methylated cytosine. By contrast, the method proposed by Viner and co-workers (with the extended IUPAC alphabet) requires to known *a priori* the methylated cytosines, that can be detected with specialized bioinformatics tools (Viner et al., 2016). For the moment, the study of methyl-sensitive TFs is limited to mono-PSSM and it should not be surprising if further studies are focused on the study of DNA feature shapes in methylated sequences.

Similarly to the mono-PSSMs, the di-PSSM models are represented in distinct formats to compute the nucleotide interdependencies, however they do not have a unified graphical representation as the logo for

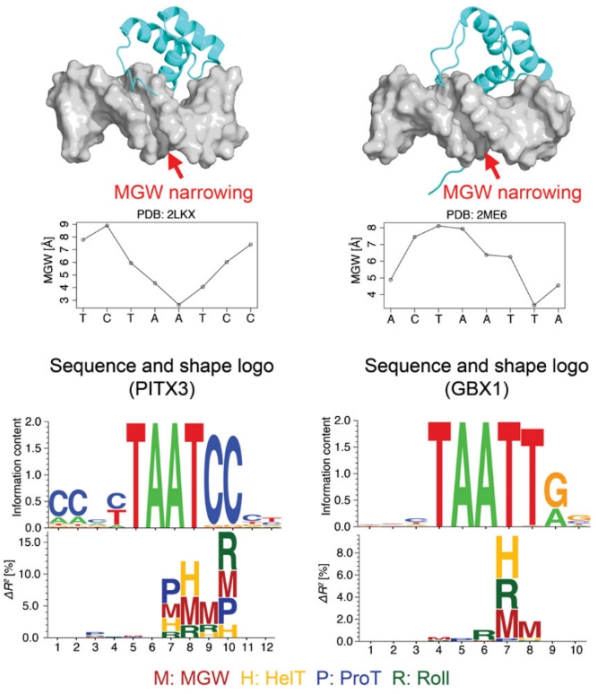
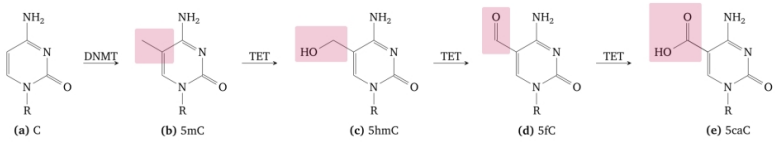


Figure 14.2: Representation of methylated motifs. (a) Stepwise epigenetic modifications of Cytosine. (b) Expanded IUPAC alphabet for the C-methylation. (c) Examples of CEBP motifs with and without the expanded IUPAC alphabet. Adapted from Viner (2016).

a)



b)

Covalent Cytosine Modification			Complement	
Abbreviation	Name	Symbol	Name	Symbol
5mC	5-Methylcytosine	m	Guanine:5-Methylcytosine	1
5hmC	5-Hydroxymethylcytosine	h	Guanine:5-Hydroxymethylcytosine	2
5fC	5-Formylcytosine	f	Guanine:5-Formylcytosine	3
5caC	5-Carboxylcytosine	c	Guanine:5-Carboxylcytosine	4

c)

Figure 14.3: DNA sequence and shape logos. The DNA bound by two TFs with similar motifs has different DNA shape properties. Minor groove width (MGW), Roll (R), propeller twist (ProT), and helix twist (HelT). Adapted from Yang (2017).

mono-PSSM in TRANSFAC, MEME or HOMER format; a standard representation should be accorded in order to make these models more user-friendly.

Regarding the novel representations of PSSMs, my observation is that most of the recent publication involving motif analysis show results based on simple methods, for example, either string-based or matrix-based on mono-nucleotide frequencies, even when there is an increasing development of novel, more complex and more precise methods (KULAKOVSKIY et al., 2013; Mathelier and Wasserman, 2013; Siebert and Johannes, 2016). One reason could be that simple algorithms are faster, more popular and could be enough informative for some purposes, as measure the enrichment of TFBSs, even it has been demonstrated that simple model are enough to model most of the TFs (Zhao, 2013) and this is the reason that not all the TFs have their di-nucleotide PSSMs in popular databases as JASPAR or HOCOMOCO (Ivan Kulakoskyi, personal communication). However, I consider that the study of one specific TF, for example modelling the binding or considering the flanking residues, that has been demonstrated that are important for the TF specificity (Jurk et al., 2016; Slattery et al., 2014) should be done using a combination of the novel models (e.g., including nucleotide interdependencies and DNA shape information). Another factor that I consider a reason because the di-PSSMs are not widely used at this time is that there is not a suite as RSAT (Medina-Rivera et al., 2015) or MEME (Bailey et al., 2015) including all the tools required for di-PSSM analysis (sequence retrieve, background models, motif discovery, motif comparison, motif scan).

14.4 Redundancy in motif databases

As more motifs are discovered and stored in databases, their redundancy becomes an issue for the motif analysis, and the clustering of motifs becomes a solution to ease the analysis for thousands of motifs in a single study, either by grouping similar motifs or by producing non-redundant collections of motifs.

The idea of reduce motif redundancy in databases should be considered when there is more than one motif for a particular TF, for example in cases when different motifs are obtained from different experiments as ChIP-seq, PBM or SELEX, because the TF binding motifs may change if the binding sites were detected *in vitro* or *in vivo*. But it should be noted that in most of the databases, the term non-redundant means that there is only one motif per TF, although motifs corresponding to TFs from the same family use to be similar (except the zinc fingers). By contrast in HOCOMOCO, a small number of TFs have two motifs associated and that might be no similar.

Having one motif per TF is useful when the user is interested in a particular TF, so using the specific motif will TF specificites as flanking residues at the TFBSs, that are underestimated in the Familial Binding Profiles.

In order to find candidate TF potentially bound to a motif resulting from motif discovery, it is normally compared with several motif collections (JASPAR, HOCOMOCO, Cis-BP, etc) that are redundant. If the motifs are clustered *a priori*, the resulting non-redundant motif clusters can substitute those stored in databases (for the purposes of comparison (Castro-Mondragon et al., 2017)). For example, if a cluster is made by ten similar motifs, then the discovered motif should be compared once and not ten times, with a trade-off between time efficiency (reducing the number of DB motifs to compare) but loss in precision (because the cluster motif does not reflect all the particularities from the individual motifs). These non-redundant collections may be obtained using RSAT *matrix-clustering*.

For the motif enrichment methods, since they are based on the detection of TFBSs and since similar motifs are expected to detect the same TFBSs (Pape et al., 2008; Vorontsov et al., 2013), a collection of non-redundant motif can be used to measure the enrichment in order to reduce the analysis time and to reduce the multitesting correction factor applied to the enrichment p-values.

14.5 Annotation of unknown TF binding motifs

The number of TFs in a genome has been estimated for several organisms as humans (Vaquerizas et al., 2009; Weirauch et al., 2014) and *Escherichia coli* (Pérez-Rueda et al., 2015), including the already known TF coding genes and putative TFs (predicted by homology of their DBD). For example in *Escherichia coli K12* there are ~300 TFs, from them ~90 are predicted and there is no knowledge about the experimental conditions under which they are active nor about their DNA binding motifs. In humans, the number of TFs is 1734 (according to Cis-BP (Weirauch et al., 2014)), higher than the number reported by a previous study of human TFs (Vaquerizas et al., 2009). However, the number of TFBMs is larger than the number of TFs. This can be explained because now it is available genome-wide information for TF binding at specific conditions (e.g., cell-type and tissue specific) which has revealed novel motifs for some TFs.

In this thesis I clustered a large compendium of vertebrate motifs (~10,000 motifs from 12 databases) using RSAT *matrix-clustering* (Castro-Mondragon et al., 2017), resulting in ~2,000 non-redundant motifs. This reduction of the number of motifs by a factor of 5 can be explained because many TFs from the same family have similar DNA affinities and the collections are redundant. In this analysis I detected hundreds of motifs that are not similar to other known motifs but in other cases tens of highly similar motifs (e.g., Hox motifs) are grouped in a cluster, which make us set the next question: how TFs elicit specific responses despite the fact that many of them have the same motifs as other TFs? The answer comes partly from the interactions between TFs in CRMs, the interaction with co-factors or condition-specific TFs (Slattery et al., 2011), but more studies are required to have a better understanding of this question. In this analysis I also showed that many motifs discovered by a tool, for example Epigram (Whitaker et al., 2015), or by consortia as the FANTOM5 project (Andersson et al., 2014) have no match (i.e., are not similar) to any known motif, this may suggest that these could be *bona fide* motifs for an uncharacterized TFs, but we should not discard the possibility that these motifs might be either artifacts (e.g., false positives) of the motif discovery tools used in these studies, or in the case of epigram, since its motifs were discovered in histone peaks only, they could not necessarily correspond to TF binding affinities.

14.6 Differences between enhancers and promoters

For many years, the definitions of enhancers and promoters have remained as a dichotomy. The promoters are defined as TSS-proximal regions that can activate gene transcription. Enhancers were originally defined as regions that can activate gene transcription distally and independently of their orientation (Banerji et al., 1981). In addition to the distance, others features have been associated specifically to enhancers and promoters, for example histone marks (Heintzman et al., 2009; Chatterjee and Ahituv, 2017), as consequence, many recent studies define the enhancers based on these marks. It is important to note that, indeed, there are specifics histone marks associated with enhancers, but enhancers should be defined by operational criteria and not by correlative observations. It is not surprising that some of these regions (i.e., enhancers defined by correlation with histone marks) are not capable to activate gene transcription when they are tested experimentally (Kheradpour et al., 2013; Vanhille et al., 2015; Inoue et al., 2016).

Given that enhancers and promoters share genomic and epigenomic similarities, recent studies have discussed the differences between enhancers and promoters (Andersson, 2015; Schaffner, 2015), but two recently developed methods to quantify enhancer activity at genome-wide scale, Starr-seq (Arnold et al., 2013) and CapStarr-seq (Vanhille et al., 2015), in flies and mammals (human and mouse), respectively, have detected thousands of regions with enhancer activity, based on the operational criteria of enhancers only (i.e., activating genes distally); these studies have found that some regions with enhancer activity overlap at promoters.

In a recent study, where I contributed, using CapStarr-seq it was demonstrated that ~2-3% of human promoters display enhancer activity (the enhancer-like regions overlapping at promoters, are actually promoters acting as enhancers), and they are able to activate surrounding promoters. This novel class of cis-regulatory sequences (promoters with enhancer activity) is denoted as Epromoters (Dao et al., 2017). These Epromoters display distinctive epigenomic features related to enhancers, are enriched for motifs for Jun, Fos, YY and IRF, and they are associated with stress-response genes.

In this study, it was also shown that some promoters display enhancer activity after a stimulus (e.g., response to interferon), which suggest that novel studies should be performed to discover cis-regulatory sequences which activity is condition-dependent. In addition, this observation is in agreement with a study that showed that some genomic regions have promoter-associated epigenomic features in one cell line and enhancer features in other cell line (Leung et al., 2015).

Genome-wide enhancer assays can be used to validate predicted enhancers in different cell lines, based on the sequences itself (Arnold et al., 2013; Vanhille et al., 2015), and although the enhancer activity of all the genome can be analyzed in a single experiments, the main limitations of these enhancer reporter assays is that the experiments are done outside their endogenous genomic context. However, the recent advent of mediated mutagenesis techniques such as CRISPR-Cas9, allows the individual study of enhancer and the effects of mutations and how they may affect the enhancer activity, for example mutation a TFBSs may affect the interaction with another cis-regulatory region (Santiago-Algarra et al., 2017). Taking advantage of these mediated mutagenesis technique will allow to study with more details the TF grammar and may confirm the existence of the *cis-regulatory code*. Another use for these assays is that they could be adapted to identify silencer at genome-wide scale.

14.7 Integrating analysis of TF binding regions with other (epi)genomic features

It is important to remember that the transcriptional regulation driven by TFs is only one part of the gene regulation mechanisms, the TF binding *per se* is not necessarily enough to determine whether the target gene will be active or inactive. For this reason, in order to infer gene regulatory networks or test hypothesis about regulatory mechanisms, the detection of TFBSs (by experimental or *in silico* methods) should be complemented with information of open chromatin regions (e.g., searching ChIP-seq peaks overlapping DNAseI sensitive sites) or chromatin interactions.

Combining these information, the false positive rate of computational methods to detect TFBSs can be reduced and give robustness to the detected TFBSs. In addition, the location and quantification of histone marks related with gene expression or RNAPII positioning can be used to annotate or detect cis-regulatory regions as enhancer or promoters.

It should be considered that a predicted site can be a "true positive" (i.e., a functional site) in one tissue at one developmental stage, and a "false positive" in another tissue, or in the same tissue at a different stage. For this reason I think that "binding" is not an intrinsic property of a location in the genome (the "site"), but is context-dependent.

Another feature that is more evident are the physical interaction between genomic regions given by the 3D conformation of the chromatin, in other words, regions that linearly seems to be distal, physically can interact via chromatin loops.

Combining these evidences gives us a clearer but yet incomplete vision of transcriptional regulation, two of the missing parts are the following:

- Time: most of the results are obtained at a fixed moment, giving us the idea of statics when in reality the regulatory elements and the genome as well are highly dynamics. More time-series studies should be done to understand the dynamics of the regulatory elements.
- TF-Target gene association: Evidences of TF binding at distal locations do not bring additional information about the regulated gene or the change of expression of target genes after the regulatory interactions. In addition to locate the TF binding (e.g., with ChIP-seq), additional experiments detecting chromatin conformation (e.g., interactions between RNAPII and enhancers) and gene expression are helpful to detect the TF-Target gene interactions (Aerts et al., 2010), although this additional information is not always available for most of the genomes and the chromatin accessibility changes according the cell types.

Bibliography

Aerts, S. (2012). *Computational Strategies for the Genome-Wide Identification of cis-Regulatory Elements and Transcriptional Targets*, volume 98. Elsevier Inc., 1 edition.

Aerts, S., Loo, P. V., Thijs, G., and Moreau, Y. (2003). Computational detection of cis -regulatory modules. 19:5–14.

Aerts, S., Quan, X. J., Claeys, A., Sanchez, M. N., Tate, P., Yan, J., and Hassan, B. A. (2010). Robust target gene discovery through transcriptome perturbations and genome-wide enhancer predictions in drosophila uncovers a regulatory basis for sensory specification. *PLoS Biology*, 8(7).

Aerts, S., Thijs, G., Dabrowski2, M., Moreau, Y., and Moor, B. D. (2007). Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC genomics*, 8:130.

Ambrosini, G., Dreos, R., Kumar, S., Bucher, P., Park, P., Furey, T., Dunham, I., Kundaje, A., Aldred, S., Collins, P., Davis, C., Doyle, F., Epstein, C., Frietze, S., Harrow, J., Kaul, R., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M., Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., Zhang, J., Leleu, M., Lefebvre, G., Rougemont, J., Langmead, B., Trapnell, C., Pop, M., Salzberg, S., Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X., Rougemont, J., Amzallag, A., Iseli, C., Farinelli, L., Xenarios, I., Naef, F., Beauparlant, C., Lamaze, F., Deschenes, A., Samb, R., Lemacon, A., Belleau, P., Bilodeau, S., Droit, A., Barta, E., Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y., Laslo, P., Cheng, J., Murre, C., Singh, H., Glass, C., Ji, H., Jiang, H., Ma, W., Johnson, D., Myers, R., Wong, W., Ye, T., Krebs, A., Choukrallah, M., Keime, C., Plewniak, F., Davidson, I., Liu, T., Ortiz, J., Taing, L., Meyer, C., Lee, B., Zhang, Y., Shin, H., Wong, S., Ma, J., Lei, Y., Schmid, C., Bucher, P., Dreos, R., Ambrosini, G., Perier, R., Bucher, P., Ambrosini, G., Praz, V., Jagannathan, V., Bucher, P., Auton, A., Brooks, L., Durbin, R., Garrison, E., Kang, H., Korbel, J., Marchini, J., McCarthy, S., McVean, G., Abecasis, G., Jee, J., Rozowsky, J., Yip, K., Lochovsky, L., Bjornson, R., Zhong, G., Zhang, Z., Fu, Y., Wang, J., Weng, Z., Landt, S., Marinov, G., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B., Bickel, P., Brown, J., Cayting, P., Schmid, C., Bucher, P., Pjanic, M., Schmid, C., Gaussin, A., Ambrosini, G., Adamcik, J., Pjanic, P., Plasari, G., Kerschgens, J., Dietler, G., Bucher, P., Pepke, S., Wold, B., Mortazavi, A., Tateno, Y., Saitou, N., Okubo, K., Sugawara, H., Gojobori, T., Barrett, T., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I., Tomashevsky, M., Marshall, K., Phillippy, K., Sherman, P., Holko, M., Schones, D., Cui, K., Cuddapah, S., Roh, T., Barski, A., Wang, Z., Wei, G., Zhao, K., Boyle, A., Davis, S., Shulha, H., Meltzer, P., Margulies, E., Weng, Z., Furey, T., Crawford, G., Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Cunningham, F., Amode, M., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Kapitonov, V., Jurka, J., Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E., Sirotkin, K., Rosenbloom, K., Armstrong, J., Barber, G., Casper, J., Clawson, H., Diekhans, M., Dreszer, T., Fujita, P., Guruvadoo, L., Haeussler, M., Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T., Langmead, B., Salzberg, S., Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S., Pollard, K., Hubisz, M., Rosenbloom, K., Siepel, A., Karolchik, D., Hinrichs, A., Kent, W., Bucher, P., Bryan, B., Orenstein, Y., Shamir, R., McLean, C., Bristor, D., Hiller, M., Clarke, S., Schaar, B., Lowe, C., Wenger, A., Bejerano, G., Boeva, V., Lermine, A.,

Barette, C., Guillouf, C., Barillot, E., Ma, W., Noble, W., Bailey, T., Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Ambrosini, G., Dreos, R., Bucher, P., Wilbanks, E., Facciotti, M., Halachev, K., Bast, H., Albrecht, F., Lengauer, T., Bock, C., Chen, T., Li, H., Lee, C., Gan, R., Huang, P., Wu, T., Lee, C., Chang, Y., Tang, P., Goecks, J., Nekrutenko, A., Taylor, J., Halbritter, F., Kousa, A., Tomlinson, S., David, F., Delafontaine, J., Carat, S., Ross, F., Lefebvre, G., Jarosz, Y., Sinclair, L., Noordermeer, D., Rougemont, J., Leleu, M., Kim, R., Smith, O., Wong, W., Ryan, A., Ryan, M., Aladjem, M., Lan, X., Bonneville, R., Apostolos, J., Wu, W., and Jin, V. (2016). The ChIP-Seq tools and web server: a resource for analyzing ChIP-seq and other types of genomic data. *BMC Genomics*, 17(1):938.

Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., Wasserman, W. W., and Odeberg, J. (2008). In silico detection of sequence variations modifying transcriptional regulation. *PLoS computational biology*, 4(1):e5.

Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, 37(3):314–323.

Andersson, R., Chen, Y., Core, L., Lis, J. T., Sandelin, A., and Jensen, T. H. (2015). Human Gene Promoters Are Intrinsically Bidirectional. *Molecular Cell*, 60(3):346–347.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, a. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. a., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Forrest, A. R. R., Carninci, P., Rehli, M., and Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–61.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, 339(6123):1074–1077.

Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T., and Stark, A. (2016). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nature Biotechnology*, 35(2).

Badis, G., Berger, M. F., Philippakis, A. a., Talukder, S., Gehrke, A. R., Jaeger, S. a., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R., and Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science (New York, N.Y.)*, 324(5935):1720–3.

Bagchi, D. N. and Iyer, V. R. (2016). The Determinants of Directionality in Transcriptional Initiation. *Trends in Genetics*, 32(6):322–333.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Computational Biology*, 9(11):5–12.

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics (Oxford, England)*, 27(12):1653–9.

Bailey, T. L., Bodén, M., Whitington, T., and Machanick, P. (2010). The value of position-specific priors in motif discovery using MEME. *BMC bioinformatics*, 11(1):179.

Bailey, T. L. and Elkan, C. (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–W49.

Bailey, T. L. and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. *Nucleic acids research*, 40(17):e128.

Ballester, B., Medina-Rivera, A., Schmidt, D., Gonzàlez-Porta, M., Carlucci, M., Chen, X., Chessman, K., Faure, A. J., Funnell, A. P. W., Goncalves, A., Kutter, C., Lukk, M., Menon, S., McLaren, W. M., Stefflova, K., Watt, S., Weirauch, M. T., Crossley, M., Marioni, J. C., Odom, D. T., Flicek, P., and Wilson, M. D. (2014). Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife*, 3:e02626.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of alpha-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 PART 1):299–308.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-y., Schones, D. E., and Wang, Z. (2007). Resource High-Resolution Profiling of Histone Methylations in the Human Genome. pages 823–837.

Beck, L. L., Smith, T. G., and Hoover, T. R. (2007). Look, no hands! Unconventional transcriptional activators in bacteria. *Trends in Microbiology*, 15(12):530–537.

Berger, M. F. and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. 4(3):393–411.

Berger, S., Omidi, S., Pachkov, M., Arnold, P., Kelley, N., Salatino, S., and van Nimwegen, E. (2016). Crunch : Completely Automated Analysis of ChIP-seq Data. *bioRxiv*.

Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. S. P. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, 23(3):415–20.

Boeva, V. (2016). Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. 7(February).

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309.

Boyle, A. P., Song, L., Lee, B. K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., and Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464.

Branco, M. R., Ficz, G., and Reik, W. (2011). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics*, 13(1):7–13.

Brohée, S., Janky, R., Abdel-Sater, F., Vanderstocken, G., André, B., and van Helden, J. (2011). Unraveling networks of co-regulated genes on the sole basis of genome sequences. *Nucleic acids research*, 39(15):6340–58.

Broin, P. Ó., Smith, T. J., and Golden, A. A. (2015). Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach. *BMC bioinformatics*, 16(1):22.

Browning, D. F. and Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638–650.

Bucher, P. and Bryan, B. (1984). Signal search analysis: a new method to localize and characterize functionally important DNA sequences. 12(1):287–305.

Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2016). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. pages 1–10.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218.

Cann, J. R. (1998). Theoretical studies on the mobility-shift assay of protein-DNA complexes. *Electrophoresis*, 19(2):127–143.

Castro-Mondragon, J., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *bioRxiv*, page 065565.

Castro-mondragon, J. A., Rioualen, C., Contreras-moreira, B., and Helden, J. V. (2016). RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes. 1482:297–322.

Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.-c., Helden, J. V., Méchali, M., Tagc, U., F, M., and Ciml, I. D. M.-l. (2015). The chromatin environment shapes DNA replication origin organization and defines origin classes. pages 1–13.

Chatterjee, S. and Ahituv, N. (2017). Gene Regulatory Elements , Major Drivers of Human Disease. (March):1–19.

Choi, I.-g., Kwon, J., and Kim, S.-h. (2004). Local feature frequency profile : A method to measure structural similarity in proteins. 101(11).

Collings, C. K., Waddell, P. J., and Anderson, J. N. (2013). Effects of DNA methylation on nucleosome stability. *Nucleic Acids Research*, 41(5):2918–2931.

Comnish-bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences. 13(9):3021–3030.

Contreras-moreira, B., Castro-mondragon, J. A., Rioualen, C., Cantalapiedra, C. P., and Helden, J. V. (2016). RSAT::Plants: Motif Discovery Within Clusters of Upstream Sequences in Plant Genomes. 1482.

Crooks, G., Hon, G., Chandonia, J., and Brenner, S. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14:1188–1190.

Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., Alomairi, J., Martin, D., Torres, M., Fernandez, N., Soler, E., van Helden, J., Puthier, D., and Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, 10.

Defrance, M., Janky, R., Sand, O., and van Helden, J. (2008). Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nature protocols*, 3(10):1589–603.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., and Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–394.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment on transcription factor binding across diverse protein families. *Genome Res*, pages 1268–1280.

Dror, I., Rohs, R., and Mandel-Gutfreund, Y. (2016). How motif environment influences transcription factor search dynamics: Finding a needle in a haystack. *BioEssays*, 38(7):605–612.

Ehret, G. B., Reichenbach, P., Schindler, U., Horvath, C. M., Fritz, S., Nabholz, M., and Bucher, P. (2001). DNA binding specificity of different STAT proteins: Comparison of in vitro specificity with natural target sites. *Journal of Biological Chemistry*, 276(9):6675–6688.

Ell, B. and Kang, Y. (2013). Transcriptional control of cancer metastasis. *Trends in Cell Biology*, 23(12):603–611.

Encode Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Engel, K. L., Mackiewicz, M., Hardigan, A. A., Myers, R. M., and Savic, D. (2016). Decoding transcriptional enhancers: Evolving from annotation to functional interpretation. *Seminars in Cell and Developmental Biology*, 57:40–50.

Erives, A. and Levine, M. (2004). Coordinate enhancers share common organizational features in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3851–6.

Feuerborn, A. and Cook, P. R. (2015). Why the activity of a gene depends on its neighbors. *Trends in Genetics*, 31(9):483–490.

Fletez-Brant, C., Lee, D., McCallion, A. S., and Beer, M. A. (2013). kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Research*, 41(W1):W544–W556.

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Lassmann, T., Itoh, M., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. a., Carninci, P., and Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–70.

Frith, M. C., Fu, Y., Yu, L., Chen, J. F., Hansen, U., and Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Research*, 32(4):1372–1381.

Frith, M. C., Saunders, N. F. W., Kobe, B., and Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Computational Biology*, 4(5).

Furey, T. S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics*, 13(12):840–52.

Galas, D. J. and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*, 5(9):3157–3170.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Moral-Chávez, V. D., Rinaldi, F., and Collado-Vides, J. (2015). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, (5):1–11.

Gilchrist, D. A., Fargo, D. C., and Adelman, K. (2009). Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods*, 48(4):398–408.

Gord??n, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L. (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Reports*, 3(4):1093–1104.

Grainger, D. C. and Busby, S. J. W. (2008). *Chapter 4 Global Regulators of Transcription in Escherichia coli:. Mechanisms of Action and Methods for Study*, volume 65. Elsevier Masson SAS.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018.

Grau, J., Posch, S., Grosse, I., and Keilwagen, J. (2013). A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21).

Griffon, A., Barbier, Q., Dalino, J., van Helden, J., Spicuglia, S., and Ballester, B. (2015). Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic acids research*, 43(4):e27.

Grubert, F., Zaugg, J. B., Steinmetz, L. M., Snyder, M., Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., and Martin, A. R. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions Article Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, 162(5):1–15.

Gupta, S., Stamatoyannopoulos, J. a., Bailey, T. L., and Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, 8(2):R24.

Habib, N., Kaplan, T., Margalit, H., and Friedman, N. (2008). A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS computational biology*, 4(2):e1000010.

Hao, B., Naik, A. K., Watanabe, A., Tanaka, H., Chen, L., Richards, H. W., Kondo, M., Taniuchi, I., Kohwi, Y., Kohwi-Shigematsu, T., and Krangel, M. S. (2015). An anti-silencer– and SATB1-dependent chromatin hub regulates Rag1 and Rag2 gene expression during thymocyte development. *The Journal of Experimental Medicine*, 212(5):809–824.

Hardison, R. C. and Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, 13(7):469–483.

Hartonen, T., Sahu, B., Dave, K., Kivioja, T., and Taipale, J. (2016). PeakXus: comprehensive transcription factor binding site discovery from ChIP-Nexus and ChIP-Exo experiments. *Bioinformatics*, 32(17):i629–i638.

He, H. H., Meyer, C. A., Hu, S. S., Chen, M.-W., Zang, C., Liu, Y., Rao, P. K., Fei, T., Xu, H., Long, H., Liu, X. S., and Brown, M. (2013). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nature Methods*, 11(1):73–78.

He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(4):395–401.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. a., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. a., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–8.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4):576–589.

Helden, J. V. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Research*, 28(4):1000–1010.

Herold, M., Bartkuhn, M., and Renkawitz, R. (2012). CTCF : insights into insulator function during development. *Development (Cambridge, England)*, 1057(6):1045–1057.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289.

Hsieh, C.-L., Fei, T., Chen, Y., Li, T., Gao, Y., Wang, X., Sun, T., Sweeney, C. J., Lee, G.-S. M., Chen, S., Balk, S. P., Liu, X. S., Brown, M., and Kantoff, P. W. (2014). Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(20):7319–24.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G. L., Wade, H., Song, H., Qian, J., and Zhu, H. (2013). DNA methylation presents distinct binding sites for human transcription factors. *eLife*, 2013(2):1–16.

Huerta, a. M., Salgado, H., Thieffry, D., and Collado-Vides, J. (1998). RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res*, 26(1):55–59.

Hughes, J. D., Estep, P. W., Tavazoie, S., and Church, G. M. (2000). Computational Identification of Cis -regulatory Elements Associated with Groups of Functionally Related Genes in Saccharomyces cerevisiae.

Igg, T., Fab, D., and Our, G. (2007). 3 domains, which we show to be critically involved in Fab arm exchange. Elucidat- ing the contribution of specific C. *Framework*, 317(September):1557–1560.

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., Mcmanus, M. T., Ahituv, N., and Shendure, J. (2016). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity . pages 38–52.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P., and Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature Methods*, (November 2016):1–12.

Istrail, S. and Davidson, E. H. (2005). Logic functions of the genomic cis-regulatory code. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4954–4959.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356.

Janky, R. and van Helden, J. (2008). Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC bioinformatics*, 9:37.

Jayaram, N., Usvyat, D., and R. Martin, A. C. (2016). Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics*, (i):1–12.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpa, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. pages 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–39.

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–8.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–492.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic acids research*, 36(16):5221–31.

Jurk, M., Helabad, M. B., Dror, I., Lebars, I., Kieffer, B., Scho, S., Imhof, P., Rohs, R., Vingron, M., Thomas-chollier, M., and Meijsing, S. H. (2016). Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity ¨.

Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology*, 1(1):40–51.

Kankainen, M. and Löytynoja, A. (2007). MATLIGN: a motif clustering, comparison and matching tool. *BMC bioinformatics*, 8:189.

Kaplun, A., Krull, M., Lakshman, K., Matys, V., Lewicki, B., and Hogan, J. D. (2016). Establishing and validating regulatory regions for variant annotation and expression analysis. *BMC Genomics*, 17(S2):393.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., Hong, M. Y., Karczewski, K. J., Huber, W., Weissman, S. M., Gerstein, M. B., Korbel, J. O., and Snyder, M. (2010). Variation in Transcription Factor Binding Among Humans. *Science*, 328(5975):232–235.

Keilwagen, J. and Grau, J. (2015). Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Research*, pages 1–12.

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7):990–999.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, 23(5):800–811.

Kheradpour, P. and Kellis, M. (2013). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42(5):2976–2987.

Kim, T.-k. and Shiekhattar, R. (2015). Review Architectural and Functional Commonalities between Enhancers and Promoters. *Cell*, 162(5):948–959.

Kolovos, P., Knoch, T. a., Grosveld, F. G., Cook, P. R., and Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics & Chromatin*, 5(1):1.

Köster, J. and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics (Oxford, England)*, 28(19):2520–2.

KULAKOVSKIY, I., LEVITSKY, V., OSHCHEPKOV, D., BRYZGALOV, L., VORONTSOV, I., and MAKEEV, V. (2013). From Binding Motifs in Chip-Seq Data To Improved Models of Transcription Factor Binding Sites. *Journal of Bioinformatics and Computational Biology*, 11(01):1340004.

Kulakovskiy, I. V., Boeva, V. a., Favorov, a. V., and Makeev, V. J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics (Oxford, England)*, 26(20):2622–3.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., Soboleva, A. V., Kasianov, A. S., Ashoor, H., Ba-Alawi, W., Bajic, V. B., Medvedeva, Y. A., Kolpakov, F. A., and Makeev, V. J. (2016). HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Research*, 44(D1):D116–D125.

Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2015). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, 48(2):206–213.

Kuttippurathu, L., Hsing, M., Liu, Y., Schmidt, B., Maskell, D. L., Lee, K., He, A., Pu, W. T., and Kong, S. W. (2011). CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments. *Bioinformatics (Oxford, England)*, 27(5):715–7.

Lande-Diner, L., Zhang, J., Ben-Porath, I., Amariglio, N., Keshet, I., Hecht, M., Azuara, V., Fisher, A. G., Rechavi, G., and Cedar, H. (2007). Role of DNA methylation in stable gene repression. *Journal of Biological Chemistry*, 282(16):12194–12200.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. a., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–31.

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics*, 32(1):42–56.

Lecellier, C.-H., Wasserman, W. W., Rohs, R., and Mathelier, A. (2016). Human enhancers associated with immune response harbour specific sequence composition, activity, and genome organization. *bioRxiv*.

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., and Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, 47(8):955–61.

Lee, D. J., Minchin, S. D., and Busby, S. J. W. (2012). Activating transcription in bacteria. *Annual review of microbiology*, 66:125–52.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*, 13(4):233–245.

Lercher, L., McDonough, M. a., El-Sagheer, A. H., Thalhammer, A., Kriaucionis, S., Brown, T., and Schofield, C. J. (2014). Structural insights into how 5-hydroxymethylation influences transcription factor binding. *Chem. Commun.*, 50(15):1794–1796.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., and Yen, C.-a. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518(7539):350–354.

Levo, M. and Segal, E. (2014). In pursuit of design principles of regulatory sequences. *Nature reviews. Genetics*, 15(7):453–68.

Lihu, A. and Holban, . (2015). A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Briefings in Bioinformatics*, 16(6):964–973.

Liu, Z., Widlak, P., Zou, Y., Xiao, F., Oh, M., Li, S., Chang, M. Y., Shay, J. W., and Garrard, W. T. (2006). A Recombination Silencer that Specifies Heterochromatin Positioning and Ikaros Association in the Immunoglobulin ?? Locus. *Immunity*, 24(4):405–415.

Lu, Z., Brigitte, H., Vollmers, C., DuBois, R., and Schmitz, R. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic acids research*, 45(2):846–860.

Luehr, S., Hartmann, H., and Söding, J. (2012). The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences. *Nucleic Acids Research*, 40(W1):104.

Ma, Q., Zhang, H., Mao, X., Zhou, C., Liu, B., Chen, X., and Xu, Y. (2014a). DMINDA: An integrated web server for DNA motif identification and analyses. *Nucleic Acids Research*, 42(W1):12–19.

Ma, W., Noble, W. S., and Bailey, T. L. (2014b). Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nature protocols*, 9(6):1428–50.

Ma, X., Kulkarni, A., Zhang, Z., Xuan, Z., Serfling, R., and Zhang, M. Q. (2012). A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Research*, 40(7).

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics (Oxford, England)*, 27(12):1696–7.

Macintyre, G., Bailey, J., Haviv, I., and Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics (Oxford, England)*, 26(18):i524–30.

Mahony, S., Auron, P. E., and Benos, P. V. (2007). DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Computational Biology*, 3(3):e61.

Mahony, S. and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research*, 35(Web Server issue):W253–8.

Mahony, S. and Pugh, B. F. (2015). Protein-DNA binding in high-resolution. *Critical reviews in biochemistry and molecular biology*, 50(4):269–83.

Maston, G. a., Landt, S. G., Snyder, M., and Green, M. R. (2012). *Characterization of enhancer function from genome-wide analyses.*, volume 13.

Mathelier, A., Shi, W., and Wasserman, W. W. (2015). Identification of altered cis-regulatory elements in human disease. *Trends in Genetics*, 31(2):67–76.

Mathelier, A. and Wasserman, W. W. (2013). The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9):e1003214.

Mathelier, A., Xin, B., Chiu, T.-p., Yang, L., Rohs, R., and Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*, pages 1–9.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, 42(Database issue):D142–7.

Matys, V. (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378.

McLeay, R. C. and Bailey, T. L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC bioinformatics*, 11:165.

Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J., and van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic acids research*, 39(3):808–24.

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D. M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., and Van Helden, J. (2015). RSAT 2015: Regulatory sequence analysis tools.

Meireles-Filho, A. C. and Stark, A. (2009). Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Current Opinion in Genetics and Development*, 19(6):565–570.

Metzler, R. (2009). Keeping up with the noise. *Physics*, 2(36):36.

Meyer, C. A. and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11):709–721.

Muerdter, F., Boryń, Ł. M., and Arnold, C. D. (2015). STARR-seq — Principles and applications. *Genomics*, 106:1–6.

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., and Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics*, 36(12):1331–9.

Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., Albu, M., Weirauch, M. T., Radovani, E., Kim, P. M., Greenblatt, J., Frey, B. J., and Hughes, T. R. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol*, 33(5):555–562.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R. S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M. J., Akey, J. M., Bender, M. a., Groudine, M., Kaul, R., and Stamatoyannopoulos, J. a. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90.

Ngo, V. and Wang, W. (2016). Finding De novo methylated DNA motifs. pages 1–15.

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., and Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, 26(8):1023–1033.

Ong, C.-T. and Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nature reviews. Genetics*, 15(4):234–46.

Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends in Biochemical Sciences*, 25(3):99–104.

Ou, J. and Zhu, L. J. (2012). motifStack guide Examples of using motifStack plot a DNA sequence logo with different fonts and. pages 1–9.

Pape, U. J., Rahmann, S., and Vingron, M. (2008). Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics (Oxford, England)*, 24(3):350–7.

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80.

Parker, D. S., White, M. A., Ramos, A. I., Cohen, B. A., Barolo, S., Barolo, S., Posakony, J. W., Ashe, H. L., Briscoe, J., Wang, Q. T., Holmgren, R. A., Lum, L., Beachy, P. A., Alexandre, C., Jacinto, A., Ingham, P. W., Aza-Blanc, P., Ramírez-Weber, F. A., Laget, M. P., Schwartz, C., Kornberg, T. B., Méthot, N., Basler, K., Müller, B., Basler, K., Vokes, S. A., Ji, H., McCuine, S., Tenzen, T., Giles, S., Zhong, S., Longabaugh, W. J. R., Davidson, E. H., Wong, W. H., McMahon, A. P., Vokes, S. A., Ji, H., Wong, W. H., McMahon, A. P., Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., Taipale, J., Dessaud, E., McMahon, A. P., Briscoe, J., Morimura, S., Maves, L., Chen, Y., Hoffmann, F. M., Ohlen, T. V., Hooper, J. E., Shea, M. A., Ackers, G. K., Buchler, N. E., Gerland, U., Hwa, T., Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., Gaul, U., Gertz, J., Siggia, E. D., Cohen, B. A., He, X., Samee, M. A., Blatti, C., Sinha, S., Janssens, H., Hou, S., Jaeger, J., Kim, A. R., Myasnikova, E., Sharp, D., Reinitz, J., Fakhouri, W. D., Ay, A., Sayal, R., Dresch, J., Dayringer, E., Arnosti, D. N., Furriols, M., Bray, S., Winklmayr, M., Schmid, C., Laner-Plamberger, S., Kaser, A., Aberger, F., Eichberger, T., Frischauf, A. M., Hersh, B. M., Carroll, S. B., Kwon, C., Hays, R., Fetting, J., Orenic, T. V., Jiang, J., Levine, M., Gaudet, J., Mango, S. E., Rowan, S., Siggers, T., Lachke, S. A., Yue, Y., Bulyk, M. L., Maas, R. L., Nguyen, V., Chokas, A. L., Stecca, B., Altaba, A. R. i., Barolo, S., Carver, L. A., Posakony, J. W., Barolo, S., Castro, B., Posakony, J. W., Johnson, L. A., Zhao, Y., Golden, K., Barolo, S., Rubin, G. M., Spradling, A. C., Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (2011). The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Science signaling*, 4(176):ra38.

Pennacchio, L. a., Bickmore, W., Dean, A., Nobrega, M. a., and Bejerano, G. (2013). Enhancers: five essential questions. *Nature reviews. Genetics*, 14(4):288–95.

Pérez-Rueda, E., Tenorio-Salgado, S., Huerta-Saquero, A., Balderas-Martínez, Y. I., and Moreno-Hagelsieb, G. (2015). The functional landscape bound to the transcription factors of Escherichia coli K-12. *Computational Biology and Chemistry*, 58:93–103.

Plass, C., Pfister, S. M., Lindroth, A. M., and Bogatyrova, O. (2013). Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nature Publishing Group*, 14(11):765–780.

Qin, J. Y., Zhang, L., Clift, K. L., Hulur, I., Xiang, A. P., Ren, B. Z., and Lahn, B. T. (2010). Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS ONE*, 5(5):3–6.

Qu, K., Zaba, L. C., Giresi, P. G., Li, R., Longmire, M., Kim, Y. H., Greenleaf, W. J., and Chang, H. Y. (2015). Individuality and Variation of Personal Regulomes in Primary Human T Cells. *Cell Systems*, 1(1):51–61.

Rahman, S., Zorca, C. E., Traboulsi, T., Noutahi, E., Krause, M. R., Mader, S., and Zenklusen, D. (2017). Single-cell profiling reveals that eRNA accumulation at enhancer–promoter loops is not required to sustain transcription. *Nucleic acids research*, 45(2):846–860.

Rao, S. S. P., Huntley, M. H., Durand, N. C., and Stamenova, E. K. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development*, 43:73–81.

Rhee, H. S., Bataille, A. R., Zhang, L., and Pugh, B. F. (2014). Subnucleosomal structures and nucleosome asymmetry across a genome. *Cell*, 159(6):1377–1388.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419.

Roider, H. G., Manke, T., O'keeffe, S., Vingron, M., and Haas, S. A. (2009). PASTAA: Identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics*, 25(4):435–442.

Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature reviews. Molecular cell biology*, 16(3):129–143.

Sandelin, A. and Wasserman, W. W. (2004). Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics. *Journal of Molecular Biology*, 338(2):207–215.

Santiago-Algarra, D., Dao, L. T., Pradel, L., Espa?a, A., and Spicuglia, S. (2017). Recent advances in high-throughput approaches to dissect enhancer function. *F1000Research*, 6(0):939.

Schaffner, W. (2015). Enhancers, enhancers – from their discovery to today's universe of transcription enhancers. 396(4):311–327.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D. T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)*, 328(5981):1036–40.

Schmidt, H. G., Sewitz, S., Andrews, S. S., and Lipkow, K. (2014). An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding. *PLoS ONE*, 9(10).

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431.

Schones, D. E., Cui, K., and Cuddapah, S. (2011). Genome-wide approaches to studying yeast chromatin modifications. *Methods in Molecular Biology*, 759(march):61–71.

Schones, D. E., Sumazin, P., and Zhang, M. Q. (2005). Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics (Oxford, England)*, 21(3):307–13.

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., and Adelman, K. (2015). Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell*, 58(6):1101–1112.

Sebastian, A. and Contreras-Moreira, B. (2014). FootprintDB: A database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, 30(2):258–265.

Sérandour, A. A., Avner, S., Mahé, E. A., Madigou, T., Guibert, S., Weber, M., and Salbert, G. (2016). Single-CpG resolution mapping of 5-hydroxymethylcytosine by chemical labeling and exonuclease digestion identifies evolutionarily unconserved CpGs as TET targets. *Genome Biology*, 17(1):56.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928):379–423.

Shi, W., Fornes, O., Mathelier, A., and Wasserman, W. W. (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Research*, 44(21):gkw691.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, 15(4):272–86.

Siebert, M. and Johannes, S. (2016). Higher-order models consistently outperform PWMs at predicting regulatory motifs in nucleotide sequences. 44(13):1–23.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J., and Mann, R. S. (2011). Cofactor binding evokes latent differences in DNA binding specificity between hox proteins. *Cell*, 147(6):1270–1282.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9):381–399.

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626.

Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. *Bioinformatics*, 5(2):89–96.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature*, 528(7580):147–51.

Starick, S. R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M. I., Chung, H.-R., Vingron, M., Thomas-Chollier, M., and Meijsing, S. H. (2015). ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. *Genome research*, 25(6):825–35.

Stegmaier, P., Kel, A., Wingender, E., and Borlak, J. (2013). A Discriminative Approach for Unsupervised Clustering of DNA Sequence Motifs. *PLoS Computational Biology*, 9(3):e1002958.

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics*, (October 2015):1–14.

Stevens, T. J., Lando, D., Basu, S., Liam, P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., Shaughnessy-kirwan, A. O., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M. G. S., Lehner, B., Croce, L. D., Wutz, A., and Hendrich, B. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):1–21.

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23.

Takahashi, K. and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. 2:663–676.

Tanaka, E., Bailey, T., Grant, C. E., Noble, W. S., and Keich, U. (2011). Improved similarity scores for comparing motifs. *Bioinformatics (Oxford, England)*, 27(12):1603–9.

Tantin, D., Gemberling, M., Callister, C., and Fairbrother, W. G. (2008). High-throughput biochemical analysis of in vivo location data reveals novel distinct classes of POU5F1(Oct4)/DNA complexes. *Genome Res*, 18(4):631–639.

Tarca, A. L., Carey, V. J., Chen, X.-w., Romero, R., and Drăghici, S. (2007). Machine Learning and Its Applications to Biology. *PLoS Computational Biology*, 3(6):e116.

Taudt, A., Colomé-Tatché, M., and Johannes, F. (2016). Genetic sources of population epigenomic variation. *Nature Reviews Genetics*, 17(6):319–332.

Tautz, D. and Pfeifle, C. (1989). A non-radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene hunchback. *Chromosoma*, 98(September 1989):81–85.

Telorac, J., Prykhozhij, S. V., Schöne, S., Meierhofer, D., Sauer, S., Thomas-Chollier, M., and Meijsing, S. H. (2016). Identification and characterization of DNA sequences that prevent glucocorticoid receptor binding to nearby response elements. *Nucleic Acids Research*, page gkw203.

Teng, M., Ichikawa, S., Padgett, L. R., Wang, Y., Mort, M., Cooper, D. N., Koller, D. L., Foroud, T., Edenberg, H. J., Econs, M. J., and Liu, Y. (2012). Regsnps: A strategy for prioritizing regulatory single nucleotide substitutions. *Bioinformatics*, 28(14):1879–1886.

Teytelman, L., Thurtle, D. M., Rine, J., and Oudenaarden, A. V. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. pages 2–7.

Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D., and van Helden, J. (2012a). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nature protocols*, 7(8):1551–68.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D., and van Helden, J. (2011a). RSAT 2011: regulatory sequence analysis tools. *Nucleic acids research*, 39(Web Server issue):W86–91.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012b). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic acids research*, 40(4):e31.

Thomas-Chollier, M., Hufton, A., Heinig, M., O'Keeffe, S., Masri, N. E., Roider, H. G., Manke, T., and Vingron, M. (2011b). Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nature protocols*, 6(12):1860–9.

Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohée, S., and van Helden, J. (2008). RSAT: regulatory sequence analysis tools. *Nucleic acids research*, 36(Web Server issue):W119–27.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. a., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–44.

Touzet, H. and Varré, J.-S. (2007). Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for molecular biology : AMB*, 2:15.

Tran, N. T. L. and Huang, C.-H. (2014). A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data. *Biology direct*, 9(1):4.

Tsankova, N., Renthal, W., Kumar, A., and Nestler, E. J. (2007). Epigenetic regulation in psychiatric disorders. 8(May):355–367.

Tuerk, C. and Gold, L. (1990). Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science*, (8).

Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature protocols*, 3(10):1578–88.

van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., and van Steensel, B. (2016). Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology*, 35(2):145–153.

van Helden, J. (2003). Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, 31(13):3593–3596.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, 281(5):827–842.

Van Helden, J., André, B., and Collado-Vides, J. (2000). A web site for the computational analysis of yeast regulatory sequences. *Yeast*, 16(2):177–187.

van Helden, J., Rios, a. F., and Collado-Vides, J. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic acids research*, 28(8):1808–18.

Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T., Fernandez, N., Ballester, B., Andrau, J. C., and Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, 6:6905.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. a., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, 10(4):252–63.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., Turner, J. M. A., Bertelsen, M. F., Murchison, E. P., Flicek, P., and Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566.

Viner, C., Johnson, J., Walker, N., Shi, H., Sjöberg, M., Adams, D. J., Ferguson-Smith, A. C., Bailey, T. L., and Hoffman, M. M. (2016). Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. *bioRxiv*, pages 0–29.

Vorontsov, I. E., Kulakovskiy, I. V., and Makeev, V. J. (2013). Jaccard index based similarity measure to compare transcription factor binding site models. *Algorithms for molecular biology : AMB*, 8(1):23.

Waddington, C. H. (1942). The epigenotype. *International journal of epidemiology*, 41(1):10–13.

Wang, L., Chen, J., Wang, C., Uusk??la-Reimand, L., Chen, K., Medina-Rivera, A., Young, E. J., Zimmermann, M. T., Yan, H., Sun, Z., Zhang, Y., Wu, S. T., Huang, H., Wilson, M. D., Kocher, J. P. A., and Li, W. (2014). MACE: model based analysis of ChIP-exo. *Nucleic acids research*, 42(20):e156.

Wang, T. and Stormo, G. D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380.

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, 5(4):276–87.

Weirauch, M., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H., Lambert, S., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J., Govindarajan, S., Shaulsky, G., Walhout, A., Bouget, F.-Y., Ratsch, G., Larrondo, L., Ecker, J., and Hughes, T. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443.

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., Bussemaker, H. J., Morris, Q. D., Bulyk, M. L., Stolovitzky, G., and Hughes, T. R. (2013). Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–34.

Weiss, V., Medina-Rivera, A., Huerta, A. M., Santos-Zavaleta, A., Salgado, H., Morett, E., and Collado-Vides, J. (2013). Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database*, 2013:1–15.

Whitaker, J. W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nat Methods*, 12(3):265–272.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S., Trinklein, N. D., Myers, R. M., Weng, Z., Lemon, B., Tjian, R., Butler, J., Kadonaga, J., Hannenhalli, S., Zhou, Q., Wong, W., Zhu, Z., Shendure, J., Church, G., Pape, U., Klein, H., Vingron, M., Bulyk, M., Frith, M., Li, M., Weng, Z., Frith, M., Fu, Y., Yu, L., Chen, J., Hansen, U., Weng, Z., Tompa, M., Li, N., Bailey, T., Church, G., Moor, B., Eskin, E., Favorov, A., Frith, M., Fu, Y., Kent, W., Makeev, V., Mironov, A., Noble, W., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z., Elnitski, L., Jin, V., Farnham, P., Jones, S., Hawkins, J., Grant, C., Noble, W., Bailey, T., Stormo, G., Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., Pontoglio, M., Kheradpour, P., Stark, A., Roy, S., Kellis, M., Myers, R., Tilly, K., Maniatis, T., Trinklein, N., Aldred, S., Saldanha, A., Myers, R., Cooper, S., Trinklein, N., Anton, E., Nguyen, L., Myers, R., Landolin, J., Johnson, D., Trinklein, N., Aldred, S., Medina, C., Shulha, H., Weng, Z., Myers, R., Matys, V., Kel-Margoulis, O., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A., Wingender, E., Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W., Lenhard, B., Xie, X., Mikkelsen, T., Gnirke, A., Lindblad-Toh, K., Kellis, M., Lander, E., Phillips, J., Corces, V., Handoko, L., Xu, H., Li, G., Ngan, C., Chew, E., Schnapp, M., Lee, C., Ye, C., Ping, J., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W., Ruan, Y., Wei, C., Bell, A., West, A., Felsenfeld, G., Ohlsson, R., Renkawitz, R., Lobanenkov, V., Vostrov, A., Quitschke, W., Fu, Y., Sinha, M., Peterson, C., Weng, Z., Filippova, G., Fagerlie, S., Klenova, E., Myers, C., Dehner, Y., Goodwin, G., Neiman, P., Collins, S., Lobanenkov, V., Lamarko, K., Thompson, C., Byers, B., Walton, E., McKnight, S., Yang, Z., Mott, S., Rosmarin, A., Yu, S., Zhao, D., Jothi, R., Xue, H., Ristevski, S., O'Leary, D., Thornell, A., Owen, M., Kola, I., Hertzog, P., Yu, M., Yang, X., Schmidt, T., Chinenov, Y., Wang, R., Martin, M., Orkin, S., Omichinski, J., Trainor, C., Evans, T., Gronenborn, A., Clore, G., Felsenfeld, G., Molkentin, J., Ohneda, K., Yamamoto, M., Pedone, P., Omichinski, J., Nony, P., Trainor, C., Gronenborn, A., Clore, G., Felsenfeld, G., Suzuki, M., Shimizu, R., Yamamoto, M., Tsai, J., Tong, Q., Tan, G., Chang, A., Orkin, S., Hotamisligil, G., Yamamoto, M., Ko, L., Leonard, M., Beug, H., Orkin, S., Engel, J., Wakil, A. E., Francius, C., Wolff, A., Pleau-Varet, J., Nardelli, J., Caramori, G., Lim, S., Ito, K., Tomita, K., Oates, T., Jazrawi, E., Chung, K., Barnes, P., Adcock, I., Shureiqi, I., Zuo, X., Broaddus, R., Wu, Y., Guan, B., Morris, J., Lippman, S., Nevins, J., Trimarchi, J., Lees, J., Cam, H., Dynlacht, B., Attwooll, C., Denchi, E., Helin, K., Reimer, D., Sadr, S., Wiedemair, A., Goebel, G., Concin, N., Hofstetter, G., Marth, C., Zeimet, A., Rabinovich, A., Jin, V., Rabinovich, R., Xu, X., Farnham, P., Wells, J., Boyd, K., Fry, C., Bartley, S., Farnham, P., Dimova, D., Dyson, N., Christensen, J., Cloos, P., Toftegaard, U., Klinkenberg, D., Bracken, A., Trinh, E., Heeran, M., Stefano, L. D., Helin, K., Lee, B., Bhinge, A., Iyer, V., Helin, K., Wu, C., Fattaey, A., Lees, J., Dynlacht, B., NGWU, C., Harlow, E., Cao, A., Rabinovich, R., Xu, M., Xu, X., Jin, V., Farnham, P., Levy, D., Darnell, J., Darnell, J., Bowman, T., Garcia, R., Turkson, J., Jove, R., Yang, J., Stark, G., Horvath, C., Hartman, S., Bertone, P., Nath, A., Royce, T., Gerstein, M., Weissman, S., Snyder, M., Qureshi, S., Salditt-Georgie, M., Darnell, J., Li, X., Leung, S., Qureshi, S., Darnell, J., Stark, G., Xiao, W., Lindner, D., Kalvakolanu, D., Martinez-Moczygemba, M.,

Gutch, M., French, D., Reich, N., Ghislain, J., Wong, T., Nguyen, M., Fish, E., Shi, Y., Lee, J., Galvin, K., Shi, Y., Seto, E., Chang, L., Shenk, T., Kim, J. D., Faulk, C., Kim, J., Lee, T., Shi, Y., Schwartz, R., Shrivastava, K. A., Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., Weng, Z., Cheng, Y., Handwerger, S., Wilberding, J., Castellino, F., Collins, P., Kobayashi, Y., Nguyen, L., Trinklein, N., Myers, R., Zhang, X., Odom, D., Koo, S., Conkright, M., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., Kadam, S., Ecker, J., Emerson, B., Hogenesch, J., Unterman, T., Young, R., Montminy, M., Bannert, N., Avots, A., Baier, M., Ing, E. S., Kurth, R., Guzman, R. D., Martinez-Yamout, M., Dyson, H., Wright, P., Honda, K., Yanai, H., Negishi, H., Asagiri, M., Sato, M., Mizutani, T., Shimada, N., Ohba, Y., Takaoka, A., Yoshida, N., Taniguchi, T., Odom, D., Zizlsperger, N., Gordon, D., Bell, G., Rinaldi, N., Murray, H., Volkert, T., Schreiber, J., Rolfe, P., Ord, D. G., Fraenkel, E., Bell, G., Young, R., Gotea, V., Visel, A., Westlund, J., Nobrega, M., Pennacchio, L., Ovcharenko, I., Wang, J., Zhuang, J., Iyer, S., Lin, X., Eld, T. W., Greven, M., Pierce, B., Dong, X., Kundaje, A., Cheng, Y., Rando, O., Birney, E., Myers, R., Noble, W., Synder, M., Weng, Z., Martin, D., Pantoja, C., Miñán, A., Valdes-Quezada, C., Moltó, E., Matesanz, F., Bogdanović, O., de la Calle-Mustienes, E., Domínguez, O., Taher, L., Furlan-Magaril, M., Alcina, A., Cañón, S., Fedetz, M., Blasco, M., Pereira, P., Ovcharenko, I., Recillas-Targa, F., Montoliu, L., Manzanares, M., Guigó, R., Serrano, M., Casares, F., Gómez-Skarmeta, J., Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., Liu, Y., Siepel, A., Pollard, K., Haussler, D., Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E., Stark, A., Lin, M., Kheradpour, P., Pedersen, J., Parts, L., Carlson, J., Crosby, M., Rasmussen, M., Roy, S., Deoras, A., Ruby, J., Brennecke, J., Curators, H., Project, B., Hodges, E., Hinrichs, A., Caspi, A., Paten, B., Park, S., Han, M., Maeder, M., Polansky, B., Robson, B., Aerts, S., van Helden, J., Hassan, B., Gilbert, D., Eastman, D., Rice, M., Weir, M., Hahn, M., Park, Y., Dewey, C., Pachter, L., Kent, W., Haussler, D., Lai, E., Bartel, D., Hannon, G., Kaufman, T., Eisen, M., Clark, A., Smith, D., Celniker, S., Gelbart, W., Kellis, M., Blanchette, M., Sinha, S., Cooper, G., Stone, E., Asimenos, G., Green, E., Batzoglou, S., Sidow, A., Kasowski, M., Grubert, F., el Nger, C. H., Hariharan, M., Asabere, A., Waszak, S., Habegger, L., Rozowsky, J., Shi, M., Urban, A., Hong, M., Karczewski, K., Huber, W., Weissman, S., Gerstein, M., Korbel, J., Snyder, M., McDaniell, R., Lee, B., Song, L., Liu, Z., Boyle, A., Erdos, M., Scott, L., Morken, M., Kucera, K., Battenhouse, A., Keefe, D., Collins, F., Willard, H., Lieb, J., Furey, T., Crawford, G., Iyer, V., Birney, E., Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W., Liu, X., Fu, Y., Weng, Z., Rosenbloom, K., Dreszer, T., Pheasant, M., Barber, G., Meyer, L., Pohl, A., Raney, B., Wang, T., Hinrichs, A., Zweig, A., Fujita, P., Learned, K., Rhead, B., Smith, K., Kuhn, R., Karolchik, D., Haussler, D., Kent, W., Zheng, L., Baumann, U., Reymond, J., Schneider, T., Stormo, G., Gold, L., Ehrenfeucht, A., Lin, J., Collins, P., Trinklein, N., Fu, Y., Xi, H., Myers, R., Weng, Z., Pruitt, K., Tatusova, T., Maglott, D., Hsu, F., Kent, W., Clawson, H., Kuhn, R., Diekhans, M., Haussler, D., Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C., Chrast, J., Lagarde, J., Gilbert, J., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S., Guigo, R., McLean, C., Bristor, D., Hiller, M., Clarke, S., Schaar, B., Lowe, C., Wenger, A., Bejerano, G., Bowie, M., Troch, M., Delrow, J., Dietze, E., Bean, G., Ibarra, C., Pandiyan, G., Seewaldt, V., Grandvaux, N., Servant, M., TenOever, B., Sen, G., Balachandran, S., Barber, G., Lin, R., and Hiscott, J. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13(9):R50.

Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241.

Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic acids research*, 41(Database issue):D165–70.

Worsley Hunt, R., Mathelier, A., Del Peso, L., and Wasserman, W. W. (2014). Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC genomics*, 15(1):472.

Worsley Hunt, R. and Wasserman, W. W. (2014). Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome biology*, 15(7):412.

Xu, M. and Su, Z. (2010). A novel alignment-free method for comparing transcription factor binding site motifs. *PloS one*, 5(1):e8797.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., and Taipale, J. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell*, 154(4):801–13.

Yáñez-Cuna, J. O., Kvon, E. Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. *Trends in Genetics*, 29(1):11–22.

Yang, L., Orenstein, Y., Jolma, A., Yin, Y., Taipale, J., Shamir, R., and Rohs, R. (2017). Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Molecular Systems Biology*, 13(2):910.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W. W., Gordân, R., and Rohs, R. (2014). TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic acids research*, 42(Database issue):D148–55.

Yeo, G. W., Van Nostrand, E. L., and Liang, T. Y. (2007). Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genetics*, 3(5):814–829.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C., and Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337).

Zabidi, M. A. and Stark, A. (2016). Regulatory Enhancer???Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in Genetics*, 32(12):801–814.

Zhang, S., Zhou, X., Du, C., and Su, Z. (2013). SPIC: a novel similarity metric for comparing transcription factor binding site motifs based on information contents. *BMC systems biology*, 7 Suppl 2(Suppl 2):S14.

Zhao, S. (2013). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. 29(6).

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 112(15):4654–4659.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic acids research*, 41(Web Server issue):W56–62.

Zhu, J., Yamane, H., and Paul, W. (2010). Differentiation of effector CD4 T cell populations. *Annu Rev Immunol.*, 28(1):445–489.

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70.