Integrative approaches to study TF-DNA interactions

Thesis for the Philosophiae Doctor (Ph.D.)

University of Oslo, 2020

Marius Gheorghe

Computational Biology & Gene Regulation Group
Centre for Molecular Medicine Norway
Faculty of Medicine
University of Oslo

Cancer Genome Variation
Department of Cancer Genetics
Institute for Cancer Research
Oslo University Hospital

Faculty of Medicine
University of Oslo

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo

ISBN 978-82-8377-622-5

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Contents

Acknowledgements

The present work has been carried out at the Centre for Molecular Medicine Norway, Faculty of Medicine, University of Oslo and in collaboration with the Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital.

Firstly, I would like to express my gratitude towards Dr. Anthony Mathelier, my principal supervisor. I thank him for the opportunity, support, and the patience he showed throughout my entire Ph.D. Secondly, I would like to thank Dr. Vessela Kristensen, my co-supervisor. I am grateful for her support, understanding, and experience shared throughout our collaboration.

I would also like to thank the members of the Computational Biology and Gene Regulation Group, namely Dr. Aziz Khan for being such a dedicated and inspiring person always willing to share and spread his knowledge and experience, Dr. Jaime Castro for the fruitful discussions and exchange of knowledge, and Dr. Roza Berhanu Lemma for her ever joyful attitude and for sharing her Ph.D. dissertation and defence experience. I also thank Dr. Xavier Tekpli and Dr. Thomas Fleischer from the Cancer Genome Variation Group for sharing their biological insights on cancer research. I also thank Georgios Magklaras and George Marselis for systems support as well as all the people that took part in the collaborations developed throughout my Ph.D.

Most importantly, I would like to acknowledge my parents, Doina and Tudor Gheorghe, and my sister Ligia Gheorghe, for believing in me and supporting me in any way imaginable. They are the only ones who truly know me and I thank them for trusting in me. This is for you.

Gheorghe Marius.

List of papers

Paper I

ReMap2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments.
Chèneby, J., **Gheorghe, M.**, Artufel, M., Mathelier, A., and Ballester, B. (2018). *Nucleic Acids Research*, 46(D1):D267–D275.

Contribution: design and development of the data processing pipeline and processing of one third of the data available for this publication/online services. Contributed to and revised the manuscript.

Paper II

JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.
Khan, A.†, Fornes, O.†, Stigliani, A.†, **Gheorghe, M.**, Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A.. (2018), *Nucleic Acids Research*, 46(D1):D260–D266.

Contribution: provided processed ChIP-seq data that was used for the database update and processed new data generating new and updated transcription factor binding profiles. Contributed to and revised the manuscript.

Paper III

MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations.
Fornes, O.†, **Gheorghe, M.**†, Richmond, P. A., Arenillas, D. J., Wasserman, W. W., and Mathelier, A. (2018), *Scientific Data*, 5:180141.

Contribution: provided and processed data to predict transcription factor binding sites. Contributed to and revised the manuscript.

Paper IV

A map of direct TF-DNA interactions in the human genome
Gheorghe, M., Sandve, G. K., Khan, A., Chèneby, J., Ballester, B., and Mathelier, A. (2019), *Nucleic Acids Research*, 47(4):e21–e21.

Contribution: design and development of the data processing pipeline, data gathering, development of all programming scripts used within the project, result assessment, all figure and table generation. Wrote and revised the manuscript.

Paper V

TF-regulons: identifying direct targets of transcription factors
Gheorghe, M. and Mathelier, A.
Manuscript

Contribution: design and development of the processing workflow, provided data and data gathering, development of programming scripts, result assessment, generating all figures and all tables. Wrote and revised the drafted manuscript.

Paper VI

Identifying key TFs driving ER positive and ER negative breast cancer subtypes
Gheorghe, M., Tekpli X., Fleischer, T., Kristensen, V., Mathelier, A.
Manuscript

Contribution: design and development of the processing workflow, data gathering and data sharing, development of programming scripts, result assessment, generating figures. Wrote and revised the drafted manuscript.

Abstract

Transcription of DNA into RNA is mainly regulated through a complex interplay between proteins and chromatin at *cis*-regulatory elements (CREs). Two critical types of CREs are promoters and enhancers, which are involved in turning on or off gene transcription. Promoters are regions of DNA ensuring the transcription of genes and are located around transcription start sites (TSSs). Enhancers are regulatory regions that are located linearly distal on the genome with respect to the genes they are regulating but in close proximity in the three-dimensional space of the nucleus of a cell. Transcription factors (TFs) are key proteins that bind to promoters and enhancers in order to ensure transcription at appropriate rates in the correct cell types. They interact with DNA at their TF binding sites (TFBSs) in a sequence specific manner, recognizing specific DNA motifs. Through their binding, they play an essential role in the development and physiology of an organism. Therefore, genome-wide identification of TFBSs is a critical step to decipher transcriptional regulation and how this process is altered in diseases. Our understanding of the mechanisms controlling gene expression is still limited. Advantageously, consortia and individual laboratories provide a milestone with the generation of large-scale data sets for the identification, collection, and categorization of CREs.

When focusing on TFBSs, the current most common practice to locate them *in vivo* is to perform chromatin immunoprecipitation (ChIP) followed by massive parallel DNA sequencing (ChIP-seq). Unfortunately, it has been recurrently shown that these assays are prone to produce experimental artifacts. Thus, delineating *bona fide* TF-bound regions from experimental noise is still an ongoing problem. This affects not only primary measurements but also the ability to compare data from multiple studies or to perform integrative analyses across multiple data types. Together with whole genome sequencing and expression quantification data, accurately predicting the regulatory factors ruling gene transcription will allow us to identify functional CREs and ultimately assess how alterations occurring in these regions contribute to disease onset.

To address this problem, my research has focused primarily on improving our

capacity to predict and analyse direct TF-DNA interactions at a genome-wide scale. As TFs recognize their binding site through a complex interplay between base/nucleotide and DNA shape readout, computational models have been instrumental in the prediction and characterization of TF-DNA interactions. To acquire large amounts of ChIP-seq data, we participated in the latest update of the ReMap database, which provides an atlas of ChIP-seq peaks in the human genome. Using the extended ReMap data collection, we participated in the latest update of the JASPAR database, which hosts high quality TF binding motifs. In a first attempt to predict TFBSs in the human genome, we combined both ReMap ChIP-seq peaks and JASPAR binding motifs to update the MANTA database, which predicts TFBSs genome-wide and assesses the impact of single nucleotide variants at TFBSs. Subsequently, we developed *ChIP-eat*, a uniform data processing pipeline, from raw ChIP-seq data to high confidence direct TF-DNA interactions. The *ChIP-eat* pipeline is centered around an entropy-based, non-parametric, data-driven algorithm allowing automatic identification of direct TF-DNA interactions supported by strong computational and experimental evidence. The predictions were *a posteriori* validated using *in vitro* assay data. This work led to the creation of the publicly available *UniBind* database in an effort to provide the community with a critical resource that will enable an array of studies aiming at better understanding transcriptional regulation. UniBind hosts the complete set of TFBS predictions from almost two thousand ChIP-seq data sets processed using four different computational models.

A second aim of my research was the identification of genes that are specific targets of TFs. In the past years, the focus has been mainly on identifying DNA regions bound by TFs. Unfortunately, TF binding is not necessarily associated to regulatory function. To better understand the functional impact of TF-DNA interactions, methods have to be developed to identify not only potential TFBSs but to characterize the ones that are regulatory functional. Besides the genome-wide TFBS predictions obtained using the *ChIP-eat* pipeline, we employed several other layers of genomic information, such as: TSSs, promoter and enhancer locations, sequence conservation scores, TF binding affinity scores, and gene-enhancer associations in a ranked list approach to better model the prediction of TF-specific target genes. The impact of the features upon the predictions was assessed by means of overlap with known sets of associated genes and by gene ontology term similarities.

The third part of my research consisted of identifying key TFs driving es-

trogen receptor positive (ER+) and estrogen receptor negative (ER-) breast cancers. As ER- breast cancers have poor or nonexistent response to hormone based therapies, as opposed to ER+, it is crucial to identify the TFs responsible for disruptions in the gene regulatory program leading to carcinogenesis in the two cancer subtypes. Making use of the high confidence TFBS predictions hosted in *UniBind* and donor samples from TCGA, we were able to identify candidate TFs that can ultimately serve as input in the development of targeted therapies for the two subtypes of breast cancer.

Abbreviations

ATAC-seq - assay for transposase-accessible chromatin followed by sequencing
BEM - binding energy model
BRE - B recognition element
CAGE - cap analysis of gene expression
ChIP - chromatin immunoprecipitation
ChIP-seq - chromatin immunoprecipitation followed by sequencing
CRE - *cis*-regulatory element
CRM - *cis*-regulatory module
DBD - DNA binding domain
DE - differentially expressed
DiMO - discrete motif optimizer
DNA - deoxyribonucleic acid
DNAse - deoxyribonuclease
ECDF - empirical cumulative distribution function
EMSA - electrophoretic mobility shift array
ER- - estrogen receptor negative
ER+ - estrogen receptor positive
eRNA - enhancer RNA
EZ - enrichment zone
GO - gene ontology
IC - information content
IHC - immunohistochemistry
Inr - initiator element
MITOMI - mechanically induced trapping of molecular interactions
mRNA - messenger RNA
ncRNA - non-coding RNA
NGS - next-generation sequencing
PBM - protein binding microarray
PCR - polymerase chain reaction
PFM - position frequency matrix
PIC - pre-initiation complex
PPM - position probability matrix
PSSM - position specific scoring matrix

PWM - position weight matrix
qPCR - quantitative PCR
RNA - ribonucleic acid
RNA-seq - RNA sequencing
RNAP - RNA polymerase
SELEX - systematic evolution of ligands by exponential enrichment
SNV - single nucleotide variant
TAD - topologically associating domain
TF - transcription factor
TFBM - transcription factor binding motif
TFFM - transcription factor flexible models
TFBS - transcription factor binding site
TSS - transcription start site

1

Introduction

Most living organisms consist of cells, the base units of life. Within each cell, a collection of molecules allows it to live and proliferate. Throughout the past century, scientists have tried to grasp how a single cell can give rise to a fully developed and fully functional multicellular organism. In multicellular organisms, how is it possible to obtain from one fertilized egg such a collection of different morphologies and different functionalities? Starting in the embryonic stage, the stem cell divides into new cells with virtually the same genome, and at the end of the developmental process the assembly of cells forms an intricate pattern of outstanding complexity and precision. All this process is attributed to the genome itself, but how does the cell determine its final pattern?

Decades ago, the flow of genetic information was described by three different biochemical processes: (i) replication, the process by which the DNA generates copies of itself, (ii) transcription, where DNA is copied into RNA, and (iii) translation, the process by which RNA is synthetized into proteins (Figure 1.1). Nevertheless, special cases of information transfer can also occur under special circumstances, such as RNA to RNA or RNA to DNA due to virus infected cells (Crick, 1970). Around the mid-twentieth century, Barbara McClintock revealed that to obtain such cellular and morphological diversity, gene expression regulation plays a central role (McClintock, 1950). Yet, the discovery of the gene regulatory mechanism was provided by François Jacob and Jaques Monod, who showed that protein synthesis is regulated by a distinct class of proteins termed *repressors*, which mediate gene activity through their binding to short sequences of DNA termed *operators* (Jacob and Monod, 1961). This led to the advent of transcriptional regulation research as a subfield of molecular biology and subsequently it was established as a level of gene expression regulation. A couple of decades later, the *activators* were discovered (McKay and Steitz, 1981). This new class of regulatory

proteins was shown to positively regulate gene expression, as opposed to *repressors*. These two classes of proteins are commonly termed transcription factors (TFs). As pioneer studies were performed in bacteria, it was found that TFs bind DNA at *promoters*, specific regions situated upstream from the gene body. Generalizing the studies in metazoa and plants, it was found that transcriptional regulation is coordinated through the interplay of several regulatory elements (Biłas et al., 2016; Riethoven, 2010).

Although the principle of passing genetic information seems straightforward, the ability of the genome to develop complex cellular states that lead to the formation of different tissues, which achieve specific functions, should not be underestimated. This in turn raises the question: how is gene expression regulated during cellular differentiation and developmental stages? It has been shown in multiple studies over the years how gene expression regulation plays a role in a wide range of biological processes, such as T-cell differentiation (Zhu et al., 2010) or cell reprogramming (Takahashi and Yamanaka, 2006) and also how gene dysregulation may lead to carcinogenesis (Ell et al., 2013) or disease in general (Mathelier et al., 2015).

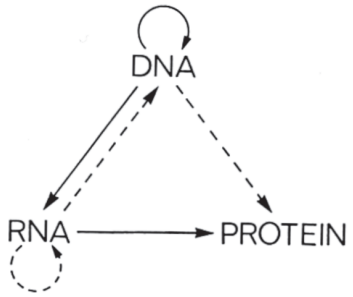

Figure 1.1: How the genetic material is replicated and biological information passed. The solid lines illustrate probable transfers (supported by evidence) with RNA to RNA being possible due to RNA viruses and the dashed lines possible transfers (no experimental evidence or theoretical requirement). Figure from Crick (1970) .

1.1 Generalities of transcriptional regulation

The proper development of each cell in an eukaryotic organism is highly dependent on the tightly regulated mechanism of gene expression. This ensures that cells can evolve into diverse cell types, achieve different functionalities, and respond to their environment and stress. This phenomenon is attained by expressing only a specific subset of genes in a certain cell type, at a certain

developmental stage, and only at specific levels (Lelli et al., 2012). There are two main levels at which gene expression regulation is achieved: (i) through transcription, which converts DNA into RNA and (ii) through translation, which converts RNA into proteins. In this work, the focus will be only on gene expression regulation achieved through transcriptional regulation.

Eukaryotic gene expression can be categorized into two main classes: (i) basal, associated with the housekeeping genes and (ii) activator-dependent or inducible gene expression, the latter being subject to differentiation and developmental constraints arising from context-specific stimuli (Thomas and Chiang, 2006; Weake and Workman, 2010). Such a complex process requires a tight regulation that is achieved by the coordinated and combined action of regulatory elements and RNA polymerase (Barrett et al., 2012). Besides regulatory elements, such as general or sequence specific transcription factors (TFs) and co-regulators, chromatin remodelers play an important role in the transcription process by altering chromatin. Chromatin structural changes will either support transcription by allowing for protein interactions if the chromatin is in an open state or repress transcription if in a closed state (Li et al., 2007).

1.2 The organization of the genome

One of the most important molecules within a cell is the deoxyribonucleic acid (DNA). This double helix shaped molecule consists of two complementary strands, each representing a sequence of nucleic acids called nucleotides. More specifically, a strand of DNA is composed of a series of four different nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T) coupled by covalent phosphodiester linkages. The four nucleotides are pairwise complementary, A to T and G to C, and due to the double stranded nature of DNA, each pair of nucleotides, also called a base pair, is connected by hydrogen bonds between the two strands (Figure 1.2 (a)). Moreover, the two strands have an antiparallel orientation, each starting from the 5'-end of the first base (i.e., the 5th carbon of the sugar backbone), which has a phosphate group, to the 3'-end of the last base (i.e., the 3rd carbon of the sugar backbone), which has a hydroxyl group (Figure 1.2 (b)). This structure allows for strand directionality, and base pairing is possible due to the opposite

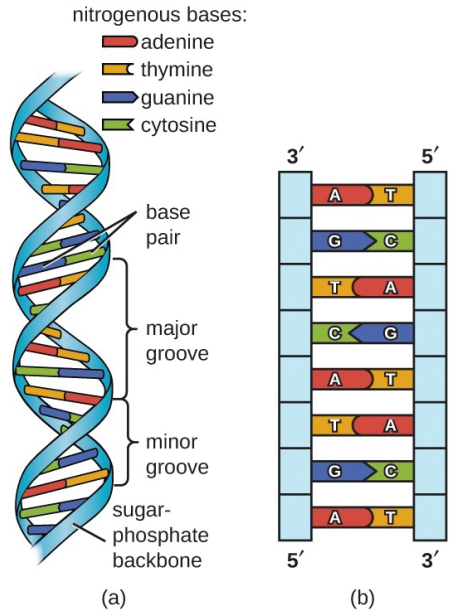

Figure 1.2: The structure of the DNA. The sugar phosphate backbone of the two strands are connected by covalent bonds forming base pairs (a). The two strands have an anti parallel orientation allowing for directionality, which is essential in replication and transcription. Figure adapted from OpenStax CNX.

orientation of the two strands. Such conformation is therefore essential for the replication and transcription of genetic information.

The complete set of genetic material, crucial for the development and functionality of a cell, is encoded within the DNA and is called the genome. In eukaryotic organisms, the genome is hosted within the nucleus of the cell. In humans, the genetic information is distributed across 23 pairs of chromosomes, 22 of the pairs are called autosomes and the last pair represents the allosomes, or the sex chromosomes.

1.2.1 Chromatin organization within the nucleus

The length of a DNA molecule is around two meters and consists of ~3.3 billion base pairs (bp). To fit such a quantity of genetic material in the

nuclear space of an eukaryotic cell, not larger than a few micrometres, DNA is compacted through folding and *via* interactions with specific proteins called histones (Hulton et al., 1990). The complex formed between histone proteins and the DNA during the compaction process is called the chromatin. The base unit of the chromatin is the nucleosome, which consists of eight positively charged histone proteins (Kornberg, 1974), two of each of H2A, H2B, H3, and H4 (Thomas and Kornberg, 1975), around which negatively charged DNA strips of ~147 bp wrap ~1.7 times (Davey et al., 2002; Hansen, 2002) (Figure 1.3). To reach two full turns around the histone octamer, an additional 20 bp DNA is wrapped by a linker histone (Simpson, 1978; Kepert et al., 2003). This represents the first level of DNA compaction.

Within each chromosome, multiple nucleosomes are inter-connected *via* linker DNA to form arrays of nucleosomes under a *beads-on-a-string* structure (Figure 1.3). Subsequently, linker histones form higher order chromatin structures through the folding of several nucleosomes into a 30 nm chromatin fiber, which in turn folds through 300 nm loops. These loops are further compressed into a supercoil struture, eventually resulting in a chromatid that forms the *arms* of the chromosome (Woodcock and Ghosh, 2010) (Figure 1.3). This multiple level compression of the nucleosomal DNA results in a seven-fold size reduction after the first level, followed by a 40-50 fold compression through nucleosome-nucleosome interactions mediated by histone H1 (Thoma et al., 1979). The individual 30 nm solenoid fibers resulting from the second compression step are tail-associated and further condensed to form the chromosome *arms* or chromatid.

1.2.2 Genomic compartments and topologically associating domains

Chromatin conformation capture methodologies such as 3C and Hi-C allow for the identification of interacting chromatin regions by determining the frequency of two DNA loci being in close proximity and/or physical contact (Naumova et al., 2012; van Berkum et al., 2010). A closer look at genome-wide chromatin interaction maps suggests that intra-chromosomal regions segregate by preferential interactions into two distinct compartments, denoted A and B (Lieberman-Aiden et al., 2009) (Figure 1.5). These ~ 5M bp regions alternate along the genome, and it has been shown that these

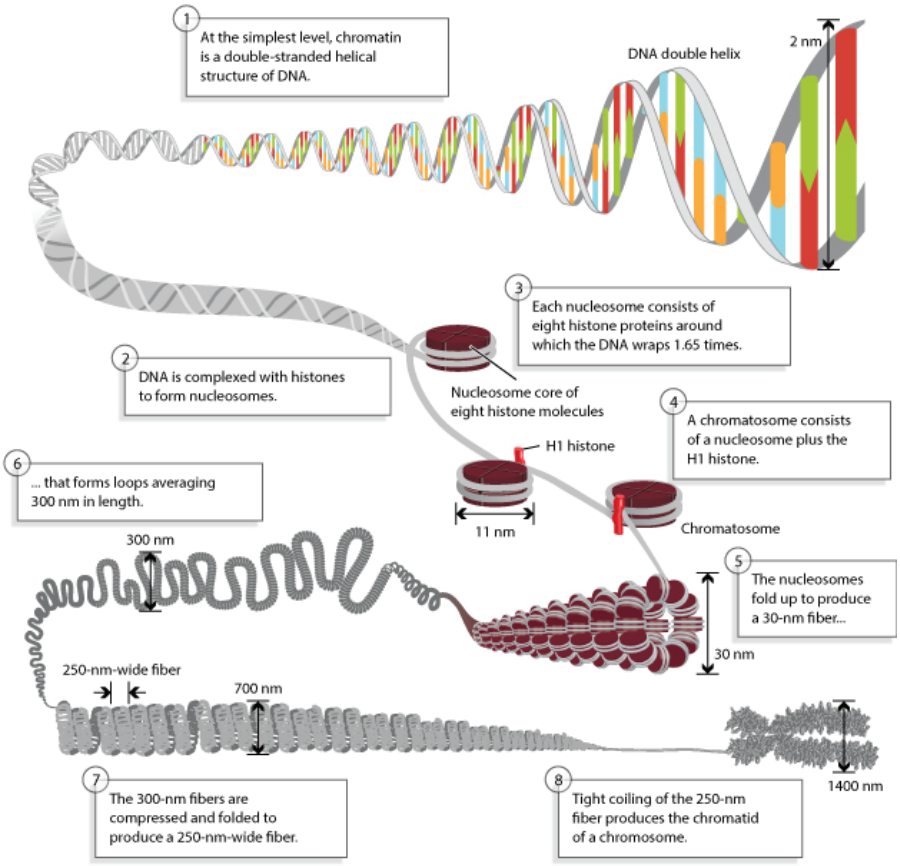

Figure 1.3: The different levels of nucleosomal DNA compaction. The naked DNA is wrapped around histone octamers called nucleosomes, which in turn are folded into a 30nm solenoid fiber. Further, the chromatin fiber loops and becomes supercoiled to finally for the chromatid of the chromosomes. Figure adapted from Pierce (2012).

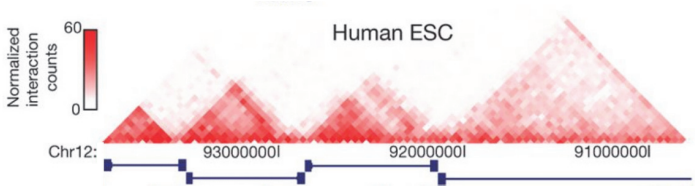

Figure 1.4: The topologically associating domains or TADs. An interaction map of the 12th chromosome in human embryonic stem cells. The blue lines under each triangular shape define the borders of each TAD. Figure adapted from Dixon et al. (2012).

chromatin domains associate with other domains of similar activity levels (Imakaev et al., 2012). Therefore, compartment A is generally associated with active transcription, whilst compartment B is mainly classified as inactive (Bonev and Cavalli, 2016). It was shown that what allows for the separation of the two regions is the preferential activity of transcriptional regulators, the density of genes, and the DNA sequence composition (i.e., the differential frequency of the nucleotides, such as GC content) (Gibcus and Dekker, 2013).

Within the A/B compartments, smaller subunits of folded DNA were identified. These regions were called topologically associating domains (TADs) and they cover between 0.5 and 1 mega bps (Figure 1.4). These are chromosomal regions where an unusually large number of chromatin interactions occur (Dixon et al., 2012; Dekker et al., 2013). Importantly, the A/B segregation is distinct from the TADs. In contrast to the TAD specific conservation across cell types and tissues (Nora et al., 2012; Dixon et al., 2012), the A/B compartments are tissue dependent and gene expression dependent (Xie et al., 2017). Within TADs, chromatin forms smaller loops on the order of hundreds of kilo bps. The TADs are defined in such a way that the number of intra-TAD loci interactions is much higher compared to the inter-TAD interactions (Dixon et al., 2012). Consequently, this structure facilitates interactions between genomic regions situated within the same TAD and isolates the genomic regions situated between two TADs (Symmons et al., 2014). Recent studies have shown that inter-TAD interactions also occur and lead to spatial chromatin reorganization at higher levels within the nucleus (Gonzalez-Sandoval and Gasser, 2016; Paulsen et al., 2019). The TAD borders are partly defined by genetic elements. Nora et *al.* showed that if this border is removed at the inactivation center of chromosome X, a partial blending of adjacent TADs

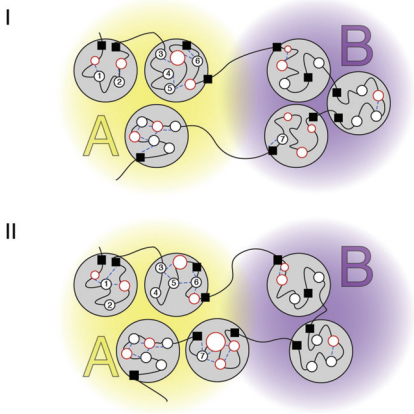

Figure 1.5: A representation of the A and B genomic compartments. These higher order structures host TADs that are grouped based on their activity. Depending on gene activity and their regulators, the compartments are reorganized but not the TADs. Compartment A is associated with active transcription, whilst compartment B is generally transcriptionally inactive. Figure adapted from Gibcus and Dekker (2013).

takes place (Nora et al., 2012). Moreover, it was also shown that disruptions in TAD boundaries can cause dysregulation in gene expression and lead to developmental disorders (Lupiáñez et al., 2016).

1.2.3 *Cis*-regulatory elements

The genome can be divided into coding and non-coding regions. The coding genome hosts the transcriptional units that are sequences of DNA replicated into RNA and subsequently translated into proteins. This accounts for roughly 2% of the entire genome. Each chromosome contains hundreds to thousands of such protein coding DNA sequences. The non-coding genome is the ensemble of DNA sequences that do not encode for proteins and represents >98% of the entire genome. Part of the non-coding sequences are transcribed into non-coding RNA molecules (ncRNA). Some of the ncRNAs can become functional units of the genome; however, they are not translated into proteins but regulate other classes of RNAs in eukaryotes (Mattick and Makunin, 2006). Another crucial class of non-coding DNA sequences are the *cis*-regulatory elements (CREs), which are involved in the transcriptional regulatory process (Wittkopp and Kalay, 2012).

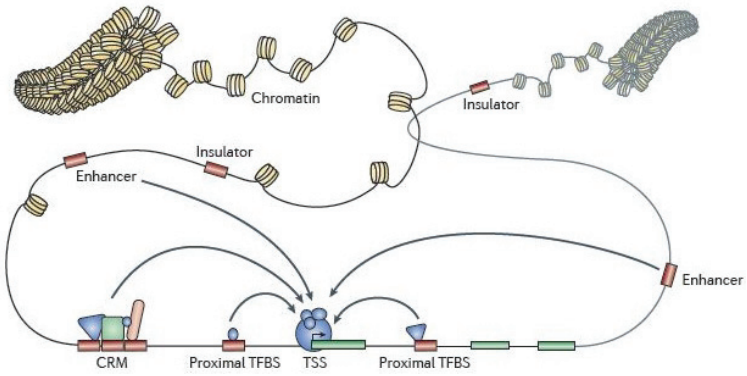

Figure 1.6: The cis-regulatory elements in metazoan transcriptional regulation and their relative positioning with respect to the transcription start site (TSS) of a gene. Figure adapted from Lenhard et al. (2012).

Genomic compartmentalization and chromatin interactions coordinate the interactions between CREs. CREs are genomic regions such as *promoters*, *enhancers*, *silencers*, and *insulators* that come in close 3D proximity through DNA looping during transcription (Lenhard et al., 2012) (Figure 1.6). In brief, there are two types of promoter regions: promoters that are capable of recruiting the RNA-polymerase (RNAP) complex without the help of other regulatory elements, called *strong* promoters, and *weak* promoters that need other regulatory elements to stabilize the RNAP and initiate transcription (Qin et al., 2010). Other regulatory regions were discovered at a later stage, and they represent distal DNA sequences with respect to the promoter: the *enhancer* regions allow the activation and/or amplification of target gene expression and the *silencer* regions suppress the expression of target genes (Banerji et al., 1981). The *insulator* regions represent the boundary between open and closed chromatin, and they reduce gene expression by inhibiting the enhancer activity (Kolovos et al., 2012). In the scope of this work, only promoters and enhancers will be detailed.

Promoters. For the transcription of a gene to occur, the transcription machinery must be recruited at CREs situated upstream from the gene transcription start site (TSS), called the *core promoter*. The transcription of a gene is initiated once the RNAP complex is loaded and stabilized. The core promoter is a ~50 bp region located on the same DNA strand as the gene to be transcribed and contains short specific regions, such as the TATA-box

(Lifton et al., 1978), the initiator (Inr) (Smale and Baltimore, 1989), BREs (Lagrange et al., 1998; Littlefield et al., 1999), downstream core promoters (DPE or DCE) (Burke and Kadonaga, 1997), and RNAP components (Figure 1.7).

An individual promoter does not necessarily contain all these regions (Lenhard et al., 2012), but the Inr is the most common (Xi et al., 2007). The TATA-box allows the recruitment of the *TATA-binding protein* and is present in the core promoter of 10 to 20 percent of the protein coding genes, whilst Inr is present in 40 to 60 percent of promoters (Yang et al., 2007). Characteristic to the core promoters is the variability of their constituents across gene types and across species (Todeschini et al., 2014). Accordingly, a core promoter alone can rarely ensure the transcription of a gene. Usually the binding of general TFs is needed at the *proximal promoter* region located immediately upstream of the core promoter region (Sainsbury et al., 2015). Here, the term promoter will be used to represent both core promoters and proximal promoters.

A gene can have multiple TSSs and therefore multiple promoters. Recruitment of the transcription machinery can thus occur at different alternative promoters. In turn, this enables the production of an increased variety of RNA transcripts. Overall, promoters contain short sequences of DNA or motifs that are recognized and bound by sequence specific proteins called transcription factors (TFs), involved in the initiation of transcription (Cooper et al., 2006). Transcription was thought to take place in an unidirectional manner, occuring downstream from the promoter and the gene TSS, but it has been shown that in metazoa, bidirectional transcription is very common (Andersson et al., 2015). It is not yet clear if this is a consequence of spatially close promoters or the presence of RNAP in open chromatin regions with a high concentration of TFs. Moreover, this phenomenon occurs equally in *enhancers* and it is thought to be driven by the overrepresentation of TFs, specific histone modifications, or extended regulatory regions facilitating the binding of additional TFs, among others (Bagchi and Iyer, 2016; Ibrahim et al., 2018). Nevertheless, during bidirectional transcription occurring at promoters of protein-coding genes, only one transcript produces a stable mRNA (Wei et al., 2011) (Figure 1.8). This is not the case for enhancers, where transcripts in both directions are found to be unstable (Andersson et al., 2014b).

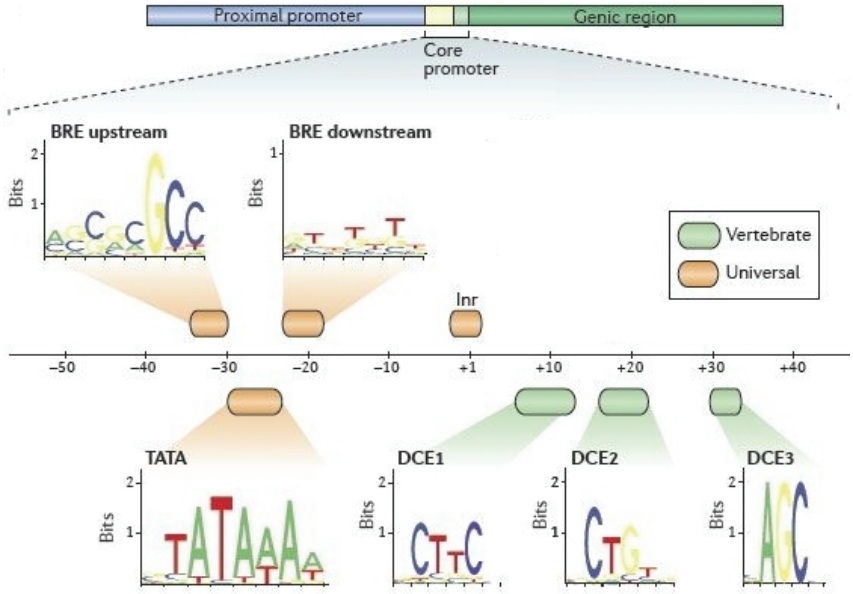

Figure 1.7: The elements of the core promoter in metazoa and specific to vertebrates. The TATA-box is flanked by the B-recognition element (BRE) and the initiator (Inr) downstream from it, followed by the downstream core promoter elements (DCE). Figure adapted from Lenhard et al. (2012).

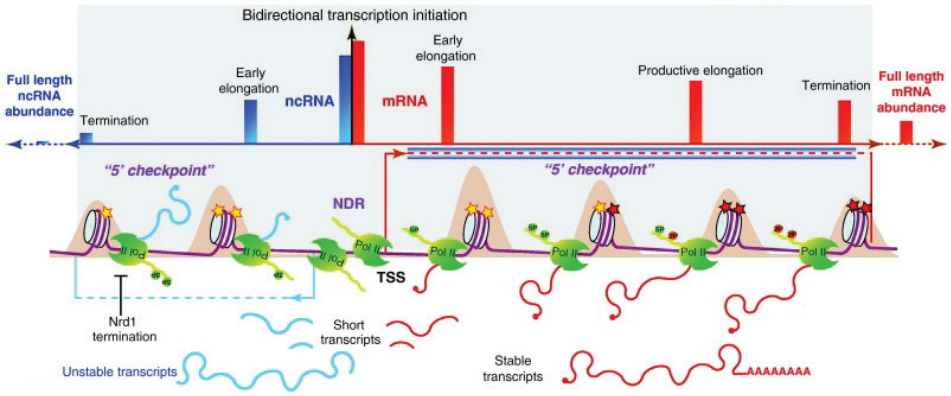

Figure 1.8: The bidirectionality of transcription at the promoter level. The RNA transcript generated upstream from the TSS is unstable as opposed to the one occurring downstream from the TSS. Figure adapted from Wei et al. (2011).

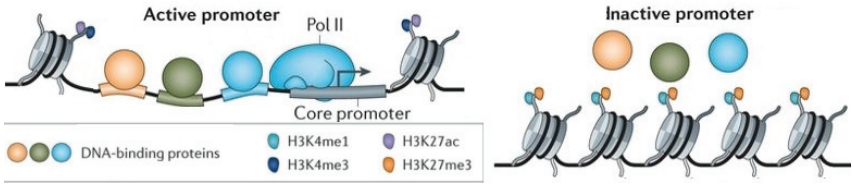

Figure 1.9: A schematic of an active promoter (left) and an inactive one (right), and their epigenetic environments. Figure adapted from Shlyueva et al. (2014).

Importantly, active promoters are located in nucleosome depleted, open chromatin regions flanked by promoter-associated nucleosomes. Epigenetic marks, such as H3K4me3 and H3K27ac histone modifications, are enriched at active promoters, while H3K27me3 is usually associated with inactive promoters (Tserel et al., 2010; Lawrence et al., 2016) (Figure 1.9).

Enhancers. The first enhancer was described during an experiment aiming at cloning the DNA sequence of a human virus (SV40). It was observed that the expression of the targeted gene was considerably increased, and its enhancement was associated to the 72 bp repeated sequence situated in the beginning of the viral gene (Banerji et al., 1981). The increase in gene expression was observed regardless of the viral sequence orientation or the distance to the gene TSS. Currently, enhancers are widely studied and numerous types were discovered *in vivo* and in different cell types. Enhancers are defined as short DNA sequences that have the capacity to increase the expression of their target genes (Blackwood and Kadonaga, 1998) regardless of location and orientation. These CREs can be located close to the target gene, within the gene itself, or distal to the target gene (Figure 1.10).

The size of the enhancer regulatory sequence varies in length, from 10 bp to 1000 bp and contains from a couple to tens of binding motifs for a wide range of TFs (Blackwood and Kadonaga, 1998; Yáñez-Cuna et al., 2013). Once bound to the enhancer, TFs recruit co-activators such as *p300*. Importantly, their regulatory action is irrespective to the orientation of the target gene. A distinct characteristic of enhancers is their ability to interact directly (physically) or indirectly (through other TFs) with their target gene(s) or other CREs (Kolovos et al., 2012) (Figure 1.11). Briefly, in the *tracking* model (Figure 1.11 (upper left)) the regulatory proteins are "charged" at the enhancer level and travel along the chromatin to reach the promoter; the *linking* model (Figure 1.11 (lower left)) implies that the regulatory proteins undergo poly-

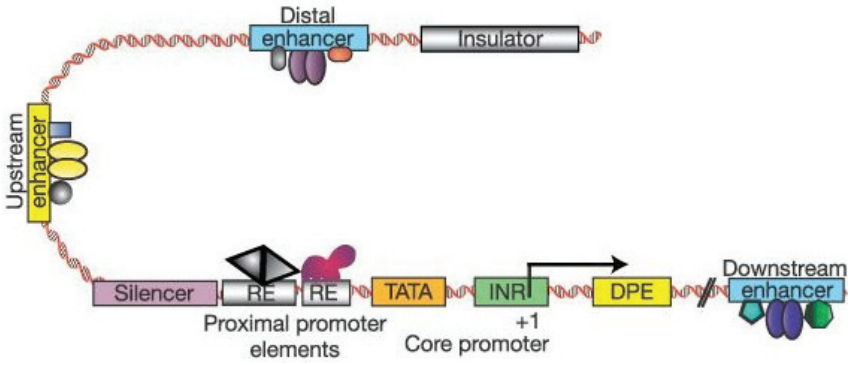

Figure 1.10: The positioning of enhancer regions relative to the target gene and its promoter. Figure adapted from Levine and Tjian (2003).

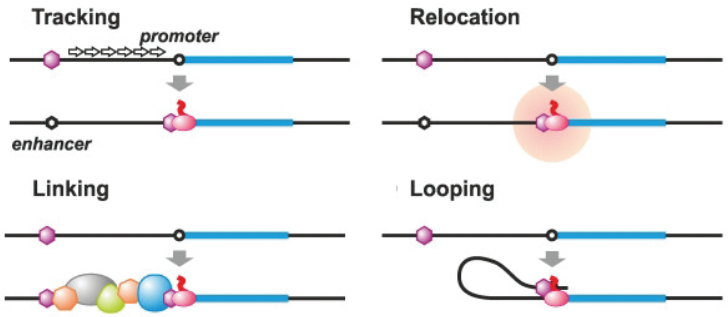

Figure 1.11: Different proposed models of physical interaction between enhancers and the promoters of their target gene(s). Figure adapted from Kolovos et al. (2012).

merization in the direction of the promoter; the *relocation* model implies that a gene relocates to a nuclear compartment that favors the enhancer-promoter interaction; finally, the *looping* model describes the situation in which an enhancer comes into proximity with a promoter *via* protein-protein interactions. These interactions are important factors in the transcriptional regulation process (Shlyueva et al., 2014). Moreover, enhancers have a low nucleosome occupancy and they are hypersensitive to DNaseI, an enzyme able to cleave exposed DNA (Zentner et al., 2011).

Together with DNase I hypersensitivity, epigenetic modifications (e.g., histone modifications) can serve as markers to identify enhancers genome-wide; however, these are not sufficient as the variability in the epigenetic landscape is higher compared to other CREs such as promoters (Heintzman et al., 2007).

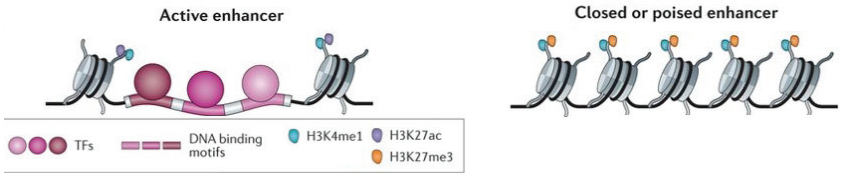

Figure 1.12: A schematic of an active enhancer (left) and an inactive one (right), and their epigenetic environments. Figure adapted from Shlyueva et al. (2014).

In fact, information about specific TFs that preferentially bind DNA at enhancer regions, such as *p300*, could be paired with specific histone marks to increase the reliability of genome-wide enhancer inference. A better method to identify enhancers genome-wide is to use Cap Analysis of Gene Expression (CAGE), which captures the bidirectional transcription at active enhancers (Kodzius et al., 2006; Andersson et al., 2014a). CAGE-predicted enhancers are three times more likely to be validated using reporter assays compared to epigenetically predicted enhancers.

One enhancer can be associated with multiple promoters, and one promoter with multiple distinct enhancers, depending on the biological condition and/or cell differentiation stage. The recruitment of a large number of TFs, as well as other regulatory proteins, requires that the chromatin is in an open state. The chromatin state is cell type dependent and, as a consequence, not all enhancers are active at the same time (Andersson et al., 2014a).

Another property, similar to promoters, is that enhancers are capable of carrying out bidirectional transcription (as captured by CAGE) due to the recruitment of the RNAP complex and the high number of TFs. This produces a different class of RNAs called enhancer RNA (eRNA). Studies have proposed that eRNAs can be used in labelling active enhancers (Andersson et al., 2014a), while others suggest that it is not a representative feature for enhancer activity (Hah et al., 2013). However, eRNAs have been associated with co-factor recruitment and stability of enhancer-promoter looping (Hsieh et al., 2014).

Altogether, it is clear that enhancers, as promoters, play an essential role in modulating gene expression, and they are necessary in biological processes, such as development and cell differentiation (Bonn et al., 2012). The dif-

ference in enhancer activity across different tissues allows for better understanding of the mechanisms controlling the diversity of cellular types that share the same genome. Nevertheless, the difference between enhancers and promoters is not easy to establish. The only apparent distinction lies in the different classes of RNAs they produce, but their context-dependent functionality suggests that they are likely to represent the two extremes of a gradient of CRE functions (Andersson, 2015). Also, enhancers were found to be more cell type specific compared to promoters (Heinz et al., 2015).

1.3 Transcription factors

TFs are proteins that control transcription through sequence specific DNA binding (reviewed in (Lambert et al., 2018)). This class of proteins consists of ~1600 members characterized by a DNA-binding domain (DBD) and a transactivation domain (Wingender et al., 2013; Lambert et al., 2018). The DBD facilitates the binding of the protein in a sequence-specific manner to DNA, whereas the transactivation domain provides the activation potential of the protein (Brivanlou and Darnell, 2002; Vaquerizas et al., 2009). By their binding at CREs (e.g., promoters or enhancers) they can activate or repress gene transcription. Through regulatory DNA sequence recognition, they control the level of gene expression by recruiting the transcription machinery to CREs (Sikorski and Buratowski, 2009). TFs may play different roles through their binding to DNA. They can induce basal transcription by interacting with general TF complexes (Sikorski and Buratowski, 2009) or, together with co-activators and specific enzymes, they can alter the chromatin structure through histone modifications and initiate or repress gene transcription (Brivanlou and Darnell, 2002; Li et al., 2007; Venters and Pugh, 2009). Moreover, a hierarchical model of TFs binding to DNA has been proposed, with some TFs altering the chromatin conformation and others responding to changes in chromatin state (Sherwood et al., 2014).

1.3.1 Functional classification

Activators vs. repressors. Depending on their impact on transcription, TFs can be classified as *activators* and *repressors*. *Activator* TFs ensure gene

transcription and contain at least a DBD and an activation domain. In contrast, *repressor* TFs, as the name states, inhibit gene expression by masking the transcriptional activation sequences (i.e., where RNAP binds) either by competing with *activator* TFs, by direct interaction with other TFs, or by affecting the chromatin structure (Payankaulam et al., 2010). For decades, this dichotomy was used to separate TFs into these two distinct groups (Jacob and Monod, 1961; McKay and Steitz, 1981; Busby and Ebright, 1999), but recently it has been shown that this partition is not straightforward. Specifically, some TFs can act as both *activators* and *repressors* in given biological conditions (Lee et al., 2012; Slattery et al., 2014).

Pioneers vs. settlers vs. migrants. The majority of DNA is wrapped around nucleosomes and thus inacessible to TFs due to the presence of histones. Higher-order chromatin structures and repressor complexes also contribute to DNA inacessibility (Symmons et al., 2014). Nevertheless, regulatory events still occur through cooperative binding of several TFs to the target site of a gene, activating gene expression. For regulatory events to occur at the chromatin level, *pioneer* TFs are required. This special class of TFs can access the closed chromatin independently of other factors and facilitate the binding of other TFs by altering the chromatin conformation (Zaret and Carroll, 2011; Young, 2011). *Pioneer* TFs are also necessary when sequential binding of several TFs over time is required (Young, 2011). Two distinct roles define *pioneer* TFs: shaping the chromatin landscape for other TFs (active role) and enhancing transcription as a direct consequence of their initial binding to chromatin (passive role) (Zaret and Carroll, 2011; Sherwood et al., 2014). Interestingly, it was shown that some *pioneer* TFs are directional, asymmetrically opening the chromatin (Kundaje et al., 2012; Sherwood et al., 2014).

Even though *pioneer* TFs do not bind a high fraction of their available binding motifs, a different class of TFs, termed *settler* TFs, binds all the genomic motifs found in a chromatin accessible region. This class of TFs follows a simple rule: binding DNA if it is in an open chromatin state. Thus, settler TFs solely rely on the ability of the *pioneer* TFs to open the chromatin and therefore their binding can be determined based on chromatin accessibility data (Sherwood et al., 2014). Other than the small fraction defined as *settlers*, the majority of non-pioneer TFs are classified as *migrants*. These TFs bind only a subset of their available genomic motifs, even if found in an open chromatin region. Therefore, their selectivity is likely dependent on specific

co-factor interactions (Sherwood et al., 2014). From a chromatin-based perspective, TF binding follows a hierarchical model; *pioneer* TFs opening the chromatin, which is in turn populated by *settler* TFs and combinations of *migrant* TFs, as the latter two TF classes do not have the capacity to evict nucleosomes (Sherwood et al., 2014; Slattery et al., 2014).

1.3.2 Structural classification

Since the discovery of TFs, several attempts have been made to classify them, either by function or by structure. As the DBD is characteristic to TFs, it was used to structurally classify the TFs. Consequently, different repertoires of human TFs have been generated based on the similarity of their DBD (Harrison, 1991; Wingender, 1997; Vaquerizas et al., 2009). For instance, a detailed classification of human TFs can be found in the resource TFClass (Wingender et al., 2018). Other such resources exist, specific to a certain organism (Ishihama et al., 2016) or multiple organisms (Portales-Casamar et al., 2009). Further stratification was obtained based on functional criteria of TFs (Wingender, 1997) (Table 1.1). The main role of the DBD is to act as an initiator of weak interactions with DNA (i.e., unspecific binding) during the sequence scanning process. Importantly, DBDs are able to recognize not only *monads* or continuous DNA sequences, but also *dyads* or sequences containg spacers of fixed or variable length (Helden et al., 2000). This allows for flexibility in the DNA sequence recognition. Commonly, TF classes have highly similar binding motifs. However, in *zinc fingers*, the most abundant TF class in mammals, the binding motifs are very different due to the presence of spacers (Ravasi et al., 2003).

Depending on the organism, the number of TF families varies and so does the number of members within each family. Notably, TF families can be associated to specific biological functions, such as the basic helix-loop-helix (bHLH), which is associated to neurogenic differentiation and myogenic differentiation, among others (Jones, 2004). Certain TF families can be associated to the same clade whilst others are specific to subclasses of organisms. Furthermore, a TF family can host from one member to several hundreds (Wingender et al., 2018).

Table 1.1: Structural classification of TFs and rank definitions. Table adapted from Wingender et al. (2013).

Level	Rank denomination	Definition	Example
1	Superclass	General topology of the DBD	Zinc-coordinating DBDs
2	Class	Structural blueprint of the DBD	Nuclear receptors with C4 zinc fingers
3	Family	Sequence and functional similarities	Thyroid hormone receptor-related factors (NR1)
4	Subfamily	Sequence-based subgroupings	Retinoic acid receptors (NR1B)
5	Genus	TF gene	RAR-α
6	Factor 'species'	TF polypeptide	RAR-α 1

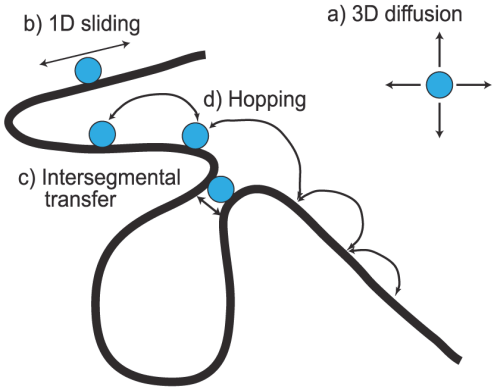

Figure 1.13: The four different TF motion models explaining the dynamics of TFBS recognition. Figure adapted from Schmidt et al. (2014).

1.3.3 Binding site recognition

As mentioned, a key feature of the TFs is their capacity to bind DNA in a sequence specific manner, as opposed to other co-factors taking part in transcriptional regulation. The DBDs of TFs can recognize short sequences of DNA, usually between 6 and 20 bp, which are termed transcription factor binding sites (TFBSs). The binding of a TF to a TFBSs is central for transcription initiation and transcriptional rate regulation. Each TF has multiple TFBSs across the genome, the number varying from a handful to even hundreds of thousands in the human genome. Identifying bound TFBSs genome-wide is a highly complex task. The mechanism by which TFs recognize their binding sites is not yet fully understood. How does a protein *scan* for such short specific sequences of DNA among billions of nucleotides in such a short time?

Four motion models have been proposed to encompass the dynamics of this

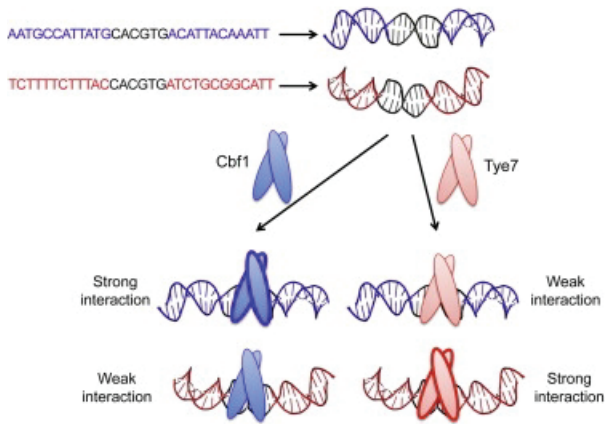

Figure 1.14: The importance of the DNA regions flanking the TF binding motif. The CACGTG enhancer box (E-box) is shown in a GC-rich environment (blue) and in an AT-rich environment (red). The nucleotide composition of the motif environment influences the binding affinity through the DNA shape. Figure adapted from Gordân et al. (2013).

process: (a) 3D diffusion, implying that the TF is freely moving within the nucleus, (b) 1D sliding, implying that the TF slides along short regions of the DNA, (c) intersegmental transfer, implying that the TF moves between two linearly distal DNA segments that are in close proximity due to looping, and (d) hopping, implying that the TF *jumps* across the DNA (Schmidt et al., 2014) (Figure 1.13). These models should be viewed as complementary and not antonymic. As TFs float freely within the nucleus (a), they can initiate weak interactions with DNA situated in an open chromatin region and slide along until they find their binding sites (b). Nevertheless, they can be dislodged from their binding site due to DNA movement or interaction with other proteins. During the sliding process (b), they can *jump* between segments of DNA or over closed chromatin regions (c) or *hop* between linearly distal regions of DNA which are in close 3D proximity. Hopping can occur even between chromosomes (Schmidt et al., 2014).

Furthermore, it has been shown that TFBSs tend to localize in DNA regions having a GC composition (i.e., guanine-cytosine content) similar to the TF binding motif (Dror et al., 2016). Therefore, TFBS flanking regions are important for TF binding affinity and specificity (Gordân et al., 2013; Schöne et al., 2016) (Figure 1.14).

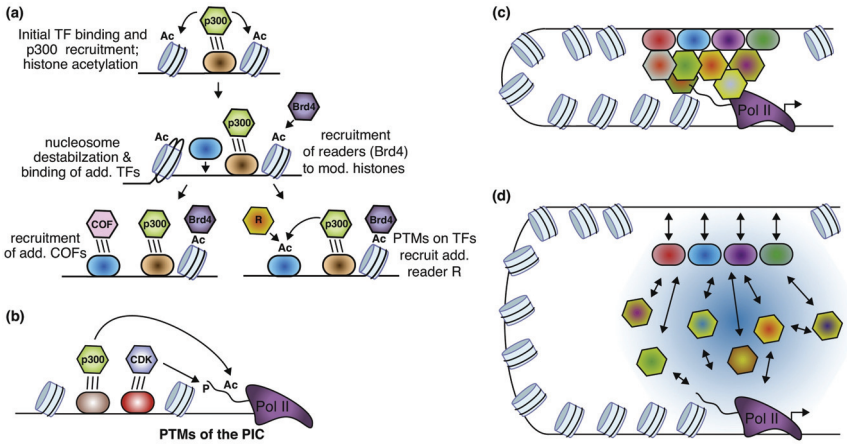

Figure 1.15: The interplay between TFs and co-factors. Schematic of the different mechanisms by which co-factors can impact gene expression: chromatin remodeling and histone modifications (a), post-translational modification of the RNAP complex (b), stabilizing RNAP with co-factors (c), and interacting with PIC (d). Figure from Reiter et al. (2017).
.

1.3.4 Combinatorics and cooperativity

In eukaryotes, TFs can combine through protein-protein interactions to form protein complexes and thus "coordinate", or act in unison to achive the required regulatory effect. The entire set of TF combinatorics defines a "dictionary" that follows a specific *grammar* (Spitz and Furlong, 2012). Although it is known that TFs may act in a cooperative way to tune transcriptional activation or repression, it is still not clear what the individual contribution of each TF to gene expresion regulation is.

Two models have been proposed to describe the effect of TF combinatorics at enhancer level on gene expression regulation. The *enhanceosome* model depicts the enhancer DNA sequence as a scaffold for other proteins to form one protein complex. This model implies high cooperative and coordinate action between enhancer-bound proteins; therefore, alterations at individual binding sites would have a drastic impact on the enhancer activity, as the output of the enhancer is binary modeled (Thanos and Maniatis, 1995) (Figure 1.16 (A)). A second model is the *billboard* model, which suggests that TFBSs within an enhancer can be individually disposed, as the enhancer-bound proteins do not act as a unit. In fact, they are considered an ensemble of

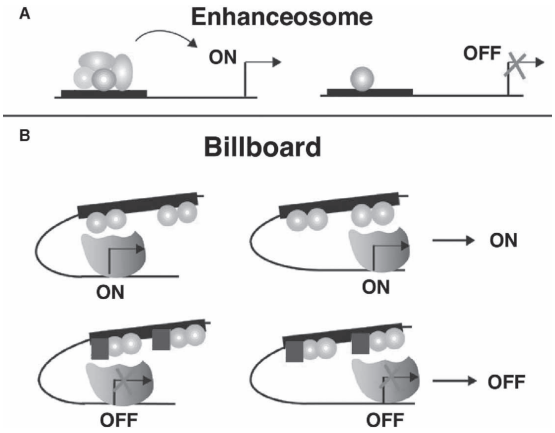

Figure 1.16: The two models of TF combinatorics at enhancer level. The *enhanceosome* model with high cooperativity among TFs (A) and the *billboard* model suggesting that TFs act as individual units (B). Figure from Arnosti and Kulkarni (2005) .

separately acting TFs, independently interacting with their targets (Arnosti, 2003) (Figure 1.16 (B)). Therefore, the *billboard* model allows for more flexibility as compared to the *enhanceosome* model, assuming individual contributions rather than an overall output summed across the proteins within a complex.

It has been shown in animals that TFs have the tendency to cluster at CREs and form so called *cis*-regulatory modules (CRMs), but how the positioning of each TF within the CRM can influence gene regulation is not fully understood (Hardison and Taylor, 2012). To add to the complexity of the problem, in addition to the presence of the TF, the relative orientation of the motif, the distance between motifs, the order of the motifs, presence of co-factors, and even biological context should be taken into account when studying the behavior of a TF or a TF combination (Spitz and Furlong, 2012; Whitfield et al., 2012).

1.4 Identification of TF-DNA interactions

To decipher transcriptional regulation, it is essential to understand how TFs regulate the expression of their target genes. A first step is to identify TF-DNA binding events, specifically TFBSs. Based on this information, regula-

Table 1.2: A classification of experimental assays to identify TF-DNA binding events. Table adapted from Geertz and Maerkl (2010).

Assay	Approach	Technique	Yield	Throughput	Resolution
EMSA	Gel shift	in vitro	around 10 sites	low	few binding sites only
BIAcore	Surface plasmon resonance	in vitro	up to 100 sites	low	few binding sites only
PICh	Reverse ChIP	in vivo	one genomic site	low	–
DNAse footprint	Gel shift	in vitro	local genomic region	low	Nucleotide resolution
MITOMI	Mechanical trapping	in vitro	100 to 1000 sites	low-high	Nucleotide resolution
SELEX, CASTing	Selection of target	in vitro	>200 000 sites	high	few high affinity binding sites
HT-SELEX, Bind-n-Seq	Selection of target coupled to NGS	in vitro	>200 000 sites	high	Nucleotide resolution feasible
PBM, CSI	Protein binding microarray	in vitro	up to 1 milion sites	high	Nucleotide resolution feasible
DIP-chip	DNA immunoprecipitation	in vitro	all genomic sites	high	between 100 and 500 bp
ChIP–chip	ChIP coupled to microarray	in vivo	all genomic sites	high	between 100 and 500 bp
ChIP-seq	ChIP coupled to NGS	in vivo	all genomic sites	high	between 100 and 500 bp
ChIP-exo/ChIP-nexus	ChIP + exonuclease + NGS	in vivo	all genomic sites	high	Nucleotide resolution feasible
DamID	TF mediated DNA methylation profiling	in vivo	all genomic sites	high	between 100 and 500 bp
DNAseI-seq	DNaseI sensitivity profiling coupled to NGS	in vivo	all genomic sites	high	Nucleotide resolution feasible
FAIRE-seq	DNaseI sensitivity profiling coupled to NGS	in vivo	all genomic sites	high	Between 500 and 1000 bp
ATAC-seq	DNaseI sensitivity profiling coupled to NGS	in vivo	all genomic sites	high	Between 200 and 600 bp

tory networks can be inferred and subsequently one can assess how disruptions in these regulatory networks can cause gene expression dysregulation. Over the years, several experimental assays have been designed in this scope, varying between the characterization of TF-DNA binding affinities to the genome-wide indentification of TFBSs for a given TF *in vivo*. Using these data, computational models have also been developed in parallel with experimental assays aiming at modeling TFBSs and predicting *bona fide* TF-DNA interactions.

1.4.1 Experimental approaches

TF-DNA interactions are identified by both *in vitro* and *in vivo* experimental assays. *In vitro* assays aim at identifying the binding specificities and affinity of a protein to a certain nucleotide sequence, while *in vivo* assays aim at identifying the binding location within the genome (i.e., TFBSs). These assays can be further classified as low-throughput and high-throughput. For instance, low-throughput *in vivo* assays can identify the exact genomic location of around a dozen TFBSs at nucleotide resolution, while high-throughput assays can identify TF binding *regions* genome-wide under certain experimental conditions. Nevertheless, subsequent computational processing is needed to determine the *bona fide* TFBSs. Table 1.2 contains a non-exhaustive list of such assays and a brief description of each.

In the following subsections, the assays that have been widely used and their characteristics will be briefly detailed.

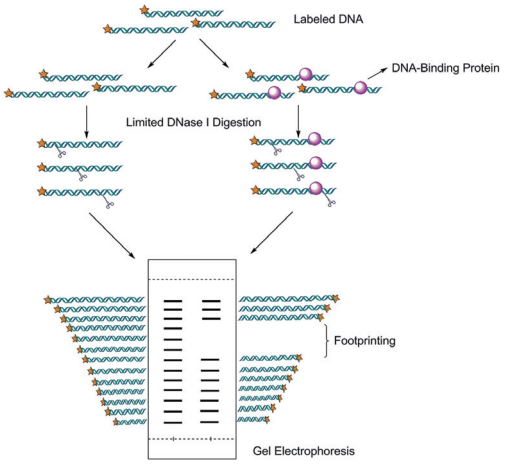

Figure 1.17: Schematic of the DNAse I footprint assay to identify the exact location of the DNA bound regions. Figure from Song et al. (2015)

1.4.1.1 Low-throughput assays

DNAse I footprinting. DNAse I footprinting is an established assay that can identify TF-DNA interactions at single nucleotide resolution (Galas and Schmitz, 1978). This *in vitro* assay is based on the molecular properties of the enzyme deoxyribonuclease (DNAse) which is able to degrade DNA fragments that are not bound/protected by a protein. Consequently, the DNA fragments where a protein (e.g., TF) is bound are preserved. The bound protein will leave a so-called *footprint* that becomes visible during the gel electrophoresis step, as opposed to the cleaved naked DNA that is used as a control (Figure 1.17). This procedure allows for the identification of the binding site of the protein on the DNA. Through polymerase chain reaction (PCR) amplification (Mullis et al., 1989), the isolated DNA fragments can be used to identify the exact genomic location of the protein binding site or the TFBS in the case of TFs (Galas and Schmitz, 1978).

In the past decades, DNAse I footprinting has been central to identifying ligand-DNA interactions, and it was also employed in drug screening and measurement of thermodynamic and kinetic properties of interactions with DNA (Brenowitz et al., 1986; Ellis et al., 2007). However, there are several downsides to the method, such as the preparation and duration of the experiment, amount of biological materials needed, and the limited yield (i.e.,

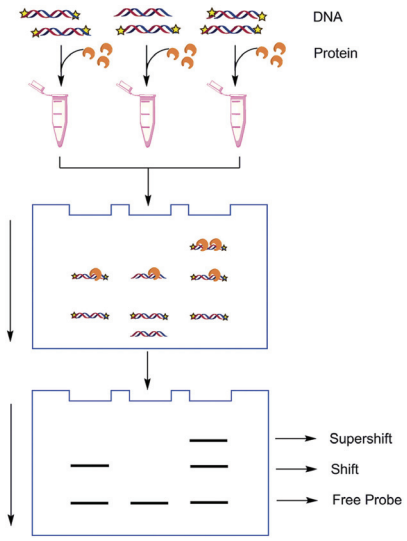

Figure 1.18: Schematic of the EMSA assay to identify the presence or absence of a protein bound to the DNA. Figure from Song et al. (2015) .

only a few binding regions per experiment). As detailed later on, modifications have been made to the protocol during the past years, dampening its downsides.

Electrophoretic mobility shift assay. Another *in vitro* low-throughput assay making use of gel electrophoresis to detect and study nucleic acid-protein interactions is the electrophoretic mobility shift assay (EMSA) (Garner and Revzin, 1986). This experimental procedure allows one to determine the presence or absence of a protein in a given genomic sequence but does not detect the exact location of the bound region. In other words, it allows one to study which proteins, such as TFs, have preferential binding to specific DNA sequences. As in DNAse footprinting, the naked DNA (i.e., DNA not bound by any protein) is used as a control in the gel electrophoresis step, by using its molecular weight as a signature pattern in the gel. If a protein is bound to a DNA sequence, it will increase the molecular weight of the sequence, thus changing the speed at which it travels through the gel and consequently changing the gel pattern (Figure 1.18).

Moreover, the binding affinity of a protein to a DNA sequence can be classified as weak or strong, also based on the gel shifting patterns (Cann, 1998).

This can help determine the preferred DNA binding motif of the protein or the protein complex. The set of DNA sequences with the highest binding affinity for a given protein can be thus identified, as the experiment allows for multiple sequences (but still very few) to be evaluated in a single run. Nevertheless, prior knowledge of the DNA sequence is required.

Mechanically induced trapping of molecular interactions. A different approach for the identification of TF-DNA interactions introduces microfluidics to enable mechanically induced trapping of molecular interactions (MITOMI) (Maerkl and Quake, 2007). This assay represents the transition from low-throughput to high-throughput experimental assays (Geertz and Maerkl, 2010). In brief, it allows one to generate the binding energy landscape of a protein based on the binding interactions occurring at equilibrium. Thousands of micro-arrayed DNA sequences that are printed on microfluidic chips allow measurement of the affinity of every interaction occurring between the protein of interest and the micro-arrayed DNA spots. Even though this approach in theory reduces the experimental design time and the study materials needed, its design and setup can become cumbersome, see Maerkl and Quake (2007) for a schematic and a description of the protocol.

1.4.1.2 High-throughput assays

Protein binding microarrays. Another widely used *in vitro* technique for the identification of TF binding affinities is the protein binding microarray (PBM) assay (Mukherjee et al., 2004). This was the first technique initially designed to be high-throughput and it is able to identify the binding events of a given protein in a genome-independent manner (Berger and Bulyk, 2009). The principle of PBMs is as follows: all oligonucleotides (or aptamers) of a chosen length k (usually 8 or 10, denoted k-mers) are represented as de Bruijn sequences and segmented into sub-sequences overlapping a given number of bases, each sub-sequence having different flanking regions. Next, these single-stranded sequences are fixed on the microarray and become double-stranded *via* primer extension. After the TF(s), tagged with an epitope are added, only the sequences bound by the TF(s) are kept, *via* immunodetection. This implies the use of specific fluorescent antibodies which allows calculation of the signal intensities of each k-mer sequence, which in turn is used to score the binding affinities to the sequences. Figure 1.19 shows a schematic of the

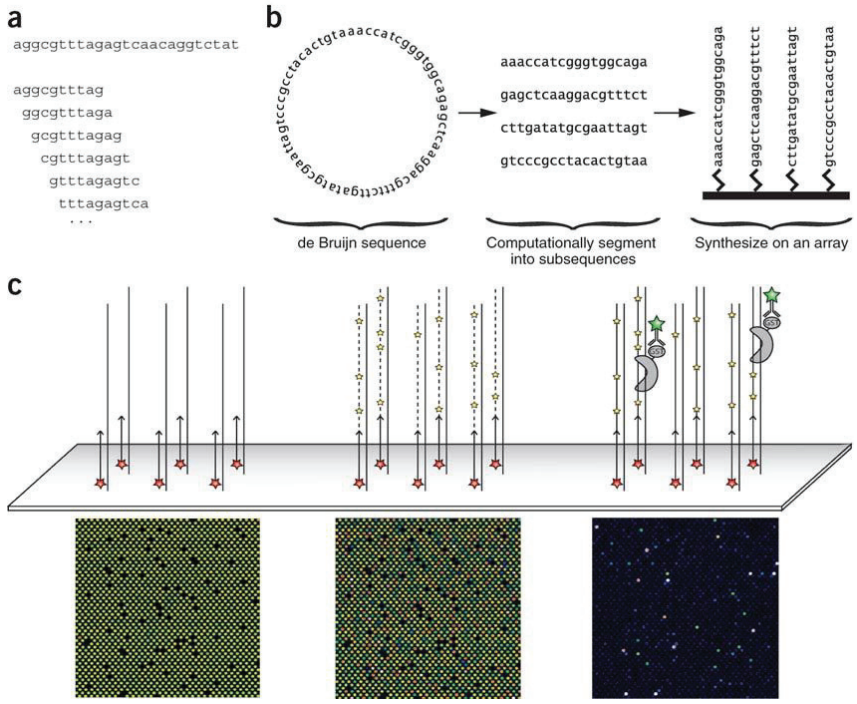

Figure 1.19: Overview of the protein binding microarray assay. A complete set of overlapping k-mers is generated (a) and converted into a de Bruijn sequence (b). Next, the de Bruijn sequence is partitioned into sub-sequences that overlap by two bases and have different flanking regions, which are synthesized on the microarray (b). The fluorescent signal intensities provide an affinity score for the bound sequences (c). Figure from Berger et al. (2006) .

PBM workflow.

Evaluating all the possible k-mer combinations allows the detection of both strong and weak binding events, as opposed to the SELEX technique which favors the strong affinity events (see below). The complete set of scored sequences can then be used to infer a preferred binding sequence for a given TF (Figure 1.20). Besides the clear benefits of detecting both strong and weak binding sites and working regardless of the genome (sequenced or not), the PBM technique has a couple of important limitations: the exact genomic location of the inferred binding motifs cannot be determined and the inferred binding sites cannot be tested. This technique was developed to solely study the binding affinity of a given protein to a given nucelotide sequence.

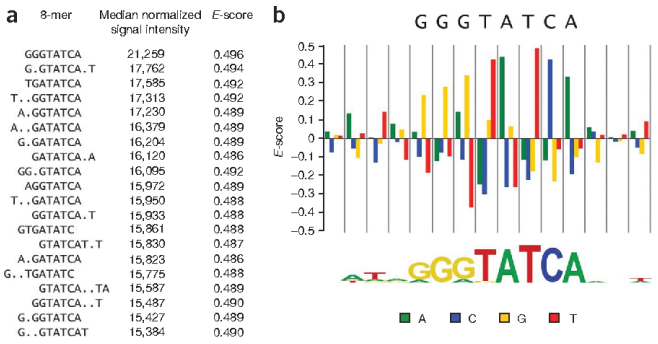

Figure 1.20: Inferring the genomic sequence for the TF binding motif. All the k-mers that were scored, including "wildcard" nucleotides (i.e., each dot within the k-mer sequence) are assigned an enrichment score (i.e., E-score) (a) and a consensus motif representative for the TF binding sequence is built (b). Figure from Berger and Bulyk (2009) .

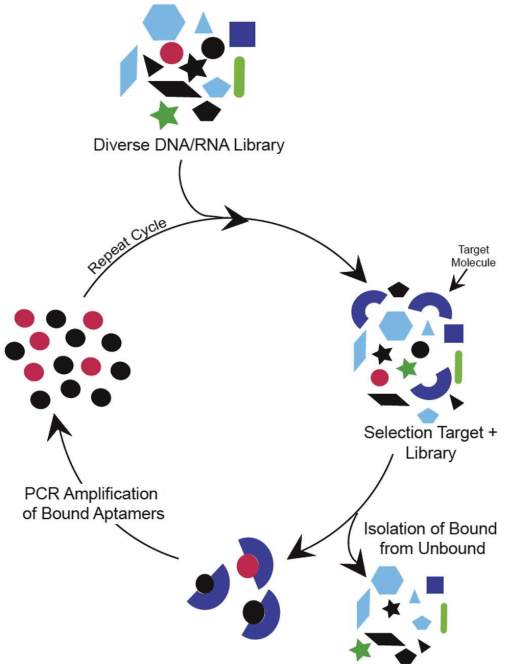

Figure 1.21: The SELEX cycle. A schematic of all the steps for *in vitro* identification of preferentially bound sequences by a target protein or molecule. Figure adapted from Wu and Kwon (2016) .

Systematic evolution of ligands by exponential enrichment. A technique generally used to determine binding preferences *in vitro* is SELEX (*S*ystematic *E*volution of *L*igands by *EX*ponential enrichment) (Tuerk and Gold, 1990). This assay is based on the following principle: the protein or molecule of interest, such as a TF, is incubated together with a large pool of randomly generated aptamers (or oligonucleotides). After a washing step involving immunoprecipitation, in which the unbound aptamers are removed, the remaining sequences are amplified through PCR, and the entire cycle is repeated several times. As a result, the aptamers that present the highest binding affinity are preserved and enriched (Figure 1.21).

Initially designed to work with only one protein or molecule of interest at a time and with a relatively limited yield (Table 1.2), more recent variants of this technique have bypassed these limitations: (i) high-throughput SELEX (HT-SELEX) is a massively parallelized version of SELEX, able to identify the binding affinities of hundreds of proteins in a single assay (Jolma et al., 2010); (ii) SELEX followed by sequencing (SELEX-seq) was designed to identify the binding sites of TF complexes (Riley et al., 2014); (iii) consecutive affinity purification SELEX (CAP-SELEX) allows selection of aptamers that are bound by two different TFs at the same time (Jolma et al., 2015).

Even if SELEX performs generally well and its setup is straightforward, two major drawbacks are to be taken into account: (i) due to its design, strong binding sites are favored over weak binding sites, which are discarded; (ii) enriched aptamers may represent DNA sequences that do not exist within the genome, as they are randomly generated.

Chromatin immunoprecipitation based methods. Chromatin immunoprecipitation (ChIP) is the experimental technique of choice when investigating protein-DNA interactions *in vivo* (O'Neill and Turner, 1996). ChIP is commonly used to map the cistrome, which is the genome-wide TFBS locations and/or post-translationally modified histones and histone variants (Collas, 2010). A schematic representation of the typical ChIP workflow is depicted in Figure 1.22. In brief, DNA and proteins are covalently bound *in vivo* through formaldehyde cross-linking, which ensures that the proteins of interest, such as TFs, are directly binding the DNA. It is worth noting at this step that using formaldehyde alone is not suitable to study indirect protein-DNA interactions (Zeng et al., 2006). After the cross-linking step, the DNA is fragmented through sonication or DNAse digestion which results

in ~500 bp long DNA segments on average. Using an antibody specific to the protein of interest, the protein-DNA complexes are immunoprecipitated from the chromatin, which results in pulling down the bound fragments of DNA. The precipitated DNA is washed and the cross-linking is reversed, owing to the heat-reversible properties of formaldehyde, thus removing the bound protein. As a last step of the experimental assay, the precipitated ChIP-enriched DNA is purified and ready for analysis (Figure 1.22 (1-4)).

A good practice when performing ChIP assays, is to carry out a control experiment, in which no protein is immunoprecipitated. This serves as background when investigating the enrichment of DNA fragments. While ChIP-based methods have a clear advantage over other assays such as PBM or SELEX, a drawback is the high number of false positives that arise due to experimental material quality, unspecific protein-DNA binding, or other artefacts (Teytelman et al., 2013). Using a control ChIP experiment will help reduce the number of false positives but not completely remove it.

The ChIP experimental yield can be analyzed *via* numerous approaches to identify the genomic locations where protein-DNA interactions occur, such as PCR, quantitative PCR (qPCR), microarrays, labeling and hybridization, or high-throughput sequencing. As we will further see, the results of these ChIP-based methods are the basis for catalogues hosting genome-wide CRE/TFBS annotations.

ChIP-chip. The ChIP-chip (or ChIP-on-chip) technique complements the ChIP assay by identifying enriched genomic regions using microarray chips (Ren et al., 2000). In addition to the ChIP steps (Figure 1.22), the precipitated and purified DNA fragments are denatured to single strand DNA, tagged with a fluorescent bead, and hybridized on a microarray chip. Next, the microarray chip is read, and the intensities of each DNA fragment serve as an entry point for further data analysis. For a detailed description and a schematic of the workflow, please refer to Buck and Lieb (2004). ChIP-chip presents several disadvantages: (i) an extra hybridization step is required before reading the sequences, (ii) the set of DNA sequences is limited due to the design of the microarray chip, (iii) a higher DNA quantity is required, and (iv) the method yields lower resolution with respect to the binding region size.

ChIP-seq. With the advent of next generation sequencing (NGS), high-throughput methods have become fundamental for genome-wide analyses.

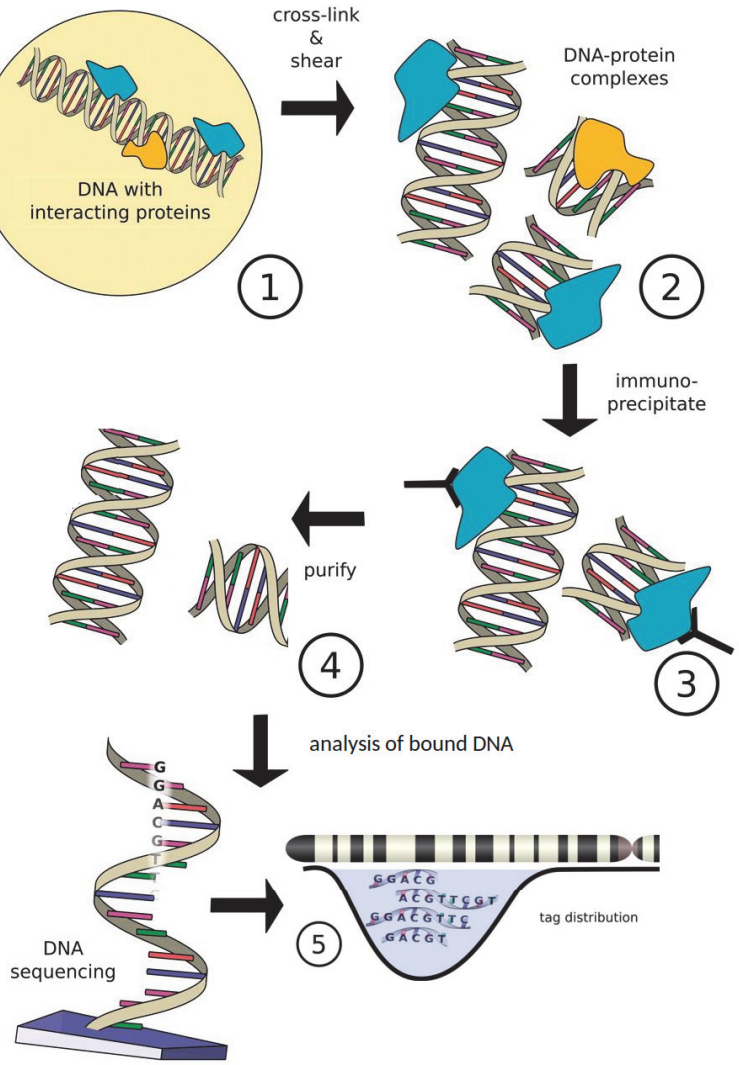

Figure 1.22: Overview of the chromatin immunoprecipitatiopn (ChIP) assay (1-4), followed by DNA sequencing and mapping to the genome (5). Figure adapted from Szalkowski and Schmid (2011).

To date, the most popular technique to identify genomic regions where TF-DNA interactions occur *in vivo*, in a genome-wide manner, is ChIP followed by sequencing (ChIP-seq) (Johnson et al., 2007). The experimental setup for this assay is identical to the one for ChIP (Figure 1.22) with extra steps added for the preliminary processing of the precipitated and purified DNA fragments. The DNA fragments are sequenced using a short-read sequencer and the resulting reads are mapped to a reference genome (Figure 1.22). Next, the genomic regions where a significantly higher number of reads (or tags) as compared to the control, the so-called *ChIP-seq peaks*, are identified using a *peak-caller* algorithm (Pepke et al., 2009).

While ChIP-seq presents many advantages over the methods described so far, it has of course its own drawbacks. The most important of which is the relatively high number of false positives that arise either from experimental or computational artefacts. Experimental artefacts, such as antibody quality or the DNA fragmentation not being equal among the samples, result in an uneven distribution of the reads, while computational artefacts can arise from repetitive DNA sequences appearing as enriched genomic regions or due to the diversity of the data analysis tools that exist, with no standardized parameter settings (Park, 2009; Bailey et al., 2013).

Nevertheless, ChIP-seq remains the method of choice when characterizing genome-wide TF-DNA interactions *in vivo*. Other derivatives of the ChIP assay have been designed and they will be briefly discussed here in comparison with ChIP-seq.

ChIP-exo. The ChIP-exo technique extends the ChIP assay with the addition of an exonuclease digestion step. The exonuclease degrades the DNA from 5' to 3' end, leaving out the fragments bound by the protein. The sequencing and peak-calling steps remain the same as for ChIP-seq. Due to the exonuclease trimming, the ChIP-exo peaks are much narrower and able to reach single nucleotide resolution for TF-DNA interaction identification (Rhee and Pugh, 2011). Therefore, ChIP-exo peaks can correspond to individual TFBSs, which allows for the identification of binding events, or the organization of histones (Rhee et al., 2014; Mahony and Pugh, 2015). For more details about this technique, please refer to Rhee and Pugh (2011).

ChIP-nexus. This technique is highly similar to ChIP-exo with the difference that an auto-circularization and fragment amplification step is added prior to sequencing. This addition allows for better coverage of genomic se-

quences when compared to ChIP-exo. For more details about this technique, please refer to (He et al., 2015).

In comparison with ChIP-seq, the ChIP-exo and ChIP-nexus techniques (i) do not allow for use of a control due to the exonuclease degrading the "naked" DNA; thus, the peak-calling methodology differs (Wang et al., 2014), (ii) the cost and labour is higher compared to ChIP-seq, (iii) due to the additional washing steps, the DNA libraries are more limited compared to ChIP-seq. For a more comprehensive comparison, please refer to (Mahony and Pugh, 2015; Hartonen et al., 2016). Nevertheless, they represent the most accurate techniques for the identification of TF-DNA binding *in vivo*.

1.4.2 Computational approaches to model and predict TF-DNA interactions

Despite the wealth of available ChIP-seq data, computational models are necessary to infer TFBSs, as experiments cannot be performed in all cell types and in all biological conditions. TFs are known to be sequence specific DNA-binding proteins. While ChIP-based assays provide the binding regions of TFs, there is a need to refine those data and extract the actual binding sites that correspond to each TF on a genome-wide scale. As binding preferences of TFs differ between TF classes, a plethora of computational methodologies and tools have emerged aiming at addressing this issue.

Identifying TFBSs implies two separate issues: (i) discovering the *motif* to which a given TF has the highest binding affinity and (ii) identifying instances of these *motifs* genome-wide. For the former, the sequence of nucleotides that is the most representative for a given TF (i.e., the TF binding *motif*, TFBM) needs to be identified. These sequences can be derived from *in vitro* assays or through *de novo* motif discovery, both of which are designed to identify protein binding specificities. The identified *motifs* are used in a pattern matching approach to identify all instances genome-wide. A common approach to computationally derive TFBSs through assays such as ChIP-seq is to use ChIP-seq *peaks*. The short DNA fragments resulting from sequencing, also called *reads*, are mapped to the genome and they pile up at specific genomic locations forming so-called *peaks*. These peaks represent the genomic regions to where more reads map as compared to background (e.g., control

Table 1.3: The IUPAC nomenclature for nucleotide combinations. Table adapted from IUPAC (1985).

Symbol	Bases	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

or random expectation). Such *peaks*, have an average length between 300 and 400 bp and are where the 6 to 20 bp long TFBS is located.

1.4.2.1 TFBS representation

The TFBSs of the same TF are generally very similar, but not identical. The sequence of nucleotides may vary from one TFBS to another, but a consensus binding sequence can be inferred using computational approaches (Stormo et al., 1982). This sequence simply summarizes the collection of DNA sequences bound by the TF in question (Figure 1.23 (b)). This representation makes use of the International Union of Pure and Applied Chemistry (IUPAC) nucleotide code which encompasses all the possible combinations of one or more nucleotides and follows precise rules (Cavener, 1987). As such, each position within the TFBM corresponds to one IUPAC nomenclature based on the nucleotides observed at that position during the alignment of TFBSs.

For instance, Y corresponds to T or C, R corresponds to G or A, while N corresponds to any nucleotide (Table 1.3 and Figure 1.23 (a-b)). Nevertheless, if a given nucleotide is observed at a given position in the majority of TFBS sequences, the other observed nucleotides are sometimes ignored,

and thus the most frequent nucleotide becomes representative for that position within the TFBS (Wasserman and Sandelin, 2004) (Figure 1.23 (a-b)). Even though this representation is simple and straightforward, it assumes equiprobability among nucleotides at a given position within the TFBM. It encodes the information if a nucleotide is present or absent at a given position. The consensus sequence can be determined for a large number of TFs and its accuracy depends on the number of available TFBSs for a given TF (Wasserman and Sandelin, 2004). However, the individual nucleotides represented by a consensus do not equally contribute to the TF binding. In fact, the most conserved nucleotides in the consensus (i.e., the largest letters in the sequence logo (Figure 1.23 (e))) contribute the most in the DNA-binding processs (Stormo, 1990).

To overcome this limitation, a more sensitive approach was developed through the position frequency matrix (PFM) (Stormo, 2000). This matrix is built by counting the occurrences of each nucleotide at each position within the aligned TFBSs (Figure 1.23 (a-c)). By dividing the number of occurrences by the number of sequences, the position probability matrix is obtained (PPM). This matrix will give the probability of each nucleotide to be present at each position within the TFBM (Wasserman and Sandelin, 2004). In case of a very limited number of identified TFBSs, these probabilities are corrected to avoid having null values. One classic way of converting a PFM into a PPM is as follows:

$$p(n,i) = \frac{f(n,i) + s(n)}{N + \sum_{n' \in [A,C,G,T]} s(n')}$$

with $p(n,i)$ being the probability of nucleotide n to be present at position i within TFBSs, $f(n,i)$ the frequency of nucleotide n at position i, $s(n)$ being the *pseudocount* which allows to correct for null values by adding a small value, and N the number of identified TFBSs (Figure 1.23 (a)). Once the PPM is calculated, the position weight matrix (PWM) can be derived (Figure 1.23 (d)) by dividing the probabilities from the PPM by the expected probability of each nucleotide to be at each position within the TFBS. One common way to convert the PPM into a PWM is by calculating the log-likelihood ratio:

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B	R	M	C	W	A	W	H	R	W	G	G	B	M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

d Position weight matrix (PWM)

A	–1.93	0.79	0.79	–1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	–1.93	–1.93	0.79
C	0.45	–1.93	0.79	1.68	–1.93	–1.93	–1.93	0.45	–1.93	–1.93	–1.93	–1.93	0.00	0.79
G	0.00	0.45	–1.93	–1.93	–1.93	–1.93	–1.93	–1.93	0.66	–1.93	1.30	1.68	1.07	–1.93
T	0.15	0.66	–1.93	–1.93	1.07	0.66	0.79	0.00	0.00	0.79	–1.93	–1.93	–0.66	–1.93

e

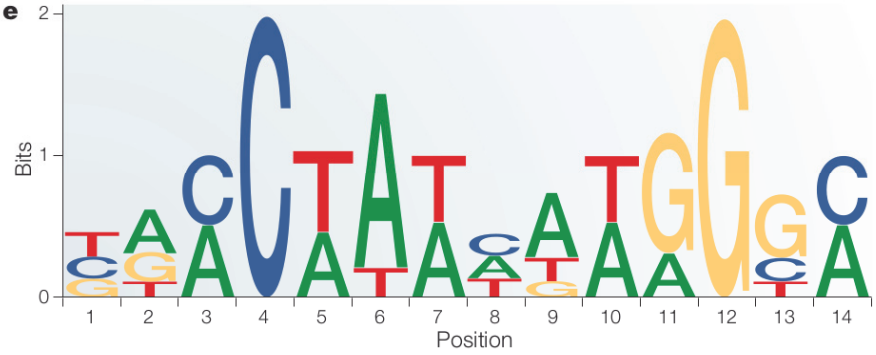

Figure 1.23: The aligned TFBSs (a) can be used to generate a consensus sequence (b) or a position frequency matrix (PFM) (c) to preserve information about the nucleotide frequency at each position within the TFBS. The PFM can be converted to a PWM (d) used to score each observed TFBS. The PFM can be visualized under a logo representation (e). Figure adapted from Wasserman and Sandelin (2004).

$$w(n,i) = log2\left(\frac{p(n,i)}{p(n)}\right)$$

with $p(n)$ being the expected (i.e., background) probability of nucleotide n. Commonly, the *background* probability of a nucleotide is set to be 0.25 (i.e., equiprobability among the four nucleotides). These weights represent how different the observed frequencies are compared to what is expected by chance. Positive weights translate to more frequent observations of a given nucleotide at a given position within the TFBSs. Negative weights translate to less frequent observations of a given nucleotide at a given position within the TFBSs. In other words, positive weights translate to gain within the overall score and negative weights act as penalties. This model captures the heterogeneity of nucleotide probabilities. However, this representation can be hard to visually interpret. To facilitate visualisation, the sequence logo representation was invented (Schneider and Stephens, 1990) (Figure 1.23 (e)). The sequence logo displays the PWM with information content (IC), which is the relative importance of each nucleotide at each position within the TFBM. The IC is calculated based on Shannon's uncertainty theory separating the binding site from the *background* probabilities (Shannon, 1948).

Even though the classic PWM representation of TFBMs is the most widely used, one drawback is that it assumes inter-nucleotide independence. That is, the frequency/probability of each nucleotide at each position is independent of the frequency/probabilities of other nucleotides within the TFBM. While this is the case for a large number of TFs, others rely on internucleotide dependencies and allow gaps or even variable length in their binding motif (Stormo, 2013). To address this limitation, alternatives to the PWM model have been developed that take into consideration nucleotide interdependencies and allow for spacing within the TFBM. It is important to note that these alternatives are possible due to the vast amounts of data generated in the past years (e.g., ChIP-seq experiments), as more sophisticated computational models generally require parameter learning and tuning, which in turn require large amounts of data for training.

Briefly, among those alternatives are (i) the binding energy model (BEM) which includes energy parameters for adjacent nucleotides (Zhao et al., 2012; Stormo, 2013), (ii) the di-position-specific scoring matrix (di-PSSM) which takes into consideration di-nucleotide dependencies and has extended the

Table 1.4: An overview of popular binding motif databases. The organism column refers to Homo Sapiens (HS), Mus Musculus (MM), Arabidopsis Thaliana (AT), and Multi for multiple organisms and/or from different taxonomies.

Database	Motifs	Organism	Content	DOI
JASPAR	1404	Multiple	Manually curated TFBMs	10.1093/nar/gkx1126
HOCOMOCO	1302	HS + MM	Manually curated TFBMs	10.1093/nar/gkx1106
Cis-BP	11491	Multiple	Collection from other DBs	10.1016/j.cell.2014.08.009
FootprintDB	7032	Multiple	Manually curated inferred TFBM	10.1093/bioinformatics/btt663
RegulonDB	3560	E. coli	Regulatory units and network	10.1093/nar/gkv1156
UniPROBE	594	Eukaryotes	Motifs derived from PBM data	10.1093/nar/gkn660
ENCODE	2065	Multiple	Motifs derived from ChIP-seq data	10.1101/gr.139105.112
HOMER	331	HS	Motifs compiled from several resources	10.1016/j.molcel.2010.05.004
Cistrome	862	AT	Motifs discovered from DAP-seq data	10.1016/j.cell.2016.04.038
HumanTF	818	HS	TFBMs from HT-SELEX and ChIP-seq	10.1016/j.cell.2012.12.009

classic PSSM (Kulakovskiy et al., 2013), (iii) the transcription factor flexible models (TFFM) which take into consideration di-nucleotide dependencies as well as variable TFBM length (Mathelier and Wasserman, 2013), Bayesian networks (Barash et al., 2003), and Bayesian Markov models (BaMM) that can be extended to a larger order, thus taking into account non successive inter-nucleotide dependencies (Xing et al., 2003; Siebert and Söding, 2016). Another type of relevant information to include in the prediction model is DNA shape and structural information. It has been shown that for some TF families, including information such as the helix twist, minor groove width, propeller twist, and roll can improve the prediction performance (Mathelier et al., 2016). Even if these models have been shown to outperform the classic PWM in various scenarios (Mathelier and Wasserman, 2013; Siebert and Söding, 2016; Mathelier et al., 2016), additional TF information can be relevant when choosing between a higher order computational model and the classic PWM in order to avoid overfitting. Table 1.4 contains a list of popular TFBM databases derived from experimental assay and computational modeling of TFBSs and Table 1.5 contains a timeline of the different computational approaches.

Table 1.5: A timeline of different computational approaches to model TFBSs.

TFBS modeling methods	Features integrated	References
PWM (position weight matrix)	NA (not applicable)	Stormo et al., 1982; Schneider and Stephens, 1990

Table 1.5: A timeline of different computational approaches to model TFBSs. *(continued)*

TFBS modeling methods	Features integrated	References
HMDM (hidden Markov Dirichlet-multinomial)	Positional dependencies	Xing et al., 2003
DWM (dinucleotide weight matrix)	Dinucleotides	Siddharthan, 2010
BEM (binding energy model)	Dependencies (adjacent positions) and binding affinity data	Zhao et al., 2012; Stormo, 2013
TFFM (TF Flexible Model)	Dependencies (adjacent position) and background	Mathelier and Wasserman, 2013
PIM (pairwise interaction model)	Dependencies between all positions	Santolini et al., 2014
gkm-SVM (gapped k-mer support vector machine)	k-mers supporting gaps	Ghandi et al., 2014
SeqGL	k-mer, chromatin accessibility	Setty and Leslie, 2015
MORPHEUS	Dependencies between all positions	Minguet et al., 2015
FeatureREDUCE	Dependencies between all positions	Riley et al., 2015
DeepBind	NA	Alipanahi et al., 2015
DeepSEA (deep learning-based sequence analyzer)	Integrate DNase I hypersensitivity data and histone-mark profiles	Zhou and Troyanskaya, 2015
DNAshaped TFBS	Helix twist, propeller twist, minor groove width, and rotation	Mathelier et al., 2016
Cytomod	DNA methylation	Viner et al., 2016
DWT (dinucleotide weight tensor)	Dependencies between all positions	Omidi et al., 2017
TFImpute	NA	Qin and Feng, 2017
BEESEM (short for Binding Energy Estimation on SELEX with Expectation Maximization)	NA	Ruan et al., 2017
DeFine	Integrate Hi-C data	Wang and Dynlacht, 2018

Table 1.5: A timeline of different computational approaches to model TFBSs. *(continued)*

TFBS modeling methods	Features integrated	References
DFIM (Deep Feature Interaction Maps)	Dependencies between all positions, interaction between motifs, core motif flanking region, and chromatin accessibility	Greenside et al., 2018
NRLB (No Read Left Behind)	NA	Rastogi et al., 2018
KSM model (k-mer set memory)	k-mers	Guo et al., 2018
SelexGLM	Core motif flanking region	Zhang et al., 2018

1.4.2.2 The plague of false positives

One of the main issues when inferring protein-DNA interactions is the high number of false positives present in the experimental yield (Teytelman et al., 2013; Jain et al., 2015) and in the computational predictions (Worsley Hunt and Wasserman, 2014). To address this problem, other layers of relevant information can be used on top of the simple sequence scanning approach to infer TFBSs (Aerts, 2012). Briefly, these include: (i) sequence conservation, where TFBSs with a higher phastCons score (Siepel et al., 2005) are considered more likely to be functional TFBSs, as they are conserved across evolution, (ii) clusters of TFBS (with and without sequence conservation) relies on identifying regions with a higher concentration of TFBSs (not necessarily for the same TF) known as *cis*-regulatory modules (CRM) (Schmidt et al., 2010; Ballester et al., 2014), (iii) using information about the chromatin state (e.g., histone modifications, DNAse footprinting or ATAC-seq) can reduce the search space by discarding those found in a close chromatin region, (iv) using gene expression relies on scanning for the TFBSs associated with expressed genes only, with the limitation that distal CREs are ignored, (v) the motif environment can give information about *bona fide* TFBSs as it has been shown that the flanking regions present similar GC composition (Dror et al., 2016), and (vi) using DNA structural information given by the physical interactions between nucleotides (Zhou et al., 2013; Yang et al., 2014; Mathelier et al., 2016). A schematic representation of the above methods can

Table 1.6: An overview of relevant TF binding regions and TFBS databases derived from human ChIP-seq data. The number of transcriptional regulators (TRs) and the number of regulatory regions (RR) identified is noted as well as a brief description of the content of each database

Database	TRs	CRRs	Content	DOI
UniBind	231	~8.3M	Unique TFBSs - uniformly processed	10.1093/nar/gkw951
ReMap	485	~80M	Peaks - uniformly processed	10.1093/nar/gkx1092
MANTA	225	~48M	TFBS and TFBS variants	10.1038/sdata.2018.141
GTRD	402	~445M	TFBSs - uniformly processed	10.1093/nar/gky1128
ChIP-atlas	852	~130M	Peaks - uniformly processed	10.15252/embr.201846255
Cistrome DB	-	~235M	Peaks curated from public datasets	10.1093/nar/gkw983
OregAnno	-	~8K	TFBSs manually curated	10.1093/bioinformatics/btk027

be found in Figure 1.24.

Nevertheless, none of the above methods can completely remove false positives, but a combination of them may present an improvement over the performance of individual methods. Moreover, this performance is TF family specific, as it has been shown that certain features improve the prediction accuracy for some TF classes but not for others (Aerts, 2012; Jayaram et al., 2016; Mathelier et al., 2016).

In general, the existence of false positives should always be minded and accounted for in the context of TFBS prediction. The discrepancy between false positives and true positives was referred to as the *futility theorem*, stating that the predicted TFBSs will most likely not be functional *in vivo* regardless of their binding affinity *in vitro* (Wasserman and Sandelin, 2004). Still, combining both experimental and computational evidence will help reduce the number of false positives (Worsley Hunt et al., 2014).

1.4.2.3 Inferring binding sites from ChIP-seq genome-wide

Commonly, TFBSs are inferred from ChIP-seq data through sequence *scanning* and *scoring*. The entire genomic sequence of a ChIP-seq peak is scanned in a *sliding-window* manner (in one nucleotide increments) and each subsequence within the ChIP-seq peak is scored based on its similarity to a reference motif. After all the positions within all of the sequences (e.g., ChIP-seq peaks) have been scanned, the subset of *bona fide* TFBSs has to be identified. This step has proven not trivial, as it is both data dependent and model dependent. In the case of ChIP-seq assays, the expectations are to have direct

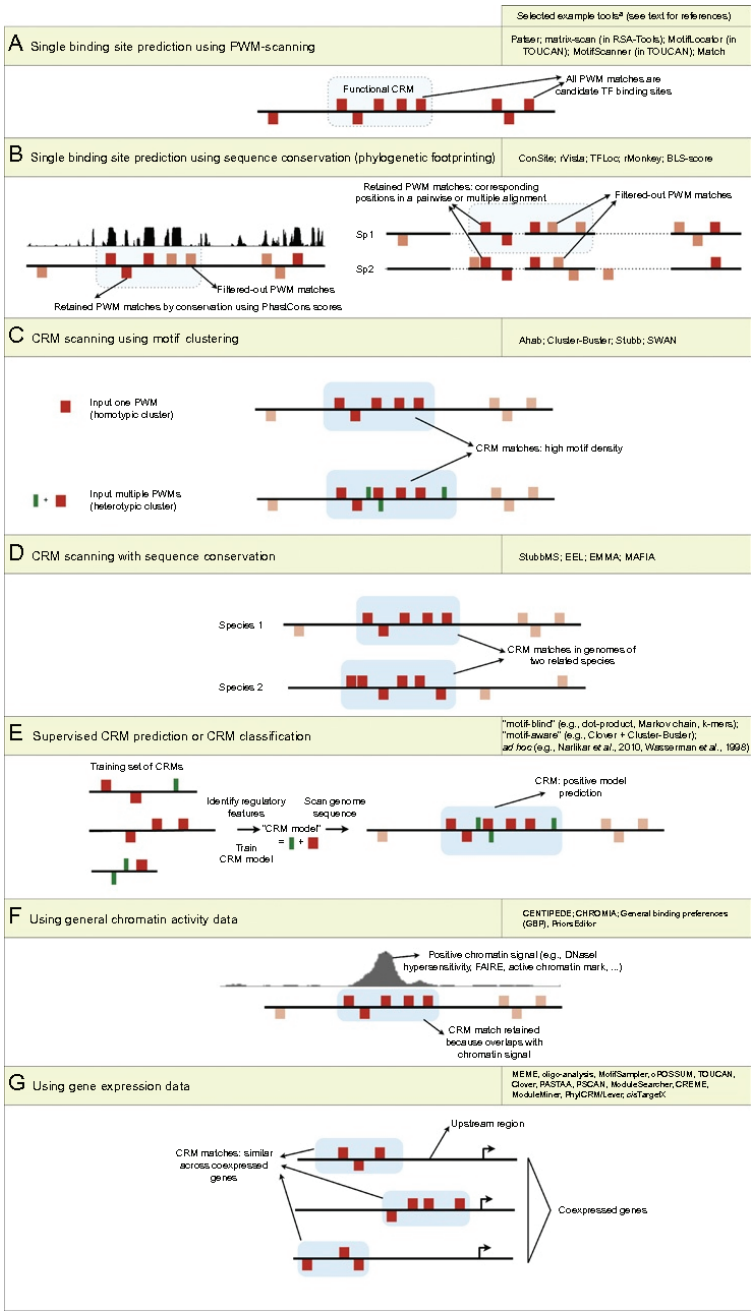

Figure 1.24: A schematic representation of the TFBS prediction approaches and the different layers of information used. Figure from Aerts (2012) .

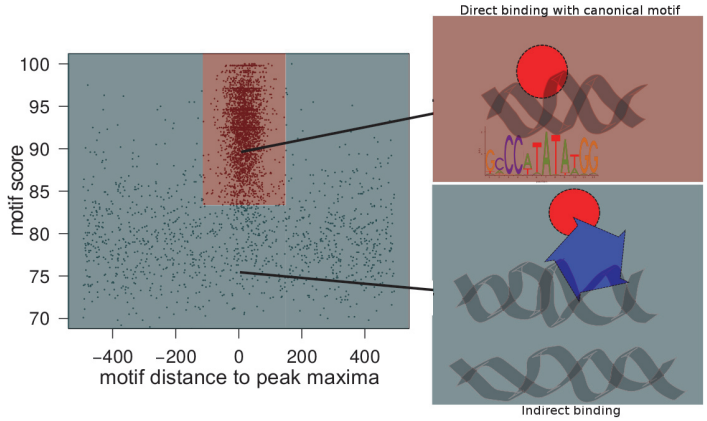

Figure 1.25: Expectations from ChIP-seq data. A landscape view, as described in Worsley Hunt et al. (2014), of the scored ChIP-seq peaks and the demarcation of an enrichment zone discriminating between direct TF-DNA binding (in red) and indirect binding (in light blue). Figure adapted from Mathelier et al. (2015) .

DNA binding, indirect binding, and unspecific binding/noise/experimental artefacts (Teytelman et al., 2013; Jain et al., 2015; Worsley Hunt and Wasserman, 2014). The goal is to discriminate between direct TF-DNA binding and the rest of the binding events. It may very well be that indirect DNA binding contains useful information, but currently there is no method to discriminate between indirect DNA binding and unspecific binding or experimental artefacts (Mathelier et al., 2015). In general, direct TF–DNA interactions are expected to be enriched at ChIP-seq peak summits (Bailey and Machanick, 2012; Kulakovskiy et al., 2010; Jothi et al., 2008) and to present high computational scores (Worsley Hunt et al., 2014) (Figure 1.25). Table 1.6 contains an overview of inferred binding regions and TFBSs derived from ChIP-seq data.

A previous study showed that ChIP-seq data falls within one of three categories: (i) enriched for the TFBS close to the ChIP-seq peak summit (where the highest number of ChIP-seq reads map), (ii) lacking enrichment for the TFBS close to the peak summit, and (iii) a combination of ChIP-seq peaks with and without the TFBS close to the peak-summit (Worsley Hunt et al., 2014). Typically, hardcoded thresholds are set on the computational model output (e.g., PWM score and/or distance to the ChIP-seq peak summit) to discriminate between true TFBSs and *background* or noise. This approach

works well in some cases, but in general, the choice of threshold is somewhat arbitrary and computational model specific, depending on the motif scoring implementation. In an attempt to automate the detection of this threshold, a heuristic method has been proposed to delineate an enrichment zone containing direct TF-DNA interactions based on the distance to the ChIP-seq peak summit and the computational score (Worsley Hunt et al., 2014). However, this method, specifically developed for simple PWM scoring, does not work with some more recent TFBS computational models (Zhao et al., 2012; Mathelier and Wasserman, 2013; Mathelier et al., 2016) and uses hard-coded parameter values.

Another approach to discriminate between *bona fide* TFBSs and the rest of binding events is the use of *p*-values. For instance, a *p*-value can be assigned based on the probability of the *background* to reach a PWM score greater than the actual motif score (Touzet and Varré, 2007). The principle behind this method is that some sub-sequences can achieve a certain PWM score more frequently than other PWM scores. To correct for that, each PWM score is given a *p*-value based on the expected distribution of all the other PWM scores. Nevertheless, as for the PWM score threshold, the *p*-value based threshold is also arbitrary and data dependent, and its value can significantly influence the amount of false positives within the set of predicted TFBSs (Touzet and Varré, 2007).

1.4.2.4 Binding motif enrichment

Another use of TFBMs is motif enrichment analyses. These analyses aim at answering the following question: given a set of genes, what are the TFs that regulate them? In other words, what are the TFs that show an overrepresentation of their TFBM at CREs associated to these genes (e.g., promoter regions) (Figure 1.26). In such analysis, two main approaches have been developed so far: (i) *foreground* versus *background* and (ii) *ranking-and-recovery*. For the former, the *background* is built based on the complete set of genomic sequences corresponding to the CREs of all genes (e.g., promoter regions) and the *foreground* is represented by the genomic sequences corresponding to CREs of the gene set of interest. The *background* allows one to calculate the expected number of TFBS occurrences for a given TF relative to the other TFs. In turn, this allows calculation of an enrichment score (e.g., a

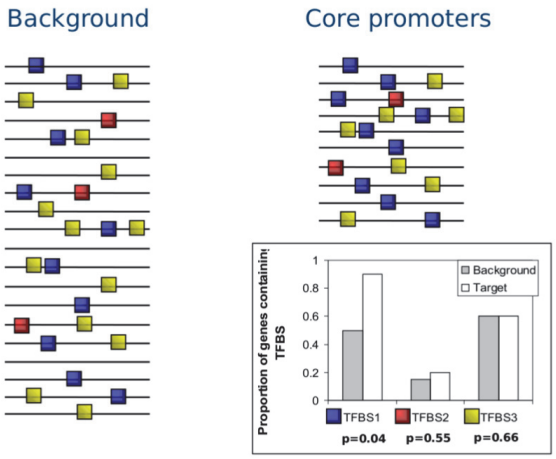

Figure 1.26: A schematic representation of a motif enrichment analysis. Figure from Kwon et al. (2012).

p-value) based on the *foreground*. For the latter, all genes are individually ranked based on each TF motif (i.e., ranking step) and each of the rankings is then tested against the gene set of interest (i.e., recovery step) and an enrichment score is calculated (e.g., area under the curve) (Herrmann et al., 2012; Janky et al., 2014). A high enrichment score indicates that a TF motif recovers a large fraction of the input genes within the top ranking. Other layers of information can be added in the model, such as CRMs or chromatin state information (Herrmann et al., 2012).

1.4.2.5 Inferring gene regulatory networks

In the same context, another question to answer is: given a TF, which are the genes it regulates? In other words, what genes represent the direct targets of TFs. This allows one to infer gene regulatory networks and therefore characterize TFBSs that are more likely to be functional. A *regulon* is the ensemble of genes regulated by the same TF, and their common characteristic is the presence of TFBSs for this TF at their CREs (Lengeler et al., 1999). Over the past years, the focus has been on identifying DNA regions bound by TFs. Unfortunately, TF binding is not necessarily associated with function and tools have to be developed to characterize the ones that are functional

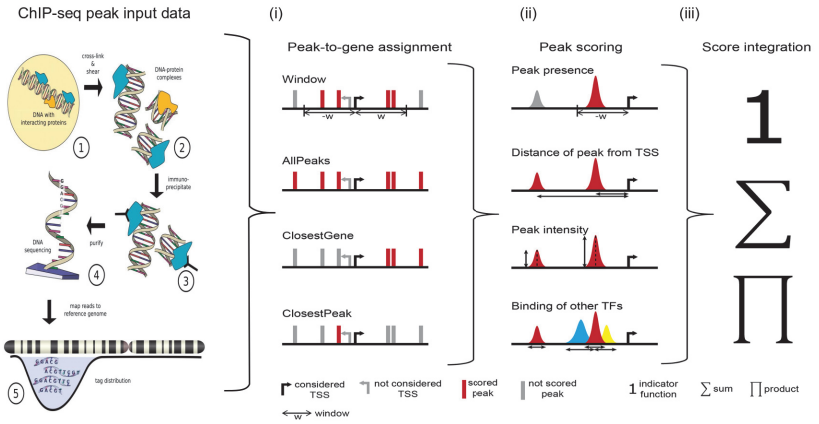

Figure 1.27: The main approaches and their steps when inferring regulons from ChIP-seq data. The ChIP-seq peaks are used as input and assigned to genes (i) and scored (ii). Lastly, a final score is assigned to a given peak-gene pair (iii). Figure adapted from Szalkowski and Schmid (2011) and Sikora-Wohlfeld et al. (2013).

(Whitfield et al., 2012). However, as previously mentioned, due to the relatively high number of false positives in experimental data and diversity of cell types and biological conditions, the predicted TFBSs will most likely not be functional *in vivo* regardless their binding affinity *in vitro* (Wasserman and Sandelin, 2004).

Due to its abundance, ChIP-seq has also become the preferred assay to infer direct TF targets. Generally, computational tools developed to predict TF regulons from ChIP-seq data follow a three-step workflow (Sikora-Wohlfeld et al., 2013): (i) assigning ChIP-seq peaks to genes, (ii) assessing the regulatory potential of a ChIP-seq peak to a gene, and (iii) integrating the corresponding scores per gene (Figure 1.27). Classically, ChIP-seq peaks are considered representative for TF binding to DNA. In their implementation, the computational models assign them to the TSS of the closest gene in linear genomic distance or to all TSSs within a certain predefined genomic window (Sikora-Wohlfeld et al., 2013). Their scoring is either dicothomic (i.e., 0 or 1) or gradient in relation to the distance to the TSS (Sikora-Wohlfeld et al., 2013; Mei et al., 2017). More recent ChIP-seq based methods use additional data, such as genome-wide binding profiles of TFs (Cheng et al., 2011; Yang et al., 2016) or correlation between histone marks and gene ex-

pression (O'Connor et al., 2017), to infer TF regulons. Commonly, regulon predictions are validated using experimental data. More specifically, sets of associated genes are generated through knock -in and -down experiments targeting specific TFs (Subramanian et al., 2005). Subsequently, the set of differentially expressed genes between the condition and control is associated to the TF. Nevertheless, there is no *gold standard* for TF regulon prediction validation.

Further improvement of TF regulon prediction methods can be achieved by including associations between enhancers and the TSSs (Fishilevich et al., 2017) and/or promoter information, which have not been included in existing methods. Also, it has been shown that CRMs are more likely to host functional TFBSs, as they represent the genomic regions where TFs cooperate (Davidson, 2006; Lambert et al., 2018). Including this information may also increase prediction accuracy. Integrating these data together with other layers of relevant information may allow for the identification of TFBSs that are most likely to be functional and to have an impact on transcriptional regulation.

Objectives of the study

Transcriptional regulation is a biological mechanism essential to cell growth and cell differentiation. Disruptions occurring in the gene regulatory program can lead to disease prone phenotypes or abnormal development of tissues or the organism as a whole. One way to gain more insight into the transcriptional regulation mechanism and have a deeper understanding of gene expression regulation is to study how key proteins involved in transcription, such as TFs, interact with DNA and to infer gene regulatory networks. This in turn, will facilitate the assessment of the impact of gene expression dysregulation caused by disruptions in transcriptional regulation.

Capitalizing on the massive amounts of data that are generated through experimental assays followed by next-generation sequencing (NGS), we can computationally derive TF-DNA interactions, their affinities, as well as their specificities. Impressive efforts have been made by consortia such as ENCODE (The ENCODE Project Consortium, 2012) and GEO (Edgar et al., 2002) to create publicly available repositories for experimental data generated from the study of the regulatory mechanism. Even though NGS is teeming with positive aspects regarding genetic research, one important ubiquitous challenge persists: making use of the data at its full potential. This implies assembling, curating, and integrating data in common frameworks and/or databases. Such databases should eventually include extensive linkage to the underlying biological processes and the associated clinical data.

The most popular biological assay aiming at identifying TF-DNA interactions is chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Johnson et al., 2007). In the past years, tens of thousands of ChIP-seq experiments were carried out in different organisms, cell lines, and under various biological conditions to study the dynamics and particularities of TF-DNA interactions. A large proportion of this data was made publicly available. Nevertheless, biological variation and experimental artefacts, as well as the varying percentage of false positives to which ChIP-seq data is prone to, impede the creation of a standardized, genome-wide library of TF-DNA interactions and ultimately an extensive regulatory network. To achieve this, one obvious aprroach is to curate the existing data, processes it in a uni-

form manner, and improve the computational models used to predict *bona fide* TF-DNA interactions. Altogether, the key to successful computational biology research is access to high quality data, for which we have a strong understanding, in order to produce dedicated processing tools.

In an attempt to fill in these gaps, the project relies on the development of new computational methods and resources that are derived from in depth analyses of experimentally-generated data developed to study gene expression regulation. Specifically, to develop computational methods, tools, and data resources to:

1. **improve our capacity to predict TF-DNA interactions and generate a genome-wide map of high confidence direct TF-DNA interactions** by

- making use of publicly available ChIP-seq data
- developing a computational pipeline to uniformly process ChIP-seq data
- integrating multiple TFBS prediction computational models
- developing a methodology to derive high confidence TFBSs
- generating a publicly available resource based on these data

2. **predict the direct TF target genes (i.e., regulons)** by

- making use of the TFBS predictions from the previous step
- integrating additional layers of relevant information
- developing a statistical framework to define TF regulons
- assessing the results based on biological relevance

3. **determine the transcriptional differences between oeastrogen receptor negative (ER-) and oestrogen receptor positive (ER+) breast cancers** by

- making use of the resources generated in the previous steps
- making use of other data available for breast cancer, such as RNA-seq
- identifying the potential key TFs that are specific to ER- and ER+

2

Summary of the papers

2.1 Papers I-IV: towards a map of direct TF-DNA interactions in the human genome

Binding of TFs to DNA occurs in a sequence specific manner (Badis et al., 2009). As TFs recognize sequence motifs, computational tools have been instrumental in the prediction and characterization of TF-DNA interactions. Classically, TFBSs are modeled using PWMs and the underlying probabilities of each nucleotide at each position within the motif are derived from a collection of TFBSs taken from experimental assays. While binding affinities are derived from *in vitro* assays, genomic binding regions are derived from *in vivo* assays such as ChIP-seq. However, ChIP-seq experiments have been recurrently shown to be prone to noise (Teytelman et al., 2013; Worsley Hunt and Wasserman, 2014; Jain et al., 2015). Hence, computational models of TF-DNA interactions can be used to highlight *bona fide* binding regions. While PWMs are usually working well, more sophisticated approaches have recently been designed to model complex features of TF-DNA interactions captured by next-generation sequencing data and to refine binding region prediction. Indeed, TFs recognize their binding sites through a complex interplay between base/nucleotide readout and DNA shape readout (Rohs et al., 2009). Computational models combining both sequence and DNA shape information have shown improvement in our capacity to predict TFBSs from ChIP-seq data (Rohs et al., 2010; Mathelier et al., 2016). However, studies have also shown that the best performing model for different TFs varies; therefore, developing a *one-fits-all* TFBS prediction model is not currently an optimal solution (Will and Helms, 2014).

Large amounts of ChIP-seq data have been generated and a vast majority are hosted in publicly available data repositories such as ENCODE (The

ENCODE Project Consortium, 2012) and GEO (Edgar et al., 2002). We combined these publicly available ChIP-seq data with manually curated TF binding profiles to improve our capacity to predict TFBSs genome-wide (Gheorghe et al., 2019) and to assess the impact of single nucleotide variants (SNVs) at TFBSs on the alternate alleles (Fornes, Gheorghe, et al., 2018). We combined several prediction models, varying from simple to complex, into one data processing pipeline, *ChIP-eat*, to improve our capacity to predict TFBSs genome-wide (Gheorghe et al., 2019). Our work culminated with >8 million TFBS predictions in the human genome, which are made available to the community through the UniBind database (https://unibind.uio.no). Following is the list of publications that led to the creation of our map of direct TF-DNA interactions in the human genome.

2.1.1 Paper I

Chèneby, J., **Gheorghe, M.**, Artufel, M., Mathelier, A., and Ballester, B. (2018). **ReMap2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments.** *Nucleic Acids Research*, 46(D1):D267–D275. See published manuscript at the end of the thesis.

To acquire large amounts of ChIP-seq data, we participated in the 2018 update of the ReMap database (Chèneby et al., 2018) (http://remap.cisreg.eu), providing an atlas of *cis*-regulatory elements (CREs) in the human genome. We processed >3,000 ChIP-seq datasets from the public repositories ENCODE (The ENCODE Project Consortium, 2012), GEO (Edgar et al., 2002), and ArrayExpress (Sarkans et al., 2005). Starting from the raw sequencing data, we mapped the reads to the latest version of the human genome (GRCh38), filtered out low quality reads, and called ChIP-seq peaks (i.e., predicted TF binding genomic regions). As a result, we obtained a total of ~80 million ChIP-seq peaks accounting for 485 transcriptional regulators (including TFs).

2.1.2 Paper II

Khan, A.†, Fornes, O.†, Stigliani, A.†, **Gheorghe, M.**, Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A. (2018). **JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.** *Nucleic Acids Research*, 46(D1):D260–D266. See published manuscript at the end of the thesis.

Using, among other resources, the entire collection of ChIP-seq peaks obtained through the ReMap2018 collection, we updated the JASPAR database (Khan et al., 2018) (http://jaspar.genereg.net). The JASPAR database is one of the most popular databases of its kind. It is an open access resource and hosts manually curated and experimentally derived TF binding profiles for around 1400 unique TFs in six taxa. This update added 322 new PFMs and updated 33. We complemented the existing collection of binding profiles using transcription factor flexible models (TFFM) (Mathelier and Wasserman, 2013) trained on ChIP-seq peaks (316 new profiles) to account for internucleotide dependencies. This collaboration provided us with an extended collection of high quality TF binding profiles, which along with the ChIP-seq peaks from ReMap2018, served as input in the computational models employed for TFBS predictions in the human genome.

2.1.3 Paper III

Fornes, O.†, **Gheorghe, M.**†, Richmond, P. A., Arenillas, D. J., Wasserman, W. W., and Mathelier, A. (2018). **MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations.** *Scientific Data*, 5:180141. See published manuscript at the end of the thesis.

In a first attempt to predict TFBSs in the human genome, we combined both the ReMap2018 (Chèneby et al., 2018) ChIP-seq peaks and the JASPAR2018 (Khan et al., 2018) binding profiles to update the Mongo database for the analysis of TFBS alterations (MANTA) database (Fornes, Gheorghe, et al., 2018) (http://manta.cmmt.ubc.ca/manta2/upload). In a nutshell, we

used the ReMap ChIP-seq peaks and JASPAR PWMs to predict TFBSs using a unique PWM score threshold for all PWMs. With this update, we expanded the database to host genome-wide TFBSs based on the intersection between the binding profiles hosted in JASPAR and the ChIP-seq peaks from ReMap. A predicted impact score for each SNV that could occur within the predicted TFBS was calculated based on a z-score computation obtained from all possible SNVs within the TFBSs. These impact scores have been found to correlate with allelic imbalance of ChIP-seq data (i.e., allele specific binding).

2.1.4 Paper IV

Gheorghe, M., Sandve, G. K., Khan, A., Chèneby, J., Ballester, B., and Mathelier, A. (2019). **A map of direct TF–DNA interactions in the human genome**. *Nucleic Acids Research*, 47(4):e21–e21. See published manuscript at the end of the thesis.

Making use of the ReMap2018 (Chèneby et al., 2018) ChIP-seq peaks and the JASPAR2018 (Khan et al., 2018) binding profiles, we developed the *ChIP-eat* data processing pipeline for high confidence prediction of direct TF-DNA interactions (https://bitbucket.org/CBGR/chip-eat/src/master/). The entire set of TFBS predictions is publicly available through the *UniBind* database (http://unibind.uio.no). Paper provided at the end of the thesis.

The ChIP-eat pipeline. With the entire collection of uniformly processed ChIP-seq data sets from ReMap2018 and the extended collection of TF binding profiles from JASPAR2018, we have high quality, uniformly processed data to serve as a base for the development of an improved methodology for TFBS prediction. We developed the *ChIP-eat* data processing pipeline that takes as input ChIP-seq peak regions and TF binding profiles and predicts high confidence TFBSs. Based on the observation that currently there is no *one-fits-all* model for TFBS prediction (Will and Helms, 2014), we developed *ChIP-eat* to support 4 different existing prediction models, from simple to complex: PWMs optimized with the discriminative motif optimizer (DiMO) (Patel and Stormo, 2014), binding energy model (BEM) (Zhao et al., 2012), transcription factor flexible models (TFFM) (Mathelier and Wasserman, 2013), and including DNA shape features (DNAshapedTFBS) (Mathelier et al., 2016). While the *ChIP-eat* pipeline has been used on these 4

models, it can be applied to any TF binding profile model. We uniformly processed 1,983 ChIP-seq data sets accounting for 231 unique TFs.

The improvement of the TFBS predictions relies on the non-parametric, entropy-based, data driven computational method we developed to automatically delineate an enrichment zone (Worsley Hunt et al., 2014) containing a subset of high confidence predicted TFBSs that are supported by both strong experimental evidence and computational evidence (Gheorghe et al., 2019). In brief, the enrichment zones highlight the TFBSs predicted with high computational score and proximity to the ChIP-seq peak summit (where most reads align). This method can be used regardless the prediction model, surpassing the limitations of a previous approach specifically designed to work with PWMs (Worsley Hunt et al., 2014). Moreover, the entropy-based approach does not require any arbitrary/hardcoded threshold, which makes it more flexible and data dependent.

The predictions were a posteriori assessed using protein binding microarray and ChIP-exo data, and were predominantly found in high quality ChIP-seq peaks. Our predictions derived co-binding TFs supported by protein-protein interaction data and defined *cis*-regulatory modules (CRMs) enriched for disease- and trait-associated SNPs.

The UniBind database. We provide our collection of >8 million of high quality TFBS predictions and *cis*-regulatory modules through the publicly available UniBind web-interface (http://unibind.uio.no). This online resource has a very simple and intuitive graphical interface that allows to search, browse, and download (individual sets or in bulk) the sets of TFBS predictions for each computational model. With this freely available resource, we empower the community with genome-wide direct TF-DNA interactions that can serve as entry point for various transcriptional regulation studies.

2.2 Paper V

Gheorghe, M., Mathelier, A., **TF-regulons: identifying direct targets of transcription factors**, *Draft manuscript at the end of the thesis*

TFs mediate gene expression through their binding to DNA. A TF can regulate several genes and this set of genes is called a *regulon*. Common to all genes within a *regulon* is that TFBSs for the same TF are found at their CREs. However, TF binding is not necessarily associated with function, and how TF-DNA interactions impact gene expression is still poorly understood. It has become obvious that to better understand the functional impact of TF-DNA interactions, methods have to be developed to identify not only the potential TFBSs, but to characterize the functional ones (Whitfield et al., 2012).

We developed *TF-regulons*, a ranked list-based statistical approach to predict TF target genes from individual ChIP-seq data sets. We employed high confidence TFBS predictions from the UniBind database (Gheorghe et al., 2019) and ChIP-seq peaks from the latest release of ReMap (Chèneby et al., 2018). Apart from the distance to transcription start sites, we integrated six additional *features* in our prediction model, such as enhancer/promoter information and gene-enhancer associations (Fishilevich et al., 2017). The performance of the model was assessed on 13 TFs by means of overlap between *TF-regulons* top scoring genes and annotated reference gene sets, and by assessing functional similarity of enriched GO terms between *TF-regulons* and annotated reference gene sets.

Our results show that using high quality TFBS predictions outperforms the predictions obtained when relying on ChIP-seq peaks and that gene-enhancer associations with additional *feature* combinations, such as sequence conservation and enhancer/promoter information, performs best in predicting TF regulons. However, predicting regulons remains a highly complex problem and we speculate that the prediction performance can be further improved by incorporating cell line specificities and/or by implementing a different prediction framework.

2.3 Paper VI

Gheorghe, M., Tekpli, X., Fleischer, T., Kristensen, V., and Mathelier, A., **Identifying key TFs driving ER positive and ER negative breast cancer subtypes**, *Draft manuscript at the end of the thesis*

Regulatory disruptions in cell growth and cell differentiation can lead to cancer. Due to its heterogenous nature, cancer accounts for a wide range of affected tissues in the human body (Wade, 2001). The broad variation in the genetic makeup of each individual adds to the complexity of this problem and renders the design of a generic cancer treatment unattainable (Burrell et al., 2013). With the advent of NGS, it has become possible to classify cancer subtypes based on their molecular signatures by analyzing gene expression profiles. Based on the expression levels of estrogen receptor (ER), breast cancer can be classified as estrogen receptor positive (ER+) or estrogen receptor negative (ER-). In contrast to ER+ breast cancers, the ER- subtype cannot be targeted by hormone therapies, and to date, the only effective treatment against it remains chemotherapy (Chavez et al., 2010).

We aimed at identifying the molecular differences between ER+ and ER- breast cancers by predicting key TFs involved in driving differential protein-coding gene expression between these two subtypes. We compared the RNA-seq profiles of 981 ER+ and ER- breast cancer donors obtained from The Cancer Genome Atlas (Weinstein et al., 2013) and identified sets of differentially expressed genes between the two subtypes. We found that genes upregulated in ER- were enriched within the set of genes that are involved in hematopoietic and lymphoid tissue, suggesting the enrichment of an immune signature in ER- breast cancer subtype.

We extracted TFBSs from the UniBind database (Gheorghe et al., 2019) that are found at promoter and enhancer regions associated with the sets of ER- or ER+ differentially expressed genes to predict TFs with enriched binding events at these CREs. The TFs with enriched TFBSs are likely regulators of the corresponding genes in the breast cancer subtypes. We found E2F4 and Myc motifs to be enriched at the CREs of genes upregulated in ER-, and RFX1, TFAP2C, and FOXA1 binding sites enriched at CREs of ER+ genes. Interestingly, most RFXs presented a negative fold change of expression in ER+, suggesting that they can act as repressors for Estrogen Receptor 1 (ESR1). Functional enrichment on the genes predicted to be targets of these

TFs indicate that DNA repair and replication associated processes, and cell cycle regulation may be affected in ER- breast cancer. For ER+ breast cancer, we found cillium and organelle assembly processes to be enriched, possibly due to a loss of RFX binding, as the majority of these TFs also show a negative fold change of expression in ER+.

3

Discussion and perspectives

Since the first whole human genome was sequenced in 2001, the sequencing costs have unceasingly decreased from USD 100 million to ~ USD 1000 (NIH, 2019). Next generation sequencing (NGS) or massively parallel sequencing has enabled the community to initiate a plethora of studies that aim at improving our understanding of the intricate mechanisms underlying cellular development and differentiation, phenotypic variation, and disease onset and progression at an unprecedented scale (Mardis, 2008). Such studies include DNA-protein interactions, DNA methylation analyses, and prediction of disease-associated genes (Zhang et al., 2011; Buermans and den Dunnen, 2014). Moreover, the development of transcriptome sequencing technologies has shown that the entire genome is actually transcribed and not only the ~2% coding for proteins (Kapranov et al., 2007). Albeit, what is defined as *functional* is still subject to discussion (The ENCODE Project Consortium, 2012; Graur et al., 2013). Lately, as observed phenotypes could not be explained by genomic variations occurring in the protein-coding regions, the focus has shifted towards dechiphering the remaining ~98% of the genome (Khurana et al., 2016).

As a consequence, massive amounts of sequencing data are constantly generated. These data are hosted in private or public repositories such as ENCODE (The ENCODE Project Consortium, 2012) or GEO (Edgar et al., 2002). Even though NGS is teeming with positive aspects regarding genetic research, one important ubiquitous challenge persists: making use of the data at its full potential. This implies assembling, curating, integrating, and analyzing data in common frameworks or databases. Such databases should include extensive linkage to the underlying biological processes and the clinical data associated. The enrichment of these databases is constantly contributed to by newly generated data and emerging analytical bioinformatics tools aimed at interpreting the impact of genomic variants on phenotypes.

The data are generally integrated into publicly available genome browsers, and computational frameworks have been developed to provide users with visualization and analytical tools (Kent et al., 2002; Zhang et al., 2011; Weinstein et al., 2013).

3.1 Quality control and resource maintenance

During the early years of the bioinformatics field, the diversity of such resources and data heterogeneity raised substantial challenges to data integration and quality assessment (Davidson et al., 1995). Over the years, standards and guidelines were proposed to ensure a certain quality level of biological data, using reference quality indicators and object modeling languages (Brazma et al., 2001; Qureshi and Ivens, 2008; The ENCODE Project Consortium, 2012). Nevertheless, due to the flexibility of the guidelines, the community has fragmented and adopted variations of these standards, making data harmonization and integration even more difficult (Burgoon, 2006). Moreover, there is no clear definition on what represents a data standard (Tenenbaum et al., 2014).

Recently, there have been several open calls for data sharing, which has spurred a wide range of reactions within the research community and journals (Tenenbaum et al., 2014; Gewin, 2016; Figueiredo, 2017; Vasilevsky et al., 2017). Importantly, it has been shown that journals with a high impact factor generally adhere less to public availability of data policies than lower impact factor journals (Alsheikh-Ali et al., 2011). In fact, the quality of scientific experiments and reliability of the findings was found to be negatively correlated with the ascent of the impact factor (Brembs, 2018). This is an alarming discovery, as the rule of thumb so far has been the more prestigious the journal, the better the methodological research and results. Therefore, it is imperative that quality indicators are put in place and that data made publicly available rises to quality standards, because in the near future these findings will serve as foundation to other studies (Cai and Zhu, 2015).

For instance, the ENCODE consortium has implemented quality control procedures and data quality standards (The ENCODE Project Consortium, 2012), as opposed to the GEO repository (Edgar et al., 2002). Throughout our work, we uniformly processed thousands of ChIP-seq data sets from both

ENCODE and GEO, and we observed that numerous ChIP-seq data sets obtained from GEO had considerably lower quality when compared to the ones obtained from ENCODE, even for the same TF.

Besides quality control, of equal importance is the maintenance of online resources. Through NGS, every research group or laboratory can generate their own resource and/or database that become the tools to use within the group (Cochrane and Galperin, 2010). An important fraction of these data resources are publicly available for the research community to employ in their studies. As most of the publicly available resources tend not to be controlled by private institutions, their longterm maintenance can become an issue due to external funding (Bastow and Leonelli, 2010). Longterm maintenance is crucial for result reproducibility.

Resource maintenance can be viewed from two perspectives: infrastructure and content. The former implies physical storage and accessibility, while the latter implies curation, technical support, and updates. These aspects can quickly become expensive for institutions that rely on government funding (Bastow and Leonelli, 2010). Moreover, every funding round (i.e., through grants) generally implies added functionalities and/or content expansion, which in turn will further increase the maintenance cost (Methods, 2016). In an effort to circumvent this phenomenon, non-profit organizations are put together to maintain public databases and resources, through manual curation, updates, and infrastructure expansion (Methods, 2016). For instance, programs such as ELIXIR (https://elixir-europe.org/) aim at implementing standards to ensure data quality and to provide scalable infrastructure for data storage and shareability across European countries. A smaller scale example is the JASPAR database for manually curated TF binding profiles, which was first created 15 years ago (Sandelin et al., 2004). Since then, it has been regularly (i.e., every two years) maintained and updated, constantly expanding and improving the quality of its TF binding profile collection (Khan et al., 2018). This process is essential to ensure high quality data and result reproducibility, as research based on this data can serve as input in other studies aimed at deciphering transcriptional regulation. This is a good example that should be followed by our new UniBind resource.

3.2 The DNA-encoded rules of transcriptional regulation

To understand disease onset and progression, it is mandatory to understand the underlying molecular mechanisms that mediate gene regulatory networks and their impact on gene expression. TFs are central to transcriptional regulation, and they subsequently control gene expression through their binding to DNA (Chen and Rajewsky, 2007). They generally interact with DNA in a sequence specific manner (Badis et al., 2009) and binding is achieved by recognizing a chracteristic binding motif (Stormo, 2013) and/or DNA shape conformation (Rohs et al., 2009). By binding DNA at CREs, such as promoters and/or enhancers, TFs control the rate of RNA transcription and subsequently gene expression, ensuring that the right genes are expressed at the right levels in the correct cell types (Nelson and Wardle, 2013; Mathelier et al., 2015). This in turn controls the developmental stages of an organism and its responses to environmental stimuli. For decades, identifying TF binding motifs, TFBSs, and TF regulons (i.e., the set of targeted genes) either *in silico*, *in vitro*, or *in vivo* has remained a challenge in understanding gene regulatory networks.

It has been suggested that the transcriptional program follows a set of predefined rules that are encoded in the DNA sequence (Meireles-Filho and Stark, 2009). These rules are in turn based on genomic organization, such as A/B compartments (Lieberman-Aiden et al., 2009; Bonev and Cavalli, 2016), TADs (Dixon et al., 2012; Dekker et al., 2013), and chromatin accessibility (Chen and Rajewsky, 2007), but also on genomic sequence composition, such as TFBS enrichment (Yan et al., 2013), CRMs (Hardison and Taylor, 2012), and/or GC composition (Kudla et al., 2006; Dror et al., 2016). The similarity or diversity of the binding motifs at CREs has also been shown to have an influence on gene expression levels (Ezer et al., 2014; Grossman et al., 2017). For instance, studies show that functionally related enhancers tend to contain motifs for the same TFs (Erives and Levine, 2004), but the order of the TFBSs within CREs is not important to mediate gene expression levels (Zinzen et al., 2009).

Nevertheless, the presence/absence of a certain TF or its binding to DNA does not necessarily associate with enhancer activity or function, as DNA-bound TFs generally act in a cooperative manner by forming complexes dur-

ing gene expression regulation (Jolma et al., 2015). In other words, the combinatorial effect of TFs at CREs is more critical than the order in which they bind the DNA (Schmidt et al., 2010). Based on their binding affinities, TFBSs were dichotomozied in *strong* (i.e., high affinity) binding sites and *weak* (i.e., low affinity) binding sites. In general, the *strongest* binding site is considered, but it has been shown that gene expression regulation is "fine-tuned" through the *weak* binding sites (Parker et al., 2011). In fact, it has been shown that a whole spectrum of binding affinities orchestrate gene regulation, through clusters of TFBSs at which TFs bind with different affinities, including very low affinities (Crocker et al., 2016). Altogether, it is obvious that to decipher the mechanisms ruling gene expression regulation and ultimately understand how disruptions in the regulatory program can lead to disease-prone phenotypes, it is crucial to improve the genome-wide identification of TFBSs.

3.3 Tackling false positives to infer *bona fide* TFBSs genome-wide

In parallel with experimental assays and NGS technologies developed to identify TF-DNA interactions *in vivo*, such as the widely used ChIP-seq (Johnson et al., 2007), computational tools have become instrumental to process the large amounts of data generated through such assays. These tools generally infer so called ChIP-seq peaks, which are expected to contain the TFBSs. Unfortunately, it has been recurrently shown that ChIP-seq is prone to experimental artefacts (Teytelman et al., 2013; Worsley Hunt and Wasserman, 2014; Jain et al., 2015), which in turn generate a varying number of false positives. Thus, delineating *bona fide* bound regions from experimental noise is still an ongoing problem. The ever-increasing number of publicly available ChIP-seq data sets provides an unprecedented opportunity to develop and evaluate computational tools designed to infer the precise locations of the TFBSs within ChIP-seq peaks by combining both computational and experimental evidence of direct TF-DNA interactions. ChIP-seq data sets fall into one of three categories: (i) data sets enriched for the TF binding motif close to the ChIP-seq peak summit, (ii) data sets lacking enrichment for the binding motif close to the peak summit, and (iii) data sets having a combination of

peaks with and without the TF canonical binding motif proximal to the peak summit (Worsley Hunt et al., 2014). Regardless of the presence or absence of a canonical binding motif enrichment close to the peak summit, useful co-binding information may be derived from ChIP-seq peaks falling into the (ii) or (iii) category if indirect binding is discriminated from experimental artefacts/noise (Teytelman et al., 2013; Worsley Hunt and Wasserman, 2014; Jain et al., 2015), although in the current work we did not address this issue. Considering the amount of already publicly available ChIP-seq data, it is also imperative to develop computational methods and tools that aim at harmonizing the data processing workflow and results storage. Up to now, no official standards have been created, but several guidelines were put in place (Landt et al., 2012; Bailey et al., 2013).

Improving existing or developing new computational methods to detect TFBS locations with high confidence is therefore necessary to reduce the high amount of false positives. In turn, this can improve downstream computational analyses, such as TF binding motif enrichment, genomic region enrichment, regulatory variant detection, or TF regulon prediction. Previous analyses have shown that up to 60% of peaks computationally inferred from ChIP-seq experiments and stored in the ENCODE (The ENCODE Project Consortium, 2012) public repository do not contain a TFBS for the targeted TF (Worsley Hunt et al., 2014). Instead, they contain genomic regions that represent clusters of TFs indirectly binding to DNA (Worsley Hunt et al., 2014; Wreczycka et al., 2019; Gheorghe et al., 2019), or are just a consequence of the open chromatin regions (Yan et al., 2013). When inferring TFBSs, the trade-off is generally between sensitivity and specificity. A common practice is to keep the best scoring binding site, whether it presents strong binding affinity or high computational score. However, this does not necessarily mean lower affinity or lower scoring sites are false positives. For instance, a TF binding site with lower affinity can still show preferential binding, which may become relevant to transcriptional regulation depending on the interactions with other TFs within CREs (Parker et al., 2011).

Other studies have shown that the number of false positives can be as high as three orders of magnitude compared to the true binding locations (Fickett, 1996). It has been speculated that this may be solely due to the suitability of the model used to infer TFBSs (Tronche et al., 1997). Indeed, *in vitro* assays detect potential binding sites, but these sites do not necessarily translate to

function *in vivo*. This huge discrepancy between false positives and true positives was referred to as the *futility theorem*, stating that the predicted TFBSs will most likely not be functional *in vivo* regardless of their binding affinity *in vitro* (Wasserman and Sandelin, 2004). This suggests that additional information should be used to increase the prediction accuracy. As shown in the results of the work presented here, by combining computational evidence, such as scoring based on binding profiles derived through *in vitro* assays and subsequently curated through literature (e.g, the JASPAR database (Khan et al., 2018)) and experimental evidence, such as proximity to the ChIP-seq peak summit (Worsley Hunt et al., 2014) we improve our prediction accuracy and confidence of inferring *bona fide* direct TF-DNA interactions (Gheorghe et al., 2019). Nevertheless, there is no hard evidence that the inferred TFBSs are functional in the given biological condition or not.

Another important aspect regarding the rate of false positives arising from ChIP-seq data is the parameter setting used in the computational tools. Depending on the aim of the ChIP-seq experiment (e.g., TFBSs or histone modifications), the *peak-caller* should be chosen accordingly, as the behavior may differ considerably. It has been shown that even slight changes in parameter tuning of the tools used in either read mapping, peak calling, or TFBSs prediction can significantly affect the results (Bailey et al., 2013; Zhang et al., 2016) and therefore the number of false positives. This is in part due to the variation in the length of the ChIP-seq peaks. For instance, the peaks identified from a ChIP-seq experiment based on TFs have an average length of 300-400 bp, while the peaks identified from histone based ChIP-seq can reach several thousand bps (Park, 2009). Depending on the parameter setting and the data quality, the number of false positives may vary wildly. Attempts to generalize the parameter settings used in the computational tools were made, but to date there is no standard for ChIP-seq data analysis (Bailey et al., 2013). In general, the parameter settings are based on somewhat arbitrary values that are either derived based on performance on the data used in model testing or on common use within the community. While some of these "hard-coded" values have become the *gold standard* in their context (e.g., a p-value < 0.05 to determine significance), others are highly dependent on the data quality used in the model, such as read depth, length, or quality (Zhang et al., 2016). When predicting TFBSs, a threshold on the computational score (e.g., PWM score) is generally used to define the set of *bona fide* TFBSs (Medina-Rivera et al., 2011). Recently, a heuristic approach

was developed to predict *bona fide* direct TF-DNA interactions based on an *enrichment zone* derived from PWM scores (i.e, computational evidence) and distance to peak summit (i.e., experimental evidence) (Worsley Hunt et al., 2014). While this method works well with PWM scores, it is not suited for more recent TFBS computational models (Mathelier and Wasserman, 2013; Mathelier et al., 2016; Zhao et al., 2012). Moreover, this method also makes use of "hard-coded", somewhat arbitrary values that are used in the model (Worsley Hunt et al., 2014). In this work, we did not assess the impact of ChIP-seq peak caller parameter settings, but we aimed at developing a method that is able to define an *enrichment zone* that is data driven and not based on pre-defined thresholds. Therefore, we have developed a non-parametric methodology that is able to automatically define this enrichment zone for each ChIP-seq data set individually (Gheorghe et al., 2019). This data driven method favors specificity over sensitivity and delineates a subset of high confidence TFBS predictions that are supported by both strong computational and experimental evidence.

3.4 Identifying regulons: still a highly complex problem

Another challenge besides the identification of *bona fide* TFBSs is to infer the subset of genes that are direct targets of TFs and, in doing so, the most likely functional TFBSs. This will allow for the generation of gene regulatory networks and subsequently help understanding and predicting the cascading effect resulting from disruptions in the transcriptional regulation machinery. A *regulon* is the ensemble of genes that are regulated by the same TF, and their common characteristic is the presence of TFBSs for this TF at their CREs (Lengeler et al., 1999). Identifying TF regulons is therefore crucial in order to understand the pathways that are affected in disease and disease progression.

As ChIP-seq data is so abundant and available through public repositories, it has also become the preferred assay to infer TF regulons. The general approach to determine if a TF regulates a gene or not is based on the distance between a ChIP-seq peak and the TSS of that gene. Subsequently, different scoring implementations assign a certain probability of the TF to

regulate a gene (i.e., the closer to the TSS, the higher the probability) (Sikora-Wohlfeld et al., 2013). In addition to the distance to TSS and gene-enhancer associations, we observed that adding other layers of information, such as promoter information or sequence conservation score, can also improve predictive power, but it varies greatly between TFs and between ChIP-seq data sets. We hypothesize that this is due to cell line and biological condition specificities. For instance, a TFBS may be found in two different cell lines but may not be functional in both cell lines. In our model, we have aggregated promoter/enhancer/CRM information across all cell lines for simplicity reasons. We suggest that using cell line-specific information, where available, can further improve prediction accuracy.

Another issue is the lack of a standardized "benchmark" to assess the performance of TF regulon prediction models. The general practice is to use the overlap with known sets of associated genes as a performance metric. These sets of genes are derived through knock -in or -down experiments, but they are subject to false positives generated through experimental variation and/or subsequent computational analysis. Here, we observed that using sets of genes that are manually curated from the literature increases prediction performance, when compared to using the experimentally derived genes directly. We considered using gene ontology (GO) term similarity to assess model performance, but due to the GO term redundancy as a consequence of the database structure itself, we can not conclude that this metric is better than using gene overlap. Ultimately, experimental validation should be used to assess the prediction performance of the model, but obviously this process is time and resource expensive.

All these aspects render TF regulon prediction a highly complex problem, due to the high feature dimensionality and lack of a *gold standard* in performance assessment. In this work, we opted for a ranked list-based prediction model and GO term similarity for result validation, but other approaches and performance metrics should be explored. For instance, the prediction model could be implemented within a Bayesian framework or in a semi supervised machine learning framework, such as neural networks. The prediction performance can be further improved by including cell line specificities or known biological processes associated to the TF in the model. Most importantly, the type and quality of the data used in the prediction model strongly affects its performance. Using precise TFBS locations instead of ChIP-seq peaks improves the prediction performance, and using manually curated gene asso-

ciations is likely to improve prediction validation.

3.5 Computationally deriving molecular specificities of cancers

Using our high confidence TFBS predictions and publicly available RNA-seq data, we were able to identify sets of genes that are dysregulated between ER+ and ER-, and subsequently infer which TFs are enriched at the CREs of these genes. This is an important first step in deciphering transcriptional regulatory specificities of the two cancer subtypes. Of course, validation of these findings is needed. One approach for validation would be through knock-in or knock-out experiments in ER-/ER+ models. The analysis flow herein presented can be used to identify molecular signatures of other breast cancer subtypes or cancers in general, provided that sufficient data is available for statistical tests to be performed. This analysis could also be applied to a case/control scenario instead of two cancer subtypes. For instance, comparing expression profiles between non-tumorous and tumorous samples to identify the subset of genes that are dysregulated in an oncogenic phenotype or any phenotype in general.

Here, we used RNA-seq data to analyse genome-wide expression profiles. Our methodology can also be applied to GRO-seq data (Lopes et al., 2017), which measures nascent RNA as opposed to steady-state RNA measured by RNA-seq. Based on data availability, the analysys workflow can be further developed to measure nascent RNA levels at different time points (i.e., as a time series) and study the dynamics of the two cancer subtypes. Another data type that can be used is ATAC-seq (Buenrostro et al., 2015), which is an assay for measurement of chromatin accessibility genome-wide. Using such data, one can study the chromatin state at CREs as an indication of enhancer and/or promoter activity (i.e., if in an open chromatin region, the CRE is more likely to be functional), and the CREs that are active or inactive in each of the two conditions (i.e., ER+ vs. ER-) can be inferred. Subsequently, TF motif and/or TFBS enrichment analysis can be performed on these sets of CREs. However, these two experimental assays are fairly new and to date, data availability is quite limited.

A potential limitation of this analysis workflow could be the set of TF binding profiles used in the motif enrichment analysis. Depending on the size and diversity of the binding motifs, the results of the motif enrichment may be hindered. The same applies to the TFBS enrichment analysis, as the results are restricted by the set of TFBSs serving as input. When performing the genomic region enrichment analysis, one could use ChIP-seq peaks instead of TFBSs, but this could lead to a high number of false positives (Teytelman et al., 2013; Worsley Hunt et al., 2014; Jain et al., 2015). Here we opted for high confidence, precise TFBS locations, favoring specificity over sensitivity.

Another layer of information that can be added to this workflow is the set of mutations for each sample. This will potentially allow one to further cluster or stratify the samples between the two conditions and perform a multilayered analysis. As such, more refined sets of differentially expressed genes can be identified based on subsets of mutations, again, provided that sufficient data is available. One can also use genome-wide DNA methylation status instead of (or complementary to) mutation information. Subsequently, more tailored targeted therapies can be developed based on the unique set of mutations of a patient and the TFs enriched within each subset. This type of approach to characterize cancer subtypes or disease in general, from a molecular perspective, has opened the door to personalized medicine. As such, the individual genetic makeup of a patient can be analysed and efficient treatments can be developed faster than a generic treatment (Wang, 2016; Nussinov et al., 2019).

3.6 Biomedical considerations for targeted cancer therapy

Understanding the causes of cellular regulatory program disruption leading to carcinogenesis is key to the development of targeted cancer therapies. There are several other factors besides genetic background that can lead to carcinogenesis, such as DNA replication errors, exposure to environmental stress, or inappropriate diet. As a result of these factors, cells may accumulate somatic mutations over time, which are DNA modifications occurring in non-germ cells. As opposed to passenger mutations that do not affect cell fitness, specific somatic mutations are considered as cancer drivers when they dys-

regulate the cell regulatory program and provide a fitness advantage to the cells carrying them (Martincorena and Campbell, 2015).

An ongoing challenge lies in discriminating between such driving events from background passenger mutations. The broad variation in the genetic makeup of each individual adds to the complexity of this problem (Burrell et al., 2013). Personalized medicine arises from the fact that each patient has a unique set of mutations (Chin et al., 2011). Recently, the focus has shifted towards the non-coding part of the genome, as mutations occurring within the protein coding region were not sufficient to explain the resulting oncogenic phenotype (Khurana et al., 2016). This translates to, among others, assessing the impact of somatic mutations occurring at CREs and thus assessing how mutations occurring at promoters and/or enhancers can disrupt gene regulatory networks, ultimately leading to carcinogenesis.

One approach to designing targeted therapies can be based on the set of mutations of each individual. The entire set of patient-specific mutations can be extracted, together with their gene expression profiles, from data portals such as ICGC (Zhang et al., 2011). This information, in combination with high confidence TF-DNA interactions, such as the ones hosted in UniBind (Gheorghe et al., 2019) can be used to determine the subset of mutations that occur at TFBSs. These mutations can be associated with gene activity within regulatory networks and their impact on gene expression assessed. Computational frameworks that allow prediction of such associations have already been developed; one an example is *xseq* (Ding et al., 2015). This tool was initially developed to analyze the effect of somatic mutations in protein-coding regions on transcription. The *xseq* tool can be adapted to work with CRE information and predict mutations that are likely to present a functional impact on gene regulation.

The set of mutations identified to be highly likely responsible for disruptions in gene regulatory programs associated with cancer can be experimentally validated using, for instance, genome editing techniques. Ultimately, these results can lead to the development of personalized approaches that would inhibit gene expression dysregulation.

3.7 Further improvement of the tools and resources

The tools and resources developed here represent an effort to improve our understanding of the intricate mechanisms of transcriptional regulation. The *ChIP-eat* pipeline was designed to predict direct TF-DNA interactions from ChIP-seq data. It is able to automatically identify a set of high confidence TFBS predictions, supported by both strong computational and experimental evidence in a non-parametric, data driven manner. Nevertheless, *ChIP-eat* can be improved by adding a ChIP-seq peak *rescanning* step. This translates to keeping the top scoring sequence per ChIP-seq peak that falls within the enrichment zone, even if it does not represent the top scoring sequence across the entire ChIP-seq peak (Gheorghe et al., 2019) (see Results subsection 2). Moreover, the UniBind database (https://unibind.uio.no/), which to our knowledge is the most comprehensive of its kind to date, can be extended by adding TFBS predictions from multiple species and/or from different ChIP-seq peak callers.

Our framework of predicting TF regulons was designed to use additional layers of relevant information, besides the distance to the closest TSS. The ranked-list based approach implemented in *TF-regulons* may not be the optimal implementation to solve this highly complex problem. Other machine learning approaches can be used to infer TF target genes with more accuracy. For instance, if unsupervised learning is the *weapon of choice*, Bayesian networks can be employed. A further improvement would also be to use hybrid Bayesian networks, which should be better at modeling the features represented as continuous variables. If a semi-supervised machine learning approach is of more interest, deep learning approaches such as neural networks can be employed and biological information used as prior knowledge. Nevertheless, the features which improve the prediction performance should be thoroughly assessed, as prediction accuracy and feature selection might prove to be highly variable among biological conditions, even for the same cell type.

As for identifying sets of TFs that are specific to different types of cancer or phenotypes in general, one important aspect is the quality and quantity of data used as input. Here, we aimed at employing high quality input data (i.e., UniBind predictions) in our analyses to identify key TFs between ER-

and ER+ breast cancers. A limitation might be the relatively small number of TFs for which we had manually curated binding profiles and subsequently predicted TFBSs. Therefore, a more comprehensive set of input data might identify a wider set of potential TF candidates. Nevertheless, this relies on the presence, diversity, and quality of the ChIP-seq data sets made publicly available that are used to derive TF-DNA interactions.

3.8 General discussion

The work presented here fits in the general context of developing, improving, and applying bioinformatics tools and resources to shed more light on the intricate molecular mechanisms governing gene expression regulation. More specifically, the focus has been on transcriptional regulation achieved through TFs and how disruptions in the gene regulatory networks they rule can explain oncogenic phenotypes. This translates to a multilayered, highly complex problem, as transcriptional regulation is governed by a complex interplay between these key proteins and DNA. To resolve this puzzle, high quality, reliable data should be employed in order to obtain reliable results. Generally, multiple analyses are used in a workflow in which the output of one processing step serves as input for the next. Therefore, using low quality data will generate low quality results, as the late computer scientist Wilf Hey coined the phrase, "garbage in, garbage out".

Computational tools are critical to reduce the search space in such contexts, where the posibilities and combinations are countless. Nevertheless, these tools have to be reliable and transparent. In other words, they should be able to reproduce the results they generate, and ideally all processing steps should be visible. The bioinformatics field is relatively new, and the tendency is to generate bits and pieces of software aiming at solving one highly specific task. This is a direct consequence of the high speed at which biological assays are developed and new data types and approaches are generated (Phan et al., 2009). As such, software developed with the same aim can be produced independently in different institutions and written in different programming languages. This in turn translates to differences in implementation and thus differences in the results obtained. In most of the cases, there is no *gold standard* for comparison when choosing what computational tool to use.

Moreover, a large portion of these tools do not have a graphical interface that allows the users to easily interact with the tool. As such, these tools require the user to launch the processing through a command line interface, which for most researchers constitutes an early roadblock (Stein, 2002).

As a consequence, the massive amounts of biological data that are generated through next-generation sequencing cannot be used at their full potential (Stein, 2002). Lately, guidelines have been put in place to ensure data quality standards and, to a lesser extent, uniformity in processing the data from different experimental assays (Bailey et al., 2013; Landt et al., 2012; Mason et al., 2010). This encourages the community to use software *suites*, which represent a collection of individual computational tools that are able to fulfill a broader set of tasks in a more standardized manner (Bailey et al., 2009; Quinlan and Hall, 2010). In parallel with the development of such software toolboxes, processing *pipelines* have been developed (Leipzig, 2017). Generally, they represent a data processing workflow able to perform an entire analysis from raw data to the final results. It is increasingly common that *suites* and *pipelines* also integrate quality checks, as the biological data quality can vary greatly due to experimental conditions, material and instruments used, etc. Large consortia, such as ENCODE, host thousands of publicly available data sets and have compiled sets of guidelines and reference quality indicators for different types of data to ensure input quality in the processing pipelines (The ENCODE Project Consortium, 2012). Moreover, *pipelines* ensure result reproducibility, as the same set of tools and data processing steps with the same parameter settings is applied on every data set (Kanwal et al., 2017).

Besides the development of computational tools, suites of tools, and data processing pipelines, the generation, update, and maintenance of databases is equally important. These are important resources for the community, enabling larger scale, higher level analyses. In some cases, the development of databases even precedes the development of tools. For instance, in the world of TFs, databases hosting reference binding motifs were generated through literature curation before tools for motif enrichment or motif discovery were developed. Two examples of such databases that appeared around the same time are TRANSFAC (Wingender et al., 1996) for eukaryotic organisms and RegulonDB (Huerta et al., 1998) for bacteria.

In general, databases hosting TFBMs, PFMs and/or TFBSs, or any CRE

information specialize in one organism or the same taxonomy. Others, such as JASPAR (Sandelin et al., 2004) host data for several organisms. For every database, the methods through which the TFBSs or PFMs were obtained, as well as their representation format, varies. Regardless, most (if not all) of the information contained in these databases comes from manual curation, which means that the TFBSs and PFMs are more likely to be trustworthy. Such curation is feasible when dealing with simpler organisms like bacteria where the regulatory network is less complex. When dealing with higher organisms, such as vertebrates, the manual curation process becomes tedious, and huge efforts are made to populate and update the databases. Moreover, as the vast majority of these databases rely on data coming from high-throughput assays and sequencing, the manual curation process becomes even more time consuming due to the large amounts of data. However, directly inferring TFBSs from high-throughput data is prone to introduce a large number of false positives, as it has been shown for the RegulonDB (Weiss et al., 2013).

Most of these databases are constantly updated and maintained, and with every update the size of the database increases considerably. This is a normal phenomenon given the huge increase in the amounts of data generated every year as sequencing costs become lower. These updates allow for the refinement of the inferred PFMs, but also increase the number of redundant motifs. However, some databases are a "side product" of large scale studies (Jolma et al., 2015; Whitaker et al., 2015) and they are not maintained nor updated. Due to the ever increasing number of databases, the question becomes which database to use in a study. Recently, some databases such as footprintDB (Sebastian and Contreras-Moreira, 2014) and Cis-BP (Weirauch et al., 2014) implemented and standardized data from multiple databases complementary to their own.

As personalized medicine is a rapidly growing field, it is crucial to rely on high quality data and dependable computational tools to develop targeted therapies based on the unique genetic makeup of each patient. With the work presented here, we improved our capacity to predict *bona fide* TFBSs genome-wide and made this resource publicly available. These data may serve as a base in numerous research projects that aim at understanding the impact of alterations occurring in the transcriptional regulatory machinery. As a direct application of our UniBind database hosting the entire set of TFBS predictions, we showed that we can infer key TFs that are specific to ER- and ER+ breast cancer subtypes. These TFs can be used to infer

associated regulons (i.e., gene regulatory networks) and ultimately develop targeted therapies replacing for instance chemotherapy as treatment for ER- breast cancers.

Bibliography

Aerts, S. (2012). Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. *Current Topics in Developmental Biology*, 98:121–145.

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., and Ioannidis, J. P. A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLOS ONE*, 6(9):e24357.

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

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

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., and et al. (2014a). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461.

Andersson, R., Refsing Andersen, P., Valen, E., Core, L. J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014b). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Communications*, 5:5336.

Arnosti, D. N. (2003). ANALYSIS AND FUNCTION OF TRANSCRIPTIONAL REGULATORY ELEMENTS: Insights from Drosophila. *Annual Review of Entomology*, 48(1):579–602.

Arnosti, D. N. and Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898.

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):1720–1723.

Bagchi, D. N. and Iyer, V. R. (2016). The Determinants of Directionality in Transcriptional Initiation. *Trends in genetics : TIG*, 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):e1003326.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37(Web Server issue):W202–W208.

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 a -globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2, Part 1):299–308.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003). Modeling Dependencies in protein-DNA Binding Sites. In *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, RECOMB '03, pages 28–37, New York, NY, USA. ACM. event-place: Berlin, Germany.

Barrett, L. W., Fletcher, S., and Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*, 69(21):3613–3634.

Bastow, R. and Leonelli, S. (2010). Sustainable digital infrastructure. *EMBO Reports*, 11(10):730–734.

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

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively

determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429.

Biłas, R., Szafran, K., Hnatuszko-Konka, K., and Kononowicz, A. K. (2016). Cis-regulatory elements used to control gene expression in plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, 127(2):269–287.

Blackwood, E. M. and Kadonaga, J. T. (1998). Going the Distance: A Current View of Enhancer Action. *Science*, 281(5373):60–63.

Bonev, B. and Cavalli, G. (2016). Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661–678.

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genetics*, 29(4):365.

Brembs, B. (2018). Prestigious Science Journals Struggle to Reach Even Average Reliability. *Frontiers in Human Neuroscience*, 12.

Brenowitz, M., Senear, D. F., Shea, M. A., and Ackers, G. K. (1986). [9] Quantitative DNase footprint titration: A method for studying protein-DNA interactions. In *Methods in Enzymology*, volume 130 of *Enzyme Structure Part K*, pages 132–181. Academic Press.

Brivanlou, A. H. and Darnell, J. E. (2002). Signal Transduction and the Control of Gene Expression. *Science*, 295(5556):813–818.

Buenrostro, J., Wu, B., Chang, H., and Greenleaf, W. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 109:21.29.1–21.29.9.

Buermans, H. P. J. and den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica Et Biophysica Acta*, 1842(10):1932–1941.

Burgoon, L. D. (2006). The need for standards, not guidelines, in biological data reporting and sharing. *Nature Biotechnology*, 24(11):1369.

Burke, T. W. and Kadonaga, J. T. (1997). The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes & Development*, 11(22):3020–3031.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345.

Busby, S. and Ebright, R. H. (1999). Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology*, 293(2):199–213.

Cai, L. and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14(0):2.

Cann, J. R. (1998). Theoretical studies on the mobility-shift assay of protein-DNA complexes. *Electrophoresis*, 19(2):127–141.

Cavener, D. R. (1987). Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. *Nucleic Acids Research*, 15(4):1353–1361.

Chavez, K. J., Garimella, S. V., and Lipkowitz, S. (2010). Triple Negative Breast Cancer Cell Lines: One Tool in the Search for Better Treatment of Triple Negative Breast Cancer. *Breast disease*, 32(1-2):35–48.

Chen, K. and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8(2):93–103.

Cheng, C., Min, R., and Gerstein, M. (2011). TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics (Oxford, England)*, 27(23):3221–3227.

Chin, L., Andersen, J. N., and Futreal, P. A. (2011). Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, 17(3):297–303.

Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Research*, 46(D1):D267–D275.

Cochrane, G. R. and Galperin, M. Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Research*, 38(Database issue):D1–4.

Collas, P. (2010). The Current State of Chromatin Immunoprecipitation. *Molecular Biotechnology*, 45(1):87–100.

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., and Myers, R. M. (2006). Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Research*, 16(1):1–10.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258):561.

Crocker, J., Preger-Ben Noon, E., and Stern, D. L. (2016). Chapter Twenty-Seven - The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. In Wassarman, P. M., editor, *Current Topics in Developmental Biology*, volume 117 of *Essays on Developmental Biology, Part B*, pages 455–469. Academic Press.

Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W., and Richmond, T. J. (2002). Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution††We dedicate this paper to the memory of Max Perutz who was particularly inspirational and supportive to T.J.R. in the early stages of this study. *Journal of Molecular Biology*, 319(5):1097–1113.

Davidson, E. H. (2006). *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, Burlington, MA ; San Diego, 1 edition edition.

Davidson, S., Overton, C., and Buneman, P. (1995). Challenges in Integrating Biological Data Sources. *Journal of Computational Biology*, 2(4):557–572.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature reviews. Genetics*, 14(6):390–403.

Ding, J., McConechy, M. K., Horlings, H. M., Ha, G., Chun Chan, F., Funnell, T., Mullaly, S. C., Reimand, J., Bashashati, A., Bader, G. D., Huntsman, D.,

Aparicio, S., Condon, A., and Shah, S. P. (2015). Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature Communications*, 6:8554.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380.

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.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210.

Ell, B., Mercatali, L., Ibrahim, T., Campbell, N., Schwarzenbach, H., Pantel, K., Amadori, D., and Kang, Y. (2013). Tumor-Induced Osteoclast miRNA Changes as Regulators and Biomarkers of Osteolytic Bone Metastasis. *Cancer Cell*, 24(4):542–556.

Ellis, T., Evans, D. A., Martin, C. R. H., and Hartley, J. A. (2007). A 96-well DNase I footprinting screen for drug–DNA interactions. *Nucleic Acids Research*, 35(12):e89.

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):3851–3856.

Ezer, D., Zabet, N. R., and Adryan, B. (2014). Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. *Computational and Structural Biotechnology Journal*, 10(17):63–69.

Fickett, J. W. (1996). Quantitative discrimination of MEF2 sites. *Molecular and Cellular Biology*, 16(1):437–441.

Figueiredo, A. S. (2017). Data Sharing: Convert Challenges into Opportunities. *Frontiers in Public Health*, 5.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D., and Cohen, D. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database: The Journal of Biological Databases and Curation*, 2017.

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

Garner, M. M. and Revzin, A. (1986). The use of gel electrophoresis to detect and study nucleic acid— protein interactions. *Trends in Biochemical Sciences*, 11(10):395–396.

Geertz, M. and Maerkl, S. J. (2010). Experimental strategies for studying transcription factor–DNA binding specificities. *Briefings in Functional Genomics*, 9(5-6):362–373.

Gewin, V. (2016). Data sharing: An open mind on open data. *Nature*, 529(7584):117–119.

Gheorghe, M., Sandve, G. K., Khan, A., Chèneby, J., Ballester, B., and Mathelier, A. (2019). A map of direct TF–DNA interactions in the human genome. *Nucleic Acids Research*, 47(4):e21–e21.

Gibcus, J. H. and Dekker, J. (2013). The Hierarchy of the 3d Genome. *Molecular Cell*, 49(5):773–782.

Gonzalez-Sandoval, A. and Gasser, S. M. (2016). On TADs and LADs: Spatial Control Over Gene Expression. *Trends in genetics: TIG*, 32(8):485–495.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. (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.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013). On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biology and Evolution*, 5(3):578–590.

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E., Mikkelsen, T. S., and Lander, E. S. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences*, page 201621150.

Hah, N., Murakami, S., Nagari, A., Danko, C. G., and Kraus, W. L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. *Genome Research*, 23(8):1210–1223.

Hansen, J. C. (2002). Conformational Dynamics of the Chromatin Fiber in Solution: Determinants, Mechanisms, and Functions. *Annual Review of Biophysics and Biomolecular Structure*, 31(1):361–392.

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

Harrison, S. C. (1991). A structural taxonomy of DNA-binding domains. *Nature*, 353(6346):715.

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, 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., 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):311–318.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3):144–154.

Helden, J. v., 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):1808–1818.

Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (2012). i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Research*, 40(15):e114–e114.

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):7319–7324.

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

Hulton, C. S., Seirafi, A., Hinton, J. C., Sidebotham, J. M., Waddell, L., Pavitt, G. D., Owen-Hughes, T., Spassky, A., Buc, H., and Higgins, C. F. (1990). Histone-like protein H1 (H-NS), DNA supercoiling, and gene expression in bacteria. *Cell*, 63(3):631–642.

Ibrahim, M. M., Karabacak, A., Glahs, A., Kolundzic, E., Hirsekorn, A., Carda, A., Tursun, B., Zinzen, R. P., Lacadie, S. A., and Ohler, U. (2018). Determinants of promoter and enhancer transcription directionality in metazoans. *Nature Communications*, 9(1):4472.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003.

Ishihama, A., Shimada, T., and Yamazaki, Y. (2016). Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Research*, 44(5):2058–2074.

IUPAC, I. (1985). Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. *European Journal of Biochemistry*, 150(1):1–5.

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

Jain, D., Baldi, S., Zabel, A., Straub, T., and Becker, P. B. (2015). Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Research*, 43(14):6959–6968.

Janky, R., Verfaillie, A., Imrichová, H., Sande, B. V. d., Standaert, L., Christiaens, V., Hulselmans, G., Herten, K., Sanchez, M. N., Potier, D., Svetlichnyy, D., Atak, Z. K., Fiers, M., Marine, J.-C., and Aerts, S. (2014). iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLOS Computational Biology*, 10(7):e1003731.

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

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, 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. *Genome Research*, 20(6):861–873.

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):384–388.

Jones, S. (2004). An overview of the basic helix-loop-helix proteins. *Genome Biology*, 5(6):226.

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):5221–5231.

Kanwal, S., Khan, F. Z., Lonie, A., and Sinnott, R. O. (2017). Investigating reproducibility and tracking provenance – A genomic workflow case study. *BMC Bioinformatics*, 18.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830):1484–1488.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, 12(6):996–1006.

Kepert, J. F., Tóth, K. F., Caudron, M., Mücke, N., Langowski, J., and Rippe, K. (2003). Conformation of Reconstituted Mononucleosomes and Effect of Linker Histone H1 Binding Studied by Scanning Force Microscopy. *Biophysical Journal*, 85(6):4012–4022.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., and Mathelier, A. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46(D1):D260–D266.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nature Reviews. Genetics*, 17(2):93–108.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nature Methods*, 3(3):211–222.

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.

Kornberg, R. D. (1974). Chromatin Structure: A Repeating Unit of Histones and DNA. *Science*, 184(4139):868–871.

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. (2006). High Guanine and Cytosine Content Increases mRNA Levels in Mammalian Cells. *PLOS Biology*, 4(6):e180.

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*, 26(20):2622–2623.

Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C. L., Raha, D., Winters, E. E., Johnson, S. M., Snyder, M., Batzoglou, S., and Sidow, A. (2012). Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, 22(9):1735–1747.

Kwon, A. T., Arenillas, D. J., Hunt, R. W., and Wasserman, W. W. (2012). oPOSSUM-3: Advanced Analysis of Regulatory Motif Over-Representation

Across Genes or ChIP-Seq Datasets. *G3: Genes|Genomes|Genetics*, 2(9):987–1002.

Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., and Ebright, R. H. (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes & Development*, 12(1):34–44.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4):650–665.

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., and et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831.

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

Lee, D. J., Minchin, S. D., and Busby, S. J. (2012). Activating Transcription in Bacteria. *Annual Review of Microbiology*, 66(1):125–152.

Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, 18(3):530–536.

Lelli, K. M., Slattery, M., and Mann, R. S. (2012). Disentangling the Many Layers of Eukaryotic Transcriptional Regulation. *Annual Review of Genetics*, 46(1):43–68.

Lengeler, J. W., Drews, G., and Schlegel, H. G. (1999). *Biology of the Prokaryotes*. Georg Thieme Verlag. Google-Books-ID: MiwpFtTdmjQC.

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

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147.

Li, B., Carey, M., and Workman, J. L. (2007). The Role of Chromatin during Transcription. *Cell*, 128(4):707–719.

Lieberman-Aiden, E., Berkum, N. L. v., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950):289–293.

Lifton, R., Goldberg, M., Karp, R., and Hogness, D. (1978). *The Organization of the Histone Genes in Drosophila melanogaster: Functional and Evolutionary Implications*, volume 42. Spring Harbor Symp Quant Biol.

Littlefield, O., Korkhin, Y., and Sigler, P. B. (1999). The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Proceedings of the National Academy of Sciences*, 96(24):13668–13673.

Lopes, R., Agami, R., and Korkmaz, G. (2017). GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression. In Napoli, S., editor, *Promoter Associated RNA: Methods and Protocols*, Methods in Molecular Biology, pages 45–55. Springer New York, New York, NY.

Lupiáñez, D. G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in genetics: TIG*, 32(4):225–237.

Maerkl, S. J. and Quake, S. R. (2007). A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*, 315(5809):233–237.

Mahony, S. and Pugh, B. F. (2015). Protein-DNA binding in high-resolution. *Critical reviews in biochemistry and molecular biology*, 50(4):269–283.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics: TIG*, 24(3):133–141.

Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255):1483–1489.

Mason, C. E., Zumbo, P., Sanders, S., Folk, M., Robinson, D., Aydt, R., Gollery, M., Welsh, M., Olson, N. E., and Smith, T. M. (2010). Standardizing the next generation of bioinformatics software development with BioHDF (HDF5). *Advances in Experimental Medicine and Biology*, 680:693–700.

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).

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*, 3(3):278–286.e4.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, 15(suppl_1):R17–R29.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–355.

McKay, D. B. and Steitz, T. A. (1981). Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. *Nature*, 290(5809):744.

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):808–824.

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L., Liu, T., Brown, M., Meyer, C. A., and Liu, X. S. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, 45(D1):D658–D662.

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

Methods, N. (2016). Database under maintenance. *Nature Methods*, 13:699.

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):1331–1339.

Mullis, K. B., Erlich, H. A., Arnheim, N., Horn, G. T., Saiki, R. K., and Scharf, S. J. (1989). Process for amplifying, detecting, and/or cloning nucleic acid sequences.

Naumova, N., Smith, E. M., Zhan, Y., and Dekker, J. (2012). Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods*, 58(3):192–203.

Nelson, A. C. and Wardle, F. C. (2013). Conserved non-coding elements and cis regulation: actions speak louder than words. *Development*, 140(7):1385–1395.

NIH, N. (2019). The Cost of Sequencing a Human Genome.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature*, 485(7398):381–385.

Nussinov, R., Jang, H., Tsai, C.-J., and Cheng, F. (2019). Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. *PLOS Computational Biology*, 15(3):e1006658.

O'Connor, T., Bodén, M., and Bailey, T. L. (2017). CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. *Nucleic Acids Research*, 45(4):e19–e19.

O'Neill, L. P. and Turner, B. M. (1996). Immunoprecipitation of chromatin. In *Methods in Enzymology*, volume 274 of *RNA Polymerase and Associated Factors, Part B*, pages 189–197. Academic Press.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680.

Parker, D. S., White, M. A., Ramos, A. I., Cohen, B. A., and Barolo, S. (2011). The cis-regulatory logic of Hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Science Signaling*, 4(176):ra38.

Patel, R. Y. and Stormo, G. D. (2014). Discriminative motif optimization based on perceptron training. *Bioinformatics (Oxford, England)*, 30(7):941–948.

Paulsen, J., Ali, T. M. L., Nekrasov, M., Delbarre, E., Baudement, M.-O., Kurscheid, S., Tremethick, D., and Collas, P. (2019). Long-range interactions

between topologically associating domains shape the four-dimensional genome during differentiation. *Nature Genetics*, 51(5):835.

Payankaulam, S., Li, L. M., and Arnosti, D. N. (2010). Transcriptional repression: conserved and evolved features. *Current biology : CB*, 20(17):R764–R771.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature Methods*, 6(11s):S22–S32.

Phan, J. H., Moffitt, R. A., Stokes, T. H., Liu, J., Young, A. N., Nie, S., and Wang, M. D. (2009). Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. *Trends in biotechnology*, 27(6):350–358.

Pierce, B. A. (2012). *Genetics a conceptual approach*. New York W.H. Freeman, 4th ed edition.

Portales-Casamar, E., Arenillas, D., Lim, J., Swanson, M. I., Jiang, S., McCallum, A., Kirov, S., and Wasserman, W. W. (2009). The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research*, 37(Database issue):D54–D60.

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):e10611.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Qureshi, M. and Ivens, A. (2008). A software framework for microarray and gene expression object model (MAGE-OM) array design annotation. *BMC Genomics*, 9:133.

Ravasi, T., Huber, T., Zavolan, M., Forrest, A., Gaasterland, T., Grimmond, S., and Hume, D. A. (2003). Systematic Characterization of the Zinc-Finger-Containing Proteins in the Mouse Transcriptome. *Genome Research*, 13(6b):1430–1442.

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

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson,

C. J., Bell, S. P., and Young, R. A. (2000). Genome-Wide Location and Function of DNA Binding Proteins. *Science*, 290(5500):2306–2309.

Rhee, H., Bataille, A., 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.

Riethoven, J.-J. M. (2010). Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. In Ladunga, I., editor, *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology, pages 33–42. Humana Press, Totowa, NJ.

Riley, T. R., Slattery, M., Abe, N., Rastogi, C., Liu, D., Mann, R. S., and Bussemaker, H. J. (2014). SELEX-seq: A Method for Characterizing the Complete Repertoire of Binding Site Preferences for Transcription Factor Complexes. In Graba, Y. and Rezsohazy, R., editors, *Hox Genes: Methods and Protocols*, Methods in Molecular Biology, pages 255–278. Springer New York, New York, NY.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, 79:233–269.

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–1253.

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., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Research*, 32(Database issue):D91–94.

Sarkans, U., Parkinson, H., Lara, G. G., Oezcimen, A., Sharma, A., Abeygunawardena, N., Contrino, S., Holloway, E., Rocca-Serra, P., Mukherjee, G., Shojatalab, M., Kapushesky, M., Sansone, S.-A., Farne, A., Rayner, T., and Brazma, A. (2005). The ArrayExpress gene expression database: a software engineering and implementation perspective. *Bioinformatics*, 21(8):1495–1501.

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):1036–1040.

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.

Schöne, S., Jurk, M., Helabad, M. B., Dror, I., Lebars, I., Kieffer, B., 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. *Nature Communications*, 7:12621.

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

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T., and Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, 32(2):171–178.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4):272–286.

Siebert, M. and Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44(13):6055–6069.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050.

Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E. G., Singaravelu, K., and Beyer, A. (2013). Assessing Computational Methods for Transcription Factor Target Gene Identification Based on ChIP-seq Data. *PLOS Computational Biology*, 9(11):e1003342.

Sikorski, T. W. and Buratowski, S. (2009). The Basal Initiation Machinery: Beyond the General Transcription Factors. *Current opinion in cell biology*, 21(3):344–351.

Simpson, R. T. (1978). Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. *Biochemistry*, 17(25):5524–5531.

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.

Smale, S. T. and Baltimore, D. (1989). The "initiator" as a transcription control element. *Cell*, 57(1):103–113.

Song, C., Zhang, S., and Huang, H. (2015). Choosing a suitable method for the identification of replication origins in microbial genomes. *Frontiers in Microbiology*, 6.

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

Stein, L. (2002). Creating a bioinformatics nation. *Nature*, 417(6885):119.

Stormo, G. D. (1990). [13] Consensus patterns in DNA. In *Methods in Enzymology*, volume 183, pages 211–221. Academic Press.

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

Stormo, G. D. (2013). Modeling the specificity of protein-DNA interactions. *Quantitative biology*, 1(2):115–130.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, 10(9):2997–3011.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L., and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Research*, 24(3):390–400.

Szalkowski, A. M. and Schmid, C. D. (2011). Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Briefings in Bioinformatics*, 12(6):626–633.

Takahashi, K. and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4):663–676.

Tenenbaum, J. D., Sansone, S.-A., and Haendel, M. (2014). A sea of standards for omics data: sink or swim? *Journal of the American Medical Informatics Association*, 21(2):200–203.

Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 110(46):18602–18607.

Thanos, D. and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–1100.

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

Thoma, F., Koller, T., and Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *The Journal of Cell Biology*, 83(2):403–427.

Thomas, J. O. and Kornberg, R. D. (1975). An octamer of histones in chromatin and free in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 72(7):2626–2630.

Thomas, M. C. and Chiang, C.-M. (2006). The General Transcription Machinery and General Cofactors. *Critical Reviews in Biochemistry and Molecular Biology*, 41(3):105–178.

Todeschini, A.-L., Georges, A., and Veitia, R. A. (2014). Transcription factors: specific DNA binding and specific gene regulation. *Trends in Genetics*, 30(6):211–219.

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

Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M., and Pontoglio, M. (1997). Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *Journal of Molecular Biology*, 266(2):231–245.

Tserel, L., Kolde, R., Rebane, A., Kisand, K., Org, T., Peterson, H., Vilo, J., and Peterson, P. (2010). Genome-wide promoter analysis of histone modifications in human monocyte-derived antigen presenting cells. *BMC Genomics*, 11:642.

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510.

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., and Lander, E. S. (2010). Hi-C: A Method to Study the Three-dimensional Architecture of Genomes. *Journal of Visualized Experiments : JoVE*, (39).

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):252–263.

Vasilevsky, N. A., Minnier, J., Haendel, M. A., and Champieux, R. E. (2017). Reproducible and reusable research: are journal data sharing policies meeting the mark? *PeerJ*, 5.

Venters, B. J. and Pugh, B. F. (2009). How eukaryotic genes are transcribed. *Critical reviews in biochemistry and molecular biology*, 44(2-3):117–141.

Wade, P. A. (2001). Transcriptional control at regulatory checkpoints by histone deacetylases: molecular connections between cancer and chromatin. *Human Molecular Genetics*, 10(7):693–698.

Wang, L., Chen, J., Wang, C., Uusküla-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–e156.

Wang, X. (2016). Gene mutation-based and specific therapies in precision medicine. *Journal of Cellular and Molecular Medicine*, 20(4):577–580.

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

Weake, V. M. and Workman, J. L. (2010). Inducible gene expression: diverse regulatory mechanisms. *Nature Reviews Genetics*, 11(6):426–437.

Wei, W., Pelechano, V., Järvelin, A. I., and Steinmetz, L. M. (2011). Functional consequences of bidirectional promoters. *Trends in Genetics*, 27(7):267–276.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics*, 45(10):1113–1120.

Weirauch, M. T., Yang, A., Albu, M., Cote, A., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., and et al. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443.

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: The Journal of Biological Databases and Curation*, 2013:bas059.

Whitaker, J. W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nature Methods*, 12(3):265–272, 7 p following 272.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13(9):R50.

Will, T. and Helms, V. (2014). Identifying transcription factor complexes and their roles. *Bioinformatics*, 30(17):i415–i421.

Wingender, E. (1997). Classification Scheme of Eukaryotic Transcription Factors. *lecular Biology*, 31(4):483–497.

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(D1):D165–D170.

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., and Dönitz, J. (2018). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research*, 46(D1):D343–D347.

Wittkopp, P. J. and Kalay, G. (2012). *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69.

Woodcock, C. L. and Ghosh, R. P. (2010). Chromatin Higher-order Structure and Dynamics. *Cold Spring Harbor Perspectives in Biology*, 2(5):a000596.

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*, 15:472.

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.

Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., Bulut, S., Tursun, B., and Akalin, A. (2019). HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Research*.

Wu, Y. X. and Kwon, Y. J. (2016). Aptamers: The "evolution" of SELEX. *Methods*, 106:21–28.

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., and Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Research*, 17(6):798–806.

Xie, W. J., Meng, L., Liu, S., Zhang, L., Cai, X., and Gao, Y. Q. (2017). Structural Modeling of Chromatin Integrates Genome Features and Reveals Chromosome Folding Principle. *Scientific Reports*, 7(1):2818.

Xing, E. P., Jordan, M. I., Karp, R. M., and Russell, S. J. (2003). A Hierarchical Bayesian Markovian Model for Motifs in Biopolymer Sequences. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 1513–1520. MIT Press.

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):801–813.

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M., and Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65.

Yang, C.-C., Andrews, E. H., Chen, M.-H., Wang, W.-Y., Chen, J. J. W., Gerstein, M., Liu, C.-C., and Cheng, C. (2016). iTAR: a web server for identifying target genes of transcription factors using ChIP-seq or ChIP-chip data. *BMC genomics*, 17(1):632.

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):D148–155.

Young, R. A. (2011). Control of Embryonic Stem Cell State. *Cell*, 144(6):940–954.

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

Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, 25(21):2227–2241.

Zeng, P.-Y., Vakoc, C. R., Chen, Z.-C., Blobel, G. A., and Berger, S. L. (2006). In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. *BioTechniques*, 41(6):694–698.

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Research*, 21(8):1273–1283.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L., and Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database: The Journal of Biological Databases and Curation*, 2011.

Zhang, Q., Zeng, X., Younkin, S., Kawli, T., Snyder, M. P., and Keleş, S. (2016). Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. *BMC Bioinformatics*, 17.

Zhao, Y., Ruan, S., Pandey, M., and Stormo, G. D. (2012). Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, 191(3):781–790.

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. E. (2010). Differentiation of Effector CD4 T Cell Populations. *Annual Review of Immunology*, 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.

I

ReMap2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments

Chèneby, J., **Gheorghe, M.**, Artufel, M., Mathelier, A., and Ballester, B.*

2018, *Nucleic Acids Research*, 46(D1):D267–D275.

Published online 8 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D267–D275***
doi: 10.1093/nar/gkx1092

ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments

Jeanne Chèneby[1,2], Marius Gheorghe[3], Marie Artufel[1,2], Anthony Mathelier[3,4] and Benoit Ballester[1,2,*]

[1]INSERM, UMR1090 TAGC, Marseille F-13288, France, [2]Aix-Marseille Université, UMR1090 TAGC, Marseille F-13288, France, [3]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway and [4]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 15, 2017; Revised October 19, 2017; Editorial Decision October 20, 2017; Accepted October 20, 2017

ABSTRACT

With this latest release of ReMap (http://remap.cisreg.eu), we present a unique collection of regulatory regions in human, as a result of a large-scale integrative analysis of ChIP-seq experiments for hundreds of transcriptional regulators (TRs) such as transcription factors, transcriptional co-activators and chromatin regulators. In 2015, we introduced the ReMap database to capture the genome regulatory space by integrating public ChIP-seq datasets, covering 237 TRs across 13 million (M) peaks. In this release, we have extended this catalog to constitute a unique collection of regulatory regions. Specifically, we have collected, analyzed and retained after quality control a total of 2829 ChIP-seq datasets available from public sources, covering a total of 485 TRs with a catalog of 80M peaks. Additionally, the updated database includes new search features for TR names as well as aliases, including cell line names and the ability to navigate the data directly within genome browsers via public track hubs. Finally, full access to this catalog is available online together with a TR binding enrichment analysis tool. ReMap 2018 provides a significant update of the ReMap database, providing an in depth view of the complexity of the regulatory landscape in human.

INTRODUCTION

Transcription factors (TFs), transcriptional coactivators (TCAs) and chromatin-remodeling factors (CRFs) drive gene transcription and the organization of chromatin through DNA binding. TFs specifically bind to DNA sequences (TF binding sites) to activate (activators) or repress (repressors) transcription, TCAs enhance gene transcription by binding to activator TF. While CRFs modify the chromatin architecture to allow DNA access for transcription machinery proteins. In recent years, the development of high-throughput techniques like chromatin immunoprecipitation followed by sequencing (ChIP-seq) (1) has allowed to experimentally obtain genome-wide maps of binding sites across many cell types for a variety of DNA-binding proteins. The popularity of ChIP-seq has led to a deluge of data in current data warehouses (2,3) for TFs, TCAs and CRFs, collectively named transcriptional regulators (TRs). The rapid accumulation of ChIP-seq data in public databases provides a unique and valuable resource for hundreds of TR occupancy maps. There is a strong need to integrate these large-scale datasets to explore the transcriptional regulatory repertoire. Unfortunately, the heterogeneity of the pipelines used to process these data, as well as the variety of underlying formats used, challenge the analysis processes and the underlying detection of TF binding sites (TFBSs). Integrative studies would offer significant insights into the dynamic mechanisms by which a TF selects its binding regions in each cellular environment.

ReMap has been the first large scale integrative initiative to study these data, offering significant insights into the complexity of the human regulatory landscape (4). The ReMap 2015 resource created a large catalog of regulatory regions by compiling the genomic localization of 132 different TRs across 83 different human cell lines and tissue types based on 395 non-ENCODE datasets selected from Gene Expression Omnibus (2) and ArrayExpress (3). This catalog was merged with the ENCODE multi-cell peaks (5), generating a global map of 13M regulatory elements for 237 TRs across multiple cell types. However, since the 2015 publication of ReMap, an even greater number of ChIP-seq assays has been submitted to genomic data repositories.

*To whom correspondence should be addressed. Tel: +33 4 91 82 87 39; Fax: +33 4 91 82 87 01; Email: benoit.ballester@inserm.fr

Here, we introduce the ReMap 2018 update, which includes the integration of 2829 quality controlled ChIP-seq datasets for TFs, TCAs and CRFs. The new ChIP-seq datasets ($n = 1763$, defined as 'Public' for non-ENCODE) as well as the latest ENCODE ChIP-seq data ($n = 1066$) have been mapped to the GRCh38/hg38 human assembly, quality filtered and analyzed with a uniform pipeline. In this update, we propose a unified integration of all public ChIP-seq datasets producing a unique atlas of regulatory regions for 485 TRs across 346 cell types, for a total of 80M DNA binding regions. Each experiment introduced in this release has been assessed and manually curated to ensure correct meta-data annotation. Our ReMap database provides DNA-binding locations for each TR, either for each experiment, at cell line or primary cell level, or at the TR level in a non-redundant fashion across all collected experiments. This update represents a 2-fold increase in the number of DNA-binding proteins, 7-fold in the number of processed datasets, 4-fold in the number of cell lines/tissue types and 6-fold in the number of identified ChIP-seq peaks. While the first version of the ReMap catalog covered 26% (793 Mb) of the human genome, the regulatory search space for ReMap 2018 covers 46% (1.4Gb).

Finally, we give the community access to various options to visualize and browse our catalog, allowing users to navigate and dissect their genomic loci of interest co-occupied by multiple TRs in various cell types. Browsing the ReMap 2018 catalog using the Public Track hub, IGV data sever, Ensembl or UCSC sessions clearly exposes the abundance and intricacy of combinatorial regulation in cellular contexts.

This report presents the extensive data increase and regulatory catalog expansion of ReMap as a result of our large-scale data integration and genome-wide analysis efforts. The manual curation specific to the ReMap initiative offers a unique and unprecedented collection of TR binding regions. These improvements, together with several novel enhancements (search bars, data track displays, format and annotation), constitute a unique atlas of regulatory regions generated by the integration of public resources.

MATERIALS AND METHODS

Available datasets

ChIP-seq datasets were extracted from the Gene Expression Omnibus (GEO) (2), ArrayExpress (AE) (3) and ENCODE (5) databases. For GEO, the query '('chip seq' OR 'chipseq' OR 'chip sequencing') AND 'Genome binding/occupancy profiling by high-throughput sequencing' AND 'homo sapiens'[organism] AND NOT 'ENCODE'[project]' was used to return a list of all potential datasets, which were then manually assessed and curated for further analyses. For ArrayExpress, we used the query (Filtered by organism 'Homo sapiens', experiment type 'dna assay', experiment type 'sequencing assay', AE only 'on') to return datasets not present in GEO. Contrary to other similar databases (chip-atlas http://chip-atlas.org, (6,7)), ReMap meta-data for each experiment are manually curated, annotated with the official gene name from the HUGO Gene Nomenclature Committee (8) (www.genenames.org) and BRENDA Tissue Ontologies (9) for cell lines (www.ebi.ac.uk/ols/ontologies/bto). Datasets involving polymerases (Pol2 and Pol3), and some mutated or fused TFs (e.g. KAP1 N/C terminal mutation, GSE27929) were filtered out. A dataset is defined as a ChIP-seq experiment in a given GEO/AE/ENCODE series (e.g. GSE37345), for a given TF (e.g. FOXA1), and in a particular biological condition (e.g. LNCaP). Datasets were labeled with the concatenation of these three pieces of information (e.g. GSE37345.FOXA1.LNCAP).

A total of 3180 datasets were processed (Supplementary Table S1). Specifically, we analyzed 2020 datasets from GEO (1862) and ArrayExpress (158) repositories (July 2008 to May 2017). We define these non-ENCODE datasets as the 'Public' set, in opposition to ENCODE datasets (1160) (full list of experiments in Supplementary Tables S2 and 3).

ReMap 2015 contained the multi-cell peak calling processed from ENCODE release V3 (August 2013). For the ReMap 2018 update, we re-analyzed, starting from the raw data, all ENCODE ChIP-seq experiments for TFs, transcriptional and chromatin regulators, following the same processing pipeline as the Public set. We retrieved the list of ENCODE data as FASTQ files from the ENCODE portal (https://www.encodeproject.org/) using the following filters: Assay: 'ChIP-seq', Organism: 'Homo sapiens', Target of assay: 'TF', Available data: 'fastq' on 21 June 2016. Metadata information in JSON format and FASTQ files were retrieved using the Python *requests* module. We processed 1160 datasets associated to 161 TRs and 87 cell lines. We removed 2 TRs (POLR2A, POLR3G), and renamed TR aliases into official HGNC identifiers (e.g. p65 into RELA, see Supplementary Table) leading to a final list of 279 TRs from ENCODE.

ChIP-seq processing

Both ENCODE and Public datasets were uniformly processed and analyzed. Bowtie 2 (version 2.2.9) (10) with options –end-to-end –sensitive was used to align all reads on the human genome (GRCh38/hg38 assembly). For Public datasets, adapters were removed using TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), trimming reads up to 30 bp. Polymerase chain reaction duplicates were removed from the alignments with samtools rmdup (11). For the ENCODE data, the adapter trimming step was not employed, as this data already passed certain quality assessment steps (https://www.encodeproject.org/data-standards/). TR binding regions were identified using the MACS2 peak-calling tool (version 2.1.1.2) (12) in order to follow ENCODE ChIP-seq guidelines (13), with stringent thresholds (MACS2 default thresholds, *P*-value: 1e-5). Input datasets were used when available. All peak-calling files are available to download. Among the 80M peaks identified, 99.5% of peaks (79 753 407) were below 1.5 kb in size (mean size: 286 bp, median size: 231 bp) and only 376 017 peaks were above 1.5 kb in size (mean size: 2209 bp, median size: 1859 bp).

Quality assessment

As raw data are obtained from various sources, under different experimental conditions and platforms, data quality differs across experiments. Since the ReMap 2015 release, our ChIP-seq pipeline assesses the quality of all

datasets, unlike similar databases (chip-atlas http://chip-atlas.org, (6,7)), (Supplementary Table S4). We compute a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE consortium (13) (Supplementary Figure S1). Descriptions of the ENCODE quality coefficients can be found on the UCSC Genome portal (http://genome.ucsc.edu/ENCODE/qualityMetrics.html). Our pipeline computes the normalized strand cross-correlation coefficient (NSC) as a ratio between the maximal fragment-length cross-correlation value and the background cross-correlation value, and the relative strand cross-correlation coefficient (RSC), as a ratio between the fragment-length cross-correlation and the read-length cross-correlation. The same methods and quality cutoffs were applied as in ReMap 2015 (4). Datasets not passing the QC were not included in the catalog of peaks available for download (http://remap.cisreg.eu).

DNA constraint scores

We provide the conservation profiles at the nucleotide level for each of the 485 TRs present in our catalog. We assessed the DNA constraint for each base pair by considering ±1 kb around the summit of each non-redundant peak (see below). Genomic Evolutionary Rate Profiling scores (GERP) were used to calculate the conservation of each nucleotide in a multi-species alignment (14). The computed GERP scores were obtained from the 24-way amniota vertebrates Pecan (15) multi-species alignment, and extracted from the Ensembl Compara database release v89 (16).

Genome coverage, non-redundant peak sets and CRMs

Genome coverages were computed using the BedTools suite (17) (version 2.17.0) using the 'genomecov' function with the option -max 2 that combines all positions with a depth ≥2 binding locations. Full details of the ReMap 2015 and 2018 genome coverage are available in Supplementary Table S5. ReMap also provides a catalog of discrete, non-redundant binding regions for each TR, a specificity not found in other databases (chip-atlas http://chip-atlas.org, (6,18)). We used BedTools to merge overlapping peaks (with at least 1 bp overlap) identified in different datasets for the same TR. The summit of the resulting peaks was defined as the average position of the summits of the merged peaks. Those peaks made of at least two or more peaks for a given factor are defined as non-redundant peaks. We observed a mean variation of 77 bp between the summits of the non-redundant peaks and the individual peak summits (Supplementary Figure S2). Similarly, to obtain the *cis*-regulatory modules (CRMs) in the genome, overlapping peaks of all TRs were merged using BedTools. Regions bound by several TRs are called CRMs, whereas regions bound by only one TR are labeled as singletons.

Roadmap human epigenome annotations

Two sets of chromatin accessibility data were used to better characterize the ReMap atlas. We employed BedTools for overlap analyses allowing a minimum of 10% overlap. The NIH Roadmap Epigenomics Mapping Consortium (19) data were downloaded from the roadmap data portal (http://egg2.wustl.edu/roadmap). Delineation of DNaseI-accessible regulatory regions were accessed from http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation. BED files with coordinates of each region type for each epigenome separately are available for 81 232 promoter regions (1.44% of genome), 2 328 936 putative enhancer regions (12.63% of genome) and 129 960 dyadic promoter/enhancer regions (0.99% of genome). The core 15-state model of chromatin combinatorial interactions between different chromatin marks was downloaded from http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Chromatin state definitions and abbreviations are: 1 Active TSS (TssA), 2 Flanking active TSS (TssAFlnk), 3 Transcr. at gene 5′ and 3′(TxFlnk), 4 Strong transcription (Tx), 5 Weak transcription (TxWk), 6 Genic enhancers (EnhG), 7 Enhancers (Enh), 8 ZNF genes + repeats (ZNF/Rpts), 9 Heterochromatin (Het), 10 Bivalent/poised TSS (TssBiv), 11 Flanking bivalent TSS/Enh (BivFlnk), 12 Bivalent enhancer (EnhBiv), 13 Repressed Polycomb (ReprPC), 14 Weak repressed Polycomb (ReprPCWk) and 15 Quiescent/low (Quies).

DATA COLLECTION AND CONTENT

Integration of data sources

The 2018 release of the ReMap database reflects significant advances in the number of binding regions, the number of TFs, transcriptional co-activators, chromatin regulators and overall the total number of datasets integrated in our catalog. We initially selected, processed and analyzed 3180 ChIP-seq datasets against TRs from GEO, AE and ENCODE. To ensure consistency and comparability, all datasets were processed from raw data, through our uniform ChIP-seq workflow that included read filtering, read mapping, peak calling and quality assessment based on ENCODE quality criterions. As the quality of ChIP-seq experiments vary significantly (20,21), we incorporated a critical data quality filtering step in our pipeline—not implemented in other databases (chip-atlas http://chip-atlas.org (6,7,18)). Specifically, we considered four quality metrics, two metrics independent of peak calling for assessing signal-to-noise ratios in a ChIP-seq experiment and two metrics based on peak properties. Following ENCODE ChIP-seq guidelines and practices (13), we used the NSC and the RSC (see 'Materials and Methods' section). Further, we used the FRiP and the number of peaks in the dataset (see 'Materials and Methods' section). After applying our quality filters based on these four ChIP-seq metrics we retained 2829 datasets (89%): 1763 datasets from GEO and ArrayExpress and 1066 from ENCODE (Figure 1A and Supplementary Figure S1). The significant increase of data is spread across almost all TFs when compared to ReMap 2015 (Figure 1B). Nevertheless, we observe TFs (e.g. AR, ESR1, FOXA1) and CRFs (e.g. BRD4, EZH2) displaying a larger data growth than other DNA-binding proteins. The majority of TRs show additional datasets integrated in ReMap 2018 (Figure 1B, dark blue bars).

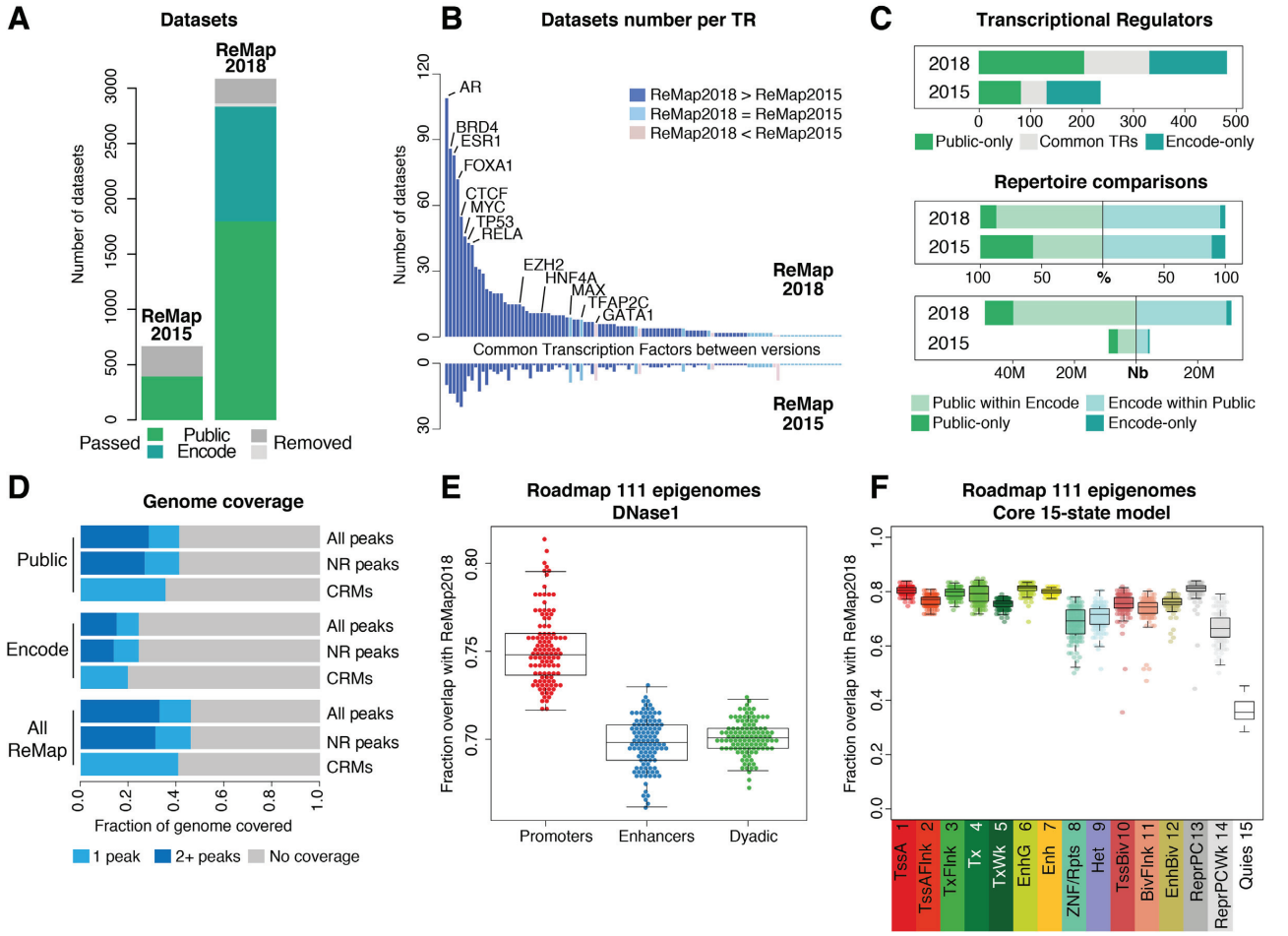

Figure 1. Overview of the ReMap database expansion. (**A**) Analyzed datasets growth in ReMap 2018 compared to ReMap 2015. (**B**) Evolution of the number of datasets per TRs, ranked across common between both ReMap versions. (**C**) Common TRs between Public and ENCODE sets of data (gray). Direct comparison of Public and ENCODE repertoire, defined as percentages (%), and as number (Nb) of peaks. (**D**) Genome coverage fraction of each ReMap dataset (NR non-redundant, CRM Cis Regulatory Modules). (**E**) Comparison of DNase I-accessible regulatory regions against the ReMap 2018, regions from the Roadmap Epigenomics Consortium defining promoter-only, enhancer-only or enhancer–promoter alternating states (Dyadic). Each dot represents the fraction overlap with ReMap 2018 for one of the 111 epigenomes. (**F**) Comparison of the Roadmap Epigenomics Consortium chromatin states annotations against the ReMap 2018 catalog, using the Core 15 chromatin states model, and a minimum overlap of 50% between regions. Each dot represents the overlap for one of the 111 epigenomes. Chromatin state definitions and abbreviations are as follows; 1 Active TSS (TssA), 2 Flanking active TSS (TssAFlnk), 3 Transcr. at gene 5′ and 3′(TxFlnk), 4 Strong transcription (Tx), 5 Weak transcription (TxWk), 6 Genic enhancers (EnhG), 7 Enhancers (Enh), 8 ZNF genes + repeats (ZNF/Rpts), 9 Heterochromatin (Het), 10 Bivalent/poised TSS (TssBiv), 11 Flanking bivalent TSS/Enh (BivFlnk), 12 Bivalent enhancer (EnhBiv), 13 Repressed Polycomb (ReprPC), 14 Weak repressed Polycomb (ReprPCWk), 15 Quiescent/low (Quies).

Regulatory catalog expansion

With all ChIP-seq data uniformly processed, the ReMap 2018 catalog displays ENCODE data down to the cell line and dataset level rather than the simpler multi-cell analysis provided by ENCODE DCC used in ReMap 2015. Our analyses produced 48 693 300 peaks for the Public-only (non-ENCODE) set across 331 TRs and 31 436 124 peaks for the ENCODE set across 279 TRs, leading to a final ReMap regulatory atlas of 80 129 705 peaks generated from 485 TRs (Figure 1C). We found 125 TRs common to the two sets, 154 proteins specific to ENCODE and 206 specific to the Public catalog (Figure 1C). We also found that 839 400 CRMs are shared between both catalogs. Taken separately, the ENCODE peaks overlaps by 96% the Public regions, and 87% of the Public peaks overlap ENCODE regions (Figure 1C). It suggests that merging both Public and ENCODE sets complements the annotation of DNA-bound regions, as it increases the number of regulatory regions in our atlas, hence improving the annotation of DNA-bound elements in the human genome (Figures 1C and 2).

Indeed, about 13% (405 Mb) of the human genome is covered by at least one feature only from the entire ReMap catalog and 33% (1.02 Gb) are covered by two or more features (Figure 1D and Supplementary Table S4). The Public-only and ENCODE-only sets cover the genome by two or more peaks by 28 and 15% respectively. The observed differences can be explained by the wide spectrum of cell lines and treatments included in the Public set (300 cell lines) compared to the ENCODE set (86 cell lines). As a comparison, the

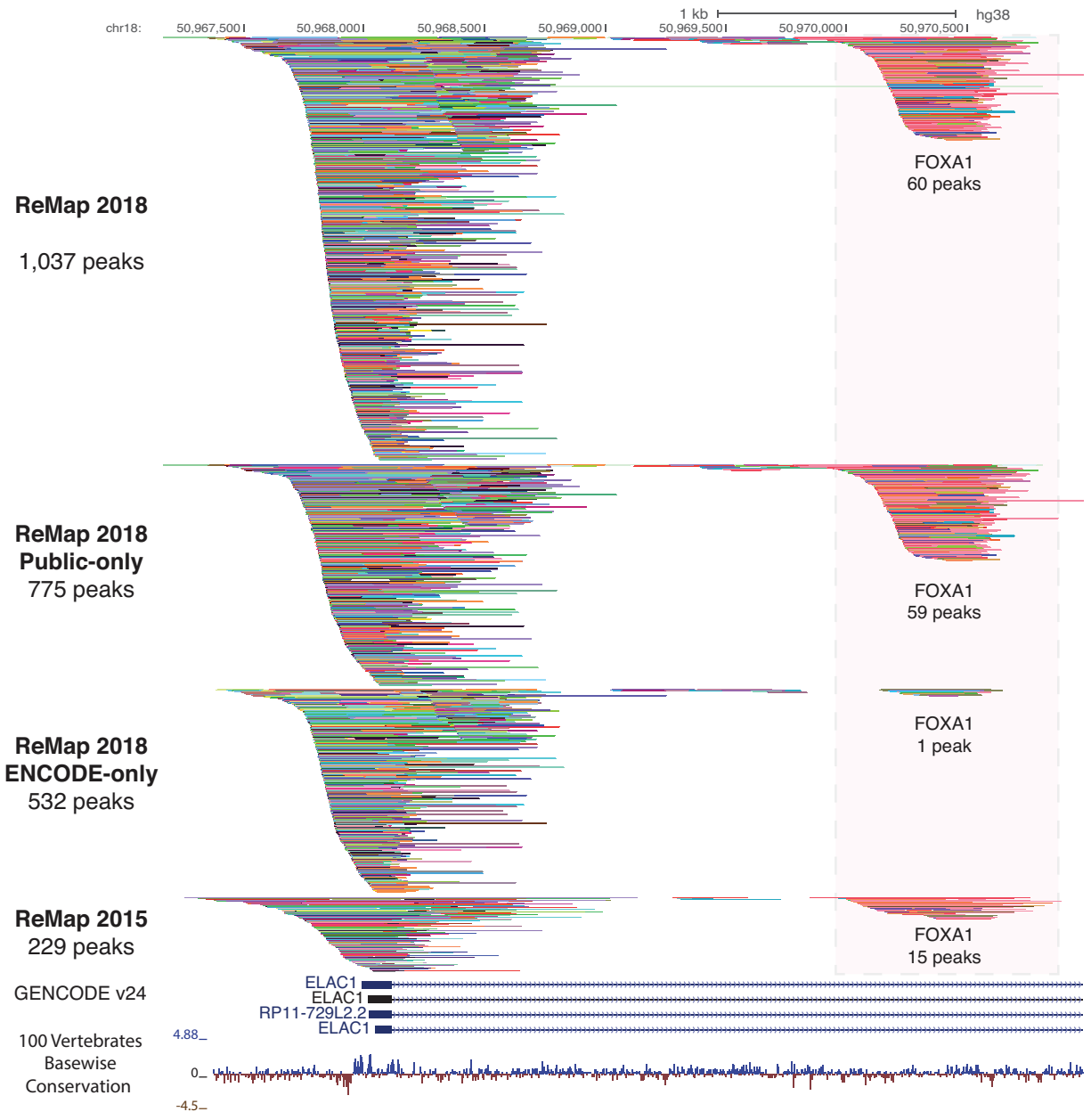

Figure 2. ReMap ChIP-seq binding pattern of 2829 datasets. A genome browser example of the ChIP-seq binding peak depth of the ReMap 2018 catalog compared to ReMap 2015 at the vicinity of the ELAC1 promoter (chr18:50,967,094-50,970,983). The tracks and peaks displayed are compacted to thin lines so the depth of ReMap 2018 bindings can be compared to ReMap 2015. A full and un-compacted screenshot is available as Supplementary Figures S2 and 3. On this location the ReMap 2018 catalog contains 1307 peaks, whereas the ReMap 2015 contains 229 peaks (ReMap 2015 lifted to GRCh38/hg38 assembly). The following genome tracks correspond to the GENCODE v24 Comprehensive Transcript Set and the 100 vertebrates base-wise conservation showing sites predicted to be conserved (positive scores in blue), and sites predicted to be fast-evolving (negative scores in red). A detailed view of the redundant peaks for a FOXA1 site is available in Figure 3.

ReMap 2015 catalog covered 10% (321 Mb) of the genome with one feature only, and 15% (471 Mb) with at least two or more features. Between the two ReMap versions, we observe that the fraction of the human genome covered by one feature remains extremely stable (+84 Mb from 2015 to 2018), whereas the fraction covered by two or more regulatory features increases by 545 Mb. With ReMap 2018, we increase the range of the regulatory space, and provide binding regions for similar TRs at a greater depth, revealing tight and dense co-localization sites (Figures 2 and 3).

Overlap with *cis*-regulatory genomic regions

Using the NIH Roadmap 111 epigenomes analyses, we asked whether the DNase I defined regions as well as the core 15 chromatin states model would better characterize

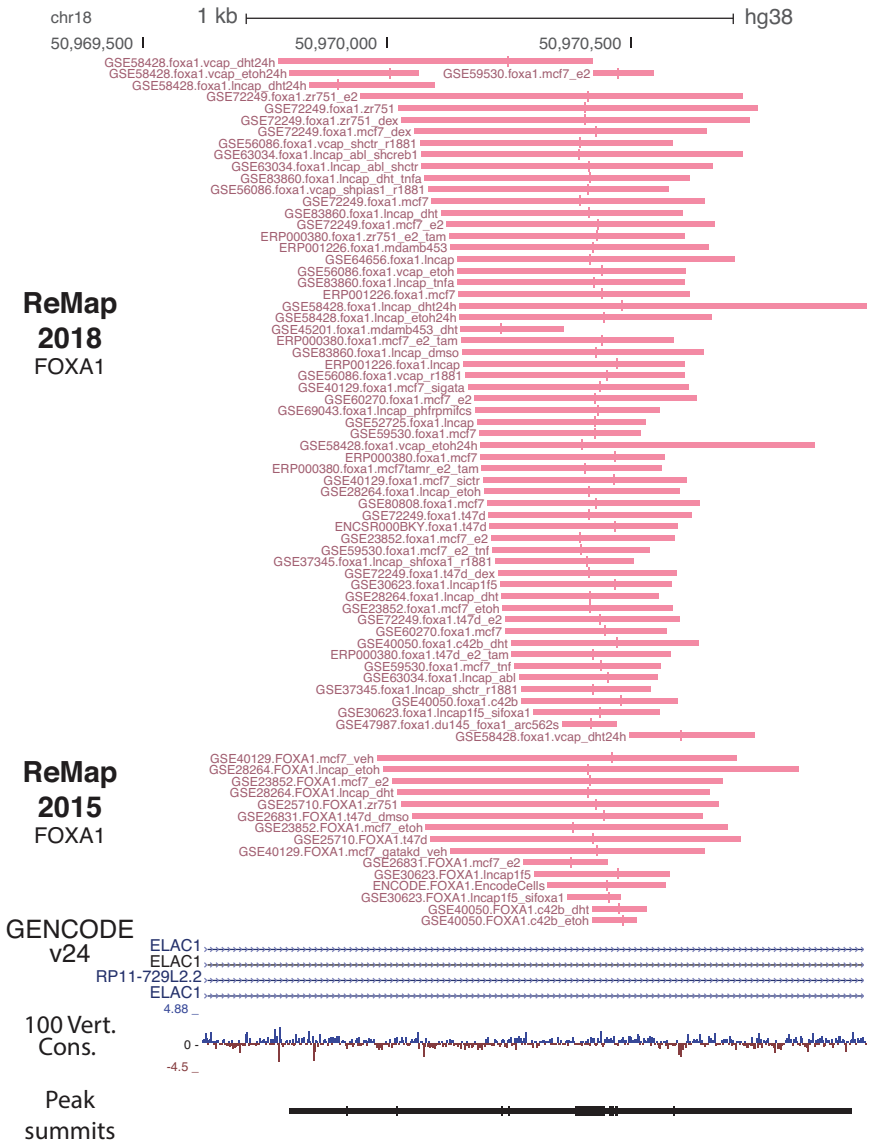

Figure 3. FOXA1 ChIP-seq peaks pattern evolution across ReMap versions. Detailed view of the FOXA1 peaks present in ReMap 2018 (60 peaks) compared to the FOXA1 peaks in ReMap 2015 (15 peaks) found at the genomic location chr18:50,969,638-50,970,931 in the first intron of the ELAC1 gene. Those 60 FOXA1 peaks are derived from GEO, ArrayExpress and ENCODE ChIP-seq across multiple cell lines. Interestingly, it can be noted that the peak summits (vertical bars) of each peak aggregate closely from each other, defining precisely the DNA binding location. Those aggregations of the FOXA1 summits are an illustration of what is globally observed for peaks of different TFs across the genome.

the ReMap atlas (Figure 1E and F). The Roadmap consortium defined a total of 3.5M DNase I-accessible regulatory regions by merging all DNase I hypersensitive regions across epigenomes, which were then annotated using the core 15-state model focusing on chromatin states for promoters, enhancers and dyadic (promoter + enhancer) ambiguous regions (see 'Materials and Methods' section). Among these three categories, the ReMap atlas could recapitulate on average 75.2% of the Roadmap promoter regions, 69.8% of enhancer regions and 70.1% of dyadic regions from the Roadmap annotation. Looking at the core 15-state model, we observe that the ReMap catalog recapitulates more than 70% of the regions covered by each state (Enhancer Genic (81%), Enhancer (80%) and TSS active (80%) states) with the exception of quiesent state (36%). Taken together, these results suggest that some promoter and enhancer activities from Roadmap may be cell type specific, as about 20–30% of those regions seem specific to Roadmap consortium cells. The ReMap initiative results from a large-scale integration of hundreds of diverse cell types, and leads to a regulatory landscape illustrating the large regulatory circuitry of those cells. The constant integration of novel data will allow for a greater definition of the regulatory space across the genome.

Large regulatory atlas

The ReMap database provides a large view of a unique regulatory landscape constituted by 80M binding regions forming 1.6M CRMs. The genomic organization of our occupancy map reveals dense co-localizations of sites forming tight clusters of heterogeneous binding sites with variable TRs complexity (Figure 2). For instance, the regulatory regions observed in the vicinity of the ELAC1 promoter illustrate the ReMap 2018 expansion ($n = 1037$ peaks). It highlights how the regulatory repertoire can be complemented by merging both Public and ENCODE sources. We observe a large cluster of peaks at the ELAC1 promoter followed by two clusters at +500 bp and 1 kb from the transcription start site. The third cluster exemplifies how integrating data from different sources improves genome annotations, as few peaks are available from ENCODE at this location. Additionally, this cluster was detailed in our previous ReMap publication (4) and consisted of 15 FOXA1 ChIP-seq peaks from different cells, antibodies, and laboratories (Figure 3). In this update, we consolidate this FOXA1 binding location with 60 peaks. The summit of each peak is represented by vertical bars aggregated closely from each other, providing an information about the putative location of the DNA binding site. The clustering of FOXA1 peaks and summits illustrates our genome-wide repertoire. However, this FOXA1 example shows overlapping sites derived from various experimental conditions, and therefore does not reflect the total number of discrete binding regions across the genome. To address redundancy between datasets, we merged binding regions for the same TR, resulting in a catalog of 35.5M peaks for all TRs combined. These merged peaks, defined as non-redundant peaks, are made of at least two or more peaks and singletons for a given factor across all experiments, and are available for download from the ReMap website. The TRs with the most merged binding regions across cell types are AR, FOXA1, CTCF and ESR1 (Supplementary Figure S6). These results indicate that most bindings are shared across different ChIP-seq experiments, either for similar or for different cell types. Overall, our ReMap update provides a unique opportunity to identify complex regulatory architectures containing multiple bound regions. We observe that by adding more cell lines, more experiments and more DNA-binding proteins, we increased the genome regulatory space and its depth (Figure 2), but also refined the current annotations of bound regions (Figure 3).

IMPLEMENTATION AND PUBLIC ACCESS

Web display

ReMap provides free public access to all data at http://remap.cisreg.eu. The results presented here provide an informative annotation for 80M ChIP-seq peaks coming from public data sources, which are derived from 485 TRs across 346 diverse cell lines. This catalog provides an unparalleled resource for dissecting site-specific TF bindings (e.g. FOXA1 in Figures 2 and 3) or genome-wide binding analyses. The ReMap web interface displays informations about the integrated TRs (description, classification, external references to Ensembl gene IDs, UniProt, RefSeq, WikiGene, JASPAR, FactorBook, TF Encyclopedia and other resources), peaks, and datasets (quality assessment, read mapping and peak calling statistics, conservation score under peaks). The interface provides a simple 'Dynamic Search' available from the TRs, Cell lines and Download pages and is the entry point for users to search for specific data. The search form allows users to narrow their searches based on gene aliases, dataset names or IDs, cell line names or ontology. For example, entering 'Oct' as search term in the 'Dynamic Search' returns three TFs POU2F2, POU2F1, POU5F1 having various 'OCT' aliases. Additionally, one could use the search box in the Cell or Download page to search for specific cell types containing the 'Colo' term for instance, or 'GSE66218' for a precise experiment from the Download page. Moreover, we provide a tool that allows the annotation of genomic regions provided by users. Those regions are compared against the ReMap catalog returning statistical enrichments of TR bindings present within user-provided input regions compared to random expectations. It allows for the study of over-represented TR binding regions.

Browsing and downloading data

Updates made in ReMap 2018 reflect significant improvement in the variety of genome navigation options. As the ReMap 2015 UCSC session was popular, we now provide more data navigation alternatives. The content of the ReMap database can be browsed through four options: (i) across two mirror sites of the UCSC Genome Browser (22) where a public session has been created (Figure 2 and Supplementary Figure S3), (ii) across three Ensembl Genome Browser mirrors (16) (Supplementary Figure S4), (iii) using the ReMap public track hub (23) or (iv) using the IGV data server (24) (Supplementary Figure S5). For each option, we provide four tracks, the full ReMap catalog containing all peaks, the Public-only peaks, the ENCODE-only peaks and a track containing only peaks above 1.5 kb. As the ReMap catalog expanded, it is crucial to allow visual exploration of regulatory regions across different platforms combined with public or user-specific genome-wide annotations. In addition, the entire ReMap 2018 catalog, as well as the Public-specific or ENCODE-specific peaks, have been compiled into BED files allowing further interpretations and computational analyses.

FUTURE DIRECTIONS

Next-generation sequencing technologies are playing a key role in improving our understanding of regulatory genomics. As ChIP-seq technology is applied to a broader set of cell lines, tissues and conditions, we will continuously maintain and update the database. In the near future, we propose on adding to the ReMap portfolio different peak-caller analyses to further consolidate the peak repertoire. Also, we aim to provide direct access to aligned reads through a FTP server, allowing users to upload and navigate aligned raw data of their choice. We plan on releasing a Bioconductor R-package for genomic region enrichment analyses for large genomic catalogs such as ReMap, which will be replacing our current web enrichment tool. In

the coming year, we would like to provide a Bioconductor R-package to search and download ReMap data for a specific study, to get genomic range objects, raw counts and/or metadata used for a specific study. Overall, determining the best approach to curate and annotate ChIP-seq data with a very broad level of submitted annotations and metadata into a simple-to-use, easy-to-analyze and up-to-date system will become a focus for the ReMap project.

CONCLUSION

The 2018 release of ReMap maintains the long-term focus of providing the research community with the largest catalog of high-quality regulatory regions by integrating all available ChIP-seq data from DNA-binding assays. The usefulness of ReMap is exemplified by the last release of the JASPAR database (25), for which ReMap ChIP-seq peaks were used to derive 45 new TF binding profiles that were incorporated in the 2018 release of the vertebrate CORE collection (Khan *et al*. 2018), providing a 9% increase from JASPAR 2016 (26) by solely relying on the ReMap 2018 catalog. Although new datasets are constantly added to repositories, we believe that our ReMap atlas will help in better understanding the regulation processes in human. In this update, we have (i) widely expanded the collection of datasets curated and analyzed from public sources with now 485 TFs, transcriptional co-activators and chromatin regulators; (ii) uniformly processed and integrated the ENCODE ChIP-seq data; (iii) enhanced the website usability by allowing dynamic search of TRs, aliases, cell lines and experiments, (iv) expanded the genome browsing experience by integrating ReMap in all UCSC and Ensembl Genome Browsers mirror sites and provided a Track Hub for data integration in other platforms; (v) improved the capacity to download all ReMap files in bulk or individually.

AVAILABILITY

ReMap 2018 can be accessed through a web interface at http://remap.cisreg.eu. Downloads are available in BED format for the entire ReMap catalog, the Public-only peaks, the ENCODE-only peaks, and in FASTA and BED formats for each TR. In addition, UCSC and Ensembl Genome Browsers users can navigate ReMap across their mirror sites, use ReMap in UCSC public sessions, or use the public track hub. Finally, Integrative Genome Browser (IGV) users have the option of loading an IGV optimized dataset directly in the application.

FEEDBACK

The ReMap team welcomes your feedback on the catalog, use of the website and use of the downloadable files. Please contact us at benoit.ballester@inserm.fr or remap@cisreg.eu for development requests. We thank our users for their feedback to make ReMap useful for the community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

French Ministry of Higher Education and Research (MESR) PhD Fellowship (to J.C.); Norwegian Research Council (to A.M., M.G.); Helse Sør-Øst (to A.M., M.G.); University of Oslo (to A.M., M.G.). Funding for open access charge: Institut national de la santé et de la recherche médicale (INSERM).

Conflict of interest statement. None declared.

REFERENCES

1. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
2. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
3. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
4. 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 Res.*, **43**, e27.
5. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
6. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
7. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
8. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
9. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
10. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
11. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
12. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
13. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
14. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
15. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
16. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
17. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
18. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.

19. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
20. Mendoza-Parra,M.-A., Saleem,M.-A.M., Blum,M., Cholley,P.-E. and Gronemeyer,H. (2016) NGS-QC generator: a quality control system for ChIP-Seq and related deep sequencing-generated datasets. *Methods Mol. Biol.*, **1418**, 243–265.
21. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, **4**, 209–223.
22. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
23. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
24. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
25. Khan,A., Fornes,O., Stigliani,A., Gheorghe,F.N., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles andits web framework. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1126.
26. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.

JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Khan, A.[†], Fornes, O.[†], Stigliani, A.[†], **Gheorghe, M.**, Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A.[*], Vandepoele, K., Lenhard, B.[*], Ballester, B., Wasserman, W. W.[*,] Parcy, F., and Mathelier, A.[*]

(2018), *Nucleic Acids Research*, 46(D1):D260–D266.

II

D260–D266 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 13 November 2017*
doi: 10.1093/nar/gkx1126

JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Aziz Khan[1,†], Oriol Fornes[2,†], Arnaud Stigliani[3,†], Marius Gheorghe[1], Jaime A. Castro-Mondragon[1], Robin van der Lee[2], Adrien Bessy[3], Jeanne Chèneby[4,5], Shubhada R. Kulkarni[6,7,8], Ge Tan[9,10], Damir Baranasic[9,10], David J. Arenillas[2], Albin Sandelin[11,*], Klaas Vandepoele[6,7,8], Boris Lenhard[9,10,12,*], Benoît Ballester[4,5], Wyeth W. Wasserman[2,*], François Parcy[3] and Anthony Mathelier[1,13,*]

[1]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [2]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 28th Ave W, Vancouver, BC V5Z 4H4, Canada, [3]University of Grenoble Alpes, CNRS, CEA, INRA, BIG-LPCV, 38000 Grenoble, France, [4]INSERM, UMR1090 TAGC, Marseille, F-13288, France, [5]Aix-Marseille Université, UMR1090 TAGC, Marseille, F-13288, France, [6]Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium, [7]VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium, [8]Bioinformatics Institute Ghent, Ghent University, Technologiepark 927, 9052 Ghent, Belgium, [9]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK, [10]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W12 0NN, UK, [11]The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, DK2200 Copenhagen N, Denmark, [12]Sars International Centre for Marine Molecular Biology, University of Bergen, N-5008 Bergen, Norway and [13]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 25, 2017; Revised October 17, 2017; Editorial Decision October 18, 2017; Accepted October 27, 2017

ABSTRACT

JASPAR (http://jaspar.genereg.net) is an open-access database of curated, non-redundant transcription factor (TF)-binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups. In the 2018 release of JASPAR, the CORE collection has been expanded with 322 new PFMs (60 for vertebrates and 262 for plants) and 33 PFMs were updated (24 for vertebrates, 8 for plants and 1 for insects). These new profiles represent a 30% expansion compared to the 2016 release. In addition, we have introduced 316 TFFMs (95 for vertebrates, 218 for plants and 3 for insects). This release incorporates clusters of similar PFMs in each taxon and each TF class per taxon. The JASPAR 2018 CORE vertebrate collection of PFMs was used to predict TF-binding sites in the human genome. The predictions are made available to the scientific community through a UCSC Genome Browser track data hub. Finally, this update comes with a new web framework with an interactive and responsive user-interface, along with new features. All the underlying data can be retrieved programmatically using a RESTful API and through the JASPAR 2018 R/Bioconductor package.

INTRODUCTION

Transcription factors (TFs) are sequence-specific DNA-binding proteins involved in the transcriptional regulation of gene expression (1). TFs bind to DNA through their DNA-binding domain(s) (DBDs), which are used for TF classification (2). DNA regions at which TFs bind are defined as TF-binding sites (TFBSs) and can be identified

*To whom correspondence should be addressed. Tel: +47 228 40 561; Email: anthony.mathelier@ncmm.uio.no
Correspondence may also be addressed to Albin Sandelin. Tel: +45 2245 6668; Fax: +45 3532 2128; Email: albin@binf.ku.dk
Correspondence may also be addressed to Boris Lenhard. Tel: +44 20 8383 8353; Email: b.lenhard@imperial.ac.uk
Correspondence may also be addressed to Wyeth W. Wasserman. Tel: +1 604 875 3812; Fax: +1 604 875 3840; Email: wyeth@cmmt.ubc.ca
†These authors contributed equally to the paper as first authors.

Table 1. Overview of the growth of the number of PFMs in the JASPAR 2018 CORE collection compared to the JASPAR 2016 CORE collection

Taxonomic group	Non-redundant PFMs in JASPAR 2016	New non-redundant PFMs in JASPAR 2018	Updated PFMs in JASPAR 2018	Total PFMs (non-redundant) in JASPAR 2018	Total PFMs (all versions) in JASPAR 2018
Vertebrates	519	60	24	579	719
Plants	227	262	8	489	501
Insects	133	0	1	133	140
Nematodes	26	0	0	26	26
Fungi	176	0	0	176	177
Urochordata	1	0	0	1	1
Total	**1082**	**322**	**33**	**1404**	**1564**

in vivo by methods such as chromatin immunoprecipitation (ChIP) or *in vitro* by methods based on binding of large pools of DNA fragments (e.g. Systematic evolution of ligands by exponential enrichment (SELEX) or protein-binding microarrays (PBM)) (reviewed in (3)). Analysis of TFBSs for a given TF provides models for its specific DNA-binding preferences, which in turn can be used to predict TFBSs in DNA sequences (4). This is important as experiments can only identify TFBSs that are bound in the cell and state analyzed.

The computational representation of TF binding preferences has evolved over the years, from simple consensus sequences to position frequency matrices (PFMs). A PFM summarizes experimentally determined DNA sequences bound by an individual TF by counting the number of occurrences of each nucleotide at each position within aligned TFBSs. Such matrices can be converted into position weight matrices (PWMs), also known as position-specific scoring matrices, which are probabilistic models that can be used to predict TFBSs in DNA sequences (reviewed in (5)).

PFMs/PWMs have been the standard models for describing binding preferences of TFs for many years. The JASPAR database is among the most popular and longest maintained databases for PFMs and a standard resource in the field. In particular, the JASPAR CORE collection of the database, which is the most used, stores non-redundant TF binding profiles, providing a single representative DNA binding model per TF decided by expert curators. Exceptionally, multiple TF-binding profiles are associated to a TF when it is known to interact with DNA with multiple distinct sequence preferences, due to differential splicing for example (6,7). JASPAR was created and persists under three guiding principles: (i) unrestricted open-access; (ii) manual curation and non-redundancy of profiles; and (iii) ease-of-use. The 2016 release of the JASPAR CORE collection stored 1082 non-redundant and manually curated TF-binding profiles as PFMs for TFs from six different taxonomic groups (vertebrates, plants, insects, nematodes, fungi and urochordata) (8).

An intrinsic limitation to PFMs/PWMs is that they ignore inter-nucleotide dependencies within TFBSs (9–13). TF–DNA interaction data derived from next-generation sequencing assays has improved the computational modeling of TF binding (14–19). For example, the TF flexible models (TFFMs) (14), based on first-order hidden Markov models, capture dinucleotide dependencies within TFBSs and were introduced in the 2016 release of the JASPAR database.

In this report, we describe the seventh release of JASPAR (8,20–24), which comes with a major expansion and update of the CORE collection of TF-binding profiles as PFMs and TFFMs. These models have been manually assessed by expert curators who reconciled recent high-throughput data with available literature and linked the models to the classification of their TF DBDs from TFClass (2). The CORE collection expansion is supported by a range of new functionalities and resources, including PFM clustering, genome-wide UCSC tracks of predicted TFBSs and fully redesigned user and programming interfaces.

EXPANSION AND UPDATE OF THE JASPAR CORE COLLECTION

In this 2018 release of the JASPAR database, we added 355 new PFMs for TFs from plants (270), vertebrates (84) and insects (1) to the JASPAR CORE collection (Table 1). Specifically, we added 322 PFMs (262 for plants, a 118% increase and 60 for vertebrates, an 11% increase) for TF monomers and dimers that were not previously present in JASPAR and updated 33 (8 in plants, 3% of JASPAR 2016, 24 in vertebrates, 5% of JASPAR 2016 and 1 in insects). The PFMs were manually curated using independent external literature supporting the candidate TF-binding preferences, as previously described in (23). The curated PFMs were derived from ChIP-seq (from ReMap (25) and (26–30)), DAP-seq (31), SMiLE-seq (32), PBM (33) and HT-SELEX (34) experiments. The JASPAR CORE collection now includes 1404 non-redundant PFMs (579 for vertebrates, 489 for plants, 176 for fungi, 133 for insects, 26 for nematodes and 1 for urochordata) (Table 1).

We continued with the incorporation of TFFM models, initiated in JASPAR 2016. In this release of JASPAR, we introduced 316 new TFFMs for vertebrates (95), plants (218) and *Drosophila* (3), which represents a 243% increase in the number of non-redundant TFFMs stored in the JASPAR CORE collection.

HIERARCHICAL CLUSTERING OF TF-BINDING PROFILES

While the non-redundancy of binding profiles is one of the guiding principles of JASPAR, TFs with similar DBDs often have similar binding preferences (35,36). To facilitate the exploration of similar profiles in the JASPAR CORE collection, we performed hierarchical clustering of PFMs using the RSAT matrix-clustering tool (37). Specifically, the tool was applied to PFMs in each taxon independently as

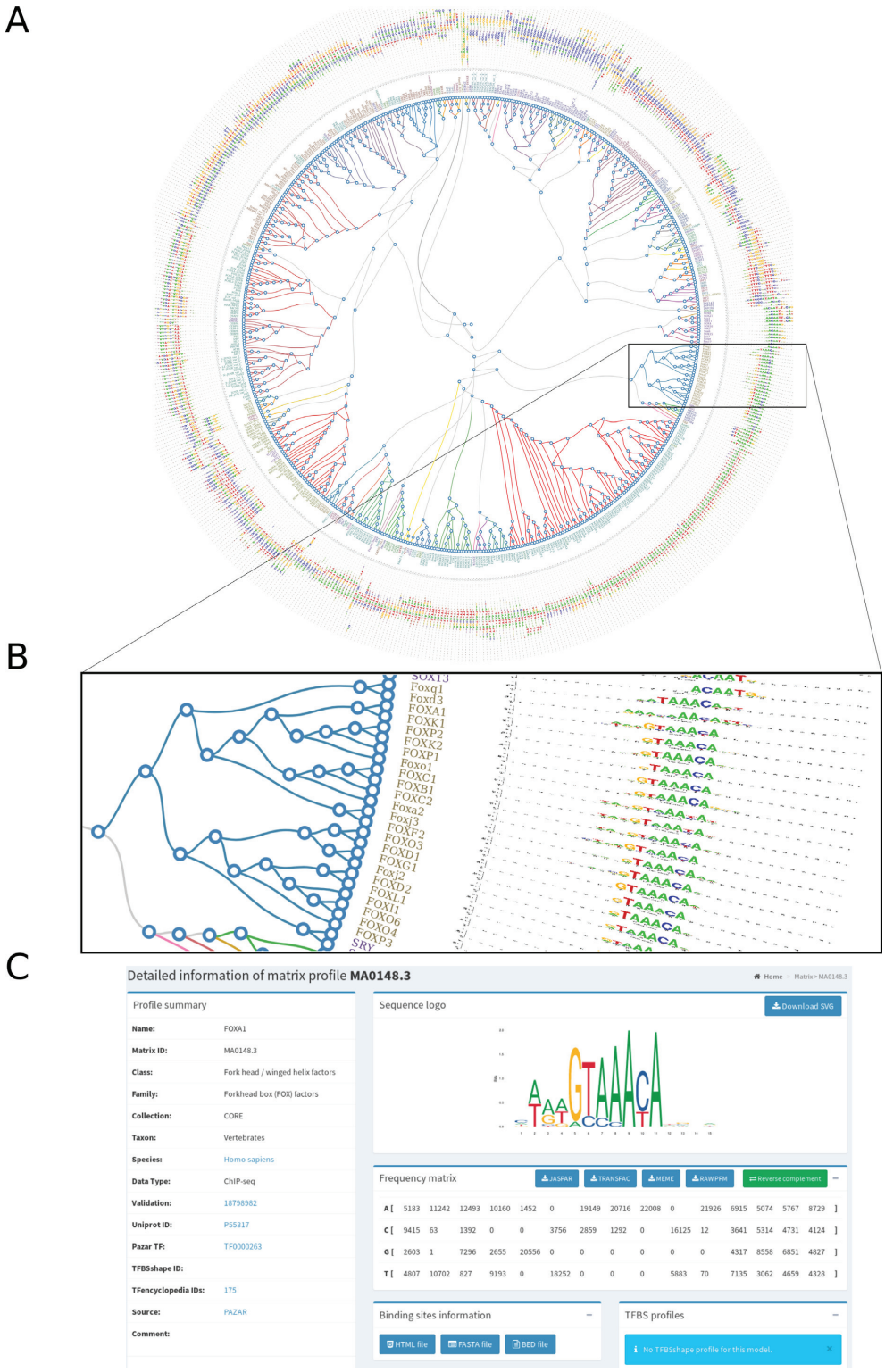

Figure 1. JASPAR PFM clustering. (**A**) Radial tree representing the clusterization of the JASPAR CORE vertebrate PFMs. (**B**) Zoom in view of the radial tree where the predicted clusters are highlighted at the branches and the TF classes are indicated with different colors at the leaves. (**C**) Clicking on a leaf in the radial tree will open a link to the corresponding motif description page on the JASPAR website (the MA0148.3 profile associated to FOXA1 is provided here as an example).

Figure 2. Overview of the JASPAR 2018 new web interface with interactive searching activity. (**A**) A quick and detailed search feature on the homepage. (**B**) A responsive table lists the searched profile(s), which can be further selected and added to the cart listed on the right panel for users to perform their own analyses. (**C**) A detailed page for the GATA3 matrix profile, which is divided into sub-panels including the profile summary, sequence logo, PFM, TF-binding information, external links, version information, ChIP-seq centrality, TFFM and other details. (**D**) The PFM for the GATA3 profile (MA0037.2) is downloaded in MEME format using the RESTful API.

well as in each TF class per taxon. The clustering results are provided as radial trees (Figure 1), which can further be explored through dedicated web pages (http://jaspar.genereg.net/matrix-clusters).

JASPAR UCSC TRACKS FOR GENOME-WIDE ANALYSES OF TFBSs

A typical application of JASPAR TF-binding profiles in gene regulation studies is the identification of TFBSs in DNA sequences for further analyses. Although, we recognize that genome-wide PWM-based predictions contain a high number false positives, we believe that they are a powerful resource for the research community in the context of a variety of genomic information, including transcription start site activity, DNA accessibility, histone marks, evolutionary conservation or *in vivo* TF binding (38–46). To facilitate such integrative analyses, we have performed TFBS predictions on the human genome using the JASPAR CORE vertebrate PFMs (see Supplementary Data for details on the computation). The predicted TFBSs are publicly available through a UCSC Genome Browser data hub (47) containing tracks for the human genome assemblies hg19 and hg38 (http://jaspar.genereg.net/genome-tracks/).

A NEW, POWERFUL AND USER-FRIENDLY WEB INTERFACE

A new web interface

The JASPAR 2018 release comes with a completely redesigned web interface that meets modern web standards. This interactive web framework is implemented using Django, a model-view-controller based web-framework for Python. We used MySQL as a backend database to store profile metadata and Bootstrap as a frontend template engine. We have greatly improved the visibility and usability of existing functionality, created easier navigation with semantic URLs, and enhanced browsing and searching. On the homepage, we provide a dynamic tour of JASPAR 2018, walking users through the main features of the new website. A video of the tour is available at http://jaspar.genereg.net/tour. The database can be browsed for individual collections by using the navigation links on the left sidebar. Moreover, it can be searched for each of the six different taxonomic groups included in the JASPAR CORE collection using the tabs available on the homepage (Figure 2). TF-binding profiles can be further filtered through the case insensitive search option available on the homepage. In addition, through the 'Advanced Options', the search criteria can be further restricted (Figure 2A). Search results are presented in a responsive and paginated table along with sequence logos of the PFMs, which can be selected for download or to perform a variety of analyses available on the right panel (Figure 2B). All information in the tables can be downloaded as comma-separated value files. Profile IDs and sequence logos can be clicked to view the detailed profile pages (Figure 2C). PFMs can be downloaded in several formats including JASPAR, TRANSFAC and MEME (Figure 2D). Furthermore, we have incorporated new features to the web interface, such as 'Add to Cart', where users can add TF profiles of interest for download or further analyses (Figure 2B). Finally, we have introduced semantic URLs to facilitate external linking to the detailed pages of individual profiles (e.g. http://jaspar.genereg.net/matrix/MA0059.1/). We have implemented a URL redirection mechanism to correctly direct the links pointing to previous JASPAR URL patterns from external resources.

RESTful API

In previous releases, the underlying data could be retrieved as flat files or by using programming language-specific modules. Associated with this release, we introduced a RESTful API to access the JASPAR database programmatically (see https://www.biorxiv.org/content/early/2017/07/06/160184 for details). The RESTful API enables programmatic access to JASPAR by most programming languages and returns data in seven widely used formats: JSON, JSONP, JASPAR, MEME, PFM, TRANSFAC and YAML. Further, it provides a browsable interface and access to the JASPAR motif inference tool for bioinformatics tool developers. The RESTful API is implemented in Python using the Django REST Framework and is freely accessible at http://jaspar.genereg.net/api/. The source code for the website and RESTful API are freely available at https://bitbucket.org/CBGR/jaspar under GPL v3 license.

CONCLUSION AND PERSPECTIVES

In this seventh release of the JASPAR database, we continue our commitment to provide the research community with high-quality, non-redundant TF-binding profiles for TFs in six taxa. As in previous releases, we have greatly expanded the number of available profiles in the database, both for PFMs and TFFMs. We also greatly improved user experience through a new easy-to-use website and a RESTful API that grants universal programmatic access to the database. Moreover, for the PFMs in the JASPAR CORE collection, we provide a hierarchical clustering and genome-wide TFBS predictions for the hg19 and hg38 human genome assemblies as UCSC tracks.

During the curation process, hundreds of PFMs were discarded because our curators failed to find any support from existing literature. As new experiments and data become available, binding preferences for these TFs will be considered for JASPAR incorporation. For instance, we re-examined data from (34) to incorporate seven previously excluded PFMs into JASPAR 2018. In the future, we would like to engage the scientific community in the curation process to increase our capacity to introduce new TF-binding profiles in JASPAR. We plan to dedicate a specific section of the website to hosting the profiles that were not introduced into JASPAR, to encourage researchers to perform experiments and/or point us to literature that our curators missed in order to support these profiles. We believe that the engagement of the scientific community to support JASPAR will further improve our capacity to expand the collection of high quality TF-binding profiles.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the scientific community for performing experimental assays of TF–DNA interactions and for publicly releasing the data. We thank Georgios Magklaras and his team for IT support. We thank José Manuel Franco for sharing the plant PBM data, Jens De Ceukeleire for help with plant ChIP-seq data processing and José Luis Villanueva-Cañas for sharing the *Drosophila* TFFMs prior to publication. We thank Rachelle Farkas for proofreading the manuscript.

FUNDING

Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., M.G., A.K.); Genome Canada and Canadian Institutes of Health Research (On-Target Grants) [255ONT and BOP-149430 to W.W.W., O.F., R.v.d.L., D.J.A.]; Natural Sciences and Engineering Research Council of Canada (Discovery Grant) [RGPIN-2017–06824 to W.W.W.]; Weston Brain Institute [20R74681 to O.F.]; Agence Nationale de la Recherche [ANR-10-LABX-49–01 to F.P., A.S.]; IDEX graduate schoool (to A.S.); CNRS (to A.B., F.P.); Research Foundation–Flanders Grant [G001015N to S.R.K.]; French Ministry

of Higher Education and Research (MESR) PhD Fellowship (to J.A.C.-M.); Lundbeck Foundation (to A.S.); Independent Research Fund Denmark (to A.S.); Innovation Fund Denmark (to A.S.); Elixir Denmark (to A.S.); Wellcome Trust [106954 to G.T., D.B., B.L.]; Biotechnology and Biological Sciences Research Council [BB/N023358/1 to G.T., D.B., B.L.]; Medical Research Council UK [MC_UP_1102/1 to G.T., D.B., B.L.]. The open access publication charge for this paper has been waived by Oxford University Press - *NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.
Conflict of interest statement. None declared.

REFERENCES

1. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
2. Wingender,E., Schoeps,T., Haubrock,M. and Dönitz,J. (2015) TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.*, **43**, D97–D102.
3. Xie,Z., Hu,S., Qian,J., Blackshaw,S. and Zhu,H. (2011) Systematic characterization of protein-DNA interactions. *Cell. Mol. Life Sci.*, **68**, 1657–1668.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
5. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant. Biol.*, **1**, 115–130.
6. Stormo,G.D. (2015) DNA motif databases and their uses. *Curr. Protoc. Bioinformatics*, **51**, 1–6.
7. 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. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
8. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
9. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
10. Bulyk,M.L., Johnson,P.L.F. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
11. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
12. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
13. Chin,F. and Leung,H.C.M. (2008) DNA motif representation with nucleotide dependency. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 110–119.
14. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
15. Zellers,R.G., Drewell,R.A. and Dresch,J.M. (2015) MARZ: an algorithm to combinatorially analyze gapped n-mer models of transcription factor binding. *BMC Bioinformatics*, **16**, 1–14.
16. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 1–15.
17. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
18. 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 Syst.*, **3**, 278–286.
19. Omidi,S., Zavolan,M., Pachkov,M., Breda,J., Berger,S. and van Nimwegen,E. (2017) Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. *PLoS Comput. Biol.*, **13**, e1005176.
20. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
21. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
22. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
23. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D1010.
24. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D1427.
25. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2017) ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seqexperiments. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1092.
26. Eveland,A.L., Goldshmidt,A., Pautler,M., Morohashi,K., Liseron-Monfils,C., Lewis,M.W., Kumari,S., Hiraga,S., Yang,F., Unger-Wallace,E. *et al.* (2014) Regulatory modules controlling maize inflorescence architecture. *Genome Res.*, **24**, 431–443.
27. Verkest,A., Abeel,T., Heyndrickx,K.S., Van Leene,J., Lanz,C., Van De Slijke,E., De Winne,N., Eeckhout,D., Persiau,G., Van Breusegem,F. *et al.* (2014) A generic tool for transcription factor target gene discovery in Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiol.*, **164**, 1122–1133.
28. Li,C., Qiao,Z., Qi,W., Wang,Q., Yuan,Y., Yang,X., Tang,Y., Mei,B., Lv,Y., Zhao,H. *et al.* (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell*, **27**, 532–545.
29. Cui,X., Lu,F., Qiu,Q., Zhou,B., Gu,L., Zhang,S., Kang,Y., Cui,X., Ma,X., Yao,Q. *et al.* (2016) REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in Arabidopsis. *Nat. Genet.*, **48**, 694–699.
30. Birkenbihl,R.P., Kracher,B. and Somssich,I.E. (2017) Induced genome-wide binding of three Arabidopsis WRKY transcription factors during early MAMP-triggered immunity. *Plant Cell*, **29**, 20–38.
31. O'Malley,R.C., Huang,S.-S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.
32. 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. *Nat. Methods*, **14**, 316–322.
33. Franco-Zorrilla,J.M., López-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
34. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
35. Weirauch,M.T., Yang,A., Albu,M., Cote,A., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

36. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
37. Castro-Mondragon,J.A., 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. *Nucleic Acids Res.*, **45**, e119.
38. Kwon,A.T., Arenillas,D.J., Worsley Hunt,R. and Wasserman,W.W. (2012) oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3*, **2**, 987–1002.
39. Mathelier,A., Lefebvre,C., Zhang,A.W., Arenillas,D.J., Ding,J., Wasserman,W.W. and Shah,S.P. (2015) Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.*, **16**, 1–17.
40. Verfaillie,A., Imrichova,H., Janky,R. and Aerts,S. (2015) iRegulon and i-cistarget: reconstructing regulatory networks using motif and track enrichment. *Curr. Protoc. Bioinformatics*, **52**, 1–39.
41. Arenillas,D.J., Forrest,A.R.R., Kawaji,H., Lassmann,T. and FANTOM ConsortiumFANTOM Consortium, Wasserman,W.W. and Mathelier,A. (2016) CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. *Bioinformatics*, **32**, 2858–2860.
42. Shi,W., Fornes,O., Mathelier,A. and Wasserman,W.W. (2016) Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.*, **44**, 10106–10116.
43. Arner,E., Daub,C.O., Vitting-Seerup,K., Andersson,R., Lilje,B., Drabløs,F., Lennartsson,A., Rönnerblad,M., Hrydziuszko,O., Vitezic,M. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.
44. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
45. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
46. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
47. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations

Fornes, O.[†], **Gheorghe, M.**[†], Richmond, P. A., Arenillas, D. J., Wasserman, W. W.[*], and Mathelier, A.[*]

2018, *Scientific Data*, 5:180141.

III

SCIENTIFIC DATA

OPEN

Data Descriptor: MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations

Received: 4 December 2017
Accepted: 27 April 2018
Published: 24 July 2018

Oriol Fornes[1,*], Marius Gheorghe[2,*], Phillip A. Richmond[1], David J. Arenillas[1], Wyeth W. Wasserman[1] & Anthony Mathelier[2,3]

Interpreting the functional impact of noncoding variants is an ongoing challenge in the field of genome analysis. With most noncoding variants associated with complex traits and disease residing in regulatory regions, altered transcription factor (TF) binding has been proposed as a mechanism of action. It is therefore imperative to develop methods that predict the impact of noncoding variants at TF binding sites (TFBSs). Here, we describe the update of our MANTA database that stores: 1) TFBS predictions in the human genome, and 2) the potential impact on TF binding for all possible single nucleotide variants (SNVs) at these TFBSs. TFBSs were predicted by combining experimental ChIP-seq data from ReMap and computational position weight matrices (PWMs) derived from JASPAR. Impact of SNVs at these TFBSs was assessed by means of PWM scores computed on the alternate alleles. The updated database, MANTA2, provides the scientific community with a critical map of TFBSs and SNV impact scores to improve the interpretation of noncoding variants in the human genome.

Design Type(s)	data integration objective • database creation objective • transcription factor bindining site prediction objective
Measurement Type(s)	transcription factor binding site
Technology Type(s)	prediction
Factor Type(s)	cell line • antibody • transcription factor
Sample Characteristic(s)	Homo sapiens

[1]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 W 28th Ave, Vancouver, BC V5Z 4H4, Canada. [2]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway. [3]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to W.W.W. (email: wyeth@cmmt.ubc.ca) or to A.M. (email: anthony.mathelier@ncmm.uio.no).

Background & Summary

Understanding the relationship between DNA sequence variation (genotype) and observable traits and diseases (phenotype) is one of the central paradigms of the post-genomics era. While most analyses have focused on the ~2% of the genome that codes for proteins, genome-wide association studies have shown that up to 88% of disease- and trait-associated variants are located in the 98% of the genome that is noncoding[1]. Several computational tools, such as SIFT[2] and Polyphen[3], are well established for the assessment of the deleterious impact of coding variation on protein functions yet interpreting the functional impact of noncoding variants continues to be challenging[4].

Recently, bioinformatics methods have been developed for scoring the impact of noncoding variants based on their pathogenicity and regulatory capacity (Table 1). These methods vary both in their algorithmic approaches and the underlying genomic features used. For instance, evolutionary conservation[5] can be used to evaluate nucleotides under purifying selection, and experimental data such as histone modifications[6], chromatin accessibility[7,8], and DNA methylation[9] are used to identify biochemically active DNA, which is indicative of regulatory capacity.

Transcription factors (TFs) are sequence-specific DNA-binding proteins that regulate gene transcription[10]. Genomic locations at which TFs interact with DNA are defined as TF binding sites (TFBSs). They are typically short (6–10 bp) and often exhibit degeneracy. Chromatin immunoprecipitation combined with sequencing (ChIP-seq)[11] provides *in vivo* TF-DNA interactions at ~200–300 bp resolution. These ChIP-seq regions are expected to encompass the 6–10 bp fragments corresponding to TF-DNA interactions (TFBSs). The ReMap database[12] is a publicly available resource providing an atlas of such regions obtained from 2,829 uniformly processed human ChIP-seq data sets.

The DNA sequences bound by a given TF can be represented as position frequency matrices (PFMs), which count the number of occurrences of each nucleotide within the TFBSs for that TF[13]. PFMs can be converted into probabilistic computational models, namely position weight matrices (PWMs), which can be used to predict TFBSs on any DNA sequence (reviewed by Wasserman and Sandelin[14]). Several databases of PFMs exist[15], including the recently updated JASPAR database[16], which stores manually-curated and non-redundant DNA-binding profiles such as PFMs for TFs across six taxonomic groups.

With most noncoding variants associated with complex traits and disease residing in regulatory sequences[17], it is expected that some will alter the binding of TFs to DNA[18,19]. Therefore, it is imperative to develop methods that prioritize noncoding variants based on their impact on TF-DNA interactions. In 2015, we developed MANTA, a Mongo database for the analysis of TFBS alterations, to study the impact of regulatory mutations in B-cell lymphomas[20]. The database stores TFBSs in ChIP-seq regions predicted using PWMs derived from the JASPAR database, as well as the potential impact on TF binding of all possible single nucleotide variants (SNVs) that could occur at these TFBSs (Fig. 1). Building on the recent updates of both the JASPAR and ReMap databases, we have largely expanded MANTA. This second release of the database, MANTA2, hosts over 48 million TFBS predictions within ChIP-seq regions of 225 human TFs, covering about 8% of the human genome, together with computed impact scores for all

Method	Designed for	Algorithmic approach	Genomic features	PMID
CADD	pathogenicity	support vector machine	conservation, epigenomic annotations	24487276
CpGenie	impact on methylation	deep neural network	conservation, epigenomic annotations, TFBS alterations	28334830
DANN	pathogenicity	deep neural network	conservation, epigenomic annotations	25338716
DeepSEA	regulatory potential	deep neural network, logistic regression classifier	conservation, epigenomic annotations, TFBS alterations	26301843
deltaSVM	regulatory potential	support vector machine	epigenomic annotations, TFBS alterations	26075791
Eigen	pathogenicity	spectral clustering	conservation, epigenomic annotations	26727659
FATHMM	pathogenicity	hidden Markov model	conservation, epigenomic annotations	28968714
fitCons	fitness consequence	generative probability, genome partitioning	conservation, epigenomic annotations	25599402
FunSeq2	cancer pathogenicity	feature-based scoring, PWM scoring, somatic hotspots	conservation, epigenomic annotations, TFBS alterations	25273974
GWAVA	pathogenicity	random forest	conservation, epigenomic annotations, TFBS alterations	24487584
LINSIGHT	regulatory potential	linear regression, generative probability	conservation, epigenomic annotations, TFBS alterations	28288115
MANTA	regulatory potential	PWM scoring	TFBS alterations	25903198
RegulomeDB	regulatory potential	feature-based scoring, PWM scoring	conservation, epigenomic annotations, TFBS alterations	22955989
ReMM	pathogenicity	random forest	conservation, epigenomic annotations	27569544
RVSP	regulatory potential	random forest	conservation, epigenomic annotations	27406314
SNP2TFBS	regulatory potential	PWM scoring	TFBS alterations	27899579

Table 1. List of published tools with the capacity to evaluate the impact of noncoding variants. For each "Method", we describe its "Intended use", "Algorithmic approach", underlying "Genomic features" and PubMed ID ("PMID") of the corresponding publication.

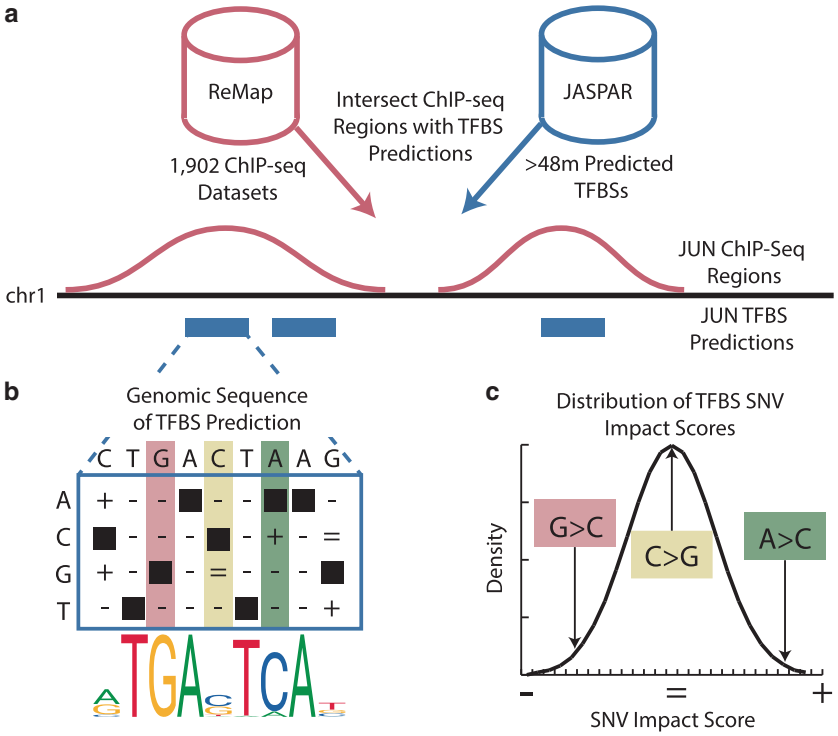

Figure 1. Overview of MANTA2. a) Intersection of the ReMap ChIP-seq regions with JASPAR TFBS predictions to produce a set of TFBSs with both experimental and computational evidence of TF binding. A mock example of JUN is given for a region on chromosome one. **b**) A matrix representing the difference in PWM score for all possible SNVs compared to the reference sequence at that TFBS, including negative impact (−), positive impact (+), and no change (0) of score. Black boxes indicate that nucleotides of the reference TFBS sequence are not stored in the database. The sequence logo for JUN is provided below the matrix where the information content is proportional to the size of the nucleotide letters. **c**) Mock distribution of TFBS SNV impact scores when considering all possible SNVs in the TFBS. The distribution is annotated with examples of decreased TF binding capacity (red), no change in TF binding capacity (yellow), and increased TF binding capacity (green).

possible overlapping SNVs. Hence, MANTA2 provides the scientific community with a critical map of TFBSs and SNV impact scores for the interpretation of noncoding variants in the human genome.

Methods

Transcription factor binding site predictions

From ReMap[12], we retrieved 1,902 uniformly processed ChIP-seq data sets (*i.e.* sets of ChIP-seq regions) for 227 human TFs for which we had binding profiles in JASPAR[16]. Each ChIP-seq data set was paired with one or more PFMs associated to the ChIP'ed TFs from the JASPAR CORE vertebrates collection (see Supplementary Table 1). For each pair, we intersected the ChIP-seq regions with the corresponding TFBSs predicted for the ChIP'ed TF using bedtools intersect[21] with "-wa -wb" options to preserve the original coordinates. The PWM-based TFBS predictions are publicly available as part of the JASPAR human genome track at http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/hg38/tsv/. The intersection resulted in 48,512,399 TFBSs for 225 TFs, covering 255,918,025 bp of the human genome (Fig. 1a). No overlap was found for 2 TFs between the ChIP-seq regions and PWM-based TFBS predictions. Note that all data relates to the build 38 of the Genome Reference Consortium human genome (hg38).

Computation of SNV impact scores

For each TFBS, we computed the impact on TF binding of all possible overlapping SNVs as described in the manuscript describing MANTA[20] (Fig. 1b). First, both strands of the $2n-1$ bp region centered around each possible SNV, where n is the length of the considered PWM, were scanned with the corresponding PWM using the TFBS Perl module[22] (version 0.7.1) to identify the best PWM score on the alternate allele. Note that we only kept the best match per SNV. We then computed the distribution of PWM scores for all these SNVs and calculated the corresponding mean, m, and standard deviation, sd.

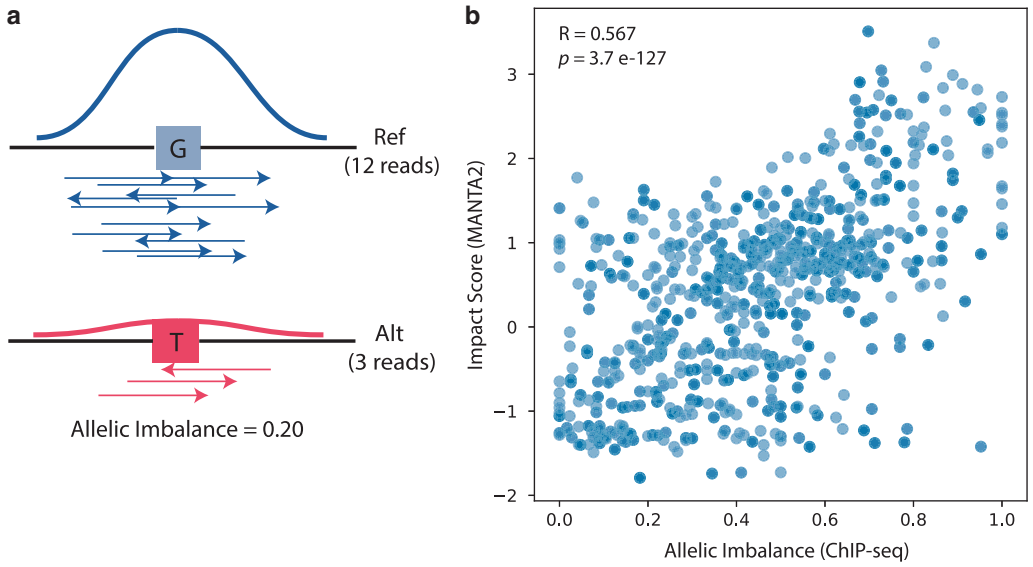

Figure 2. Assessing MANTA2 impact scores with heterozygous TF-binding events. a) Allelic imbalance is calculated as the number of ChIP-seq reads mapped on the alternate allele divided by the total number of reads mapped at heterozygous sites. **b**) MANTA2 impact scores correlate with allelic imbalance of ChIP-seq data. Events (blue dots) are plotted with respect to their allelic imbalance of ChIP-seq reads (*x*-axis) and impact scores from MANTA2 (*y*-axis). The Pearson coefficient (R) and *P*-value (*p*) of the correlation between allelic imbalance and impact score are provided in the plot.

For each SNV, the final impact score was calculated as the Z-score of its TFBS score, *S*, within the distribution of alternate PWM scores at that TFBS (*i.e.* $(S-m)/sd$). Users can refer to the webinar video describing the original MANTA database (http://www.cisreg.ca/Webinars/JASPAR_BioPython_-MANTA.flv). Therefore, for each SNV, MANTA stores its associated reference and alternate TFBS PWM scores and locations, along with the computed impact score.

Validation using heterozygous TF-binding events

We downloaded ChIP-seq data for 35,703 TF-binding events at heterozygous sites in GM12878 and HeLa cells for 36 different TFs[18]. For each event, allelic imbalance was calculated as the number of ChIP-seq reads mapped on the alternate allele divided by the total number of reads mapped at that position (Fig. 2a). The coordinates from the original publication refer to the hg19 version of the human genome; we used the liftOver tool from the UCSC Genome Browser[23] to convert them to the hg38 assembly (the conversion process failed for 12 coordinates).

Code availability

MANTA2 is freely distributed as a GitHub repository at https://github.com/wassermanlab/MANTA2.

Data Records

The Mongo database dump of MANTA2, is deposited as a tarball on Zenodo (Data Citation 1).

Technical Validation

The quality and technical validation of the ChIP-seq data and TFBS predictions is described in the 2018 manuscripts of ReMap[12] and JASPAR[16], respectively, and is summarised below.

ReMap ChIP-seq data

ReMap ChIP-seq datasets were uniformly processed using a well-established pipeline[12]. ChIP-seq reads were aligned to the human genome using bowtie2 (ref. 24) (version 2.2.9) using options "-end-to-end" and "-sensitive". When necessary, reads were trimmed and polymerase chain reaction duplicates were removed from the alignments with samtools rmdup[25]. ChIP-seq regions were identified using the MACS2 peak-calling tool[26] (version 2.1.1.2) with default parameters. The quality of all ChIP-seq datasets was assessed based on metrics developed by the ENCODE consortium[27].

JASPAR TFBS predictions

JASPAR TFBSs were predicted by scanning the human genome using two different methods[16]: the TFBS Perl module[22] (version 0.7.1) and FIMO[28], as distributed within the MEME suite[29] (version 4.11.2).

FIMO is one of the best performing tools for scanning DNA sequences with PWMs to predict TFBSs[30]. To scan the human genome with the BioPerl TFBS module, PFMs were converted to PWMs and predictions with a relative score ≥0.8 were kept. In preparation for the FIMO scan, PFMs were reformatted to MEME motifs and motifs that matched with a *P*-value < 0.05 were kept. For quality control, TFBS predictions that were not consistent between the two methods were filtered out. Such consistency ensures, for instance, technical validation for the coordinates of the TFBS predictions.

MANTA2

The technical validation of MANTA2 involved assessing data quality and database integrity controls. A spot check data quality control was performed using the UCSC Genome Browser[23]. For 15 randomly selected TFBSs (of different TFs) from MANTA2 we manually checked that: 1) the TFBS overlapped a ReMap ChIP-seq region associated with that TF; 2) the JASPAR PFM matched the start, end, and strand stored for that TFBSs; and 3) the stored SNVs for that TFBS had the expected impact on TF binding. Moreover, we assessed the usefulness of MANTA2 impact scores on an external dataset of heterozygous TF-binding events[18]. As expected, the allelic imbalance calculated for ChIP-seq reads (see Methods) significantly correlated with the impact scores from MANTA2 (Pearson correlation coefficient = 0.567, *P*-value = 3.7e-127; Fig. 2b). Additionally, we checked the database integrity for MANTA2 by dumping and restoring the database on common operating systems and workstations. Finally, we tested the command line and web interface access to MANTA2 (see Usage Notes section) to interpret variant files in VCF, GFF, and BED format.

Usage Notes

MANTA2 can be accessed either programmatically or via its web interface. To access the database programmatically, users must first clone (*i.e.* "git clone https://github.com/wassermanlab/MANTA2.git") or download MANTA2 from GitHub (see Code availability in the Methods section). The script "search_manta2.py" provides programmatic access to MANTA2. It requires the following inputs:

- The name of the MANTA2 database in the MongoDB system (option "-d")
- The name of the server where the MongoDB system is hosted (option "-H")
- A user with "read" privileges to the MANTA2 database (option "-u")
- The password for that user (option "-p")
- A file containing a list of variants in "VCF", "BED" or "GFF" format (option "-i")

Non-mandatory options include:

- The format of the input variant file (option "-t"; by default the script tries to identify the input format automatically)
- The name of a file to output the results (option "-o"; by default is set to the standard output stream (stdout))

As a usage example, the MANTA2 database hosted by the Wasserman lab can be accessed as follows: "python search_manta2.py -d manta2 -H manta.cmmt.ubc.ca -u manta_r -p mantapw -i <variant file>".

A variant file can be obtained by executing the shell script: "bash ./examples/get_VCF_example.sh".

The resulting VCF file (*i.e.* "chr20.vcf") contains high-confidence SNP, small indel, and homozygous reference calls on chromosome 20 from the Genome in a Bottle (version 3.3.2) sample HG001 (ref. 31). In response, "search_manta2.py" returns all TFBS predictions potentially impacted by these variants as tab-separated values. For each TFBS alteration, the script provides the variant information along with the associated wild-type (reference) and mutated (alternative) TFBS information, including:

- the chromosome and position of the variant;
- the reference and alternative alleles at that genomic location;
- the mutation ID (if the input file format allowed for it, otherwise the field is displayed as ".");
- the TF name and associated JASPAR profile ID;
- the start, end and strand, as well as the absolute (raw) and relative scores for both the reference and alternative TFBSs;
- and the impact score.

Users who plan on performing large numbers of searches should create a local build of the MANTA2 database. Instructions are provided in the "README.md" file of the GitHub repository.

The MANTA2 database hosted by the Wasserman lab can also be accessed via a dedicated web server at http://manta.cmmt.ubc.ca/manta2. Similar to the "search_manta2.py" script, the server requires as input a list of variants in VCF, GFF, or BED format (see help page), and it returns all TFBS predictions

potentially impacted by these variants as a tab-separated values table. The table can be sorted on any column by clicking on the column header.

References

1. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
2. Ng, P. C. & Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **11,** 863–874 (2001).
3. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
4. Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends Genet.* **31,** 67–76 (2015).
5. Bejerano, G. Ultraconserved Elements in the Human Genome. *Science* **304,** 1321–1325 (2004).
6. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146,** 1016–1028 (2011).
7. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
9. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23,** 555–567 (2013).
10. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–263 (2009).
11. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316,** 1497–1502 (2007).
12. Chèneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46,** D267–D275 (2018).
13. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quantitative Biology* **1,** 115–130 (2013).
14. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5,** 276–287 (2004).
15. Stormo, G. D. DNA Motif Databases and Their Uses. *Curr. Protoc. Bioinformatics* **51,** 2.15.1–6 (2015).
16. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46,** D260–D266 (2018).
17. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16,** 197–212 (2015).
18. Shi, W., Fornes, O., Mathelier, A. & Wasserman, W. W. Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.* **44,** 10106–10116 (2016).
19. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45,** D139–D144 (2017).
20. Mathelier, A. *et al.* Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.* **16,** 84 (2015).
21. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics* **47,** 11.12.1–34 (2014).
22. Lenhard, B. & Wasserman, W. W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18,** 1135–1136 (2002).
23. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45,** D626–D634 (2017).
24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
26. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
27. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–1831 (2012).
28. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).
29. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).
30. Jayaram, N., Usvyat, D. & R. Martin, A. C. Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* (2016); doi:10.1186/s12859-016-1298-9.
31. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–251 (2014).

Data Citation

1. Fornes, O., Gheorghe, M., Richmond, P. A., Arenillas, D. J., Wasserman, W. W. & Mathelier, A. *Zenodo* http://doi.org/10.5281/zenodo.1044747 (2017).

Acknowledgements

We thank Alice M. Kaye and Rachelle Farkas for proofreading the manuscript, and Georgios Magklaras and his team for IT support. We acknowledge the support provided by WestGrid (https://www.westgrid.ca/) and Compute Canada/Calcul Canada (https://www.computecanada.ca/). A.M. and M.G. were supported by funding from the Norwegian Research Council, Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM), which is part of the Nordic European Molecular Biology Laboratory Partnership for Molecular Medicine. D.J.A., O.F., P.A.R., and W.W.W. were supported by funding from Genome Canada and the Canadian Institutes of Health Research (OnTarget grants 255ONT and BOP-149430), the Natural Sciences and Engineering Research Council of Canada (discovery grant RGPIN-2017-06824), the Weston Brain Institute (20R74681), and the BC Children's Hospital Foundation and Research Institute (UBC:17W33804 award to P.A.R.).

Author Contributions

A.M. and M.G. provided the ReMap ChIP-seq regions, and O.F. the JASPAR TFBS predictions. M.G. intersected the two data sets and A.M. generated the TFBS SNV scores. O.F. assessed the MANTA2 scores against heterozygous TF-binding events. D.J.A. updated MANTA and created the web server. P.A.

R. generated the figures and tables, and reviewed the most relevant literature. A.M., O.F., and W.W.W. devised the project and wrote the manuscript with input from all authors.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/sdata

Competing interests: The authors declare no competing interests.

How to cite this article: Fornes, O. *et al*. MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. *Sci. Data* 5:180141 doi: 10.1038/sdata.2018.141 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A map of direct TF-DNA interactions in the human genome

Gheorghe, M., Sandve, G. K., Khan, A., Chèneby, J., Ballester, B., and Mathelier, A.*

2019, *Nucleic Acids Research*, 47(4):e21–e21.

IV

Published online 4 December 2018

Nucleic Acids Research, 2019, Vol. 47, No. 4 ***e21***
doi: 10.1093/nar/gky1210

A map of direct TF–DNA interactions in the human genome

Marius Gheorghe [1], Geir Kjetil Sandve [2], Aziz Khan [1], Jeanne Chèneby[3], Benoit Ballester [3] and Anthony Mathelier [1,4,*]

[1]Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway, [2]Department of Informatics, University of Oslo, Oslo, Norway, [3]Aix Marseille Université, INSERM, TAGC, Marseille, France and [4]Department of Cancer Genetics, Institute for Cancer Research, Radiumhospitalet, Oslo, Norway

Received August 18, 2018; Revised October 31, 2018; Editorial Decision November 18, 2018; Accepted November 20, 2018

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the most popular assay to identify genomic regions, called ChIP-seq peaks, that are bound *in vivo* by transcription factors (TFs). These regions are derived from direct TF–DNA interactions, indirect binding of the TF to the DNA (through a co-binding partner), nonspecific binding to the DNA, and noise/bias/artifacts. Delineating the *bona fide* direct TF–DNA interactions within the ChIP-seq peaks remains challenging. We developed a dedicated software, ChIP-eat, that combines computational TF binding models and ChIP-seq peaks to automatically predict direct TF–DNA interactions. Our work culminated with predicted interactions covering >4% of the human genome, obtained by uniformly processing 1983 ChIP-seq peak data sets from the ReMap database for 232 unique TFs. The predictions were *a posteriori* assessed using protein binding microarray and ChIP-exo data, and were predominantly found in high quality ChIP-seq peaks. The set of predicted direct TF–DNA interactions suggested that high-occupancy target regions are likely not derived from direct binding of the TFs to the DNA. Our predictions derived co-binding TFs supported by protein-protein interaction data and defined *cis*-regulatory modules enriched for disease- and trait-associated SNPs. We provide this collection of direct TF–DNA interactions and *cis*-regulatory modules through the UniBind web-interface (http://unibind.uio.no).

INTRODUCTION

The transcription of DNA into RNA is mainly regulated through a complex interplay between proteins and the chromatin at *cis*-regulatory regions such as promoters and enhancers. Transcription factors (TFs) are key proteins specifically binding short DNA sequences, known as TF binding sites (TFBSs), to ensure transcription at appropriate rates in the correct cell types (1). Therefore, genome-wide identification of TFBSs is a critical step to decipher transcriptional regulation, and how this process is altered in diseases (2).

Classically, genome-wide *in vivo* TF binding regions are identified through the chromatin immunoprecipitation followed by sequencing (ChIP-seq) assay (3). The genomic regions obtained with ChIP-seq, the so-called ChIP-seq peaks, are usually a few hundred base pairs (bp)-long and should encompass the TFBSs (∼10 bp-long), where direct TF–DNA interactions occur. However, ChIP-seq peaks derive from either direct TF–DNA interactions, protein-protein interactions with other regulators such as co-factors, or unspecific binding. Moreover, ChIP-seq experiments are prone to artifacts and delineating *bona fide* TF-bound regions is still an ongoing challenge (4–6) (Wreczycka *et al.*, bioRxiv, 10.1101/107680).

As TFs specifically recognize DNA sequence motifs, computational tools have been instrumental in the prediction and characterization of direct TF–DNA interactions (7). TFBSs are commonly modelled with position weight matrices (PWMs), which represent the probability of each nucleotide to be present at each position within bona fide TFBSs (7). While PWMs work well (8), more sophisticated approaches have recently been designed to model complex features of TF–DNA interactions captured by next-generation sequencing data (e.g. (9–13)). However, the best performing model varies for different TFs or TF families (8,14,15).

While multiple resources collecting TF binding regions derived from ChIP-seq exist (16–19), a limited number store genome-wide identification of TFBSs (17,20,21). The TFBS Conserved Track of the UCSC Genome Browser combined phylogenetic sequence conservation and PWMs to identify TFBSs (22) while the MANTA resource (23) integrated ChIP-seq peaks from ReMap (16) with PWMs from JASPAR (24) for TFBS predictions. A strong limitation of these approaches is that they use the same pre-defined score

*To whom correspondence should be addressed. Tel: +47 228 40 561; Email: anthony.mathelier@ncmm.uio.no

thresholds for all PWMs and all data sets. The ORegAnno database provides TFBSs obtained through literature curation (21), but the number of TFBSs available for human is limited to ∼8000.

A previous study showed that ChIP-seq data sets fall within one of three categories: (i) data sets enriched for the TF canonical binding motif close to the ChIP-seq peak summit (where the highest number of ChIP-seq reads map), (ii) data sets lacking enrichment for the canonical binding motif close to the peak summit and (iii) data sets having a combination of peaks with and without the TF canonical binding motif proximal to the peak-summit (25). Most ChIP-seq data sets were observed in category (iii). As direct TF–DNA interactions are expected to be enriched at ChIP-seq peak summits (25–30), Worsley Hunt *et al.* developed a heuristic approach specifically based on PWMs to automatically identify, in each ChIP-seq data set, this enrichment zone. The method determines the thresholds on the PWM scores and distances to the peak summits delimiting the enrichment zone that contains direct TF–DNA interactions. However, this method does not work with some more recent TFBS computational models (15,31,32).

In this study, we mapped direct TF–DNA interactions in the human genome in a refined manner by capitalizing on uniformly processed TF ChIP-seq data sets and computational tools modelling TFBSs. We provide (i) a new software to predict direct TF–DNA interactions within ChIP-seq peaks along with (ii) genome-wide predictions of such interactions in the human genome. Using an entropy-based algorithm, we have developed ChIP-eat, a tool that automatically identifies direct TF–DNA interactions using both ChIP-seq peaks and any computational model for TFBSs. We applied ChIP-eat to 1983 human ChIP-seq peak data sets from the ReMap database (16), accounting for 232 distinct TFs. The set of predicted direct TF–DNA interactions derived from PWMs covers >4% of the human genome. To make this resource available to the community, we have created UniBind (http://unibind.uio.no/), a web-interface providing public access to the predictions. We validated *a posteriori* these TFBS predictions using protein binding microarray (33) and ChIP-exo (34) data, and multiple ChIP-seq peak-callers. We used these TFBSs to (i) confirm that hotspots of ChIP-seq peaks (also known as high occupancy target regions (35)) are likely not derived from direct TF–DNA interactions, (ii) predict co-binding TFs and (iii) define *cis*-regulatory modules, which are enriched for disease- and trait-associated SNPs.

MATERIALS AND METHODS

ChIP-seq data

The ChIP-seq data sets considered were retrieved, processed, and classified as part of the last update (2018) of the ReMap database (16) (Supplementary Figure S1).

TF binding profiles

For 1983 ChIP-seq data sets used in the last ReMap update, we were able to manually assign TF binding profiles corresponding to the ChIP'ed TFs as position frequency matrices (PFMs) from the JASPAR (2018) database (24).

Training data sets

To train the TFBS computational models (see below), we considered 101 bp sequences centered around the peak summits as positive training sets. When required for training, negative training sets were obtained by shuffling the positive sequences using the *g* subcommand of the BiasAway (version 0.96) tool to match the %GC composition (25).

TFBS computational models

Position weight matrices. JASPAR PFMs were converted to PWMs as previously described in (36). For each ChIP-seq data set, PWMs were optimized using DiMO (version 1.6; default parameters with a maximum of 150 optimization steps) using the corresponding training sets (37). For TFBS predictions, we considered PWM *relative* scores, which were computed as $relative\ score = 100 \times (absolute\ score - min)/(max - min)$ where *absolute score* corresponds to the PWM absolute/raw score and *min* and *max* to the minimal and maximal absolute/raw PWM scores, respectively.

Binding energy models. JASPAR PFMs were converted to binding energy models (BEMs; (32)) using the implementation from the MARS Tools (https://github.com/kipkurui/MARSTools; Kibet and Machanick, bioRxiv, doi:10.1101/065615). We modified the implementation to return a BEM score corresponding to 1 – (original score) to consider the best site of the DNA sequence as the one with the highest BEM score (instead of the lowest one).

Transcription factor flexible models. First-order transcription factor flexible models (TFFMs) (version 2.0) were initialized with the DiMO-optimized PFMs and trained with default parameters (https://github.com/wassermanlab/TFFM; (31)) on the positive training sets.

DNAshapedTFBS models. The DNA shape-based models were trained on the training sets using the DNAshapedTFBS tool (version 1.0; https://github.com/amathelier/DNAshapedTFBS/; (15)). We trained three types of DNAshapedTFBS models with the following features: (i) DiMO-optimized PWM + DNA shape, (ii) first-order TFFM + DNA shape and (iii) 4-bits encoding + DNA shape following (15). We considered the first and second order DNA shape features helix twist, propeller twist, minor groove width, and roll with values extracted from GBShape (38).

Landscape plots

Each TFBS computational model was applied to each ChIP-seq data set independently. Following the strategy described in (25), we considered 1001 bp sequences centered around the peak summits, obtained using the bedtools (version 2.25) *slop* subcommand (39). The trained computational models were used to extract the best (maximal score) site per 1001 bp ChIP-seq peak region. For each ChIP-seq data set, landscape plots were constructed from the corresponding sites following the TFBS_Visualization tool (25). These scatter plots were also converted into heat maps using the *kde2d* function from the MASS R package (40).

Automated identification of the enrichment zone

To define the enrichment zone for each landscape plot, we automatically identified the thresholds for the TFBS computational model scores and distances to peak summits using the entropy-based algorithm from (41). The algorithm aims at identifying two classes of elements. Given a histogram, the algorithm selects the threshold that maximizes the within-class sum of the Shannon entropies for the elements in two classes (42). The two classes of elements identified are defined by the elements with values (i) above and (ii) below the threshold, respectively. This procedure optimally separates the input elements in two classes. Given a ChIP-seq data set, we applied the algorithm to the histograms of the TFBS computational model scores and distances to peak summits, independently. The maximum entropy implementation of the algorithm available in ImageJ (43) was used with default parameters.

The source code of the ChIP-eat software used to process ChIP-seq peak data sets to predict direct TF–DNA binding events is freely available at https://bitbucket.org/CBGR/chip-eat. Specifically, ChIP-eat trains a TFBS computational model and automatically defines the enrichment zone in the landscape plots to predict the underlying direct TF–DNA interactions. The identification of the enrichment zone has been applied to each TF ChIP-seq peak data set independently, allowing for the automatic detection of the thresholds that are specific to each data set with each TFBS computational model. Note that only the best hit per ChIP-seq peak has been considered to identify the enrichment zones and for all the downstream analyses.

Assessing the robustness of the enrichment zone identification

Random noise. For each ChIP-seq data set, we sampled the set of peaks using the seqtk (version 1.0) (https://github.com/lh3/seqtk) *sample* subcommand. The sequences of the sampled peaks were shuffled using the *fasta-shuffle-letters* subcommand of the MEME suite (version 4.11.4) (44) and added to the original set of ChIP-seq peaks. The automatic thresholding algorithm was applied to this new set. We tested the addition of shuffled peaks representing 10%, 25%, and 50% of the original set peaks.

Window size variability. For each ChIP-seq data set, we considered the region around the peak summit by extending with 300, 400, and 500 bp on each side using the bedtools *slop* subcommand. We considered ChIP-seq data sets where at least one TFBS was predicted within the enrichment zones obtained for all three window sizes.

Comparison with the heuristic approach to predict the enrichment zone. ChIP-eat was compared to the heuristic approach described in (25) and implemented in the TFBS_Visualization tool https://github.com/wassermanlab/TFBS_Visualization using the default parameters. The centrality of the TFBSs within the enrichment zones predicted by ChIP-eat and TFBS_Visualization was assessed using centrality *P*-value computations as described in the CentriMo tool (27). The statistical difference between the centrality *P*-values obtained with the heuristic method and ChIP-eat was assessed using a Mann-Whitney signed-rank test.

Genome coverage. The entire set of predicted TFBSs (within enrichment zones) was concatenated and then sorted using the *cat* and *sort* commands of the Unix operating system. The resulting set of locations was merged using the bedtools *merge* subcommand with default parameters. The genome coverage of the corresponding merged and non-overlapping positions was calculated as the percentage of the total number of nucleotides covered out of the total number of nucleotides in the hg38 version of the human genome.

TF–DNA binding affinity assessment with protein binding microarray data. Protein binding microarray (PBM) (45) data were retrieved from UniProbe (http://the_brain.bwh.harvard.edu/uniprobe/; (46)) for 40 TFs with available ChIP-seq data. For each ChIP-seq data set landscape plot, we extracted the DNA sequences at the sites within and outside of the predicted enrichment zone. The binding affinity of a TF to each site was computed as the median PBM intensity value of all the de Bruijn sequences containing the site sequence. The statistical difference between the distribution of PBM binding affinities from sites within and outside the enrichment zone was assessed using a two samples Mann-Whitney U test (47) implemented in the R package *stats*. A Bonferroni correction was applied to the computed *P*-values. The *P*-value density plot in Figure 3B was generated with the *density* R function with default parameters and the corresponding computed bandwidth was used to plot Supplementary Figure S10.

ChIP-exo data. ChIP-eat was applied with DiMO-optimized PFMs to the ChIP-exo data sets from (48), which were lifted over to hg38 using the liftOver tool (20). As for ChIP-seq peaks, we considered 1 001 bp regions centered around the peak summits.

ChIP-seq peaks from HOMER and BCP peak-callers. We successfully applied the HOMER (version 4.7.2) (49) and BCP (version 1.1) (50) peak-callers to 670 ENCODE ChIP-seq data sets (Supplementary Table S1). ChIP-eat was applied to the corresponding ChIP-seq peak regions with DiMO-optimized PFMs as described above. ChIP-seq peaks predicted to contain a direct TF–DNA interaction or not (using the enrichment zones) from the three peak-callers (MACS2 (51), HOMER, and BCP) were overlapped using the bedtools *intersect* subcommand. Hypergeometric tests were performed to assess the significance of the intersections using the R *phyper* function for every combination of two peak-callers with the following contingency matrix:

number of overlapping peaks **with** TFBSs from two peak-callers - 1	number of peaks **without** TFBSs from the two peak-callers
number of peaks **with** TFBSs from the two peak-callers	number of overlapping peaks from the two peak-callers

HOT/XOT regions. The high occupancy target (HOT) and extreme occupancy target (XOT) regions in all contexts were downloaded through the ENCODE data portal at http://encode-ftp.s3.amazonaws.com/modENCODE_VS_ENCODE/Regulation/Human/hotRegions/maphot_hs_selection_reg_cx_simP05_all.bed and http://encode-ftp.s3.amazonaws.com/modENCODE_VS_ENCODE/Regulation/Human/hotRegions/maphot_hs_selection_reg_cx_simP01_all.bed. ChIP-seq peaks were overlapped with the HOT/XOT regions using the bedtools *intersect* subcommand. The enrichment for overlap was assessed with a hypergeometric test using the R *phyper* function with the following contingency matrix:

number of peaks **without** TFBSs overlapping HOT/XOT regions - 1	number of peaks **with** TFBSs
number of peaks **without** TFBSs	total number of peaks

Identification of TFs with co-localized TFBSs. For each pair of distinct TFs (TF_A, TF_B), we extracted the closest TFBS associated with TF_B for each TFBS associated with TF_A and computed the geometric mean distance between midpoints of the paired TFBSs. With this approach, the geometric mean m_{AB} for the pair (TF_A, TF_B) is different from the geometric mean of the pair (TF_B, TF_A). With 232 TFs available in our analyses, we computed geometric means for 53 592 ordered pairs of TFs.

The colocalization of TFBSs for each TF pair was assessed using a Monte Carlo-based approach as follows. The number of TFBSs per TF ranged from 1 to 404 566, with 455 as the fifth percentile. We uniformly discretized the range [455, 414 172] to consider 50 TFBS set sizes (S_i for i in [1, 50]). We chose 414 172 as the maximum value to be able to compute a P-value for the set of 404 566 TFBSs. For each set size S_i, we created 500 sets of TFBSs by randomly selecting TFBSs from the total pool. Using these random sets, we computed null distributions for 500 Monte Carlo samples of geometric mean distances for each of the 2601 set size combinations. Specifically, this computation led to 2601 distributions of 500 geometric means. For the TF pair (TF_A, TF_B) with N_A and N_B TFBSs, respectively, we extracted the Monte Carlo sample of geometric mean distances M obtained from the random sets with S_A and S_B TFBSs, where $S_A = \min(S_i)$ with $S_i > N_A$ and $S_B = \min(S_i)$ with $S_i > N_B$. The empirical P-value associated with the pair (TF_A, TF_B) was computed as the number of times we observed a geometric mean smaller than m_{AB} from M over the 500 precomputed geometric means; if no smaller geometric mean was observed, the empirical P-value is defined as <0.002 (i.e. $1/500$).

Since the expected geometric mean distance increases with a decreasing number of TFBSs, this P-value computation is conservative (under-estimated significance). The obtained P-values were corrected for multiple testing using the Benjamini–Hochberg method (52), only the TF pairs with a FDR $<5\%$ were considered significant.

The detailed null distribution values can be downloaded and reproduced at https://hyperbrowser.uio.no/geirksa_sandbox/u/gsandve/h/null-distributions-for-manuscript-a-map-of-direct-tf-dna-interactions-in-the-human-genome. These computations are based on running the static methods 'ConcatenateNullDistributionsTool.execute' and 'ComputeNullDistributionForEachCombinationFromSuiteVsSuiteTool.execute' (with argument values corresponding to parameter settings annotated in the Galaxy (53) history above) in the code provided at https://hyperbrowser.uio.no/geirksa_sandbox/static/hyperbrowser/files/div/hb.zip. The source code for the comparison with null distributions is available at https://bitbucket.org/CBGR/co-binding/.

GeneMANIA. We used the GeneMANIA software (54) to extract known protein–protein interactions from the list of TFs with significant co-localized TFBSs and plot the corresponding network.

Prediction of cis*-regulatory modules.* The TFBSs predicted by ChIP-eat were sorted and merged using the bedtools *sort* and *merge* subcommands. The CREAM tool (Madani Tonekaboni *et al.*, bioRxiv, doi:10.1101/222562) was applied to the merged TFBSs to define *cis*-regulatory modules (CRMs) as genomic regions enriched for clusters of TFBSs.

GWAS trait- and disease-associated single nucleotide polymorphism enrichment analysis. We assessed the enrichment for GWAS trait- and disease-associated single nucleotide polymorphisms (SNPs) at CRMs using the *traseR* R package (version 1.10.0 (55)). CRM genomic positions were lifted over to the hg19 version of the human genome to perform the analyses. The set of SNPs (as of 30 April 2018) considered by *traseR* combined data from dbGaP (56) and NHGRI (57) as described in the corresponding bioconductor package vignette (https://bioconductor.org/packages/release/bioc/vignettes/traseR/inst/doc/traseR.pdf).

Conservation analysis. The hg38 phastCons (58) scores for multiple alignments of 99 vertebrate genomes to the human genome were retrieved as a bigWig file at http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.phastCons100way.bw. The TFBSs predicted by ChIP-eat were sorted and merged using the bedtools *sort* and *merge* subcommands. The locations overlapping CRMs were obtained using the bedtools *intersect* subcommand. The corresponding genomic locations (for all TFBSs and TFBSs in CRMs) in BED format were decomposed into 1 bp intervals using bedops v.2.4.14 (59) with the –chop 1 option. The phastCons scores at every bp were extracted with the *ex* subcommand of the bwtool (60) using the corresponding BED and phastCons bigWig files.

The UniBind web interface. All the TFBS predictions, corresponding ReMap ChIP-seq peaks, trained TFBS computational models, and CRMs are available through the UniBind database at http://unibind.uio.no/. The UniBind web interface was developed in Python using the model-view-controller framework Django. It uses MySQL to store TFBS metadata and Bootstrap as the frontend template engine. The source code is available at https://bitbucket.org/CBGR/unibind.

Statistical analyses. All statistical analyses were performed in the R environment (version 3.4.4).

RESULTS

Predicting direct TF–DNA interactions in the human genome from ChIP-seq data

Given a set of ChIP-seq peaks and a TFBS computational model such as a PWM, one can extract the best site per peak, which corresponds to the DNA subsequence of the peak with the highest score for the model. The higher the score, the stronger the computational evidence that the site is similar to TFBSs known to be bound by the TF (36). Moreover, it has been shown that the closer the site to the peak summit, the more likely it is to represent a direct TF–DNA interaction with experimental evidence from the ChIP-seq assay (25,27,30). Hence, direct TF–DNA interactions captured by ChIP-seq are enriched for high scores and small distances to the peak summits (Figure 1A,B). These characteristics have previously been used to automatically predict direct TF–DNA interactions by selecting score and distance thresholds defining these enrichment zones using a heuristic approach (25). This approach used pre-defined parameter values and was specifically designed for PWMs, but is not applicable to more recent TFBS computational models such as binding energy models (BEMs) (32), transcription factor flexible models (TFFMs) (31), and DNA shape-based models (DNAshapedTFBS) (15).

We aimed to predict direct TF–DNA interactions (TFBSs) within ChIP-seq peaks and developed the ChIP-eat software that automatically identifies the enrichment zone for any TFBS computational model. It uses a non-parametric, entropy-based algorithm originally designed to separate background/noise from foreground/signal in image processing (41) (Supplementary Figure S2). We applied this algorithm to the distributions of site scores and distance to peak summits independently to separate direct TF–DNA interaction events from other binding subtypes and ChIP-seq artifacts (Figure 1C,D; Materials and Methods). The two thresholds define the enrichment zone, which delimits the sites that are predicted as TFBSs with both experimental and computational evidence of direct TF–DNA interactions. With this approach, we automatically adjust the enrichment zone discovery specifically for each TF ChIP-seq peak data set and for each computational model. The identified enrichment zone defines the thresholds on the TFBS computational model scores and distances to the peak summits in a data set-specific manner.

We retrieved 1983 ChIP-seq peak data sets from ReMap (16), accounting for 232 TFs with a PFM available in the JASPAR database (24). Using DiMO-optimized PWMs, we compared the enrichment zones predicted by ChIP-eat with the ones obtained with the heuristic approach developed in (25). The enrichment zones predicted with ChIP-eat were more stringent than with the heuristic algorithm (Supplementary Figure S3A,B,D,E). The corresponding TFBSs predicted in the enrichment zones were more central to the peak summits with ChIP-eat than with the heuristic method as evaluated with CentriMo (27) (Supplementary Figure S3C, F). Moreover, ChIP-eat does not require any fixed values such as a predefined bin size (25) to predict the enrichment zones. Finally, ChIP-eat is not restricted to work with PWMs only and can be used with any TFBS computational model.

We applied ChIP-eat to the 1983 human ChIP-seq data sets with four types of computational TFBS models: DiMO-optimized PWMs, BEMs, TFFMs, and DNAshapedTFBS. These models were optimized for each ChIP-seq data set, independently (see Materials and Methods). In the following analyses, we focused on the predictions obtained with the DiMO-optimized PWMs (see Materials and Methods). This set of direct TF–DNA interactions (TFBSs) extracted from the enrichment zones covers ∼4% of the human genome, encompassing 8 304 135 distinct TFBS locations.

Predicted direct TF–DNA interactions are likely *bona fide* TFBSs

Robustness of the enrichment zone identification. The robustness of the method was first evaluated by applying ChIP-eat to genomic regions of ±300, 400, and 500 bp around the peak summits. The median distance threshold to the peak summit shifted from 72 bp using ±500 bp to 64 and 55 using ±400 and 300 bp, respectively. The median PWM scores thresholds were 85, 84.6 and 83.9 with ±500, 400, and 300 bp regions, respectively (see Supplementary Figure S8 for a visual representation using the 10 most frequent ChIP'ed TFs). The variability of the predicted enrichment zone when using different window sizes is similar to the variability between ChIP-seq data sets for the same TF (see below). Further, the number of predicted TFBSs within the enrichment zones were similar when using the different region sizes (Supplementary Figure S9). These analyses confirmed the robustness of the entropy-based thresholding algorithm to the window size considered. As previously used in (25), we considered the ±500 bp regions around the peak summits in the following analyses.

Considering the ChIP-seq data sets for the 10 most frequently ChIP'ed TFs, we observed that the thresholds on the PWM scores and distances to peak summits, defining the enrichment zones, were consistent between data sets for the same TF (Figure 2A,B). Namely, the median pairwise difference between PWM score thresholds for the same TF ranged from 1.7 to 3.7 and the median distance thresholds from 12 to 35 bp. As expected, the thresholds identified for distinct TFs are different (Figure 2C, D). Taken together, these results highlight that the entropy-based algorithm allows for the identification of enrichment zones specific to each TF and ChIP-seq data set, with consistent predictions between data sets for the same TF. Results were consistent with BEM, TFFM, and DNAshapedTFBS models (Supplementary Figures S4–S6).

We further evaluated the robustness of the method to noise by adding 10%, 25%, and 50% of shuffled sequences to the initial set of ChIP-seq peaks for all ChIP-seq peak data sets (see Materials and Methods). The median threshold on the distances to peak summits shifted from 73 bp in the initial set of ChIP-seq peaks to 70 bp with 10% noise, 67 bp with 25% noise, and to 63 bp when adding 50% noise. The median PWM score threshold was 85.2 for the initial set of ChIP-seq peaks and shifted to 85 when adding 10%

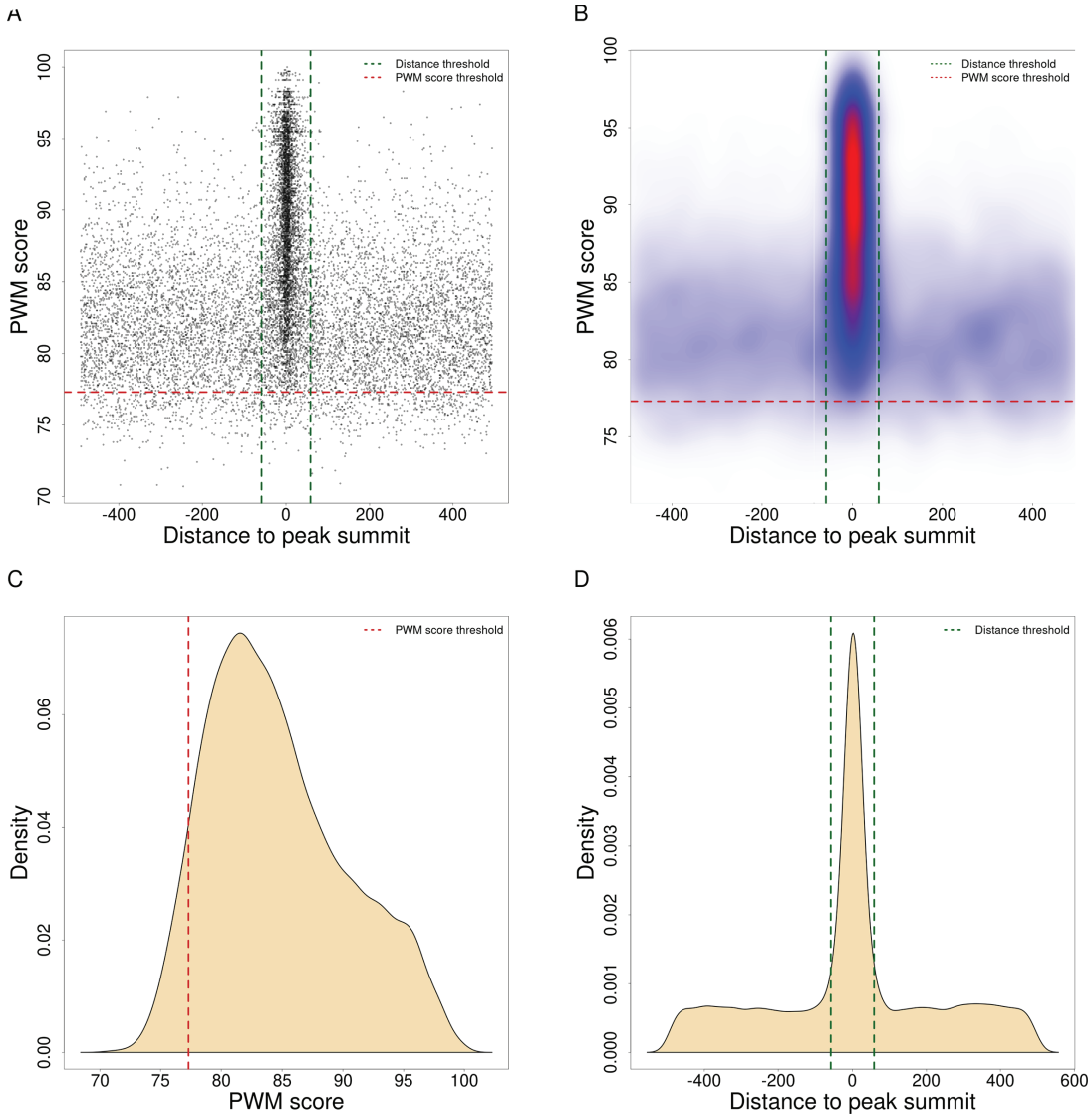

Figure 1. Automatic detection of the TFBS enrichment zone. Landscape plots (25) obtained with SRF ChIP-seq peaks using the DiMO-optimized PWM MA0083.3 from JASPAR are presented as scatter (**A**) and heatmap (**B**) plots. The enrichment zone (defined within the red and green dashed line boundaries, A-B) is automatically obtained by ChIP-eat with thresholds on PWM scores (red dashed lines; **C**) and distances to peak summits (green dashed lines; **D**). The enrichment zone provides TFBSs in ChIP-seq peaks (points in A) with supporting evidence for direct TF–DNA binding from the ChIP-seq assay (close distance to peak-summits, A-B, x-axis) and the computational model (PWM score, A-B, y-axis). Distances to peak summits in A, B and D are provided using a base pair unit.

of noise, to 84.8 when adding 25% of noise, and to 84.4 when adding 50% of noise. A visual representation for the 10 most frequently ChIP'ed TFs is available in Supplementary Figure S7. The variability of the thresholds defining the enrichment zones when adding noise is limited, within the range of variability between ChIP-seq peak data sets for the same TF (Figure 2). Taken together, these results show that the entropy-based thresholding algorithm delimiting the enrichment zones, as implemented in ChIP-eat, provides consistent results between data sets for the same ChIP'ed TF and is robust to the window sizes considered and random noise.

Validation using in vitro DNA binding affinities. To confirm *a posteriori* the high quality of our set of TFBS predictions, we assessed the TF binding affinity to DNA sequences derived experimentally from protein binding microarrays (PBM) (61). The PBM assay quantifies the binding affinity of a protein to all possible combinations of 8-mer DNA sequences. We retrieved PBM data from the UniPROBE database (46) for 40 different TFs present in our collection, corresponding to 249 ChIP-seq data sets (Supplementary Table S2). Note that the JASPAR PFMs for the ATF1, ATF3, and FOXJ2 TFs were originally derived from PBM data. For each ChIP-seq data set, we tested if the sites located in the enrichment zone presented higher binding affinity than sites outside (see Materials and Methods). The

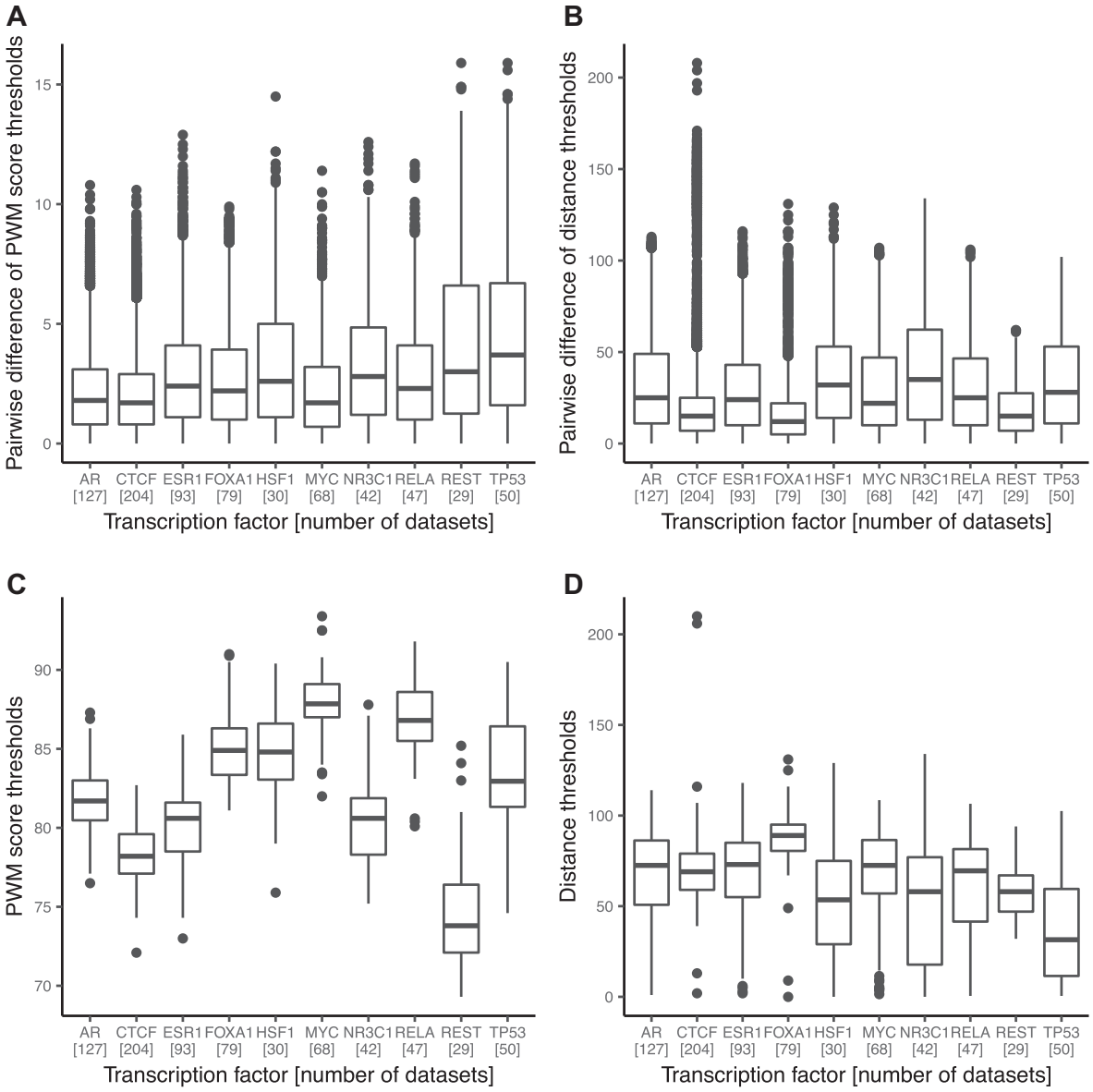

Figure 2. Assessment of the thresholds predicted by ChIP-eat across data sets. Boxplots of the pairwise differences for DiMO-optimized PWM score thresholds and distances to peak summits thresholds between ChIP-seq data sets for the same TF are provided in panels (**A**) and (**B**), respectively. Absolute variations of DiMO-optimized PWM score thresholds and distances to the peak summits within all data sets for the same TF are provided in panels (**C**) and (**D**), respectively. The ten TFs with the highest number of data sets were selected; the number of data sets for each TF is provided between brackets.

distributions of the binding affinity scores for sites within and outside the enrichment zones were compared using a Mann-Whitney U test (Figure 3A; Materials and Methods). Predicted direct TF–DNA interactions (sites within the enrichment zone) had significantly higher binding affinity than the other sites for 75% of the data sets with *P*-value <0.01 and 81% with *P*-value <0.05 (Figure 3B). Similar results were obtained when considering BEM, TFFM, and DNAshapedTFBSs computational models (Supplementary Figure S10). This analysis emphasizes that the sites predicted in the defined enrichment zones are likely to correspond to direct TF–DNA interactions.

Predicted direct TF–DNA interactions are found in high confidence ChIP-seq peaks. We hypothesized that the ChIP-seq signal at ChIP-seq peaks containing a predicted direct TF–DNA interaction were more likely to be higher than at the other peaks. To test this hypothesis, we looked at (i) the quality of the peaks based on *P*-values assigned to the peaks by the MACS2 peak-caller and (ii) the reproducibility of calling these peaks with multiple peak-callers (MACS2, HOMER, and BCP; see Materials and Methods).

We observed that the distribution of *P*-values assigned by MACS2 to the peaks containing a predicted TFBS were significantly (*P*-value < 0.01; Mann–Whitney *U* test) lower than for the rest of the peaks for 1862 (96%) data sets (Fig-

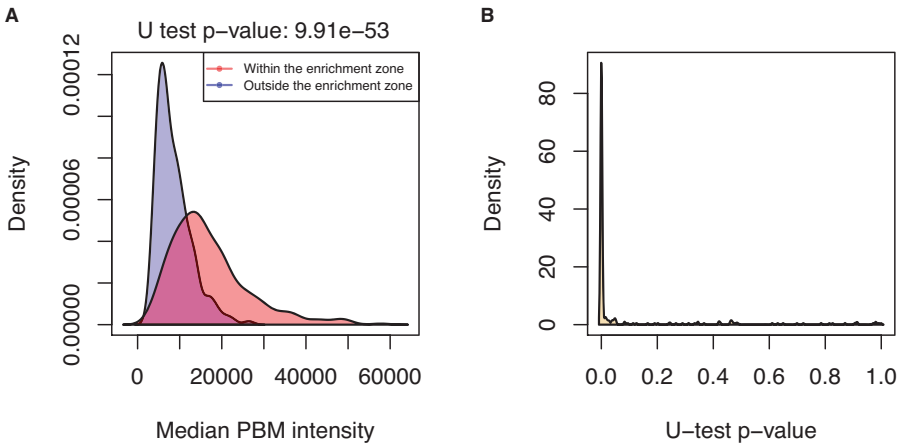

Figure 3. Binding affinity assessment for the predicted direct TF–DNA interactions. (**A**) Distribution of the median PBM intensity scores for the ENCSR000BMX GATA3 ChIP-seq data set between sequences at TFBSs (i.e. sites within the enrichment zone; in red) and sites outside the enrichment zone (in blue). (**B**) Distribution of Mann–Whitney *U* test *P*-values across the 249 data sets, showing distinct distributions of PBM intensity scores between sites within and outside the enrichment zones.

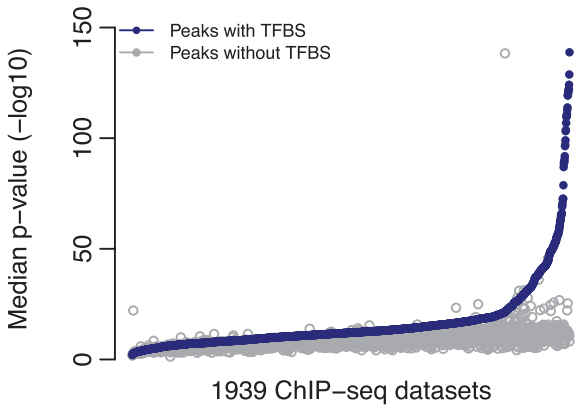

Figure 4. Quality assessment of the ChIP-seq peaks derived from direct TF–DNA interactions. Distribution of the median MACS2 *P*-values (y-axis) across all data sets. Values for peaks containing a predicted TFBS are provided in blue and values for the other peaks in grey. 1939 ChIP-seq data sets were predicted to contain direct TF–DNA interactions (x-axis).

ure 4). The other 77 data sets contained a reduced number of peaks (median of 837 compared to 18 968 for the complete set of ChIP-seq data sets), which can explain the lack of statistical significance. These results confirm that the predictions of direct TF–DNA interactions were found in ChIP-seq peaks of higher quality as assessed by MACS2.

To test ChIP-seq peak-calling reproducibility, we used two other peak-callers (HOMER and BCP) on 670 ChIP-seq data sets from ENCODE. Our choice of peak-callers was motivated by their distinct statistical approaches for peak prediction. While MACS2 and HOMER are based on an empirical model supported by a Poisson distribution, BCP uses a Bayesian approach implementing infinite-state hidden Markov models. We applied ChIP-eat to the ChIP-seq peaks to predict TFBSs. For each pair of peak-callers, we assessed whether the peaks predicted to contain a direct TF–DNA interaction were more prevalent (*P*-value < 0.01, hypergeometric test) in the set of peaks called by both peak-callers. This was observed for 63% of the data sets for MACS2 and BCP, 70% for MACS2 and HOMER, and 66% for HOMER and BCP. The data sets without significant enrichment had a median number of peaks predicted to be derived from direct TF–DNA interactions that was ∼7 fold smaller (e.g. 3358 compared to 22 499 between MACS2 and BCP) than for the data sets with significant enrichment, and a median number of peaks without TFBS ∼2 fold larger (e.g. 40 050 compared to 21 256 between MACS2 and BCP) (Supplementary Table S3). Moreover, the median quality scores assigned by the peak-callers to the peaks from the enriched data sets were significantly (*P*-value < 0.01, Mann–Whitney *U* test) higher than for the peaks in the other data sets (Supplementary Figure S11). It suggests that the data sets enriched for reproducible peaks containing predicted direct TF–DNA interactions are of better quality than the rest of the data sets.

Taken together, these results highlight that the ChIP-seq peaks in which ChIP-eat predicts direct TF–DNA interactions are of higher quality than the other peaks. Note that the ChIP-eat tool does not consider the peak quality when predicting direct TF–DNA interactions. These observations reinforce the confidence in the predicted TFBSs by ChIP-eat.

Predictions of direct TF–DNA interactions in ChIP-exo data

The ChIP-exo assay has been developed to provide a higher resolution than ChIP-seq to identify TFBSs *in vivo* (34). We aimed at assessing the performance of ChIP-eat on predicting direct TF–DNA interactions using ChIP-exo data. The ChExMix tool has recently been introduced to characterize protein-DNA binding event subtypes from ChIP-exo peak (48). ChExMix predicted different binding event subtypes for ChIP-exo data obtained for the TFs ESR1 and FOXA1, one of these subtypes corresponding to direct TF–DNA interactions (48). We applied ChIP-eat on the same ESR1 and FOXA1 ChIP-exo data sets. We compared the set of peaks identified to contain direct TF–DNA interactions

predicted by ChExMix and ChIP-eat in these two data sets. We found that 93.6% (for ESR1) and 91.3% (for FOXA1) of the peaks predicted to contain TFBSs by ChIP-eat were also predicted as direct binding events by ChExMix (Supplementary Table S4). The high overlaps between the predictions from ChExMix and ChIP-eat were confirmed by Jaccard similarity indexes of 63.7% and 68.7% for ESR1 and FOXA1, respectively. The similar results obtained with the two tools suggest that ChIP-eat, designed for the more noisy and less precise ChIP-seq data, is able to capture direct binding events from ChIP-exo data.

High-occupancy target regions are likely not derived from direct TF–DNA interactions

High-occupancy target (HOT) and extreme-occupancy target (XOT) regions are genomic regions where ChIP-seq peaks were observed for a large number of distinct ChIP'ed TFs (35,62,63). These regions are observed across species (63) and contain an unusually high frequency of ChIP-seq peaks (35,62,63). We used our set of high quality TFBS predictions to confirm that HOT/XOT regions were depleted of direct TF–DNA interactions. Indeed, we found that ChIP-seq peaks that do not contain a predicted TFBS were significantly enriched at HOT/XOT regions (odds ratio = 1.43 for HOT and 1.44 for XOT, *P*-value < 2.2e–16, hypergeometric test, Supplementary Table S5). Similar results were obtained when considering the three other computational models (BEM, TFFM, and DNAshapedTFBSs; Supplementary Table S5). This observation, combined with a previous study describing that HOT/XOT regions are likely to be derived from ChIP-seq artifacts (Wreczycka *et al.*, bioRxiv, 10.1101/107680), suggests that HOT/XOT regions are not derived from the direct binding of the ChIP'ed TFs.

Predicted direct TF–DNA interactions reveal co-binding TFs and cis-regulatory modules enriched for disease- and trait-associated SNPs

TFs are known to collaborate through specific co-binding at *cis*-regulatory modules (CRMs) to achieve their function (1,36). Hence, identifying co-binding TFs is critical to decipher transcriptional regulation of gene expression. We aimed at using our predicted direct TF–DNA interactions to reveal co-binding TFs and CRMs. We hypothesized that the distances between TFBSs of cooperating TFs are smaller than expected by chance. We tested this hypothesis for all pairs of TFs for which we predicted TFBSs (232 TFs, 53 592 pairs tested; see Materials and Methods). For each TF pair, we used a conservative Monte Carlo-based approach to compare the geometric mean of the distances between their TFBSs to the geometric mean distance expected by chance for a similar number of TFBSs randomly selected from the complete pool of TFBSs (see Materials and Methods). This approach predicted 150 pairs of TFs (accounting for 112 distinct TFs) with TFBSs closer in the genome than expected by chance (FDR < 5%; Supplementary Table S6). For 82% of the predicted TF pairs, we confirmed that the corresponding TFs physically interact using the protein-protein interaction networks from the GeneMANIA tool (54) (Supplementary Figure S12). This analysis further supports the biological relevance of the TFBSs predicted by ChIP-eat.

Next, we aimed to automatically identify CRMs, which correspond to clusters of direct TF–DNA interactions, using the clustering of genomic regions analysis method (CREAM; (Madani Tonekaboni *et al.*, bioRxiv, doi:10.1101/222562)). When considering our complete set of TFBSs, CREAM detected 61 934 CRMs in the human genome, encompassing 2 474 587 distinct TFBS locations. We found that the predicted CRMs were significantly enriched (FDR-corrected *P*-value = $2.9e^{-150}$) for disease- and trait-associated SNPs using traseR (55). Further, we observed that the TFBSs lying within the CRMs were more conserved than the TFBSs predicted outside (Supplementary Figure S13). Taken together, these results indicate a potentially functional role of the CRMs identified as clusters of direct TF–DNA interactions.

The UniBind web interface to access our collection of direct TF–DNA interactions

We catalogued the complete set of TFBS predictions from each prediction model, trained models, original ChIP-seq peaks from ReMap, and computed CRMs, and made them publicly available through UniBind at http://unibind.uio.no/. UniBind provides an interactive web interface with easy browsing, searching, and downloading for all our predictions (Figure 5). For instance, users can search for predictions for specific TFs, cell lines, and conditions.

The data can be searched by using the case insensitive search option available on the homepage. The database can be searched for each of the four TF binding models, cell/tissue type, and TF name using the 'Advanced Options', available on the homepage (Figure 5A). Search results are presented in a responsive and paginated table along with metadata information (Figure 5B), which can be clicked to view the detailed information and download TFBSs, summary plots, and ReMap ChIP-seq peaks (Figure 5C-D). All the metadata in the responsive tables can be downloaded as CSV files. UniBind displays by default the results obtained with the DiMO-optimized PWMs, but results obtained from all TFBS computational models along with the trained models are available for browsing and/or download.

DISCUSSION

To summarize, we have uniformly processed 1983 ChIP-seq peak data sets to predict high quality direct TF–DNA binding interactions in the human genome. The predictions were obtained using a non-parametric, entropy-based algorithm that automatically detects thresholds for TFBS computational model scores and distances to peak summits for each ChIP-seq data set. This new approach identified TFBSs supported by strong experimental and computational evidences for direct TF–DNA interactions. The accuracy of the predictions was *a posteriori* validated using the PBM *in vitro* assay, ChIP-exo data, and multiple ChIP-seq peak-calling algorithms. Our set of direct TF–DNA interactions confirmed that HOT genomic regions are likely not derived from direct binding of the TFs to the DNA. We used

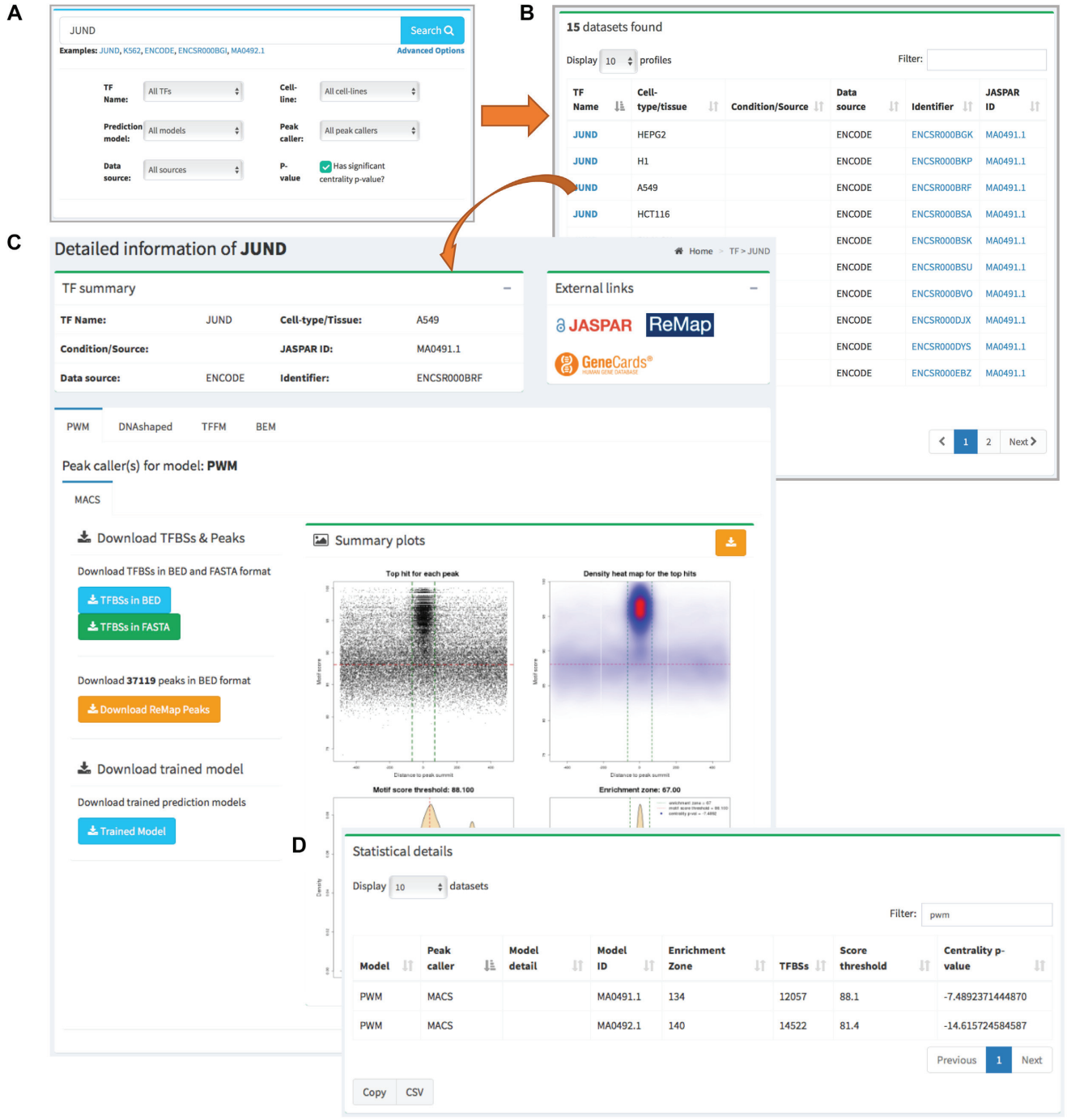

Figure 5. Overview of the UniBind user interface with interactive searching activity. (**A**) A quick and detailed search feature on the homepage. (**B**) A responsive table lists the searched data set(s), which can be clicked to view the details. (**C**) A detailed page shows the analysis for the JUND TF in cell-line A549, which is divided into sub-panels including the TF summary, external links, summary plots, and download options for each computational TFBS model. (**D**) Statistical details of the results.

our TFBSs to predict TFs with proximal binding events in the human genome, which could cooperate to achieve specific functions. Further, we defined *cis*-regulatory modules, which are clusters of TFBSs, that were enriched for disease- and trait-associated SNPs from GWAS. The complete set of predictions is publicly and freely available through the UniBind web-interface (http://unibind.uio.no/), in an effort to provide the community with an unprecedented collection of high quality direct TF–DNA interaction events in the human genome.

The output of ChIP-seq assays is generally composed of direct protein-DNA interactions, indirect binding of the protein to the DNA (through a co-binding partner), nonspecific protein binding to the DNA, and noise/bias/artifacts (4–6). Here, we specifically aimed at identifying direct TF–DNA interaction events by using an entropy-based algorithm (41). This algorithm was originally developed to discriminate between foreground and background in image processing. Hence, it assumes the presence of background (or noise) in the data. As a consequence, our approach is limited by the assumption that there is background/noise in the ChIP-seq data sets analyzed. We assume that this noise represents indirect binding of TFs, nonspecific binding, or ChIP-seq experimental artifacts. Moreover, our approach considered the best site per ChIP-seq peak (defined using TFBS computational models), which represents the best candidate. We recognize that other sites with lower scores could represent direct TF–DNA interactions. These limitations denote that our approach is stringent for the prediction of direct TF–DNA interactions, favoring specificity over sensitivity. The ChIP-seq peaks that our method did not predict to contain direct TF–DNA binding events could be further analyzed to discriminate other mechanisms for protein-DNA interactions from background noise, as proposed in the ChExMix tool established for ChIP-exo data (48).

The ChIP-eat pipeline developed for this study used four TFBS computational models to predict TF–DNA binding events. These models were specifically trained for each ChIP-seq data set to improve the quality of the predictions, as the best-performing computational model varies for different TFs or TF families (8,14,15). As a consequence, we advocate that a 'one-fits-all' TFBS prediction model is not optimal and that one should compare results from multiple models. With the predictions available through UniBind, users can assess which model would perform better for each data set. Of course, it requires to use a specific metric to compare performance. As our methods aimed at identifying enrichment zones centered around ChIP-seq peak summits, we suggest to rely on a centrality measure as implemented in the CentriMo method (27). In UniBind, we provide centrality *P*-values computed following (27) for the predictions from each model in each ChIP-seq data set. Moreover, the ChIP-eat pipeline is generalizable and users can incorporate other TFBS computational models to predict direct TF–DNA interactions and compare them to the ones already stored in UniBind.

While studies alike focus on determining where TFs directly interact with DNA, our understanding of how these TF–DNA interactions influence expression is limited. Surely, it is critical to decipher the relationship between TF–DNA interactions and transcriptional regulation (64). It is expected that a large portion of the TFBSs identified in our study are not functional, as suggested by the futility theorem (36). Nevertheless, functional TF binding events are likely to be clustered (65–68) and associated with stronger ChIP-seq peak signals (12,69). We expect that the direct TF–DNA interactions predicted in *cis*-regulatory modules and stored in UniBind are more likely to be enriched for functional events. Determining the specific set of functional TF–DNA interactions would require dedicated computational models and experiments.

DATA AVAILABILITY

Source code of the ChIP-eat software is available at https://bitbucket.org/CBGR/chip-eat and of UniBind at https://bitbucket.org/CBGR/unibind. The source code used for the identification of co-localized TFs is available at https://bitbucket.org/CBGR/co-binding. Users can browse and/or download the data through the UniBind web interface at http://unibind.uio.no/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

As research parasites (70), we would like to thank all the researchers who deposited their data. We thank Georgios Magklaras and his team for systems support, Manuela Zucknick and Andrea Cremaschi for statistical insights, Elisa Bjørgo and Ingrid Kjelsvik for management support, and Roza Berhanu Lemma, Jaime Castro-Mondragon, Oriol Fornes and Phillip Richmond for comments on the manuscript draft.

FUNDING

Norwegian Research Council (project #187615), Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., A.K., M.G.); Ph.D. fellowship from the French Ministry of Higher Education and Research (to J.C.). Funding for open access charge: Norges Forskningsråd.
Conflict of interest statement. None declared.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Mathelier,A., Shi,W. and Wasserman,W.W. (2015) Identification of altered cis-regulatory elements in human disease. *Trends Genet.*, **31**, 67–76.
3. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
4. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18602–18607.
5. Jain,D., Baldi,S., Zabel,A., Straub,T. and Becker,P.B. (2015) Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.*, **43**, 6959–6968.

6. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.*, **15**, 412.
7. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant Biol.*, **1**, 115–130.
8. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
9. 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. *J. Bioinform. Comput. Biol.*, **11**, 1340004.
10. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 375.
11. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
12. 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 Biochem. Sci.*, **39**, 381–399.
13. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.
14. 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. *Mol. Syst. Biol.*, **13**, 910.
15. 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 Syst.*, **3**, 278–286.
16. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
17. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
18. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
19. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
20. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
21. Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J.M. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
22. Kent,W.J. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
23. Fornes,O., Gheorghe,M., Richmond,P.A., Arenillas,D.J., Wasserman,W.W. and Mathelier,A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. *Sci Data*, **5**, 180141.
24. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.
25. 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**, 472.
26. Guo,Y., Mahony,S. and Gifford,D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.
27. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
28. 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*, **26**, 2622–2623.
29. 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 Res.*, **36**, 5221–5231.
30. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS One*, **5**, e11471.
31. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
32. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.
33. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
34. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
35. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
36. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
37. Patel,R.Y. and Stormo,G.D. (2014) Discriminative motif optimization based on perceptron training. *Bioinformatics*, **30**, 941–948.
38. Chiu,T.-P., Yang,L., Zhou,T., Main,B.J., Parker,S.C.J., Nuzhdin,S.V., Tullius,T.D. and Rohs,R. (2015) GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res.*, **43**, D103–D109.
39. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
40. Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S Springer*, NY.
41. Kapur,J.N., Sahoo,P.K. and Wong,A.K.C. (1985) A new method for gray-level picture thresholding using the entropy of the histogram. *Comput. Vis. Graph. Image Process.*, **29**, 140.
42. Shannon,C.E. (1948) A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, **27**, 623–656.
43. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
44. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
45. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.
46. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
47. Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
48. Yamada,N., Lai,W.K.M., Farrell,N., Pugh,B.F. and Mahony,S. (2018) Characterizing protein-DNA binding event subtypes in ChIP-exo data. *Bioinformatics*, doi:10.1093/bioinformatics/bty703.
49. 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. *Mol. Cell*, **38**, 576–589.
50. Xing,H., Mo,Y., Liao,W. and Zhang,M.Q. (2012) Genome-wide localization of protein-DNA binding and histone modification by a

Bayesian change-point method with ChIP-seq data. *PLoS Comput. Biol.*, **8**, e1002613.
51. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
52. Hochberg,Y. and Benjamini,Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.*, **9**, 811–818.
53. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Cech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
54. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W20.
55. Chen,L. and Qin,Z.S. (2015) traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics*, **32**, 1214–1216.
56. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
57. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
58. Siepel,A. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
59. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J., Thomas,S. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.
60. Pohl,A. and Beato,M. (2014) bwtool: a tool for bigWig files. *Bioinformatics*, **30**, 1618–1619.
61. Berger,M.F. and Bulyk,M.L. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.*, **338**, 245–260.
62. Xie,D., Boyle,A.P., Wu,L., Zhai,J., Kawli,T. and Snyder,M. (2013) Dynamic trans-Acting factor colocalization in human cells. *Cell*, **155**, 713–724.
63. Boyle,A.P., Araya,C.L., Brdlik,C., Cayting,P., Cheng,C., Cheng,Y., Gardner,K., Hillier,L.W., Janette,J., Jiang,L. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.
64. Whitfield,T.W., Wang,J., Collins,P.J., Christopher Partridge,E., Aldred,S., Trinklein,N.D., Myers,R.M. and Weng,Z. (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, **13**, R50.
65. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-André,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
66. Wilczyński,B. and Furlong,E.E.M. (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, **6**, 383.
67. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
68. He,Q., Bardet,A.F., Patton,B., Purvis,J., Johnston,J., Paulson,A., Gogol,M., Stark,A. and Zeitlinger,J. (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat. Genet.*, **43**, 414–420.
69. Fisher,W.W., Li,J.J., Hammonds,A.S., Brown,J.B., Pfeiffer,B.D., Weiszmann,R., MacArthur,S., Thomas,S., Stamatoyannopoulos,J.A., Eisen,M.B. *et al.* (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21330–21335.
70. Longo,D.L. and Drazen,J.M. (2016) Data sharing. *N. Engl. J. Med.*, **374**, 276–277.

TF-regulons: identifying direct targets of transcription factors

Gheorghe, M. and Mathelier, A.

Manuscript

V

Identifying key TFs driving ER positive and ER negative breast cancer subtypes

Gheorghe, M., Tekpli X., Fleischer, T., Kristensen, V., Mathelier, A.

Manuscript

VI