Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences of the

Ruprecht - Karls - University

Heidelberg

Presented by:

M.Sc. Guido Barzaghi

born in: Avellino, Italy

Oral examination: 18-06-2024

The contribution of transcription factors to chromatin accessibility at single molecule resolution

Referees:

Prof. Dr. Henrik Kaessmann

Dr. Wolfgang Huber

From a curious mind with a terminal.

Summary

Mammalian genomes evolved to be very large and complex entities. The ca. 20,000 genes there encoded are precisely regulated in time, space and dosage to instruct for the development and the environmental adaptation of an organism. Gene expression, therefore, invokes a densely interconnected network of regulatory mechanisms in which sequence-specific transcription factors (TFs) occupy a central position. TFs characteristically bind to short DNA recognition motifs enriched at *cis*-regulatory elements (CREs), such as enhancers and promoters. Thereon, they tune gene expression by recruiting co-factors such as the basal transcription machinery and chromatin modification enzymes. In eukaryotes, the nucleosomes making up chromatin pose a physical barrier for the binding of TFs to DNA.

Understanding how TFs overcome this barrier at CREs is key to disentangle mammalian gene regulation. For example, certain TFs, pioneers, have the unique ability to establish accessibility *de novo* at inactive CREs. Whether they are equally pivotal for the maintenance of chromatin accessibility at active CREs is not clear. It is also not clear how important is the combinatorial assembly of multiple TFs and how TF- and CRE-specific this competition against nucleosomes is. Finally, it is to be determined whether CREs can function modularly or not.

In this project, I quantified TF-nucleosome competition *in-vivo* at CREs across the mouse genome. I did so at the single molecule level, at high coverage and near-nucleotide resolution using SMF, single molecule footprinting. SMF uses methylation footprinting to measure the frequency of CRE accessibility in a cell population, a proxy for the CRE- and TF-specific rate at which nucleosomes are outcompeted. Through its near-nucleotide resolution SMF captures footprints for TFs and nucleosomes, a leeway to investigate their interfacing.

To capitalize on this resolution, I developed two computational tools. The first, *SingleMoleculeFootprinting*, for the robust single molecule methylation calling, is currently distributed on Bioconductor. The second, *FootprintCharter*, currently under development, is an unsupervised tool for mapping and quantifying TF and nucleosome footprints.

To isolate the contribution of individual TF instances to accessibility at CREs, I quantified the effect of perturbing their binding across the mammalian genome. To that end, I exploited the natural genetic variation among different mouse species crossed into F1 lines. Such

sequence variation often hinders the binding of one or few TFs at CREs and, therefore, their ability to outcompete nucleosomes. This highlights their effective contribution to CRE accessibility.

Among others, I observed a large degree of heterogeneity of TF-nucleosome competition across the mouse genome, with transcriptional enhancers showing the highest degree of variability. TF-nucleosome competition appears to be TF-specific, with Ctcf, Rest, Banp, Nrf1 and Nfya being most frequently successful. Pioneer TFs such as Klf4, Oct4 and Sox2 display a lesser individual contribution but assemble in context-dependent ways to outcompete nucleosomes. Nevertheless, most TF instances at CREs appear disposable for the maintenance of chromatin accessibility, highlighting pervasive CRE robustness. Finally, given how they react to perturbations, CREs seem to function in an all-or-nothing fashion rather than modularly.

With the work detailed in this dissertation, I shed some light on the contribution of transcription factors to chromatin accessibility at CREs. This work feeds into disentangling the hugely complex and vastly uncharted network of interactions that regulate eukaryotic gene expression.

Zusammenfassung

Die Genome von Säugetieren haben sich im Laufe der Evolution zu sehr großen und komplexen Einheiten entwickelt. Die ca. 20.000 darin kodierten Gene werden zeitlich, räumlich und in ihrer Dosierung genau reguliert, um die Entwicklung und die Umweltanpassung eines Organismus zu steuern. Die Genexpression beruht daher auf einem dichten Netzwerk von Regulationsmechanismen, in dem sequenzspezifische Transkriptionsfaktoren (TFs) eine zentrale Rolle einnehmen. TFs binden typischerweise an kurze DNA-Erkennungsmotive, die an cis-regulatorischen Elementen (CREs), wie Enhancern und Promotoren, angereichert sind. Dort steuern sie die Genexpression, indem sie Co-Faktoren wie die basale Transkriptionsmaschinerie und Enzyme zur Chromatinmodifikation rekrutieren. In Eukaryoten stellen die Nukleosomen, die das Chromatin bilden, eine physische Barriere für die Bindung von TFs an die DNA dar.

Zu verstehen, wie TFs diese Barriere an CREs überwinden, ist essenziell, um die Genregulation bei Säugetieren zu entschlüsseln. Bestimmte TFs, die Pioniere, haben beispielsweise die einzigartige Fähigkeit, Chromatin-Zugänglichkeit an inaktiven CREs erstmals zu etablieren. Es ist jedoch unklar, ob sie für die Aufrechterhaltung der Chromatin-Zugänglichkeit an aktiven CREs ebenso entscheidend sind. Fraglich ist auch, wie wichtig das kombinatorische Zusammenwirken mehrerer TFs ist und wie TF- und CRE-spezifisch die Konkurrenz mit Nukleosomen ist. Schließlich bleibt zu klären, ob CREs modular funktionieren können oder nicht.

In diesem Projekt habe ich die TF-Nukleosomenkonkurrenz *in vivo* an CREs im gesamten Mausgenom quantifiziert. Dies erfolgte auf der Ebene einzelner DNA-Moleküle, mit hoher Sequenziertiefe und nahezu mit Nukleotidauflösung mittels SMF (Single Molecule Footprinting). SMF verwendet Methylierungs-Footprints, um die CRE-Zugänglichkeit in einer Zellpopulation quantitativ zu bestimmen. Dies dient als Indikator für die CRE- und TF-spezifische Verdrängungsrate von Nukleosomen. Durch die hohe Auflösung auf der Ebene einzelner Nukleotide erfasst SMF die Footprints sowohl der TFs als auch der Nukleosomen und ermöglicht so die Untersuchung ihrer Wechselwirkungen.

Um diese Auflösung zu nutzen, habe Ich zwei bioinformatische Programme entwickelt. Das erste, SingleMoleculeFootprinting, für robustes Methylierungs-Calling einzelner DNA-

Moleküle, ist derzeit auf Bioconductor verfügbar. Das zweite, FootprintCharter, das derzeit entwickelt wird, ist ein Programm für das Mapping und die Quantifizierung von TF- und Nukleosom-Footprints basierend auf unsupervised Machine-Learning-Algorithmen.

Um den Beitrag einzelner TFs zur Zugänglichkeit von CREs zu bestimmen, habe ich die Auswirkungen von Perturbationen ihrer Bindung im Säugetiergenom quantifiziert. Zu diesem Zweck nutzte ich die natürliche genetische Variation zwischen verschiedenen Mausarten, die in F1-Linien gekreuzt wurden. Eine solche Sequenzvariation beeinträchtigt häufig die Bindung eines TFs an CREs und damit die Fähigkeit, Nukleosomen zu verdrängen. Dies unterstreicht den direkten Einfluss von TFs auf die Chromatin-Zugänglichkeit von CREs.

Unter anderem beobachtete ich ein hohes Maß an Heterogenität der TF-Nukleosomen-Konkurrenz im gesamten Mausgenom, wobei Enhancer den höchsten Grad an Variabilität aufweisen. Die TF-Nukleosomen-Konkurrenz scheint TF-spezifisch zu sein, wobei Ctcf, Rest, Banp, Nrf1 und Nfya die höchste Verdrängungsrate von Nukleosomen aufweisen. Pionier-TFs wie Klf4, Oct4 und Sox2 leisten einen geringeren individuellen Beitrag, wirken aber kontextabhängig zusammen, um Nukleosomen zu verdrängen. Dennoch scheinen die meisten TFs für die Aufrechterhaltung der Chromatin-Zugänglichkeit an CREs vernachlässigbar zu sein, was die weit verbreitete Robustheit von CREs unterstreicht. Und schließlich scheinen CREs angesichts der Art und Weise, wie sie auf Perturbationen reagieren, eher modular nach dem Alles-oder-Nichts-Prinzip zu funktionieren.

Mit den in dieser Dissertation beschriebenen Arbeiten leiste ich einen Beitrag zur Aufklärung des Einflusses von Transkriptionsfaktoren auf die Chromatinzugänglichkeit an CREs. Damit trägt diese Arbeit dazu bei, das äußerst komplexe und noch weitgehend unerforschte Netzwerk von Wechselwirkungen zu entschlüsseln, das die eukaryotische Genexpression reguliert.

Acknowledgements

I wish to thank all the people who contributed to this work and, more importantly, to my journey through this Ph.D.

A special mention goes to my supervisors and mentors Arnaud Krebs and Judith Zaugg, without whom today I would still be nothing more than a man with a terminal. It is thanks to them that today I can call myself a mind with a terminal…and a curious one!

I wish to thank the members of the Krebs laboratory for their support, in all of its forms. No group of people had ever gone from the abyssal depths of the Earth to the top of the tallest tree with the same great spirit.

I wish to thank "*All the single ladies*" of the Zaugg laboratory for all of their support, energy and positivity. The ones who were *"Born to be wild"* were also fantastic!

I wish to thank my TAC members, for having contributed with plenty of suggestions and stimulating discussions to this project and my professional growth.

I wish to thank with all my heart all the friends I met during these years who made my time in Heidelberg one filled with fun and joy. You have been my shell, my stage, my spice.

I wish to thank my family: "Mamma, Papà, Luca, vi giuro: questo è l'ultimo diploma. Da domani iniziamm'a faticà". [tr. "Mom, Dad, Luca, I swear: this is the last degree. From tomorrow I'll start working"]

Finally, to the most special of all…my BBABY! I can't wait for us to start our next chapter. It will be fireworks.

Table of Contents

List of Contributors

Here is a list detailing the people who contributed to the work detailed in this dissertation.

Rozemarijn Kleinendorst

Rozemarijn Kleinendorst, Laboratory Officer in charge for the Krebs laboratory, contributed to the production of the data necessary for the research project reported in this dissertation. In particular, she produced the single molecule footprinting datasets for F1 mouse ESC lines.

Laura Moniot-Perron

Laura Moniot-Perron, Postdoctoral Fellow for the Krebs laboratory, contributed to the production of the data necessary for the research project reported in this dissertation. In particular, she produced the single molecule footprinting datasets for Sox2 dTAG mouse ESC lines.

Duncan Odom laboratory

The Odom Laboratory provided the Bl6xCast F1 mouse ESC line necessary for the research project reported in this dissertation.

List of publications

- **G. Barzaghi** et al. Systematic dissection of the cooperative assembly rules used by transcription factors to establish chromatin accessibility. In preparation.
- Rauluseviciute, ..., **G. Barzaghi**, et al. 2023. Identification of transcription factor co-binding patterns with non-negative matrix factorization. BiorXiv. https://doi.org/10.1101/2023.04.28.538684
- C. Kjær, ..., **G. Barzaghi**, et al. 2023. Differential expression of the β3 subunit of voltage gated Ca2+ channel in mesial temporal lobe epilepsy. Molecular Neurobiology. https://doi.org/10.1007/s12035- 023-03426-4
- E. Kreibich, ..., **G. Barzaghi**, et al. 2023. Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. Molecular Cell. https://doi.org/10.1016/j.molcel.2023.01.017
- **G. Barzaghi** & R. Kleinendorst, et al. 2021. Genome-wide quantification of transcription factor binding at single-DNA-molecule resolution using methyl-transferase footprinting. Nature Protocols. https://doi.org/10.1038/s41596-021-00630-1
- C. Sönmezer, ..., **G. Barzaghi**, et al. 2021. Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. Molecular Cell. https://doi.org/10.1016/j.molcel.2020.11.015
- P. Rifes, ..., **G. Barzaghi**, et al. 2020. Modelling neural tube development by differentiation of human embryonic stem cells in a microfluidic WNT gradient. Nature Biotechnology. https://doi.org/10.1038/s41587-020-0525-0
- C. Kjær, **G. Barzaghi**, et al. 2019. Transcriptome analysis in patients with temporal lobe epilepsy. Brain. https://doi.org/10.1093/brain/awz265

List of abbreviations

ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
Bl6	Mus musculus domesticus
CA	chromatin accessibility
Cast	Mus musculus castaneus
ChIP-seq	chromatin immunoprecipitation sequencing
ChIP-nexus	ChIP with nucleotide resolution using exonuclease digestion
CRE	*cis*-regulatory element
CUT&RUN	cleavage under targets and release using nuclease
DNase-seq	DNase I hypersensitive sites sequencing
dTAG	Degradation tag
DKFZ	Deutsche Krebsforschungszentrum
DNMT	DNA methyl-transferase
F1	First filial generation
mESC	mouse embryonic stem cell
NDR	nucleosome depleted region
OSKM	Oct4, Sox2, Klf4, c-Myc
PIC	Pre-initiation complex
PolII	RNA polymerase II
SMF	single molecule footprinting
Spret	Mus spretus
TF	transcription factor
TFBS	transcription factor binding site
TSS	transcription start site
TKO	triple knock-out
WT	wild type

1 | Introduction

A fundamental concept in biology that has very few, albeit non negligible, exceptions is that every cell of a given organism has the same genome. Another fundamental concept, this time one which is being challenged by the complexity of our experimental observations, is that the genome encodes for the information necessary to the formation of the whole organism as well as to its adaptation to a changing environment. This seemingly simple statement implies the existence of an ungraspable amount of information encoded in a molecule, DNA, which must not only reliably store and pass on this information, but also allow for its dynamic and timely access. The study of the molecular mechanisms for the access to genomic information makes up the field of gene regulation.

Research on gene regulation was famously kicked off by François Jacob and Jacques Monod in 1961. They dissected the *Escherichia coli lac* operon, a system controlling the expression of the three genes necessary for the metabolism of lactose. The importance of this discovery was sealed shortly after, in 1965, with a Nobel Prize in Physiology, perhaps not so much for the medical relevance of this organism. The groundbreaking nature of this finding is that a very simple regulatory circuit, which includes a single repressor protein (*lacI*), is responsible for the dramatic adaptation to the sudden unavailability of glucose, the primary source of sustenance of *E.coli* (Jacob and Monod 1961).

Since the breakdown of the *lac* operon, the field has been tackling the much more intricate genomes of higher eukaryotes including mammals, the object of attention of this dissertation. In this case, genes are regulated to produce very complex cellular phenotypes that contribute, on one hand, to the formation of an organism during development and, on the other, to its response to environmental stimuli. Of the ca. 20,000 protein coding genes encoded in the human genome (Nurk et al. 2022; Venter et al. 2001), only a fraction are expressed in a given cell type, and they are so with high temporal and dosage precision. A myriad of mechanisms has been demonstrated to contribute to this regulatory machinery, including all possible biochemical species present in a living cell as well as their physical and mechanical properties.

Nevertheless, researchers have merely scratched the surface of what can be understood and, even less so, controlled about mammalian gene regulation. This dissertation, and the last five years of work, are my proud and humble contribution to the study of gene regulation.

1.1 | Eukaryotic genomes and their regulatory sequence elements

A surprising fact about the eukaryotic genome is that the order of magnitude of the number of protein coding genes is very conserved across species: ca. 20,000 in human and mouse, ca. 14,000 in fruit fly and ca. 6,000 in yeast. On the other hand, the dimension of the genome varies greatly. Most of this variation is thus associated with non-coding DNA. Less than 2% of the human genome is protein coding (International Human Genome Sequencing Consortium et al. 2001) while the estimates for what percentage has a regulatory function greatly vary from a conservative 8% (The ENCODE Project Consortium 2012) to 20% (Chi 2016) or even 40% (Stamatoyannopoulos 2012).

Several classes of genomic sequences with associated regulatory functions have been robustly identified to this date. Their classification follows both the specific function they carry (activation, repression, insulation, etc.) as well as their distance from the genes they regulate (proximal or distal).

The best studied regulatory sequences across the tree of life are proximal promoters. These elements surround the transcription start sites, TSSs, of genes and are essential to their transcription. Promoters contain several sequences, or promoter elements, that guide the directional assembly of the basal transcriptional machinery. One of these elements is the TATA box, denoted by the consensus sequence TATA(T/A)A(T/A), to which the general transcription factor TBP binds (I will discuss this in further detail in the section 1.2.1). Many promoters, so-called TATA-less, contain a much less conserved and evasive version of the TATA box (Rhee and Pugh 2011, 2012). Other sequence elements enriched at promoters in the vicinity of the TSS include BRE (TFIIB recognition element), INR (initiator element) and DPE (downstream promoter element). Elements enriched more upstream include the CCAAT box, located 60-100 bp upstream of the TSS, which mediates the binding of the nuclear factor Y (NF-Y) transcription factors and the GC box, ca. 110 bp upstream the TSS and binding motif for the TF Sp1 (Haberle and Stark 2018; Roeder 2019).

A large fraction of promoters for housekeeping genes, those genes that are constitutively expressed in most cell lines and are indispensable for the survival of the cell, are often embedded in so-called CpG islands, CGIs. CGIs are 1-2 kb wide genomic regions which are actively kept unmethylated, hence functionally active, contrary to most other instances of

CpG dinucleotides. This is thought to have led to their increased sequence conservation as compared to methylated cytosines which tend to easily mutate into thymines by spontaneous deamination (Deaton and Bird 2011; Illingworth and Bird 2009).

Additional regulatory elements have been identified that can exert their function even from considerable genomic distances. These include transcriptional enhancers, silencers and insulators. Enhancers are clusters of transcription factor motifs that can act from very large distances, in any orientation, and be located upstream or downstream of the gene they regulate or even within the gene itself (Pennacchio et al. 2013; Shlyueva, Stampfel, and Stark 2014). A prominent example is the enhancer regulating the promoter of the interferon β in response to viral infection. As I will discuss in further detail in the section 1.2.4.2, this enhancer is a beautiful example of the clockwork-like assembly of a transcription factor cluster (Panne 2008). Another notable case is the locus control region, LCR, 10-20 kb upstream of the β globin genes. This LCR drives the highly cell-specific expression of the beta globin gene cluster (Q. Li et al. 2002).

Contrary to enhancers, silencers harbor motifs for transcriptional repressors and serve to inhibit gene transcription. The expression of numerous genes, including the T cell surface protein CD4, is regulated by transcriptional silencers (Segert, Gisselbrecht, and Bulyk 2021). Finally, insulators, function as genetic boundaries, blocking the spread of both enhancer and silencer activities, thereby ensuring that these regulatory elements affect only their target genes.

Together and in way which are very much still to break down, these elements enable the complex but precisely concerted regulation of eukaryotic genomes.

1.2 | The regulation of eukaryotic transcription

As detailed in the previous chapter, eukaryotic genomes are large, complex, and they need to be regulated precisely enough to instruct for the formation of an individual during its life span, as well as for its adaptive responses to a changing environment. Given the large diversity of instructions that are needed to allow for such processes, it is not surprising that the mechanisms that regulate the actuation of such instructions, or the expression of such genes, are even more intricate. This chapter describes the components that mediate and regulate eukaryotic gene expression. I will first describe the transcriptional process itself, to then highlight our best-established knowledge on the *cis*- and *trans*-acting regulators that control it. I will describe how eukaryotes experience the added complexity of having their very large genome tightly packed into chromatin and how such compaction can prevent transcriptional regulators from interacting. Intriguingly, this compaction can be regulated enzymatically as well as physically through particular behaviors of the transcriptional regulators which I will describe in the last section of this chapter.

1.2.1 | The mechanisms of basal eukaryotic transcription

The step of the gene expression regulation cascade for which we have the best understanding among all is basal eukaryotic transcription (Cheung and Cramer 2012; Orphanides and Reinberg 2002). There, each successful transcriptional cycle produces an mRNA that can be translated into a functional protein, hence achieving the ultimate goal of the regulatory cascade. Broken down into the assembly of the transcriptional machinery, transcription initiation, elongation and termination, eukaryotic transcription can be regulated, and halted if needed, at several stages. The following sections give an overview to the process of basal eukaryotic transcription and highlight the points where regulatory activity can take place.

1.2.1.1 | RNA polymerases and the assembly of the general transcription factors at RNA polymerase II promoters

Eukaryotes have three RNA polymerases (Pol). They are structurally similar and share some subunits (Cramer et al. 2008), but they transcribe different types of genes. RNA PolI transcribes rRNAs 5.8S, 18S and 28S. RNA PolII transcribes all protein coding genes plus genes for miRNAs, snRNAs, snoRNAs and siRNAs. RNA PolIII transcribes tRNAs and genes for rRNA 5S, some snRNAs and other small RNAs (Vannini and Cramer 2012).

Gene transcription at PolII promoters requires the assembly of the general transcription factors, GTFs, to load PolII and initiate transcription. Recent *in-vitro* studies (Y. He et al. 2013, 2016) reconstructed at high resolution the order of molecular events leading to PolII loading and transcription initiation. First, the GTF TFIID, transcription factor for polymerase II D, binds to the TATA box, 25-30 nucleotides upstream of the transcription start site, TSS. TFIID is composed of TBP, TATA-binding protein, as well as several TBP-associated factors, TAFs. TAFs increase promoter selectivity, especially at TATA-less promoters. The binding of TFIID imposes a strong physical distortion in the DNA, which is thought to serve as reference point to mark active promoters. Secondly, TFIIB binds to its recognition element (BRE) upstream of TFIID. Such asymmetric binding helps to determine the right direction of transcription (Danino et al. 2015; Juven-Gershon et al. 2008). TFIIB is followed by the assembly of TFIIA, RNA PolII in complex with TFIIF, TFIIE and TFIIH. This completes the pre-initiation complex, PIC. TFIIH, the biggest and most complex of the GTFs, initiates its ATP-dependent helicase activity to unwrap the double helix of DNA at the TSSs. This exposes the DNA strand which will serve as a template for transcription. Finally, TFIIH mediates the release of PolII from the GTFs and therefore the initiation of transcription. It does so by phosphorylating the C-terminal domain, CTD, of the PolII subunit Rbp1. Specifically, the CTD is composed of 52 tandem repeats of seven amino acids, the fifth of which is the serine phosphorylated by TFIIH. Following PolII release, most GTFs detach from DNA in order to start a new transcriptional cycle starting again with PolII loading (Haberle and Stark 2018).

Transcription initiation *in-vivo* is severely complicated by the presence of chromatin (R D Kornberg 2007). Because of this physical barrier, a much greater number of proteins is required

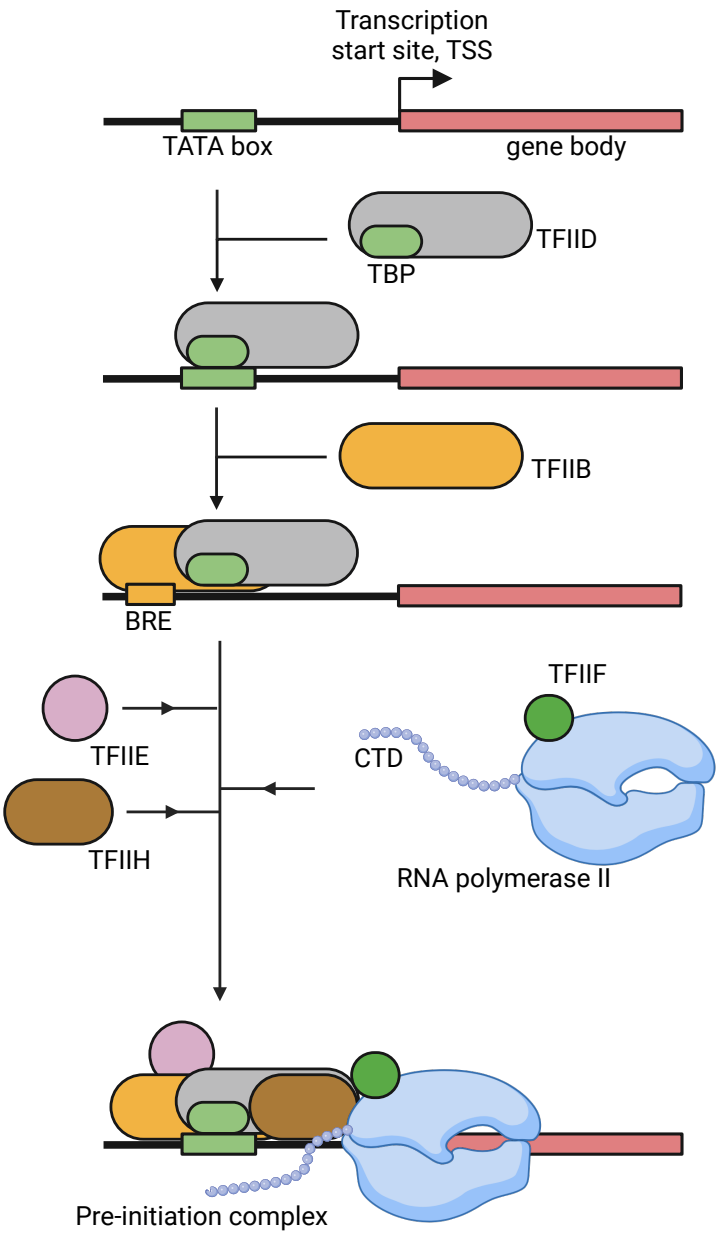

Figure 1.1. Assembly of the basal transcription machinery. First, TFIID binds to the TATA box through its subunit TBP, flagging active promoters. Secondly, TFIIB occupies the BRE element determining the direction of transcription. Then, TFIIF, TFIIE, and TFIIH, assemble together with RNA polymerase II to complete the pre-initiation complex (PIC). TFIIH is the largest and most complex of the general transcription factors (GTFs). With its ATP-dependent DNA helicase activity, it mediates the release of RNA polymerase II from the GTF and the initiation of transcription.

to effectively transcribe genes. A large class of such protein is referred to as transcriptional activators, or transcription factors (TFs). They are one of the main ways cells regulate gene expression and, being the focal point of this dissertation, they will be discussed at length in the section 1.2.2. TFs bind to DNA at specific sequences and allow the productive recruitment of PolII. Because TFs can regulate this process from very large genomic distances, a large macromolecular complex, the Mediator, allows the interactions of TFs with PolII and the GTFs, as well as with additional cofactors such as chromatin remodeling complexes, which facilitate the process.

1.2.1.2 | Transcription initiation

Owing to the helicase activity of TFIIH, RNA PolII is found is the so-called “open complex” state, in which it is bound to an opened region of DNA ca. 14 bp long referred to as the “transcription bubble”. At this stage, DNA is kept unwrapped by PolII regions called the “rudder”, “lid” and “zipper”. Notably, RNA PolII is the only RNA polymerase associated with the ATP-dependent unwrapping of DNA. Every other enzyme across the tree of life, with the exception of bacterial genes depending on the GTF σ^{54}, undergoes the spontaneous formation of the “open complex” (Nogales, Louder, and He 2017).

At this point transcription begins with a chain of energetically favorable reactions in which new incoming ribonucleoside triphosphates (NTPs) are added to the 3’ end of the growing RNA molecule. The latter is stabilized and activated for each reaction by two Mg^{2+} ions placed at the active site of RNA PolII.

Often, RNA PolII fails to produce a complete RNA molecule on the first attempts and enters the so-called “abortive initiation”, a series of abortive cycles of synthesis which results in the production of short RNA molecules, typically 2-9 nucleotides long, quickly degraded. The extent of abortive initiation can vary across promoters and it represents a potential regulatory step.

Once RNA PolII transcribed more than 9-11 nucleotides into RNA, it gets released from the promoter in a process referred to as “promoter clearance” and it begins to productively elongate in the form of “elongating complex” (Nechaev and Adelman 2011; Zhan et al. 2023).

1.2.1.3 | Transcription elongation

The elongating complex is very stably bound to DNA owing to the strength with which the two large RNA PolII subunits engage DNA. The elongation process is favored by the energy released by the polymerization and it proceeds with a speed of ca. 20-50 nucleotides per second.

Despite this efficiency, RNA PolII elongation is frequently stalled along the length gene body due either to obstructions imposed by chromatin or by specific sequence features (Gurard-Levin, Quivy, and Almouzni 2014). A notable example of the latter is termed “promoter

proximal pausing" which happens ca. 30-60 nucleotides downstream of the TSS. Promoter proximal pausing is favored by the negative elongation factors NELF and DSIF, which stabilize the paused states of RNA PolII. Promoter proximal pausing is a very widespread phenomenon in metazoan genomes and it represent an important regulatory step for gene transcription.

The action of elongation factors allows RNA PolII to resume elongation. These include, among others, the proteins ELL and P-TEFb which are part of the Super Elongation Complex. Additional proteins, including the histone chaperones FACT (facilitates chromatin transcription), Asf1 and Spt6 support the elongation process by mediating the removal and reassembly of nucleosome particles to allow the passage of RNA PolII through chromatinised DNA (Jonkers and Lis 2015; Selth, Sigurdsson, and Svejstrup 2010).

1.2.1.4 | Transcription termination

For most mRNAs a chain of adenosines, polyA, is added to the 3' of the transcript. This event is coupled with the termination of transcription, during which, the nascent polyadenylated transcript is released by RNA PolII. This instead continues transcribing a few kilobases after the transcription end site, TES, before releasing the DNA template and becoming available for the start of a new transcription cycle (Kuehner, Pearson, and Moore 2011; Porrua, Boudvillain, and Libri 2016; Santangelo and Artsimovitch 2011).

1.2.2 | Transcription factors regulate gene expression at *cis*-regulatory elements

As anticipated in the section 1.2.2.1, *cis*-regulatory elements are classified based on their functional effect on transcription regulation as well as their distance to the transcription start site of the gene they regulate. Here, I will further expand on additional experimental and computational methods that have been employed for the identification and classification of CREs, with particular focus on transcriptional enhancers. Enhancers have historically been far more elusive than promoters due to their lack of enrichment for sequence elements at fixed positions and due to the increased complexity of the regulatory interactions they engage with.

In fact, enhancers can lie anywhere on the spectrum of potency of promotion of transcriptional activation and do engage, in a cell-specific manner, with multiple genes and multiple additional CREs potentially lying at great genomic distances. For these reasons, the identification and profiling of enhancers has historically required *ad-hoc* experimental and computational approaches, of which I will give a brief, and by no means comprehensive, overview.

1.2.2.1 | Identification and classification of cis-regulatory elements

Within the initial phases of the genomic era, a diverse array of bulk assays has been systematically employed to explore transcriptional enhancers, identifying them through several biochemical features. Candidate CREs were shortlisted based on their association with chromatin accessibility using techniques such as DNase-seq, ATAC-seq, MNase-seq, and FAIRE-seq (Boyle et al. 2008; Buenrostro et al. 2013; Giresi et al. 2007; Schones et al. 2008). Furthermore, the perimeter of the genomic loci enriched for sequence-specific transcription factor binding, as well as for the activating chromatin marks H3K27ac and H3K4me1, was delineated using ChIP-seq (Johnson et al. 2007; Robertson et al. 2007) and its refined variants like ChIP-exo (Rhee and Pugh 2011). The 3D interactions necessary for enhancers to instruct the transcriptional activity at gene promoters have been revealed using various chromatin conformation capture assays, including 3C and Hi-C (Lieberman-Aiden et al. 2009). Finally, these genomic loci were probed for their characteristic production of short, transient transcripts, as detected by nascent RNA sequencing technologies such as GRO-seq, PRO-seq, or CAGE (Mahat et al. 2016; Natoli and Andrau 2012; The FANTOM Consortium et al. 2014).

The computational integration of such wealth of data through frameworks such as chromHMM generated over a million candidate regulatory elements, including enhancers, based on the concerted biochemical proprieties exhibited in their endogenous genomic context (Ernst and Kellis 2012).

More recently, single cell assays are refining the annotations achieved through their above-mentioned bulk counterparts. In particular, they are revealing the specificities of CRE usage in rarer developmental stages, in very complex and heterogeneous tissues and in homogeneous cell populations with prominent cell-to-cell variation (Buenrostro, Wu, Litzenburger, et al. 2015; Cusanovich et al. 2015; Ramani et al. 2017).

Another approach is to probe a battery of candidate DNA sequences for their ability to enhance the levels of gene transcription through so-called massively parallel reporter assays, MPRAs. There, candidate regulatory sequences are cloned in a vector together with a minimal promoter and a reporter gene. The enhancer ability of the candidate sequences is quantified based on the levels of expression of the reporter gene. MPRAs can test thousands of sequences in the same assays by leveraging candidate-specific barcodes that are sequenced together with the reporter transcripts (Gasperini, Tome, and Shendure 2020; Melnikov et al. 2012; Patwardhan et al. 2009, 2012; Shlyueva, Stampfel, and Stark 2014). A notable MPRA example is STARR-seq (Arnold et al. 2013). Peculiarly, STARR-seq allows the probing of millions of sequences by employing a construct in which the reporter gene cloned in front of the minimal promoter is the candidate CRE itself. Notably, albeit MPRA constructs are injected and tested in the nuclei of a living cells, they remain episomal, meaning that they do not integrate in the genome. This isolates CREs from their endogenous genomic context, explicitly testing for their intrinsic ability enhance transcription. The downside of this is that the measured activity might not be relevant in the endogenous context due to the local chromatin landscape as well as to interactions with other CREs (Gasperini, Tome, and Shendure 2020).

Finally, a most recent and exciting approach that has been gaining traction involves the systematic perturbation of enhancer function in their endogenous genomic context (Canver et al. 2015; Gasperini et al. 2019; Xie et al. 2017). This is possible through CRISPR assays in which a library of guide RNAs (gRNAs) targets and mutates the sequence of enhancers through indels 1-10bp long (W. Chen et al. 2019; van Overbeek et al. 2016). This library is delivered in pool to a cell population, followed by an assay that measures the transcriptional outcomes of such perturbation. Through this approach, it is possible to systematically identify the enhancers contributing to the transcription of genes. Depending on whether the whole transcriptome is probed by RNA-seq or scRNA-seq is it also possible to reveal complex regulatory interactions involving multiple genes and/or CREs (Gasperini et al. 2019; Rajagopal et al. 2016; Sanjana et al. 2016).

Differently from protein coding genes, CREs are much harder to inactivate by CRISPR-mediated sequence mutation. Almost invariably, protein coding genes form premature stop codon in response to indels, preventing the production of functional mRNAs. CREs on the other hand do not experience such constrains, and even the complete deletion of a TF motif is not granted to inactivate the CRE. To overcome these limitation, alternative CRISPR-based methods

have been more recently developed and successfully employed to deliver specific epigenetic modifications to either interfere with active elements, CRISPRi, or activate of inactive ones, CRISPRa (Domingo et al. 2024; Fulco et al. 2016, 2019; Gasperini et al. 2017; Klann et al. 2017; Simeonov et al. 2017; Thakore et al. 2015).

1.2.2.2 | Transcriptional enhancers

Promoters in mammalian genomes are often suboptimal, necessitating mechanisms to fine-tune gene expression, a role effectively filled by enhancers (Juven-Gershon, Cheng, and Kadonaga 2006). These elements, which have been defined as "endogenous, distally located DNA sequence elements that serve to enhance the transcription of a cis-located gene, in vivo, and in their native genomic context" (Gasperini, Tome, and Shendure 2020), are crucial for regulating gene activity. Unlike promoters, enhancers can influence their target genes regardless of the physical distance and orientation between them (Banerji, Rusconi, and Schaffner 1981; Moreau et al. 1981). This is achieved through their ability to be in close 3D proximity to the promoter regions of genes, despite potentially vast linear genomic distances. Enhancers are characterized by their association with open chromatin regions, bordered by nucleosomes marked specifically by H3K27ac and possibly H3K4me1, indicating active regulatory regions (Gasperini, Tome, and Shendure 2020; Shlyueva, Stampfel, and Stark 2014; Spitz and Furlong 2012).

With respect to enhancers, promoters experience a relatively simple set of regulatory interactions due to them embodying the last regulatory platform before the start of gene transcription. Enhancers, on the other end, stand more upstream in the regulatory cascade. This implies that they interact to a much larger extent with cell-type specific transcription factors, and this confers enhancers themselves a strong cell-type specificity. On feature associated with this fact is that enhancers operate in a modular fashion, meaning that they regulate genes additively (or cooperatively) and with a certain degree of redundancy (Amano et al. 2009; Shlyueva, Stampfel, and Stark 2014).

All of these features combine to the very complex nature of these regulatory elements which, from a DNA sequence perspective, are nothing other than clusters of transcription factor binding sites.

1.2.2.3 | Transcription factors

Transcription factors, TFs, form the main class of proteins that regulate gene expression. A very distinctive quality of TFs is the specificity with which they preferentially bind certain DNA sequences, often 6-12 base pairs long, as compared to the rest of the genome (Lambert et al. 2018; Wunderlich and Mirny 2009). The sequence specificity of TFs is mediated by their particular DNA-binding domain, DBD. DBDs are able to recognize specific DNA motifs by often contacting the major groove of the double helix. The specific 3D structure of the DBD and its mode of interaction with DNA can be used to classify them into distinct families (Luscombe et al. 2000; Pabo and Sauer 1992). Notably, some TF families are highly conserved across the tree of life while others appeared more recently, such as in metazoans. I will now describe the best-studied TF classes.

Helix-turn-helix, HTH, is one of the first DBD motifs to be discovered and is found across the tree of life as exemplified by the *E. coli lac* operon repressor. HTHs are characterized by the presence of 2-4 α-helixes connected though a bend (the “turn”) which imposes a fixed angle. One of these helixes is the Recognition (R) helix which accommodates into the major groove of the double DNA helix. The amino acids of the R helix that are exposed outwards specifically interact with the bases within the DNA motif. The homeodomain, HD, is a notable member of HTHs. First discovered in *Drosophila*, the HD is typical of TFs that regulate the homeotic genes which control the full body development of the animal. HDs contain 3 α-helixes plus an unstructured additional arm which contacts the minor groove of the DNA double helix for a stronger and more specific binding (Gehring, Affolter, and Bürglin 1994).

The zinc-finger motifs, first discovered in the *Xenopus laevis* TFIIIA, is characterized by the covalent binding of a zinc atom with two highly conserved cysteines and two histidines contained respectively in a β-sheet and an α-helix of the DBD. This association results in the formation of a fold that resembles the shape of a finger. Often, several zinc fingers are repeated in tandem in the structure of the TF and this increases the strength and specificity of binding to DNA (Rhodes and Klug 1993). Notable examples of mammalian TFs that belong to this family are Sp1, Kruppel-like factors (KLFs) and the insulator CTCF with its 11 zinc fingers.

Leucine zippers are dimers formed by the association of two long α-helixes. Both helixes contain a dimerization domain as well as a basic DBD, the latter of which contacts the major

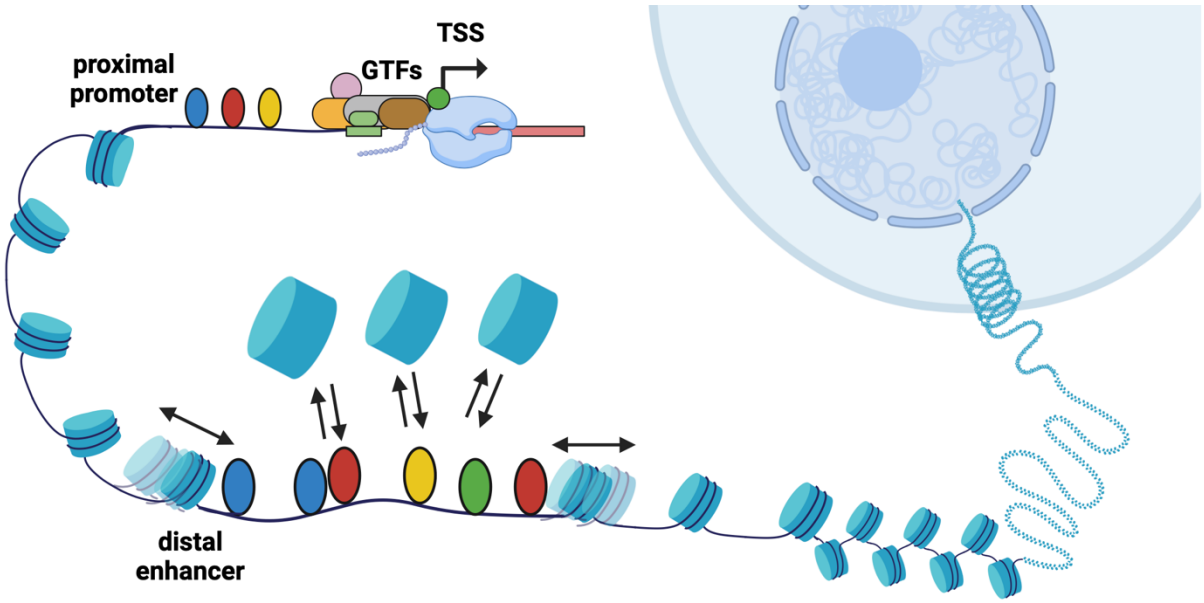

Figure 1.2. Transcription factors regulate gene expression at *cis*-regulatory elements. Transcription factors regulate gene expression by first binding to *cis*-regulatory elements (CREs), such as promoters and enhancers. Even at the cellular steady state, chromatin represents a physical barrier that transcription factors need to overcome to access their DNA recognition motifs. Such TF-nucleosome competition is an opportunity for cells to fine-tune the activity of CREs and quantitatively regulate gene expression.

groove of the DNA double helix. Because the dimerization of the two α-helixes is not specific, leucine zippers can originate both from homodimers and from heterodimers, increasing the repertoire of available combinations (McKnight 1991). Notable members of the family include the CCAAT-enhancer-binding protein C/EBP as well as AP-1, this latter being a heterodimer composed ofc-Jun and c-Fos.

Basic helix-loop-helix, bHLH, are monomeric DBDs characterized by two α-helixes connected by a loop. The longer helix contains a basic domain which mediates the interaction with DNA. Also, these domains tend to dimerize into various combinations which recognize the promoter motif E-box. Notable examples include the developmental TF Myo-D or the proto-oncogene c-Myc.

Another notable DBD family is the High mobility group (HMG) box domains. This includes TFs as well as other DNA binding proteins involved in transcription and DNA repair. The HMG is an L-shaped domain formed by three α-helixes which bind to the minor groove of the DNA double helix. Notable members of the family include the pioneer TF SRY-related HMG-box, Sox2, and the Sex-determining Region Y, SRY.

1.2.2.4 | Transcriptional enhancers as clusters of transcription factor binding sites

Differently from promoters, transcriptional enhancers are not characterized by such highly enriched sequence features as the TATAA-box or the BRE as detailed in the section 1.2.1.1. This means that enhancers, at the sequence level, are nothing other than clusters of recognition motifs for transcription factors. Each of these recognition motifs is a more or less variable short sequence, typically 6-12bp long (Lambert et al. 2018; Wunderlich and Mirny 2009), with an affinity for the specific TF which is increased with respect to the background genomic sequence.

TF motifs have been classically identified through a variety of assays which probe enriched TF binding against a background set of sequences. Notable *in-vitro* methods include protein binding microarrays (PBMs) and SELEX-based methods. In both cases, the purified TF is confronted with a pool of randomized candidate sequences. The sub-pool of sequences displaying sufficient binding affinity is analyzed to yield a "consensus" motif. Such in-vitro methods have been very powerful to identify the intrinsic sequence specificities of TFs (S. Hu et al. 2009; Jolma et al. 2013, 2015) even in the context of cooperative binding (Ibarra et al. 2020).

The *in-vivo* sequence specificities of TF binding have been explored by ChIP-based assays such as ChIP-seq or its higher resolution variants ChIP-exo and ChIP-nexus (Q. He, Johnston, and Zeitlinger 2015; Johnson et al. 2007; Rhee and Pugh 2011). These methods leverage cross-linking to enrich for TF-bound genomic sequences that are summarized computationally in "consensus" motifs. While such methods better reflect the functional sequence preferences of TFs, they suffer from the noise of skewed genomic sequences as well as the difficulty of distinguishing direct from indirect binding (Wasserman and Sandelin 2004; Worsley Hunt and Wasserman 2014).

Irrespective of the experimental approach used to enrich for TF-bound DNA, the heterogeneity of the resulting sequences is computationally summarized through so-called position weight matrixes, PWMs. PWMs are matrixes encoding the relative preference of a TF for a sequence at each nucleotide (Stormo and Zhao 2010). PWMs are typically displayed as sequence logos (Schneider and Stephens 1990), highlighting the motif preferred by the TF.

Despite the wealth of dedicated data, as well as experimental and computational approaches, predicting TF binding from sequence alone is still a highly inaccurate process.

Alternative approaches have been attempting, more or less successfully, to go beyond the analysis of the core recognition motif. One notable example is given by the analysis of DNA sequence shape that drive motif affinity (Zhou et al. 2013). In other cases, researchers have highlighted the role the sequences flanking the core motif of TFs as key determinants of binding affinity (De Almeida et al. 2022; Horton et al. 2023). More recently, deep learning methods have been approaching the problem from a functional perspective. Namely, TF motifs are predicted as those sequence instances which, in their endogenous genomic location, result as the most predictive of a molecular phenotype such as TF binding as measured it by a ChIP-based assay, gene expression or chromatin accessibility (Avsec, Agarwal, et al. 2021; Avsec, Weilert, et al. 2021; Novakovsky et al. 2021; Park et al. 2020; Zheng et al. 2021). Finally, recent works have highlighted G-quadruplexes (G4s), secondary structures formed at G-rich sequences and enriched at regulatory elements, as promoting non-canonical TF binding (Lago et al. 2021; Sirinakis et al. 2011; Spiegel et al. 2021).

1.2.2.5 | The regulation of gene expression by transcription factors

As touched upon in the previous sections, transcription factors fine-tune gene expression with cell-type, temporal and dosage specificity. This ability is prominently displayed during cell differentiation where, as documented in many cases, a handful of TFs seed the activation of *cis*-regulatory elements regulating lineage-determining transcriptional programs. Well-studied examples include the TFs FoxA and GATA-4 for the liver specification from gut endoderm (Cirillo et al. 2002; C. S. Lee et al. 2005; Lerner et al. 2023) or the TFs Oct4, Sox2, Klf4 and c-Myc establishing pluripotency in embryonic stem cells (Di Giammartino et al. 2019; M. Li and Belmonte 2017; Soufi, Donahue, and Zaret 2012).

The regulatory influence of multiple transcription factors binding at distal CREs is integrated in 3D space, and channeled to promoters, through the mediator. The mediator is a very large multi subunit complex which was shown to directly interact with both RNA polymerase II and the activation domain of transcription factors. The subunits interacting with RNA polymerase II have been shown to be indispensable for the transcription of virtually all eukaryotic genes, whereas the subunits interacting with TFs bear some degree of gene specificity (Roger D. Kornberg 2005; S. Malik and Roeder 2005, 2010). Despite its pivotal role as a link between distal regulatory elements and their target genes, the mechanisms of action

of the mediator complex are still an active area of research. As an example, it was recently proposed that the mammalian mediator and RNA polymerase II are dispensable to physically connect regulatory elements in 3D space, a function carried by architectural proteins. In this scenario, the mediator complex would integrate regulatory information in a functional way, rather than architectural (El Khattabi et al. 2019).

1.2.2.6 | Investigating the sequence specificity of transcription factor binding by natural genetic variation

As detailed in the previous sections, transcription factors (TFs) preferentially bind to specific DNA recognition motifs enriched at *cis*-regulatory elements (CREs). However, the mere presence of such motifs at CREs does not necessarily translate to TF occupancy. This implies that the determinants of TF binding *in-vivo* lie much beyond their core motif making the subject an active area of investigation to date (Avsec, Weilert, et al. 2021).

Several studies investigated the effects of sequence variation at CREs on TF binding or gene expression. They did so by leveraging the natural genetic variation between divergent mouse species both in static cell lines (Maurano et al. 2015; Wong et al. 2015; Yang et al. 2022) or across differentiation trajectories (Panten et al. 2024). Among others, these studies have highlighted how TF binding is very often influenced by genetic variation lying outside the core TF recognition motif (Deplancke, Alpern, and Gardeux 2016; Stefflova et al. 2013). Disparate mechanisms have been proposed to explain this indirect effect, with cooperative binding between transcription factors being a prominent one. This consists of the synergistic association between two or more TFs towards a more productive occupancy on DNA (Mirny 2010; Morgunova and Taipale 2017; Reiter, Wienerroither, and Stark 2017; Sönmezer et al. 2021). The dependencies among TFs that are established through cooperative binding could explain the propagation of the effects of sequence variation beyond the core TF motif. This is how previous studies have identified cooperative TF binding (Heinz et al. 2013; Stefflova et al. 2013). In particular, Stefflova et *al.* highlighted how clusters of cobound TFs tend to be more conserved and more strongly bound across divergent species, highlighting cooperativity as an evolutionary mechanism to confer stability to TF binding.

These findings brought forth a clear mechanistic framework to understand the complex sequence features driving TF occupancy *in-vivo*. However, more research is necessary to complete the picture.

1.2.3 | Chromatin is a physical barrier transcription factors need to overcome to access their DNA binding motifs

The human genome is composed by a total of 6 billion nucleotides divided into 23 chromosome pairs. If completely unwrapped, this DNA would reach a total length of two meters which needs to fit in the ca. 10μm of diameter of the cellular nucleus. Furthermore, cells need to rapidly and reliably access specific portions of their genome to regulate the transcriptional programs which not only define cellular identity, but also allow rapid response to environmental stimuli. Because of this, eukaryotic genomes are packaged in the cellular nucleus in very sophisticated ways. Several structural proteins are associated to genomic DNA to compose chromatin. I will now describe the fundamental aspect of chromatin structure and regulation to ensure proper gene expression.

1.2.3.1 | Chromatin structure

Chromatin is composed of genomic DNA, histones and non-histone proteins. Histones, among the most conserved proteins in the genome (H. S. Malik and Henikoff 2003), make up the most basic unit of chromatin: the nucleosome. Each nucleosome is an octamer made up of eight histone proteins, two each of H2A, H2B, H3 and H4. This disk-shaped histone octamer, is wrapped 1.7 times by a total of 147 DNA base pairs (Luger et al. 1997). *In-vivo*, nucleosome particles are arranged one after the other in arrays which, if experimentally unpacked and looked at through an electron microscope, famously resemble “beads on a string”. Each nucleosome particle is separated from the next by a region of “linker DNA” which can vary in length from ca. 20 to 80 nucleotides.

The interface between DNA and the histone octamer is very large and it involves 142 hydrogen bonds as well as several ionic bonds and hydrophobic interactions. The consequence

of this is a very strong and non-specific affinity which allows the stable and widespread binding of nucleosomes on DNA. Still, the positioning of nucleosomes needs to be regulated to allow for regulatory proteins such as transcription factors to access the DNA binding sites and exert their regulatory function. This is done in large part by depositing post-translational modifications on the unstructured N-terminal "tail" that each histone possesses and that protrudes outwards with respect to the nucleosome particle (Luger and Richmond 1998). PTMs are deposited on the histone tails by a series of histone modifying enzymes which I will treat in section 1.2.3.5. Another major class of enzymes involved in the regulation of the position of histones is chromatin remodelers. These ATP-dependent proteins catalyze the sliding of nucleosomes along DNA by interacting both with the histone octamer and with the DNA wrapped around it. I will treat chromatin remodelers in detail in section 1.2.3.2.

Tough very intuitive, the linear configuration of nucleosome arrays into "beads on a string" is rare *in-vivo*. There, nucleosome particles are further compacted in the so-called "30nm fiber", where the measure of 30nm refers to its diameter. The field is to date actively debating on the exact way that nucleosome particles are stacked into the 30nm chromatin fiber both *in-vitro* (Dombrowski et al. 2022; Garcia-Saez et al. 2018; Song et al. 2014) and *in-vivo* (Ohno et al. 2019; Ricci et al. 2015). The field debates two main opposing models, the "tetra-nucleosome" and the "solenoid", the detailing of which goes beyond the scope of this dissertation. It suffices to say that it is not unlikely that the 30nm chromatin fiber *in-vivo* is not organized according to one single architecture. It is rather possible that it exists in a continuum of states that are also influenced by the local variation in length of linker DNA.

The components that mediate the contacts between adjacent nucleosomes, and therefore the compaction of the fiber, are mainly the tail of histone H4 and the histone H1, the "linker histone". The latter is the largest among all histones and it is much less evolutionarily conserved. However, it is as abundant in living cells as to total of all other histones. The precise modes of action of H1 are still unclear, but it seems that it can influence the direction the DNA exiting the nucleosome particle by interacting with both DNA and the histone octamer (Robinson and Rhodes 2006).

At larger scale, chromatin is organized in two main states: euchromatin and heterochromatin. The latter represents a particularly compact state that is found at focal points in the genome such as centromeres and telomers, where gene density is particularly low. Genes that are embedded in heterochromatin are silenced and tend to be particularly resistant to

being expressed. A notable exception is the case of the X-linked escapees. For mammals, having two copies of the chromosome X would lead females to experience double the gene expression levels than males. The mechanism of X-inactivation compensates for this by compacting one of the two copies of X into heterochromatin. Escapees are genes that, through mechanisms currently under investigation, manage to be expressed despite being embedded in a heterochromatic context (Galupa and Heard 2018).

In the following section I will elaborate on the mechanisms that regulate chromatin compaction and that contribute to the control of gene expression.

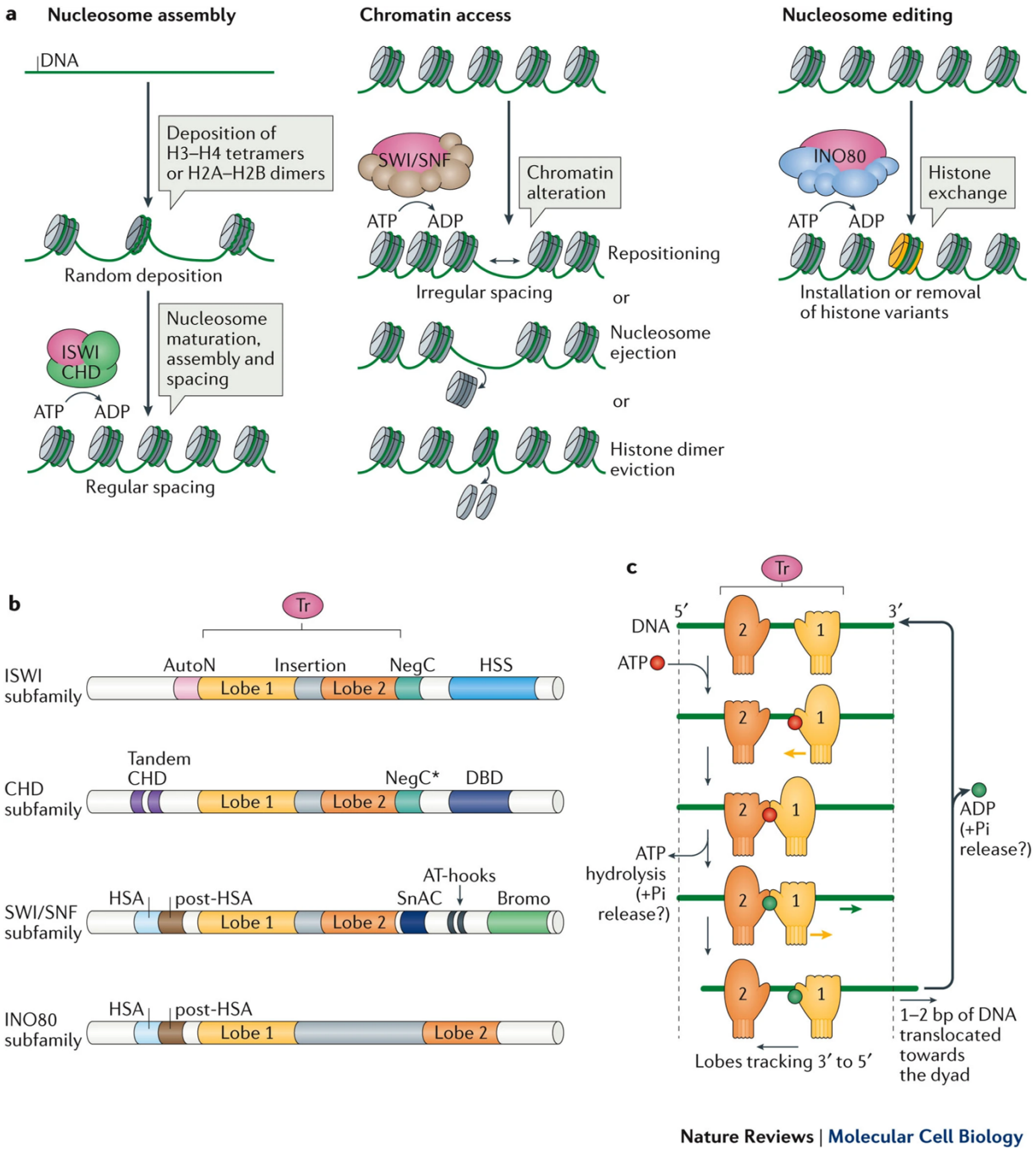

Figure 1.3. Overview of the functions and enzymatic mechanisms of chromatin remodelers. A. Chromatin remodelers are a class of ATP-dependent enzymes that mediate the reshaping of chromatin through various modalities of interaction with nucleosomes. Chromatin remodelers contribute to the assembly of nucleosomes after DNA replication, to the maintenance of chromatin accessibility at CREs by nucleosome eviction and sliding, and finally to the regulatory editing of nucleosomes. **B.** Eukaryotes share four main families of chromatin remodelers. All families share a highly conserved ATPase domain which is part of the helicase family SNF2. **C.** This domain mediates DNA-nucleosome translocation by iterative cycles of ATP-dependent association and dissociation of two RecA-like lobes with DNA (see Introduction 1.2.3.3). This figure is "*Reproduced with permission from Springer Nature*" from (Clapier et al. 2017).

1.2.3.2 | Chromatin remodelers

The presence of nucleosomes along the chromatin fiber presents a physical barrier for the binding of transcription factors and other regulatory proteins to DNA (Isbel, Grand, and Schübeler 2022). This barrier can be overcome through the action of a class of ATP-dependent enzymes globally referred to as chromatin remodelers (Barisic et al. 2019; Iurlaro et al. 2021). Chromatin remodelers, or simply remodelers, are multi-subunit complexes that have the ability to alter the structure of chromatin by either sliding nucleosome particles or by removing, replacing or relocating histone octamers in their entirety or in part.

Remodeler complexes all share an ATP-dependent subunit which is part of the DNA helicase family SNF2. Remodelers can be classified into four major subfamilies based on the similarities between their ATPase and additional subunits: imitation switch (ISWI), chromodomain helicase DNA-binding (CHD), switch/sucrose non-fermentable (SWI/SNF) and INO80 (Clapier et al. 2017; Hota and Bruneau 2016). I will now discuss these four classes and their role in chromatin and gene regulation.

The ATPase domain of ISWI is composed of two RecA-like lobes, DEXDc and HELICc, intervaled by a short insertion sequence. The ATPase domain is flanked by a HAND-SANT-SLIDE (HSS) domain which binds the adjacent linker DNA as well as the unmodified histone H3 tail. Two additional domains regulate the activity of the ATPase, namely AutoN, autoinhibitory N-terminal, and NegC, negative regulator of coupling. Functionally, most members of the ISWI subfamily regulate the assembly and regularization of nucleosome arrays to inhibit chromatin accessibility and gene expression. A notable exception is the NURF complex, nucleosome remodeling factor, which instead promotes chromatin accessibility and transcription (Clapier et al. 2017; Hota and Bruneau 2016).

Differently from ISWI, the CHD subfamily is characterized by the presence of two tandem chromodomains (CHD) in their ATPase domain. Instead of the HSS domain, they include a DNA binding domain, DBD, composed solely by the SANT and SLIDE domains. Functionally. CHD remodelers are able to assemble regularly spaced nucleosome arrays (similarly to ISWI), to promote chromatin accessibility at promoters and to edit nucleosome by inserting the histones variant H3.3. Notably, the assembly function of both ISWI and CHD complex subfamilies is exerted both after DNA replication and during gene transcription. Finally, a prominent member of the CHD subfamily, nucleosome remodeling deacetylase (NuRD), favors

the binding of transcriptional repressors to c2hromatin and itself represses genes through histone deacetylase activity (Clapier et al. 2017; Hota and Bruneau 2016).

The ATPase of the SWI/SNF subfamily includes an N-terminal helicase/SANT-associated (HAS) domain which interacts with actin and actin-related proteins (ARPs). It also includes an adjacent post-HAS domain, AT-hooks and a C-terminal bromodomain. Members of this subfamily, which notably include BRM/BRG1-associated factor (BAF) complexes, generally favor chromatin accessibility. The combinatorial assembly of SWI/SNF using further confers the remodeler tissue and developmental specificity (Clapier et al. 2017; Hota and Bruneau 2016).

INO80 is distinguished by a peculiarly long insertion sequence in between its two RecA-like lobes and, like SWI/SNF, it contains a HAS domain which interacts with ARPs. While members of the INO80 subfamily also act by spacing nucleosomes and promoting chromatin accessibility, they are peculiar for their nucleosome editing functions. The SRCAP (SWR1C, p400 and Snf2-related CBP activator protein) complex replaces canonical H2A-H2B histones with the H2A.Z-H2B variant. INO80C catalyzes the reverse reaction. The same complex also removes the variant H2A.X which might be explain its DNA repair function besides its ability to promote chromatin accessibility and transcription (Clapier et al. 2017; Hota and Bruneau 2016).

1.2.3.3 | The DNA translocation mechanism of chromatin remodelers

As detailed in section 1.2.3.2, all chromatin remodelers share the same characteristic of containing two RecA-like lobes in their ATPase domain: DEXDc and HELICc. These lobes share structure and sequence similarities to the ATPase domain of RecA, an *E. coli* DNA binding protein. Despite they often carry different functions, recent evidence (Deindl et al. 2013; Harada et al. 2016; Singleton, Dillingham, and Wigley 2007; Sirinakis et al. 2011; Velankar et al. 1999) has been highlighting how all chromatin remodelers might use the same mechanism to manage the interactions of DNA with the histone octamer.

In this model, the two RecA-like lobes act as the DNA translocation motor. They do so by sequentially, and in alternation, binding to, releasing DNA and moving 1-2 base pairs for each cycle of ATP hydrolysis. It is worth mentioning that the confirmation of this mechanism requires further evidence as a high-resolution structure of a remodeler engaged to a polymerase is lacking (Clapier et al. 2017).

1.2.3.4 | The regulatory role of chromatin remodelers

Chromatin is a physical barrier for the binding of transcription factors to DNA. Because of this chromatin itself exerts a regulatory role on gene transcription. In fact, mediating the selective access of TFs to their DNA recognition motifs at CREs has emerged as a mechanism of regulatory specificity (Isbel, Grand, and Schübeler 2022).

One recent study highlighted how the knock-out in mESCs of either BRG1 or SNF2H, ATPase domains of SWI/SNF and ISWI respectively, leads to the loss of chromatin accessibility associated with specific and non-overlapping sets of TFs (Barisic et al. 2019). Among others, the transcriptional repressor Rest, Klf5 and the pioneers Oct4 and Sox2 show selective dependency on SWI/SNF. On the contrary, TFs including Nfy, Yy1, Mafk and Ctcf show selective dependency on ISWI. This latter readily explains the dependency on SNF2H, but not BRG1, for the maintenance of chromosome folding (Barisic et al. 2019).

Similarly, BRG1 has also been implicated in the repositioning of nucleosomes at Gata1 binding sites during the differentiation of hematopoietic stem cells. This specific association has been shown to be important for the concomitant binding of Tal1 and subsequent transcriptional activation (G. Hu et al. 2011).

A more recent study explored the synergistic association between the BAF complex, RNA polymerase II and the core pluripotency regulators Oct4, Sox2, Klf4 and Nanog by chemical inhibition in mESCs (Brahma and Henikoff 2024). The authors showed that these three components engage in a positive feedback loop in which BAF mediates nucleosome displacement and eviction, while RNA polymerase II stabilizes BAF occupancy and the pluripotency TFs enhance its activity.

On a similar note, a study in yeast investigated the synergy among different classes of remodelers and how their balanced activity determines the positioning of +1 nucleosomes at TSSs (Kubik et al. 2019). This highlights the fundamental role of chromatin remodelers in gene expression dosage.

It is by now clear that these TF-remodeler dependencies exist and bare a certain degree of specificity. However, it is not as clear whether the direct interaction between these proteins is required or additional cofactors are involved. Furthermore, it remains to be investigated whether the action of remodelers is fundamental for the locus-specificity of TF binding or it contributes more to stabilizing TF-chromatin interactions (Isbel, Grand, and Schübeler 2022).

1.2.3.5 | The regulatory role of histone marks

Each histone in a nucleosome particle includes an unstructured N-terminal “tail” which is subject to a complex variety of post-translational modifications, histone marks. Such modifications are heavily involved in the regulation of chromatin and have been suggested to function as an epigenetic code. This code is maintained by a series of enzymes, often mark-specific, that can deposit (writer), remove (eraser) or interact with (reader) the particular histone mark (Lukauskas et al. 2024; Millán-Zambrano et al. 2022). Histone marks are pervasively involved in all regulatory aspects of gene expression, going from their interaction with pioneer transcription factors (Sinha et al. 2023) to regulating the transcriptional process itself (Szczurek et al. 2023; H. Wang et al. 2023). I will now give a brief overview of the most well-established histone marks.

H3K4me3 (tri-methylation of the 4^{th} lysine of histone 3) is enriched at promoters and in particular at transcription start sites. Its enrichment correlates with transcription however it does not seem to be broadly required for that. However, this mark is likely to corroborate transcription and to be involved in reducing its heterogeneity (‘noise’) across the cell population as well as in epigenetic inheritance (Benayoun et al. 2014; Kaifu Chen et al. 2015; Dahl et al. 2016; Talbert, Meers, and Henikoff 2019). Also, it is involved in a conserved negative feedback loop with the deposition of DNA methylation at CpG islands (Hanna et al. 2018; Ooi et al. 2007).

H3k4me1 (mono-methylation of the 4^{th} lysine of histone 3) and H3K27ac (acetylation of the 27^{th} lysine of histone 3) are enriched at active enhancers with cell-type specificity. However, these marks have been observed to be disposable for the activity of the enhancers themselves (Creyghton et al. 2010; Dorighi et al. 2017; Heintzman et al. 2009; Rickels et al. 2017; Zhang et al. 2020). H3k4me1 has been involved in epigenetic inheritance (Bleckwehl et al. 2021).

H3K36me3 is correlated with active transcription given the recruitment of its writer by the elongating RNA polymerase II. It appears disposable for transcriptional elongation but is involved in co- and post-transcriptional modifications of mRNAs (Bannister et al. 2005; Huang et al. 2019; Kizer et al. 2005; Meers et al. 2017; Vakoc et al. 2006).

H3K27me3, deposited by PRC2 (Polycomb repressive complex 2), seems crucial for the transcriptional silencing of chromatin into facultative heterochromatin (Pengelly et al. 2013;

Schuettengruber and Cavalli 2009). This mark can also be inherited across cell divisions in mammalian ESCs (Escobar et al. 2019).

H3k9me3 is classically associated with constitutive heterochromatin. Accordingly, it is enriched at centromeres, telomeres and silenced genes. H3k9me3 promotes the compaction of chromatin through its interaction with the self-oligomerizing HP1 (histone protein 1) (Allshire and Madhani 2018; Bannister et al. 2001; Nicetto and Zaret 2019). Genomic regions associated with the presence of this mark are not accessible for transcription factor binding (Soufi, Donahue, and Zaret 2012).

1.2.4 | Transcription factors outcompete nucleosomes through different mechanisms

1.2.4.1 | Pioneer transcription factors

Pioneer transcription factors are key regulators of gene expression with the unique characteristic ability to access their DNA recognition motifs in silenced, nucleosome occupied, chromatin. Pioneers play a pivotal role during cell differentiation where, by seeding the activation of lineage-specific CREs, induce dramatic transcriptional changes which commit the cell on the differentiation trajectory (Balsalobre and Drouin 2022; Bulyk et al. 2023; Cirillo et al. 2002; Iwafuchi et al. 2020; Roberts et al. 2021; Zaret 2020). One of the first examples of pioneers is the liver-specific FoxA. This TF has been proven to bind nucleosomes, to create nucleosome-depleted regions and, together with Gata4, to be required for hepatic induction from ectoderm (Cirillo et al. 2002; C. S. Lee et al. 2005). In the Drosophila embryo, the initial wave of zygotic gene expression is mediated by the maternal pioneer TF Zelda (Harrison et al. 2011; Liang et al. 2008). Pioneer TFs are so central to cell lineage specification that, when expressed ectopically, they can induce cell reprogramming. A prominent example is the reprogramming of differentiated fibroblasts to induced pluripotent stem cells, iPSCs, mediated by the pioneer Oct4, Sox2, Klf4 and c-Myc (Keshi Chen et al. 2020; Chronis et al. 2017; Di Giammartino et al. 2019; Soufi et al. 2015; Soufi, Donahue, and Zaret 2012; Takahashi and Yamanaka 2006; Wapinski et al. 2017).

By mechanistic definition, pioneer TFs are capable of ATP-independent displacement of nucleosomes to access the DNA recognition motif *in-vitro* (Dodonova et al. 2020a; Fernandez Garcia et al. 2019; Lerner et al. 2023; Michael et al. 2020; Yu and Buck 2019; Zhu et al. 2018a). However, it is becoming evident that the ability of pioneers to mediate the full opening and activation of CREs *in-vivo* is dependent on the recruitment of ATP-dependent chromatin remodelers, such as for the dependency of the TF Oct4 on SWI/SNF (Barral and Zaret 2024; Gong et al. 2022; King and Klose 2017; Minderjahn et al. 2020; Takaku et al. 2020), as well as on the cooperative interaction with additional non-pioneering TFs (Balsalobre and Drouin 2022; Chronis et al. 2017; Cirillo et al. 2002; K. Lee et al. 2019; Soufi et al. 2015; Soufi, Donahue, and Zaret 2012).

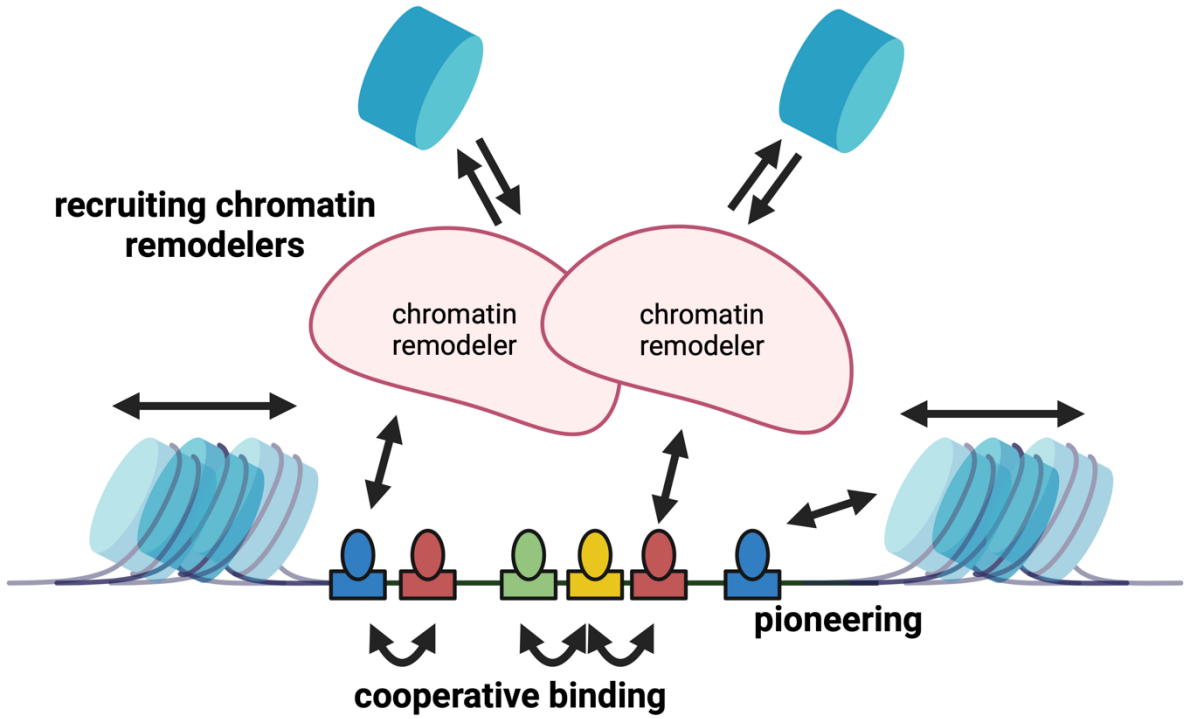

Figure 1.4. Transcription factors outcompete nucleosomes through different mechanisms. Transcription factors outcompete nucleosomes in different ways. Pioneer transcription factors have the peculiar ability to access their DNA recognition motifs at inactive, nucleosome-occupied, CREs and to seed their activation. Furthermore, TFs have long been known to engage in cooperative interaction among each other to increase their affinity for DNA. Finally, TFs dynamically and continuously recruit chromatin remodelers to mediate the ATP-dependent maintenance of chromatin accessibility at CREs.

1.2.4.2 | Transcription factor cooperative binding

The sequence specificity of transcription factor binding has been studied and catalogued (Jolma et al. 2013; Kulakovskiy et al. 2018; Mathelier et al. 2016; Weirauch et al. 2014) as extensively as to cover ca. three-quarters of the TFs encoded in the human genome (Lambert et al. 2018). Despite this wealth of data, there is still a discrepancy between motif predictions across the genome and observed TF binding. Factors such as sequence variations in adjacent motifs (Kilpinen et al. 2013; Maurano et al. 2015; Spivakov et al. 2012) or the local state of chromatin have been found to significantly influence the binding of TFs. Based on these observations, cooperativity between TFs has been proposed to be a key determinant for the ability of TFs to access their motif in the context of chromatin.

Two or more TFs that co-bind to the same regulatory region exhibit cooperative binding when there is some degree of dependence between them for a productive binding to happen. Several mechanisms have been proposed to explain TF cooperativity at the molecular level (Deplancke, Alpern, and Gardeux 2016; Khoueiry et al. 2017; Mariani et al. 2017; Morgunova and Taipale 2017; Reiter, Wienerroither, and Stark 2017; Vashee et al. 1998). For instance, the affinity of a transcription factor for a binding site can be increased by protein-protein interactions (PPI) between TFs. Alternatively, the binding of a partner TF to its own motif can cause a structural conformation change of DNA enabling the first TF to bind. This latter case is referred to as DNA-mediated cooperativity. In both of those cases the sequence syntax of the CRE is constrained by the structural requirements of a productive binding (i.e., spacing and orientation of the motifs to allow interaction between protein domains) (Ibarra et al. 2020; Jolma et al. 2015; Kerppola and Curran 1991; Kim et al. 2024). Additionally, because TFs are binding DNA in the context of chromatin, they have to compete against nucleosomes for occupancy of their binding motif (Adams and Workman 1995; Lickwar et al. 2012; Moyle-Heyrman, Tims, and Widom 2011; Polach and Widom 1996; Sönmezer et al. 2021). Two or more transcription factors can act in concert to displace the nucleosome and make a locus accessible. This mechanism has been referred to as nucleosome-mediated cooperativity (Deplancke, Alpern, and Gardeux 2016; Miller and Widom 2003; Mirny 2010). In this scenario, the arrangement of motifs within the CRE can be more flexible and be robust to syntax perturbations.

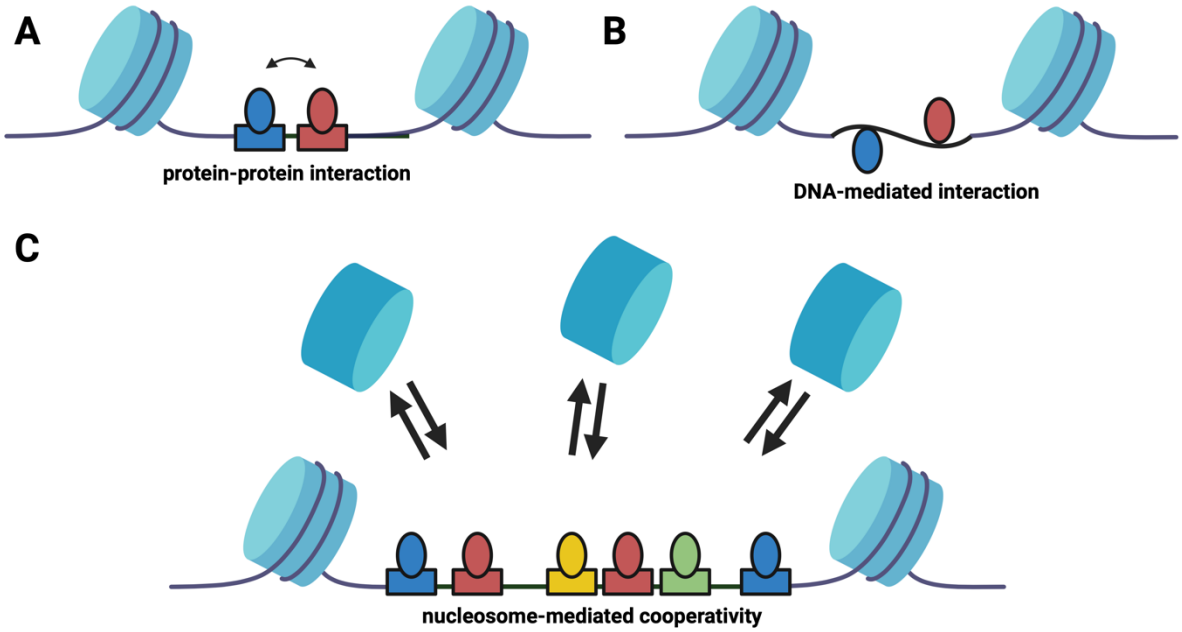

Figure 1.5. Modes of transcription factors cooperative binding. A. Transcription factors engage in protein-protein interaction to form a physical complex which has higher affinity for DNA than the individual components. **B.** Through DNA-mediated cooperativity, the binding of one TF to DNA imposes conformational changes to the local structure of the DNA double helix. This, in some cases, increases the affinity of a second TF for its DNA binding motif. No protein-protein interaction is required for this mode of cooperativity. **C.** Nucleosome-mediated cooperativity is an equilibrium-based mechanism in which multiple TFs synergize to displace nucleosomes by mass action. No protein-protein interaction, nor structural deformations of DNA are required for this mechanism. The consequence is a much less constrained arrangement of TF binding motifs at CREs.

While TF cooperative binding is observed all across the mammalian genome for most families of TFs (Lambert et al. 2018; Morgunova and Taipale 2017), it is long known to be particularly relevant for the functioning of the transcription factors involved in the establishment and maintenance of pluripotency in ESCs, most notably Oct4, Sox2 and Klf4 (Chew et al. 2005; Friman et al. 2019; M. Li and Belmonte 2017; M. Li and Izpisua Belmonte 2018; Loh et al. 2006; Michael et al. 2020; Rodda et al. 2005; Soufi et al. 2015; Soufi, Donahue, and Zaret 2012).

1.2.4.3 | Transcription factor interact with chromatin remodelers and chromatin modifying enzymes

As highlighted above, even pioneer transcription factors are dependent to a certain degree on the ATP-dependent displacement of nucleosomes by chromatin remodelers to stably interact with CREs in a chromatin context. It is becoming increasingly clear that TFs are selectively dependent on chromatin remodelers for the maintenance of local chromatin accessibility. In particular, TFs including Nfy, Yy1 and Ctcf show selective dependency on the ISWI complex. Rest, Oct4, Sox2 and Klf5 show instead dependency on SWI/SNF (Barisic et al. 2019; Isbel, Grand, and Schübeler 2022). Interestingly, the time course chemical inhibition of SWI/SNF in mESCs demonstrated that the maintenance of chromatin accessibility at the cellular steady state requires continuous remodeling activity (Iurlaro et al. 2021). This indicates that TF binding alone is not sufficient to preserve the width of nucleosome-depleted regions at CREs. Notably, the catalogue of TF-remodeler dependencies is still suffering from a certain degree of sparsity, with TFs such as Nrf1 not having been so far associated to a particular remodeler (Barisic et al. 2019; Iurlaro et al. 2021).

Less is clear about the relationship between histone marks and transcription factors. Unlike chromatin modifying enzymes, TFs show distinctly very little direct interactions with nucleosome particles (Dodonova et al. 2020a; Michael et al. 2020; Michael and Thomä 2021; Tanaka et al. 2020). This is likely due to the lack of specific histone binding domains (Andrews, Strahl, and Kutateladze 2016; Musselman et al. 2012). However, indirect interactions reportedly increase the non-specific affinity of TF-DNA contacts by stabilizing nucleosome positioning and favoring chromatin accessibility (Nishimura et al. 2020; Shoaib et al. 2021; Simon et al. 2011; Tanaka et al. 2020). Nevertheless, it is clear that there is a specificity in the interactions between co-factors and TFs and distinct types of CREs (Neumayr et al. 2022). However, more studies are required to clarify these relationships.

1.3 | The emergence of single molecule genomics and its computational tools

1.3.1 | Limitations of bulk and single cell omics technologies for the profiling of *cis*-regulatory elements

Several omics technologies have been developed and deployed to profile cis-regulatory elements genome-wide. The most well-established group of such technologies is populated by bulk assays. These, broadly speaking, probe a single molecular phenotype by obtaining, with the exception of bisulfite sequencing (Frommer et al. 1992), sequencing libraries enriched (or depleted) for DNA fragments interested by that particular molecular phenotype. Examples include ATAC-seq (Buenrostro, Wu, Chang, et al. 2015) or DNase-seq (Boyle et al. 2008) for chromatin accessibility and ChIP-seq (Solomon, Larsen, and Varshavsky 1988), ChIP-nexus (Q. He, Johnston, and Zeitlinger 2015) or CUT&RUN (Skene and Henikoff 2017) for the binding of transcription factors to DNA. The wealth of data from these technologies has been proving invaluable to advance our understanding of gene regulation from its very basic aspects to all sorts of translational contexts. One prominent example is the ENCODE (Encyclopedia of DNA Elements) project (The ENCODE Project Consortium 2012), which during the last two decades has been systematically probing, characterizing and cataloguing functional elements across the human and mouse genomes.

Despite their merits, bulk assays have a few fundamental limitations that have been stimulating researchers towards more innovative solutions. The first is that each of these bulk assays probes a single molecular phenotype at the time. This confines the study of the interactions between such phenotypes, such as DNA methylation and TF binding or histone modifications and chromatin accessibility, to a correlative space that offers limited mechanistic insights. Secondly, bulk assays quantify molecular phenomena based on enrichment (or depletion) of sequencing reads against a genomic background, inputs or control samples. This has two consequences. First, the signal across the genome is an average of the cell population, which masks a potentially insightful heterogeneity. Second, the intensity of the signal is also a function of sequencing depth and this hinders the precise estimation of the frequency of a molecular event such as TF binding.

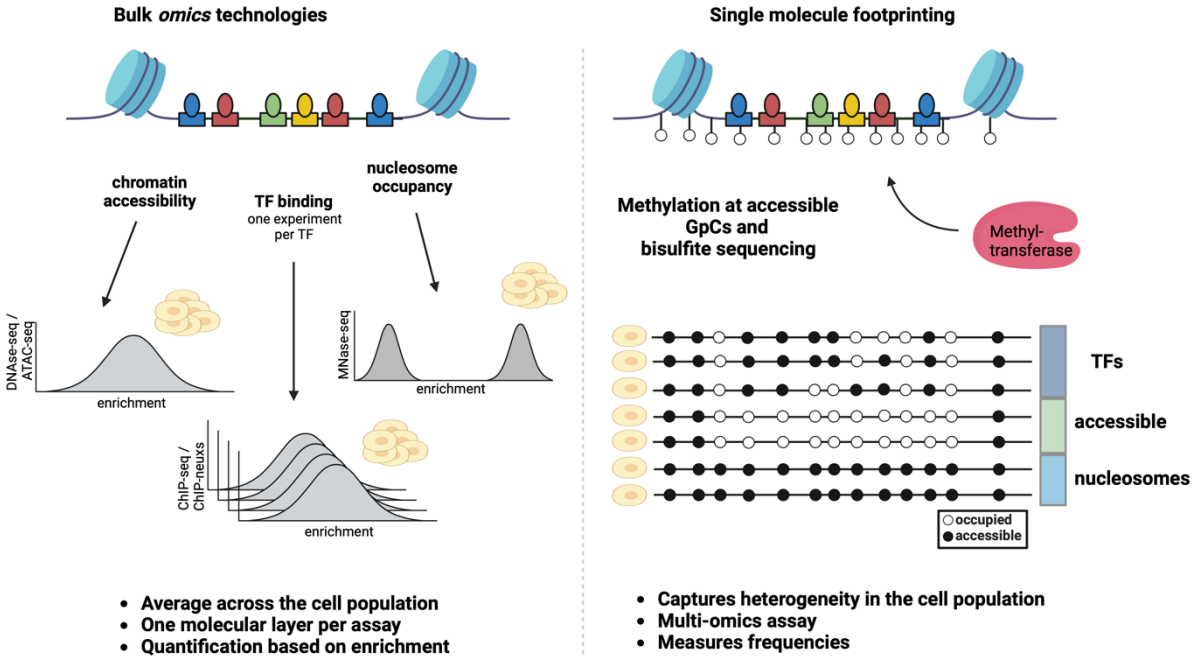

Figure 1.6. Single molecule footprinting overcomes the limitation of bulk assays for profiling *cis*-regulatory elements. When profiling *cis*-regulatory elements, bulk technologies are limited to the measurement of one molecular layer per experiment, i.e., chromatin accessibility, nucleosomes occupancy or the binding of a specific TF. This limits them to study the interactions among these layers in a correlative way. Instead, single molecule footprinting (SMF) measures the binding of multiple TFs, as well as chromatin accessibility and nucleosome occupancy in the same assay and at the single molecule level. Because single molecule footprinting sequences DNA molecules regardless of their state and without enrichments, it estimates the frequency of such molecular states within a cell population. Therefore, SMF reveals the quantitative cell-to-cell heterogeneity of CRE states.

More recently, single cell technologies in various modalities have been developed to overcome some of these limitations. For example, single cell ATAC-seq (Buenrostro, Wu, Litzenburger, et al. 2015) can reveal the cell-to-cell variability of cis-regulatory element usage across a cell population. Multi-omics single cell technologies such as sci-CAR (Cao et al. 2018), SUM-seq (Lobato-Moreno et al. 2023), scNOME-seq (Pott 2017) or scNMT-seq (Clark et al. 2018) have been developed to jointly profile multiple regulatory layers in the same assay. More specifically, sci-CAR and SUM-seq both profile chromatin accessibility and transcription while scNOME-seq profiles chromatin accessibility and CpG methylation. ScNMT-seq further combines scNOME-seq with transcriptome profiling in the attempt to bridge three regulatory layers. Single cell technologies are being pivotal in the discovery and characterization of cell types in healthy and pathological tissues. In principle, these methods could be further used to quantify the co-occurrence of regulatory events such as a CRE which is both accessible and

methylated in the same cell as well as the genes it regulates being transcribed. In practice, the sequencing depth required to do so is a limiting factor leading to extensive data sparsity. This same data sparsity also prevents the correct estimation of the frequency of molecular states in a cell population.

On the other hand, more recently developed single molecule genomics technologies are overcoming these limitations. They profile chromatin accessibility at CREs at the single molecule level at high resolution and sequencing depth. Formally, these multi-omics technologies probe CREs for the occupancy of multiple TFs and nucleosomes, as well as chromatin accessibility, in the same assay. They do so with high coverage, allowing for the dissection of the combinatorics of such regulatory layers, and by sequencing molecules from a cell population without enriching for a particular molecular state. This last point implies that the estimation of the actual frequencies of molecular states is decoupled from coverage and therefore researchers can study in great detail the heterogeneity of CRE states across a cell population.

1.3.2 | Overview of single molecule genomics technologies

Broadly speaking, all single molecule technologies follow two steps: the methylation footprinting of accessible chromatin and its readout through a sequencing method with single molecule resolution. The footprinting process can be performed with various combinations of three enzymes: M.CviPI (Ohmori, Tomizawa, and Maxam 1978) for the methylation of cytosines in the GpC context, M.SssI (Renbaum et al. 1990) for CpGs and Hia5 (Drozdz et al. 2012) or EcoGII (Murray et al. 2018) for the unspecific methylation of adenosines.

The readout has been reportedly done with short read Illumina bisulfite sequencing, a sequencing method with intrinsic single molecule resolution (Landan et al. 2012; Shipony et al. 2014). This allows to probe up to 300 bp of continuous single molecule accessibility signal, or 500 bp in case of amplicon sequencing. Two notable short read technologies include NOMe-seq (Kelly et al. 2012; Nabilsi et al. 2014) and single molecule footprinting, SMF (Kleinendorst et al. 2021; Krebs et al. 2017a). The first performs methylation footprinting at GpCs and collaterally measures endogenous CpG methylation through bisulfite sequencing. The latter

can additionally footprint at CpGs when applied to biological systems without endogenous CpG methylation such as Drosophila or DNA methyl-transferase triple knock-out (DNMT-TKO) cell lines (Domcke et al. 2015). Furthermore, SMF represented a significant leap forward with respect to NOMe-seq on account of its analytical methods developed ad-hoc to specifically and interpretably quantify single molecule protein-DNA contacts (Krebs et al. 2017a; Sönmezer et al. 2021).

Alternatively, long read sequencing with ONT nanopore sequencing (Kasianowicz et al. 1996) or PacBio sequencing (Rhoads and Au 2015) has allowed probing several kilobases, tough at lower spatial resolution on account of their reduced (Weirather et al. 2017) per-nucleotide sequencing accuracy. Early examples of the first include MeSMLR-seq (Y. Wang et al. 2019), ODM-seq (Oberbeckmann et al. 2019) and nanoNOMe-seq (I. Lee et al. 2020), all of which footprint for accessibility at GpCs and additionally at CpGs in the case of ODM-seq. Notable long read technologies which leveraged unspecific adenosine methylation include Fiber-seq (Stergachis et al. 2020), SAMOSA (Abdulhay et al. 2020) and SMAC-seq (Shipony et al. 2020), the latter with the addition of GpC and CpG footprinting. The first two technologies sequence with PacBio, the latter with ONT nanopore sequencing.

Because of the increased costs of long read sequencing as compared to the Illumina platform, many of the initial efforts in long read single molecule genomics have been confined to organisms with reduced genome sizes, such as drosophila or yeast, for which high coverages can be feasibly achieved. More recently developed ONT enrichment strategies have proven possible the nanopore based readout of footprinted loci in the human genome at high coverage (Battaglia et al. 2022).

Technology	Footprinted context	Sequencing platform	Proven enrichment	Most complex genome probed	Reference
NOMe-seq	GpC	Illumina	Yes, by PCR (specific)	Human	(Kelly et al. 2012; Nabilsi et al. 2014)
SMF	GpC, CpG	Illumina	Yes, by PCR (specific)	Mouse	(Krebs et al. 2017a; Sönmezer et al. 2021)
MeSMLR-seq	GpC	ONT	No	Yeast	(Y. Wang et al. 2019)
ODM-seq	GpC, CpG	ONT	No	Yeast	(Oberbeckmann et al. 2019)
nanoNOMe-seq	GpC	ONT	Yes, by Cas9 (specific)	Human	(Battaglia et al. 2022; I. Lee et al. 2020)
Fiber-seq	A	PacBio	No	Human	(Stergachis et al. 2020)
SMAC-seq	GpC, CpG, A	ONT	No	Yeast	(Shipony et al. 2020)
SAMOSA	A	PacBio	Yes, by MNase (unspecific)	Human	(Abdulhay et al. 2020)

Table 1. Overview of single molecule genomics technologies. Table inspired from (Hook and Timp 2023)

1.3.3 | Biological advancements mediated by single molecule genomics

To date, several studies have leveraged single molecule measurements to advance our understanding of transcriptional regulation at and across CREs.

For example, NOMe-seq and its long-read counterpart nanoNOMe-seq have been reportedly employed to investigate the cell-to-cell heterogeneity of CRE usage and epigenetic state (Kelly et al. 2012; Nabilsi et al. 2014). More long-read single molecule studies highlighted the heterogeneity of nucleosome arrays across the yeast (Y. Wang et al. 2019) and human (Abdulhay et al. 2020) genome, as well as the heterogeneity of NDR widths across yeast promoters associated with different levels of transcriptional activity (Y. Wang et al. 2019).

On the other hand, single molecule footprinting has revealed a novel aspect in the gene transcription process, namely that PolII is continuously reloaded at its pausing sites rather than being stalled until pause release (Krebs et al. 2017a). It further highlighted novel aspects of TF cooperative binding such as its role in the energy-dependent competition against nucleosomes for the occupancy of TFs at CREs (Sönmezer et al. 2021). Finally, in a more recent publication (Kreibich et al. 2023), the technology was used to highlight how the majority of CREs in the mouse genome do not depend on endogenous CpG methylation for their activity.

Fiber-seq, first applied in drosophila, leveraged its long-read measurement to reveal the distance-dependent coordination among adjacent CREs (Stergachis et al. 2020). The same paper also quantified the length of linker DNA in different genomic contexts, reporting higher values in closed chromatin with respect to the boarders of active CREs. In a later publication (Isaac et al. 2024) the same technology was adapted to study the packaging of DNA in human mitochondria, mtFiber-seq. This study revealed that human mtDNA exists in an all-or-none globally compact state, with a minority of mitochondrial chromosomes being in accessible and active state. Finally, in a recent preprint, Fiber-seq was allowed measuring the differential positioning of +1 nucleosomes upon PolII pausing (Tullius et al. 2023).

Using Cas9-enriched nanoNOMe-seq, (Battaglia et al. 2022) dissected the interplay between allele-specific endogenous CpG at an ICR, the binding of the insulator TF Ctcf and the usage of multiple enhancers for the allele specific regulation of the IGF2 gene.

Using amplicon-seq SMF as a readout, (Doughty et al. 2024) have most recently investigated the relationship between the number of TFs motifs at enhancers and the transcriptional output of promoters. They did so by developing a synthetic and inducible enhancer-promoter reporter system inserted in the AAVS1 locus in the genome of the human K562 cell line. Among other results, this set-up revealed that TFs synergize in a non-linear manner to activate transcription.

In conclusion, single molecule genomics technologies are propelling our understanding of cell-to-cell variation in CRE states and how this affects gene transcription. Furthermore, we are building a clearer picture on the relationships between chromatin components such as GTFs, sequence-specific TFs and nucleosomes. In the near future, it is foreseeable that single molecule genomics technologies will help uncover the general principles underlying transcriptional noise during cell differentiation and development.

1.3.4 | Computational tools available in single molecule genomics

To date, not many computational tools are available for the analysis of single molecule data. In particular, there is a general lack of stable and maintained tools for the general, robust and user-friendly detection and quantification of footprints at the single molecule level.

An early analytical framework that has been developed and published along the meSMLR-seq technology is NP-SMLR (nucleosome positioning from single molecule long read sequencing). NP-SMLR, adapts the Needleman and Wunsch dynamic programming algorithm to detect nucleosome positioning from meSMLR-seq data in yeast. Notably, the method assumes the independence of methylation status among GpCs, assumption which might be violated. The Perl code for this tool is distributed among the supplementary material of the publication (Y. Wang et al. 2019).

FiberHMM, an unsupervised footprint detection tool tailored for Fiber-seq data, has been recently reported. FiberHMM is an HMM based footprint caller that quantifies the single molecule occupancy of GTFs such as PolII and PolIII as well as nucleosomes around TSSs. For each genomic coordinate, FiberHMM considers the 7-mer centered at that coordinate and emits the probability of that coordinate being either accessible or inaccessible. Footprints for different molecules are distinguished by the size of the inaccessible stretch (Tullius et al. 2023). Currently the code for FiberHMM is not distributed.

(Doughty et al. 2024) also reported a new supervised molecular classifier. It is a maximum likelihood approach in which each observed molecule is annotated with one among all possible molecular states. Such states must be explicitly enumerated beforehand and are defined based on the expected TF binding combinations as well as all possible nucleosome occupancy patterns. This method is applicable to SMF data produced either for synthetic loci or for loci for which all possible protein-DNA binding events are known a-priori. When applicable, this method has the advantage of providing perfectly interpretable quantifications. Thoroughly documented code for this method is distributed on GitHub.

In addition to the above-mentioned tools, a few works (Abdulhay et al. 2020; Isaac et al. 2024) have been published with the addition of the relative GitHub repositories for the reproducibility of the analyses results.

1.4 | Aims and design of this study

1.4.1 | Aims

1.4.1.1 | Advancing the computational and analytical tools in single molecule genomics

Similar to other single molecule genomics methods, single molecule footprinting, SMF, is a relatively young technology and, as such, it suffers from a general lack of robust and documented computational and analytical tools. I reviewed the few instances of methods reported to date for specific single molecule approaches in the section 1.3.2. Furthermore, single molecule technologies produce very rich but complex datasets. This is the reason why many published works leveraged purely supervised analytical frameworks (Doughty et al. 2024; Krebs et al. 2017a; Sönmezer et al. 2021). However, *ad-hoc* unsupervised methods have the power to reveal a much larger wealth of information in a data-driven manner.

Therefore, my second aim for this study was to advance the computational tools and analytical frameworks available in single molecule genomics. Specifically, I aimed to establish a well-documented and robust software tailored for the analysis of single molecule footprinting as it has been done in previous publications (Krebs et al. 2017a; Sönmezer et al. 2021).

Furthermore, I aimed to enhance the potential of discovery by developing a computational approach that would interrogate single molecule footprinting in a data-driven manner. With an unsupervised approach, I foresaw the discovery and quantification of complex molecular states at CREs, which could not have been appreciated through supervised analysis alone.

1.4.1.2 | Dissecting the contribution of transcription factors to chromatin accessibility

Transcription factors (TFs) are fundamental for the concerted regulation of gene expression in all aspects of the life of an organism. In order to exert their function, TFs need to access and bind to their recognition DNA motifs at *cis*-regulatory elements (CREs). Because eukaryotic CREs are embedded in chromatinised DNA, TFs first need to overcome such a

physical barrier by achieving the displacement of nucleosomes (Isbel, Grand, and Schübeler 2022; Trojanowski and Rippe 2022).

With this study, I aimed to disentangle a few aspects of how transcription factors interface with nucleosomes and chromatin overall to maintain accessibility at CREs in the mammalian genome. More specifically, I wanted to quantify the extent of TF-nucleosome competition at different types of CREs such as enhancers and promoters and what are the important features of the combinatorial assembly of TFs at CREs that allow to win this competition. Finally, I sought to investigate how fundamental are individual TF instances at CREs for the maintenance of chromatin accessibility, and how is nucleosome displacement affected when TFs are prevented from engaging with CREs.

1.4.2 | Design

1.4.2.1 | *SingleMoleculeFootprinting* and *FootprintCharter*, a set of tools for the analysis of single molecule footprinting

First, I sought to develop a robust and fully documented R package, *SingleMoleculeFootprinting,* to accompany investigators through the most well-established steps of single molecule footprinting data analysis. I decided to develop it in R, in order to leverage the great wealth of tools already available for computational genomics. A large number of these tools are distributed through the open source project Bioconductor (Gentleman et al. 2004; Huber et al. 2015). Distributing *SingleMoleculeFootprinting* through Bioconductor ensured, to the best of my possibilities, the longevity and maintenance of this software. In fact, being distributed through Bioconductor involves a strict initial review process and the life of the software is followed by regular checks to ensure its continued functioning and maintenance. In order to maximize the visibility of this resource, I also published *SingleMoleculeFootprinting* alongside the end-to-end computational protocol for the analysis of SMF data (Kleinendorst et al. 2021).

Secondly, I aimed for a general tool for the unsupervised footprint detection and quantification from single molecule genomics. I wrote *FootprintCharter* to take as input any

binary matrix having a row per molecule and a footprinted nucleotide per column. This opens the possibility for its applicability to additional single molecule technologies. I developed *FootprintCharter* in compatibility with *SingleMoleculeFootprinting* in order to integrate them in a future Bioconductor release cycle. In short, I designed *FootprintCharter* to identify as many footprints for transcription factors (TFs) and nucleosomes as are experimentally measured. Differently from previous molecular classifiers, *FootprintCharter* reports footprints annotated with two key features: their frequency in the cell population and their precise genomic coordinates. Notably, all of the above implies that *FootprintCharter* should be able to distinguish and quantify very complex molecular states, including various statistical positions of nucleosome arrays as well as TF clusters of any size, a limitation from previous work I aimed to overcome (Sönmezer et al. 2021).

1.4.2.2 | Quantification of TF-nucleosome competition across the mammalian genome at single molecule resolution

In order to quantify TF-nucleosome competition in the mammalian genome, I decided to resort to single molecule footprinting, SMF. This technology probes the heterogeneity of CRE states across a cell population genome-wide and at high depth. This implies that through SMF I can observe how often each probed CRE is occupied by nucleosomes and how often it is accessible, i.e., how often nucleosomes are outcompeted. I decided to apply *FootprintCharter* to a previously reported SMF dataset for mESCs sequenced at high depth (Sönmezer et al. 2021). I decided to use the resulting unsupervised footprint quantifications to ask how frequently nucleosomes are outcompeted at different *cis*-regulatory elements, such as active promoters and enhancers or inactive elements. I also interrogated specific TFs, and different TF cluster configurations, for their association with outcompeted nucleosomes. Finally, I sought to investigate the width of nucleosome-depleted regions associated with the same genomic elements.

1.4.2.3 | Systematic perturbation of TF binding at endogenous CREs by natural genetic variation

In order to go beyond the correlative measurement of how frequently the displacement of nucleosomes is associated with certain CREs or TFs, I decided to systematically perturb the system. The way I chose to do so is by leveraging the natural genetic variation between different mice species. I chose to use for this experiment *Mus musculus castaneus* and *Mus spretus* based on their estimated evolutionary distance from the reference laboratory *Mus musculus domesticus*. Such distance of 1 and 3 million years respectively corresponds to widespread sequence variation at CREs but to very limited sequence coding variation (Keane et al. 2011;

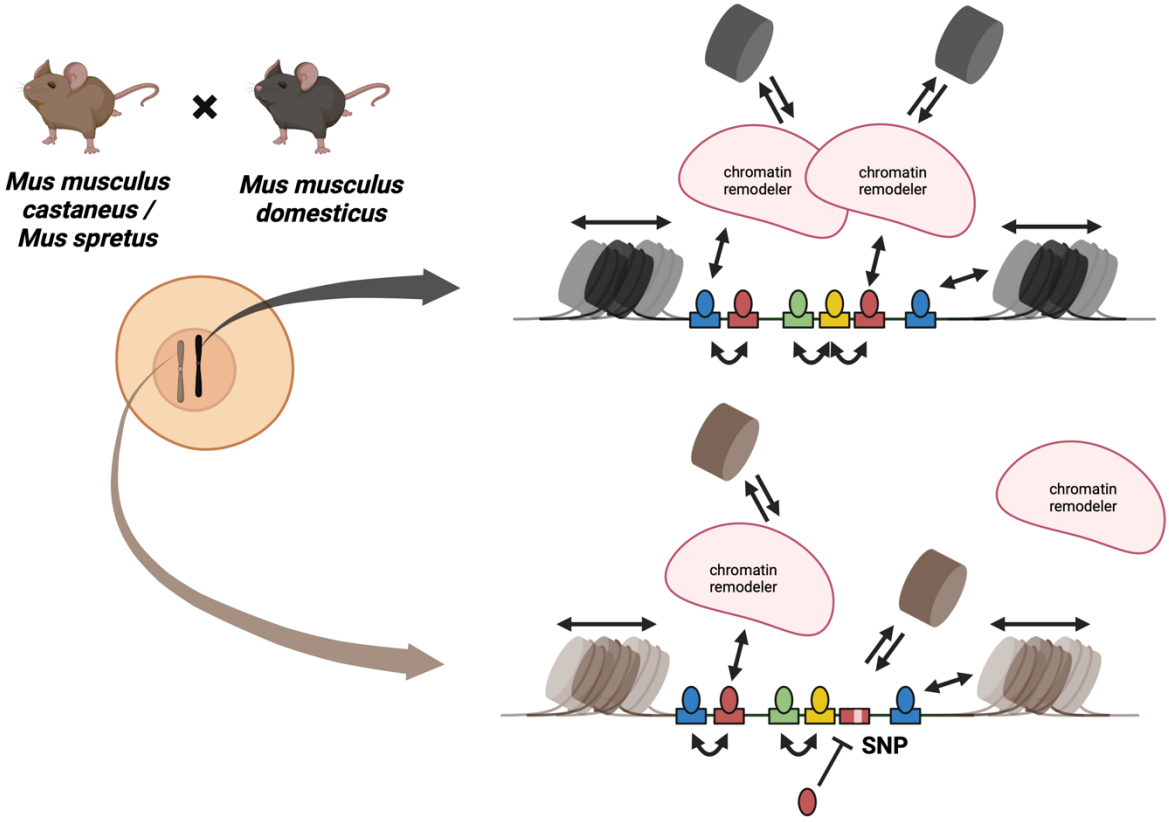

Figure 1.7. Systematic perturbation of TF binding through natural genetic variation among mouse species. In this study, I leveraged the natural genetic variation among different mouse species crossed into two F1 hybrid cell lines. The first line results from the cross of *Mus musculus domesticus* and *Mus musculus castaneus*, the second from the cross of *Mus musculus domesticus* and *Mus spretus*. Naturally occurring SNPs between these genomes occasionally reduce the affinity of TFs for their DNA binding motifs. This in turn impairs the ability of these TFs to engage in interactions with nucleosomes, chromatin remodelers and other TFs. Assessing whether the equilibrium of TFs and nucleosomes are altered by this perturbation, reports on the ability of individual TFs to maintain chromatin accessibility at CREs.

Lilue et al. 2018; Thybert et al. 2018). In order to control for experimental batch effects and to control for the effect of trans-acting variants (Floc'hlay et al. 2021; Goncalves et al. 2012; Panten et al. 2024) I opted for probing the genomes of these three mice species in two F1 mESC lines obtained, respectively, by crossing *Mus musculus castaneus* and *Mus spretus* with *Mus musculus domesticus*. Cell lines from the Bl6xCast cross were produced and provided by the Odom laboratory.

My strategy was to search for instances of sequence variation that would alter the affinity of TFs for their DNA recognition motifs, impairing TF binding and, consequently, affect the ability of individual TF instances to outcompete nucleosomes. The statistical testing for allele-specific TF-nucleosome competition would point me towards the TFs and TF cluster configurations which are necessary for the maintenance of chromatin accessibility at CREs.

1.4.2.4 | Validate the necessity of TFs for the maintenance of chromatin accessibility at CREs by rapid and acute Sox2 protein depletion

I chose to validate the dependencies I observed at CREs on certain TFs for the maintenance of chromatin accessibility by analyzing data from a protein perturbation experiment. Specifically, Laura Moniot-Perron induced the rapid and acute depletion of Sox2 at the protein level by employing a previously reported degradation tag, dTAG, system (Liu et al. 2021a; Maresca et al. 2023).

2 | Results

2.1 | *SingleMoleculeFootprinting* and *FootprintCharter*, new computational tools for Single Molecule Footprinting data

Single Molecule Footprinting, SMF, is a novel multi-omics readout first reported in (Krebs et al. 2017b). SMF has been employed to measure protein-DNA contacts at the single molecule level. More specifically, it has been used to measure multiple TF binding, the GTF, nucleosomes and chromatin accessibility. It has revealed novel insights into RNA polymerase II pausing, TF cooperative binding and TF-nucleosome competition, as well as, the epigenetic regulation of *cis*-regulatory elements (CREs) by DNA methylation.

SMF is a young technology and it therefore lacks a comprehensive set of computational guidelines and documented tools to aid users throughout the processing and analysis of this data modality. In particular, I focused on developing and documenting tools for the single molecule quantification of TF occupancy and TF co-occupancy as performed in (Sönmezer et al. 2021).

Furthermore, there was no tool available for the general and unsupervised quantification of footprints and chromatin accessibility patterns at the single molecule level. In particular, I aimed to develop a tool, hereon referred to as *FootprintCharter*, that would quantify any single molecule chromatin accessibility pattern, TF binding and nucleosome occupancy measured by SMF, regardless of prior TF motif annotation and regardless of the complexity of the molecular states at the locus.

2.1.1 | Computational guidelines and documented tools for Single Molecule Footprinting

Single Molecule Footprinting can be conceptually divided in two parts: the enzymatic footprinting of accessible chromatin and the readout of such footprinting by a sequencing

method which preserves single molecule information. Footprinting is carried by incubating purified nuclei with the methyl-transferase enzyme M.CviPI (Ohmori, Tomizawa, and Maxam 1978), which targets cytosines in the GpC context, and optionally M.SssI (Renbaum et al. 1990), which targets CpGs. The second enzyme can be used in systems without native CpG methylation such as drosophila or in systems which have been deprived of such molecular

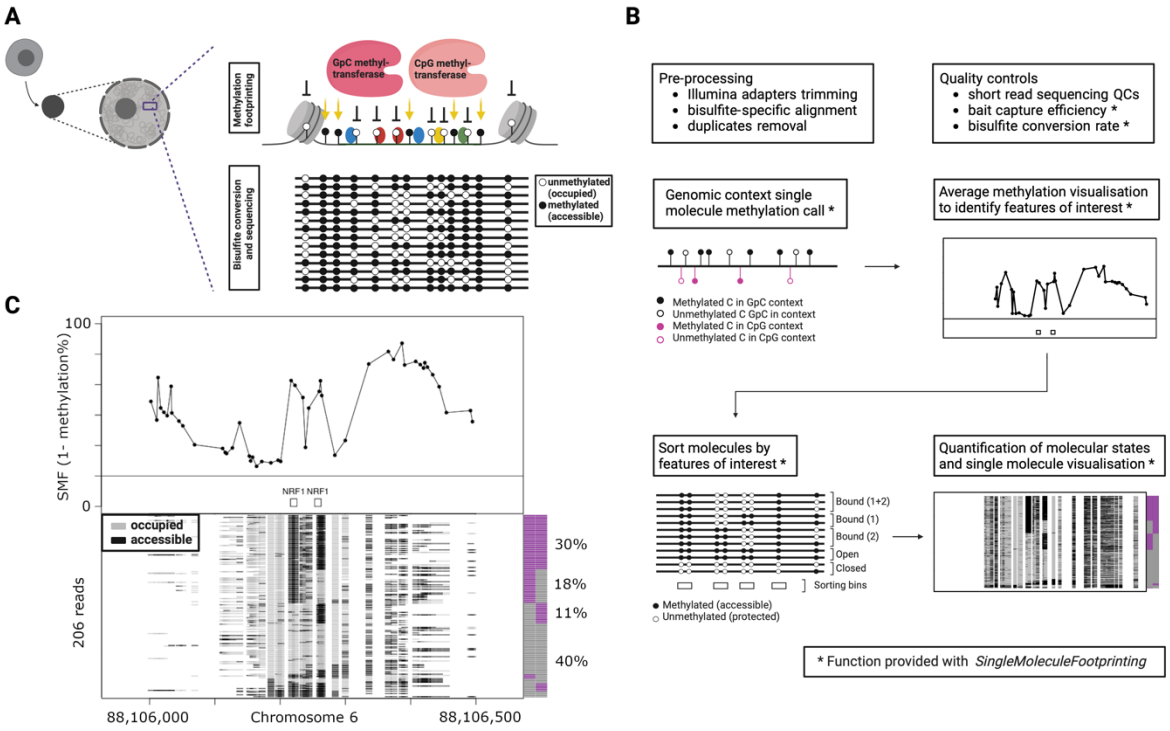

Figure 2.1. Computational guidelines and *SingleMoleculeFootprinting* Bioconductor package. A. Schematic representation of single molecule footprinting **B.** Conceptual workflow detailing the computational steps for the pre-processing and analysis of single molecule footprinting data. Those steps annotated with an asterisk "*" are supported with ad-hoc functions included in the Bioconductor package *SingleMoleculeFootprinting*. **C.** Single locus example of single molecule footprinting data as visualized using the version of *SingleMoleculeFootprinting* currently distributed through Bioconductor. Briefly, the upper panel shows on the y axis the average single molecule footprinting signal (the invers of the methylation signal) for each cytosine (black dots). The x axis, identical for the lower panel, shows genomic coordinates (ca. 500 bp in total width). Two Nrf1 binding sites are annotated with white rectangles. At the average level, SMF signal can be interpreted as a measure of cytosine occupancy. The binary heatmap in the lower panel shares, for columns, exactly the same x axis with the upper panel and has one molecule per row. Each cell is a single cytosine which is colored according to its methylation status (grey for methylated and accessible, black for unmethylated and occupied). The stacked bar-plot on the lower right displays the quantification of TF co-binding: 30% of the molecules are found to be bound by TFs at both Nrf1 sites, while 18% and 11% of the molecules, respectively, are found bound at one binding site or the other. All the panels in this figure are adaptations of Figures 1, 5 and 7 of Kleinendorst et al., 2021.

phenotype by knock-out of the three endogenous CpG methyltransferase enzymes, DNMT-TKO (Domcke et al. 2015). Such footprinted genomes are subject to bisulfite conversion, process thanks to which the methylation mark is converted into primary DNA sequence and can therefore be read by traditional Illumina sequencing (Figure 2.1A).

In order to guide new users through the analysis of SMF data, I have detailed the recommended pre-processing steps to follow to go from raw sequencing reads to aligned and deduplicated reads which can be used for single molecule methylation calling. I have included several quality controls users can perform to assess the quality of the data, what are some common sources of artifacts and how they can be prevented (Kleinendorst et al. 2021). To facilitate single molecule methylation calling, visualization and TF binding quantification, I developed *SingleMoleculeFootprinting*, the first R package dedicated to Single Molecule Footprinting which is currently distributed through Bioconductor. *SingleMoleculeFootprinting* allows users to call methylation for a locus of interest, at single molecule level and in the relevant genomic contexts with one line of R code. With equivalent ease, users can quantify the single molecule TF occupancy and co-occupancy according to the methods developed in (Sönmezer et al. 2021) and visualize single sites both at the bulk, single molecule and biologically interpreted level (Figure 2.1B-C).

The consolidation of the computational protocol for single molecule footprinting, as well as the development of a thoroughly documented R package, facilitated the spread of Single Molecule Footprinting, it gave independence to external users and collaborators and it served as a stepping stone for the development of more advanced tools for distilling information from single molecule footprinting data.

2.1.2 | Limitations of the analytical tools for single molecule footprinting

The analytical tools developed in (Krebs et al. 2017b) and (Sönmezer et al. 2021) allowed for the very precise but supervised quantification of protein-DNA interactions. In particular, (Krebs et al. 2017b) focused on the dynamics of the general transcription factors by leveraging extensive prior knowledge on the stereotypical positions at which the RNA polymerase II and the pre-initiation complex, PIC, can be found bound to DNA relative to the

transcription start site (Cianfrocco et al. 2013; C. Lee et al. 2008; C. Y. Lim et al. 2004; The modENCODE Consortium et al. 2010). On the other hand, (Sönmezer et al. 2021) focused on the study of transcription factors, such as Ctcf, Rest, Nfya and Nrf1, which occupy very predictably their DNA recognition motifs in the mouse genome. The analytical strategy used in both studies will be hereon referred to as “single molecule sorting”. This consisted in evaluating methylation at the cytosines in and around the short genomic space interested by the predicted TF binding motifs or binding sites for the general transcription factors (Figure 2.2). While this strategy has revealed very successful to study protein-DNA contacts that can be localized very predictably, it revealed more challenging for the study of TFs which bind less predictably to their motifs, such as the TFs regulating pluripotency in embryonic stem cells Oct4, Sox2, Klf4 and c-Myc. Incidentally, these TFs also make for a more interesting case study of cooperative binding, being it widely accepted that they do cooperate for their activity (Friman et al. 2019; M. Li and Izpisua Belmonte 2018; Rodda et al. 2005; Soufi, Donahue, and Zaret 2012). Single molecule sorting does not allow for the interpretable quantification of more than two TF binding events, while mammalian cis-regulatory elements (CREs) contain most often a higher number of TF binding sites, 5-6 according to some estimates (Vierstra et al. 2020). Finally, single molecule sorting does not resolve individual nucleosome particles and consequently different arrangements of nucleosome arrays and, more in general, the local structure of chromatin. Being able to extrapolate this kind of information from SMF data expands the spectrum of insights that can be gained about the mechanics of transcription factors engagement with nucleosomes (Dodonova et al. 2020b; Echigoya et al. 2020; Fernandez Garcia et al. 2019; Gadea and Nikolova 2022; Soufi et al. 2015; Zhu et al. 2018b) as well as what determines the heterogeneity of chromatin compaction across the genome and in response to perturbations (Boltengagen et al. 2023; Dombrowski et al. 2022; Oberbeckmann et al. 2024; Stergachis et al. 2020).

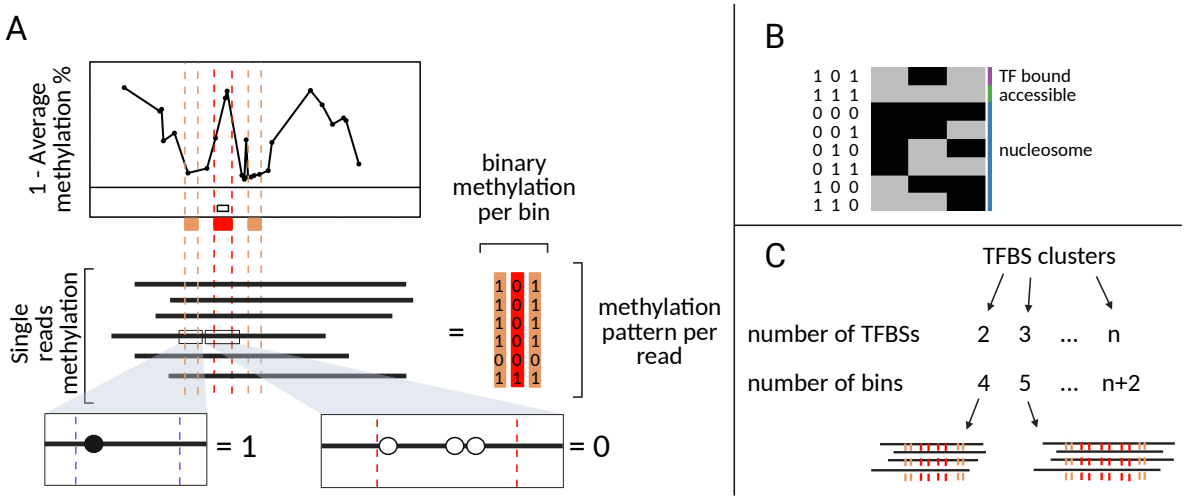

Figure 2.2. Single molecule sorting developed in Krebs et al., 2017 and Sönmezer et al., 2021. Single molecule sorting is a supervised single molecule classification approach with very high sensitivity for protein-DNA contacts that can be systematically predicted by enriched sequence features. Notable examples include RNA polymerase II and pre-initiation complex as well as transcription factors such as Ctcf, Rest and Nrf1. **A.** Single molecule sorting consists of designing bins around the sequence feature of interest, e.g., a TF binding site. The distance between consecutive bins allows enough resolution to resolve narrow TF footprints. The next step is to summarize the binary methylation signal in each bin for each molecule. In this way each molecule becomes represented by a binary sequence of as many digits as designed bins. **B.** When quantifying the footprint at a single TF binding site, three bins are designed for a total of eight possible resulting patterns. These patterns are interpreted as different molecular states. For example, all the molecules represented by the pattern "101" encode for the sequence "accessible-occupied-accessible" and are interpreted as TF bound. The molecules represented as "111" are interpreted as fully accessible. The remaining six states are interpreted as nucleosome occupied. This way of encoding single molecule accessibility patterns allows for the highly interpretable quantification of complex methylation patterns **C.** When quantifying TF footprints at clusters of binding sites, the same technique can be applied, adjusting the number of bins. However, the interpretation of the resulting patterns is hindered by the combinatorial explosion in the number of possible patterns. Also, the increasing genomic width covered by several TF bins at a certain point becomes indistinguishable from nucleosome states. This figure is adapted from Sönmezer et al., 2021 and Kleinendorst et al., 2021.

2.1.3 | *FootprintCharter*, a novel framework for the unsupervised quantification of footprints at single molecule level

To overcome the shortcomings of "single molecule sorting", I developed *FootprintCharter*, an unsupervised framework which quantifies TF and nucleosome footprints as well as chromatin accessibility patterns without relying on the prior annotation of TF motifs. Being a general footprint detection algorithm, it quantifies as many TF footprints and nucleosome array arrangements as are represented in the data (Figure 2.3).

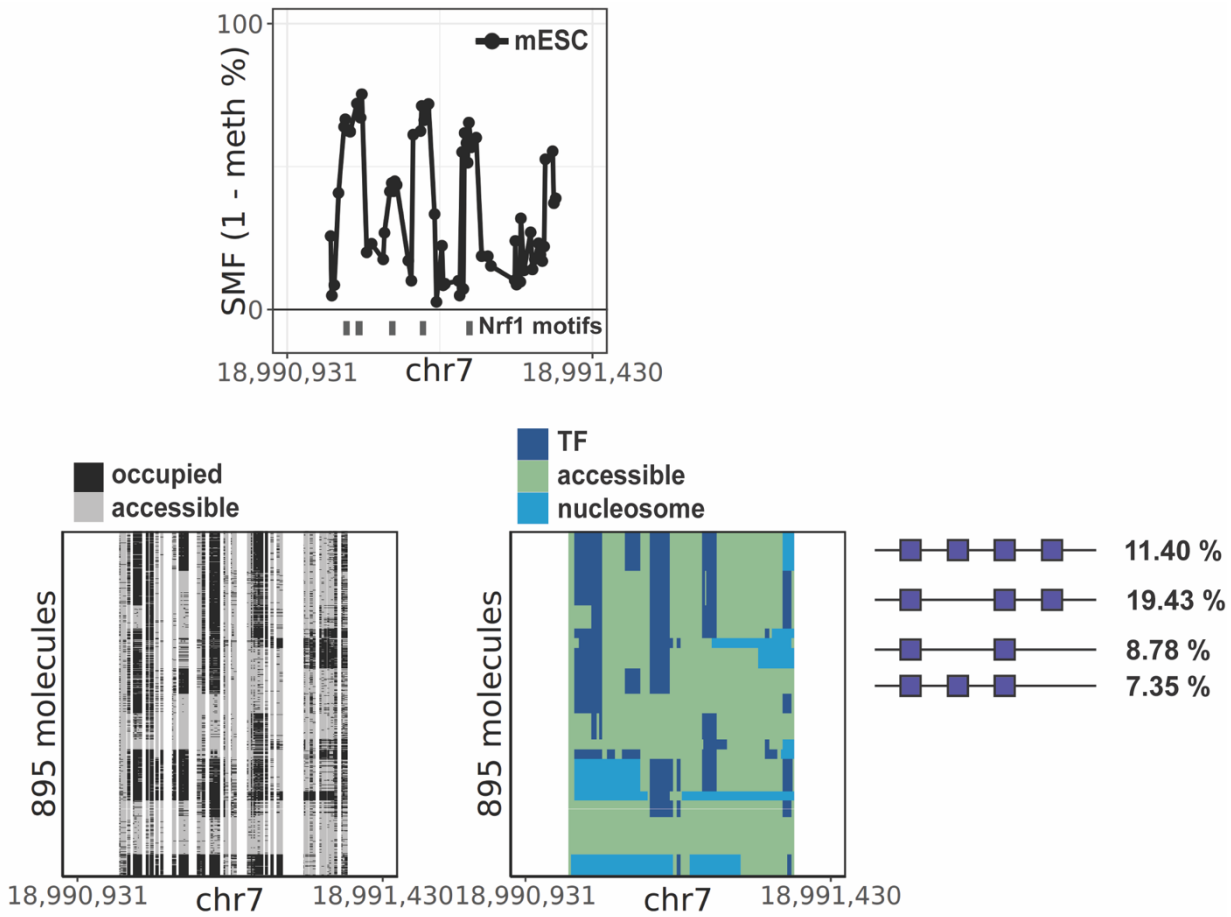

Figure 2.3. *FootprintCharter* dissects the combinatorics of co-binding at larger TF clusters. Example of a single locus occupied by a cluster of Nrf1 motifs. The upper panel shows the average SMF signal (1-methylation %) at individual GpCs and CpGs. The two lower heatmaps show the stacks of single molecules where each row is a single molecule and each column is a single cytosine (left) or nucleotide (right). Each cell is colored by the binary methylation signal (left) or by the molecular state quantified by *FootprintCharter* (right). Differently from what was possible with "classical single molecule sorting", *FootprintCharter* dissects the frequencies of co-binding among larger TF clusters.

2.1.3.1 | Overview of the FootprintCharter workflow and implementation

Differently from “single molecule sorting”, I aimed to quantify single molecule occupancy patterns independently of any annotation. This implies that I was not able to center the quantification of SMF data around previously annotated TF binding sites (Figure 2.2). Therefore, I designed *FootprintCharter* to work around generic genomic windows of interest which can simply be defined by genomic coordinates spanning any width. The expected length of the molecules for the specific dataset at hand should be considered when designing such windows of interest.

FootprintCharter takes as input single molecule methylation call matrixes as produced by the function CallContextMethylation of the *SingleMoleculeFootprinting* package. The first step is to smooth the single molecule binary methylation values. I do this by computing a sliding average over 40bp at each genomic position using custom R functions.

Secondly, I compute a matrix of all possible pairwise Euclidean distances among single molecules using the parDist function of the parallelDist R package. I perform unsupervised clustering on such distance matrix to obtain k partitions by using the *k*-medoids algorithm implemented in the function pam of the cluster R package (Figure 2.4). The number *k* of partitions is defined for each locus individually based on coverage and molecular complexity. Starting from a maximum value of *k*, I consider the clustering successful if each of the *k* resulting partitions is populated by at least *n* molecules. If this is not the case, I repeat the procedure setting the number of required partitions to *k*-1. I iteratively repeat the procedure until the number of molecules populating each partition is at least *n*. If a locus cannot be clustered in at least two partitions, I consider the procedure unsuccessful for that genomic locus. If the procedure is successful, the result is that the different molecular states represented in the data are quantified, yet still uncharacterized. The values of *k* and *n* can be provided by the user. When quantifying bait capture SMF sequenced with 150bp paired-end reads (Kleinendorst et al. 2021; Sönmezer et al. 2021), the starting value of *k*=12 is high enough to characterize most of the biological complexity that can be captured with this sequencing modality while keeping a feasible running time. Setting the minimum number of molecules per partition to 5 also resulted optimal for this sequencing modality.

To characterize the different molecular states represented in the data, I perform the footprint detection step of *FootprintCharter* using the non-smoothed binary methylation

matrix. I perform the following procedure separately for each partition resulting from the clustering step. I calculate the median methylation for each cytosine and define them as occupied if their resulting median methylation value is < 0.5. I than evaluate how many genomic base pairs are consecutively occupied or accessible and how many base pairs these stretches correspond to. Because SMF does not measure single molecule chromatin accessibility with single nucleotide resolution, and because the exact distance between consecutive cytosines varies based on the genomic sequence of each locus, whenever I detect a transition between an accessible and an occupied stretch, I extend both such stretches to the middle of the two confining cytosines in order to close the gap. At this point, each nucleotide on each single molecule has been annotated as either occupied or accessible (Figure 2.4).

To distinguish whether occupancy is explained by transcription factors or nucleosomes, I further refine this annotation by classifying occupancy in different categories based on width. I define stretches of occupancy between 5bp and 75bp in length as TF footprints, stretches longer than 120bp as nucleosome footprints. I consider footprints shorter than 5bp as noise and annotate the stretch they cover as accessible. I discard as unrecognized those footprints which are not flanked by an accessible cytosine on both sides, i.e., footprints which are positioned at the edge of molecules and for which the full width cannot be evaluated. At this point, each nucleotide on each single molecule has been annotated as either occupied by a TF, a nucleosome or as accessible.

Then, I annotate TF footprints by overlapping them with a list of predicted TF binding sites which are validated using ChIP-seq dataset as described in Methods and Materials.

FootprintCharter detects footprints separately on each partition and measuring TF-DNA interactions is associated with a certain level of biological and technical noise. Because of that, TF footprints that are biologically equivalent, i.e., occupy roughly the same genome sequence space and are annotated with the same TF identity, are often measured as heterogeneous because they do not completely overlap. Because of this, I implemented a last step in which I aggregate TF footprints that are biologically equivalent, i.e., that are annotated with the same TF identity and that overlap by at least 75% of the genomic space occupied by the shorter one. This step allows me to aggregate read count across partitions and to correctly estimate the frequency of TF occupancy across a cell population (Figure 2.4).

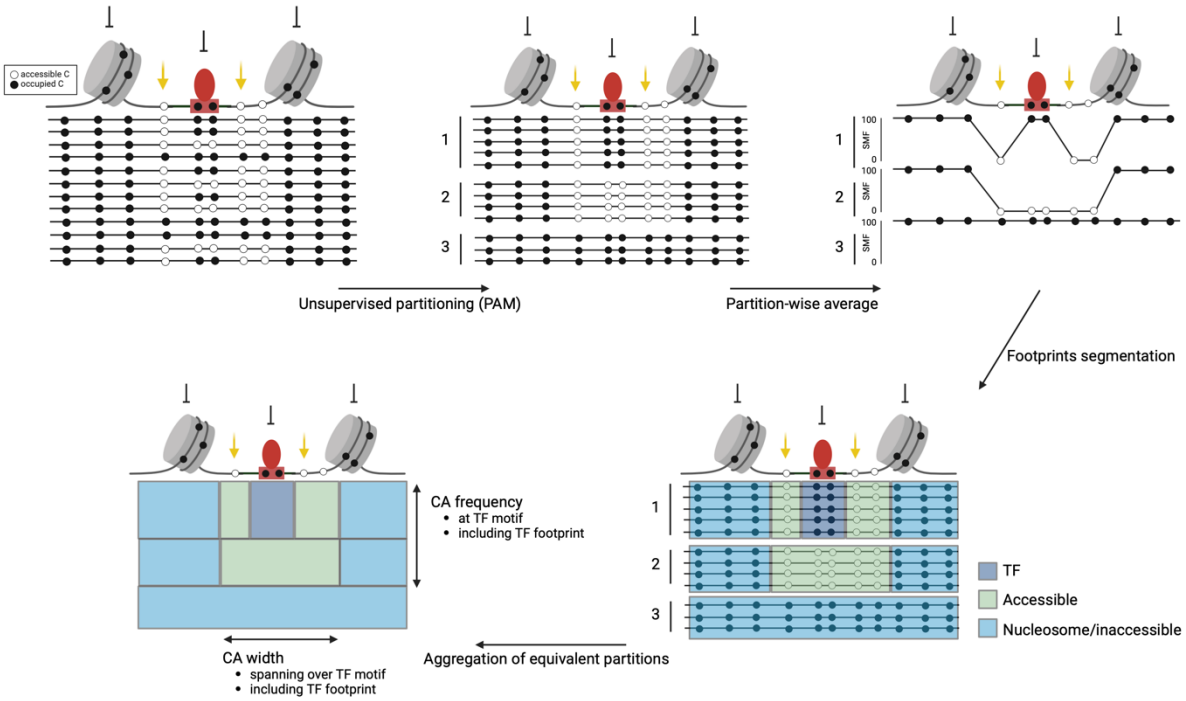

Figure 2.4. Overview of the *FootprintCharter* workflow. *FootprintCharter* is a novel computational framework for the unsupervised detection and quantification of footprints from single molecule footprinting data. The first core steps in the workflow consists of clustering single molecules into a number of partitions which adapts to the measured complexity of the *cis*-regulatory element under examination, i.e., the heterogeneity of the single molecule accessibility patterns as well as the number of molecules covering of the locus. The second core step involves the segmentation of footprints from each resulting partition. Footprints detected as such are identified as either TFs or nucleosomes based on their width. The annotation of TF footprints with predicted TF motifs aids in the aggregation of biologically equivalent footprints, i.e., footprints for the same TF that differ by a few nucleotides on account on biological and technical noise.

2.2 | Investigating the contribution of transcription factors to chromatin accessibility

In mammalian genomes, gene expression is regulated by transcription factors at CREs. In order to exert their function, TFs first need to access and occupy their binding sites and nucleosomes impose a physical barrier to this process (Bell et al. 2011; Iurlaro et al. 2021). Certain TFs, such as the pioneers Oct4, Sox2, and Klf4, are able to outcompete nucleosomes, to establish chromatin accessibility at CREs and consequently to initiate a transcriptional program that can lead to cell differentiation or reprogramming (King and Klose 2017; Zaret and Carroll 2011). TF-nucleosome competition also exists at the steady state, while cells are not undergoing differentiation. There, TFs constantly outcompete nucleosomes to maintain chromatin accessibility at CREs in order to maintain transcriptional homeostasis (Iurlaro et al. 2021; Schick et al. 2021).

Studying the contribution of individual TF instances to chromatin accessibility is hindered by the complex nature of mammalian CREs. On average, CREs contains 5-6 motifs for heterogenous TFs that can be in redundancy among each other or combine additively or cooperatively towards the maintenance of chromatin accessibility (Avsec, Weilert, et al. 2021; Kim and Wysocka 2023; Vierstra et al. 2020; Weingarten-Gabbay and Segal 2014). In particular, it is unclear how much each individual TF motif contributes to the maintenance of the nucleosome depleted region (NDR) at CREs. It is unclear how TFs combine for that and how the width of the NDR at CREs is maintained.

In this study, I applied *FootprintCharter* to precisely quantify the contribution of TFs to the maintenance of chromatin accessibility at CREs, both individually and in combination with each other. More precisely, I quantified the frequency of observing NDRs at CREs in a cell population as well as their width. To further investigate the role of TFs, Rozemarijn Kleinendorst performed SMF on two mouse F1 lines to measure how the natural genetic variation between species would systematically perturb TF binding and consequently affect their ability to maintain CREs accessible. I quantified allele-specific chromatin accessibility frequency and NDR width at CREs using *FootprintCharter*.

I observed that most TF instances are not necessary for the maintenance of chromatin accessibility at CREs. In particular, most CREs show only partial to no reduction in the frequency

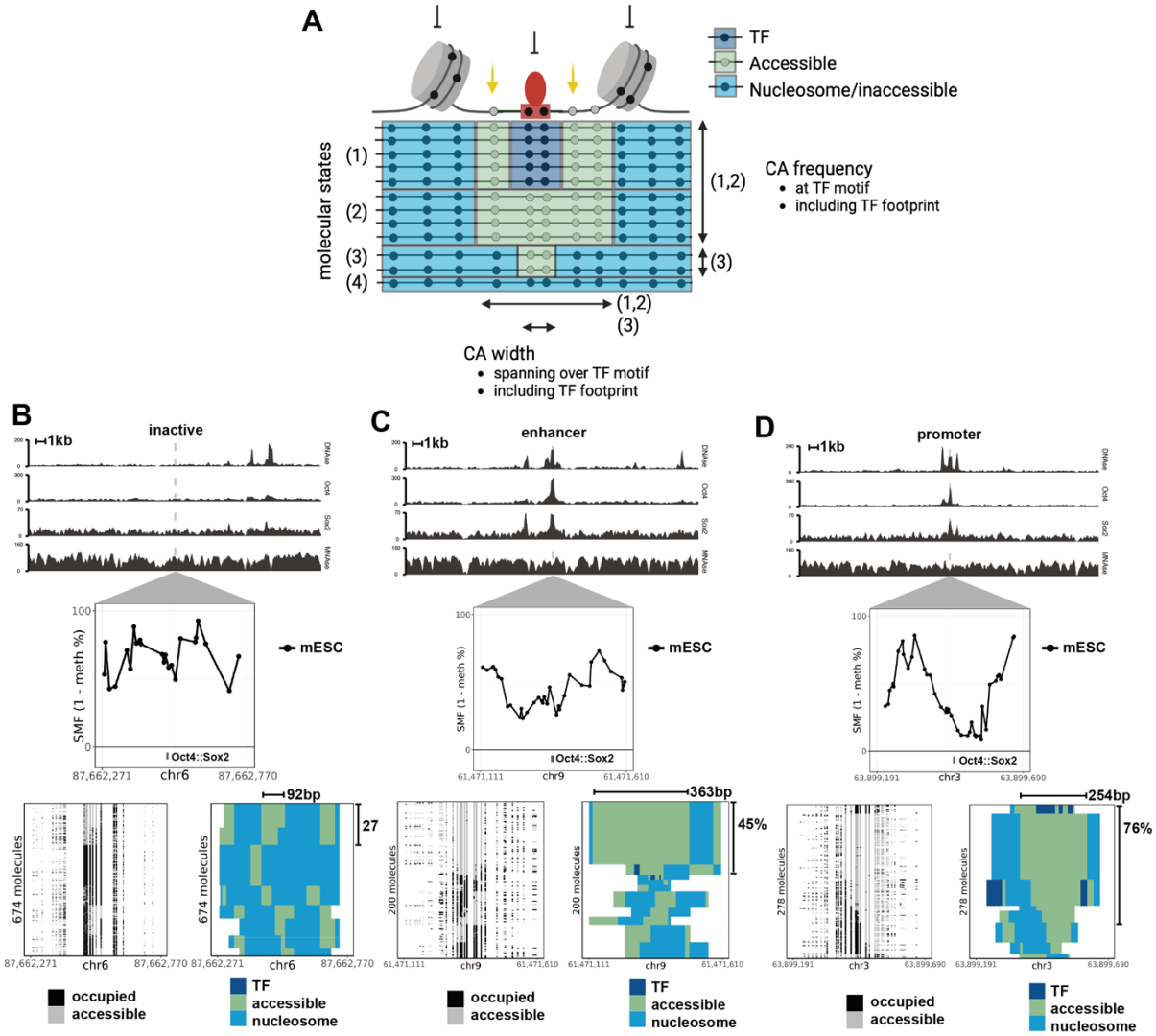

Figure 2.5. *FootprintCharter* quantifies single molecule chromatin accessibility. A. *FootprintCharter* summarizes two features of single molecule chromatin accessibility for each TF motif: the frequency of accessibility in the cell population and the single molecule widths of the nucleosome depleted region (NDR) **B.** Example of inactive single locus containing a composite motif for Oct4::Sox2 for which chromatin accessibility associated with linker DNA is detected in 27% of the cell population. The upper panel shows genomic tracks for DNase-seq, ChIP-nexus for Oct4 and Sox2 and MNase-seq around the Oct4::Sox2 motif. The middle panel shows the average SMF signal (1-methylation %) at individual GpCs and CpGs. The two lower panels show the stacks of single molecules sorted by decreasing width of NDR either as binary accessibility signal (left) or as molecular state as quantified by *FootprintCharter* (right). The maximum NDR width is annotated above the right hand stack **C.** Example of active single enhancer locus containing a composite motif for Oct4::Sox2 for which chromatin accessibility is detected in 45% of the cell population both in association with linker DNA and regulatory activity (>100bp) **D.** Example of active single promoter locus containing a composite motif for Oct4::Sox2 for which chromatin accessibility is detected in 76% of the cell population in association with regulatory activity (>100bp).

of chromatin accessibility upon loss of TF binding. However, when observed at single molecule, CREs lose their NDR space completely rather than partially shrinking. This indicates that, at each single molecule, chromatin accessibility is maintained in an all-or-nothing fashion rather than modularly. I validated these results by analyzing a SMF dataset produced by Laura Moniot-Perron upon rapid and acute Sox2 protein depletion using a previously reported degradation TAG system.

2.2.1 | *FootprintCharter* quantifies CRE-specific TF-nucleosome competition

Formally, single molecule footprinting measures chromatin accessibility at the single molecule level. This implies that it measures chromatin accessibility in its frequency within a cell population rather than in the form of enrichment of read counts against the genomic background like bulk assays such as ChIP-seq and ATAC-seq do. Because of this quality, SMF offers an upper bound to the estimated frequency of usage of a CRE in a cell population. Secondly, SMF measures the width of NDRs at CREs at the single molecule level with a high spatial resolution. For experiments involving a double enzyme footprinting (GpCs and CpGs), this resolution is on average of one measured cytosine every 7bps (Kleinendorst et al. 2021).

I employed my newly developed tool *FootprintCharter* to quantify these two metrics at annotated TF motifs in the mouse genome (Figure 2.5A). I applied *FootprintCharter* to a previously generated high-coverage SMF dataset in mESCs (Sönmezer et al. 2021). At TF motifs, I detected short (<100bp) accessibility stretches characteristic of linker DNA between nucleosomes (Figure 2.5B, C) as well as larger (>100bp) stretches of accessibility occurring at active enhancers (Figure 2.5C) or promoters (Figure 2.5D). For instance, a sequence motif for Oct4-Sox2 where no TF binding is detected by ChIP-seq, is found accessible in 27% of the molecules with a maximum estimated NDR width 92 base pairs (Figure 2.5B).

To validate that this lengths are compatible with the random exposure of the motif at linkers upon movement of unphased nucleosomes (Dombrowski et al. 2022; Fernandez Garcia et al. 2019; Ohno et al. 2019; Ricci et al. 2015; Song et al. 2014), I reanalyzed a previously published single molecule footprinting dataset upon genetic knock-out of Rest (Sönmezer et al. 2021). Rest is a strong DNA binding transcriptional repressor often found in isolation at CREs

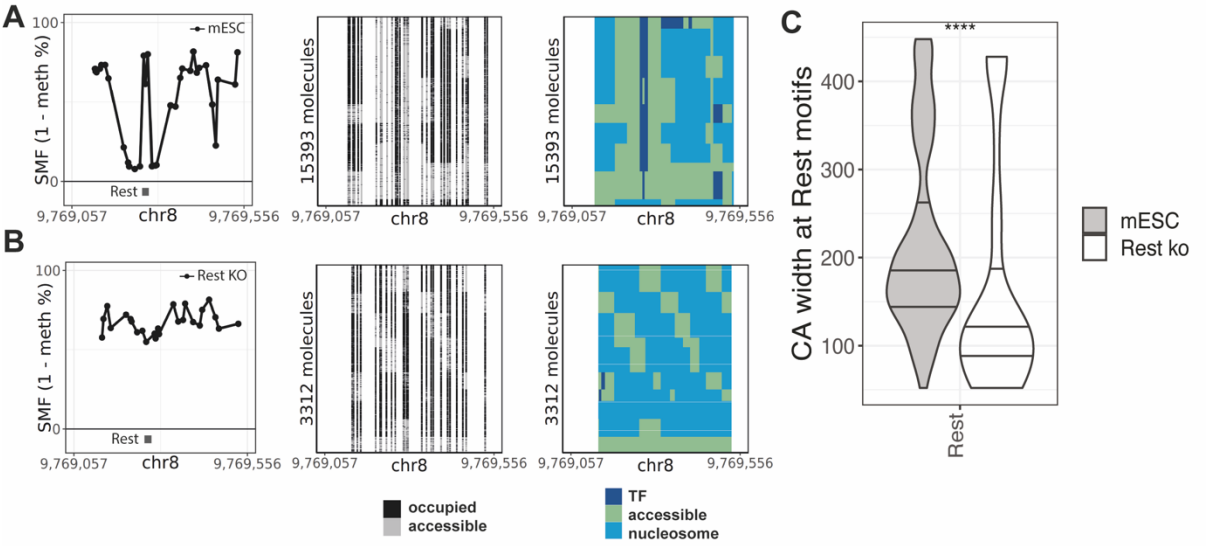

Figure 2.6. The genetic knock-out of Rest validates the distribution of linker DNA lengths quantified by *FootprintCharter*. A. Single locus example of Rest binding associated with TF binding and with larger (>100bp) widths of nucleosome depleted regions (NDRs) **B.** Same single locus example as in A. upon genetic knock-out of Rest. The locus is associated almost exclusively with nucleosome occupancy and accessibility is only associated with linker-DNA (<100bp). **C.** Distributions of NDR widths across multiple loci in wild-type mESCs and upon genetic knock-out of Rest.

(Z.-F. Chen, Paquette, and Anderson 1998). Measuring the width of chromatin accessibility at unbound Rest motifs reveals the emergence of stretches of accessibility below 100bp in width, confirming them as linker (Figure 2.6).

In contrast to an unbound Oct4-Sox2 motif (Figure 2.5B), I observed linker DNA co-occurring with larger accessibility stretches at equivalent motifs in active enhancers or promoters where TF binding is detected (Figure 2.5C,D). Specifically, I observed continuous accessibility over >100bp occurring in 45% (enhancer, Figure 2.5C) and 76% (promoter, Figure 2.5D) of the sampled DNA molecules. At the active enhancer (Figure 2.5C) I also detect linker DNA in 15% of the molecules. Consistently, I mainly detect chromatin accessibility associated with linker DNA (up to 97bp in median width, up to 37% in median frequency) at unbound Oct4 and Sox2 motifs across the genome (Figure 2.7A). Whereas, I also detect regulatory chromatin accessibility (>180bp in median width, >52% in frequency) at bound Oct4 and Sox2 in active CREs (Figure 2.7A). This suggests that unbound motifs for Oct4 and Sox2 are free of nucleosomes, and available for binding in ca. one-third of the cells at any given time, and that accessibility is increased in its frequency, and width, upon binding of the TFs. To generalise these observations, I calculated the frequency and width of chromatin accessibility at all motifs bound by any TF expressed in mESCs at various CRE types across the genome (Figure 2.7B). At

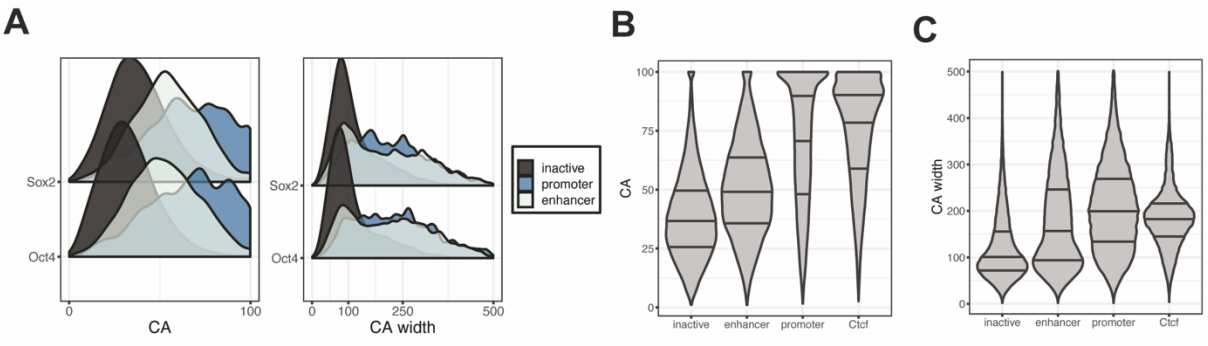

Figure 2.7. Genome-wide quantification of single molecule chromatin accessibility features validate CRE-specific TF-nucleosome competition. A. Distribution of single molecule chromatin accessibility frequency (left) and width (right) for Oct4 and Sox2 motifs across the mouse genome for different classes of CREs. **B.** Distribution of single molecule chromatin accessibility frequency for all measured TF motifs across the mouse genome for different classes of CREs. **C.** Distribution of single molecule chromatin accessibility widths for all measured TF motifs across the mouse genome for different classes of CREs. Enhancers display both kinds of widths associated one with linker DNA and the other with regulatory activity, indicating predominant TF-nucleosome competition.

inactive chromatin where minimal TF binding is expected to happen, TF motifs are accessible in 37% of the molecules in median, with short stretches of 101bp in median (Figure 2.7C). This frequency is higher at active enhancers (49%), promoters (76%) and insulators (80%), where such short stretches co-exist with larger stretches of accessibility of 157bp, 200bp and 183bp in median respectively (Figure 2.7B,C). Together, this shows that the frequency and the widths of chromatin accessibility is heterogeneous at active CREs, probably reflecting the specificity of competition between nucleosomes and multiple TFs at these loci.

2.2.2 | The combinatorial contribution of TFs to chromatin accessibility

To estimate the contribution of individual TFs to opening chromatin, I asked if the binding of specific TFs correlates with increased accessibility at their binding sites. I used publicly available ChIP-seq and ChIP-nexus (Avsec, Weilert, et al. 2021) datasets to distinguish bound from unbound motifs and calculated the frequency of chromatin accessibility for 19 TFs expressed and measurable in mESCs by SMF. I observed a strong shift in the frequency of accessibility (median >79%) at the binding sites for 5 of the tested TFs (Figure 2.8A). The

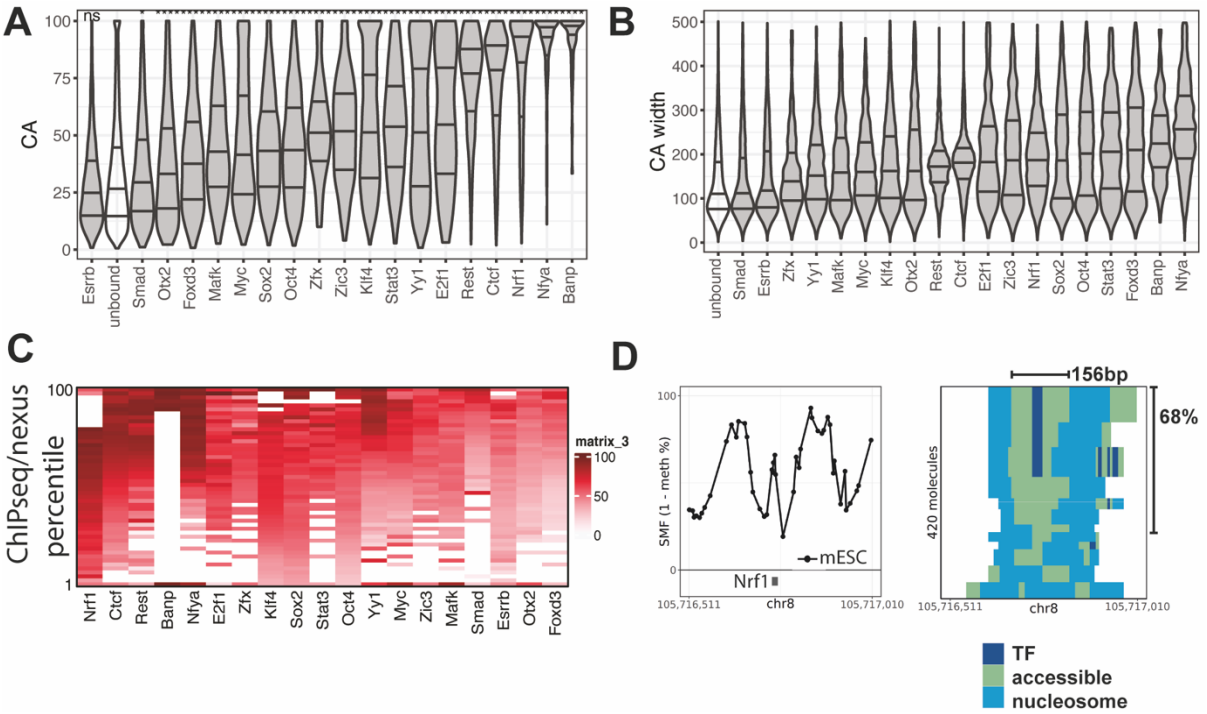

Figure 2.8. TF-specific contribution to chromatin accessibility. A. Frequencies of chromatin accessibility associated with regulatory activity (>100bp) at bound TF motifs. The number of stars stands for the significance of a t-test performed for each TF against the distribution of frequencies at unbound motifs (ns: p>0.05; *: p<=0.05; **: p<=0.01; ***: p<=0.001; ****: p<=0.0001). **B.** Widths of single molecule chromatin accessibility stretches measured at bound TF motifs **C.** Median values of chromatin accessibility frequencies associated with regulatory activity as a function of the percentile of ChIP-seq/-nexus signal at TF binding sites. **D.** Single locus example of a CRE containing one bound Nrf1 motif that is accessible in 68% of the cell population.

presence of these TFs is associated with a very high frequency of accessibility in the cell population, at a large fraction of their bound sites (i.e., more than 50% of the BANP binding sites are 100% accessible), (Figure 2.8A). These include TFs known to be able to open chromatin and bind individually (CTCF, REST and BANP), but also the histone-fold domain factor (NFY) and the general transcriptional activator NRF1 (exemplified in Figure 2.8D). The presence of most other TFs is also associated with an increased probability to find chromatin accessible (t-test p<0.01), albeit to a lesser extent (33%-58%, Figure 2.8A). These TFs include well characterised regulators of pluripotency such as Oct4, Sox2, Klf4 or Myc (OSKM), that have been shown to be critical pioneers for chromatin accessibility in this system (Chronis et al. 2017; Dodonova et al. 2020b; Soufi et al. 2015; Soufi, Donahue, and Zaret 2012). Higher occupancy by these TFs, as measured by ChIP-seq/-nexus, scales with increased frequency of

chromatin accessibility, but the maximal accessibility plateau greatly differs between TFs (Figure 2.8C). This suggests the existence of at least two classes of TFs with distinct quantitative contributions to chromatin accessibility.

In terms of width, the ranking of TFs changes according to their mode of competition against nucleosomes. Ctcf and Rest display an enrichment in chromatin accessibility width at ca. 180bp. This indicates their tendency to bind in isolation and to largely and autonomously outcompete nucleosomes. The distributions for Nrf1, Banp and Nfya are largely shifted towards widths larger than 200bp (Figure 2.8B). This indicates that they bind to regions in which sequence features beyond the TF motif might be needed to largely outcompete nucleosomes. All other TFs, including OSKM, show multimodal distribution that include chromatin accessibility widths associated with linker DNA showcasing the prominence of TF-nucleosome competition at these regions.

Because only a minority of TFs are associated with high frequencies (>75%) of open chromatin, I hypothesised that the remaining TFs might synergise towards full chromatin accessibility by combinatorial assembly. To that end, I tested whether CREs harbouring an increasing number of motifs for each TF have a higher probability of being accessible (Figure 2.9A). At Klf4 bound CREs, the percentage of accessible molecules increases from a median of 58% in case of a single Klf4 motifs to a median of 83% when more than 3 Klf4 motifs are present (wilcox test $p<0.01$, Figure 2.9A, B). This is in contrast with NRF1, where accessibility already reaches 87% in median upon binding of a single TF and increases more moderately at sites with multiple binding sites (wilcox test $p<0.01$, Figure 2.9A, B). This suggests that Klf4 binding sites strongly synergise at CREs, and that more than 3 motif instances are required to match the chromatin accessibility frequency associated with a single Nrf1 motif (Figure 2.9A, B). This contrast is further evident when inspecting individual CREs containing 1, or 6 motif instances bound by Klf4 (Figure 2.9C, D) or 1 and 2 motif instances bound by Nrf1 (Figure 2.8D, Figure 2.9C). Furthermore, the combinatorial assemble of TFs also leads to an increase in NDR width. This is most evident for Nrf1, Klf4 and Rest. (Figure 2.9F).

Together these correlative analysis suggests individual instances of most TFs, including pioneer TFs, lead to a modest increase in chromatin accessibility and that combinatorial assembly of TFs motifs is essential to explain chromatin accessibility observed at enhancers and promoters.

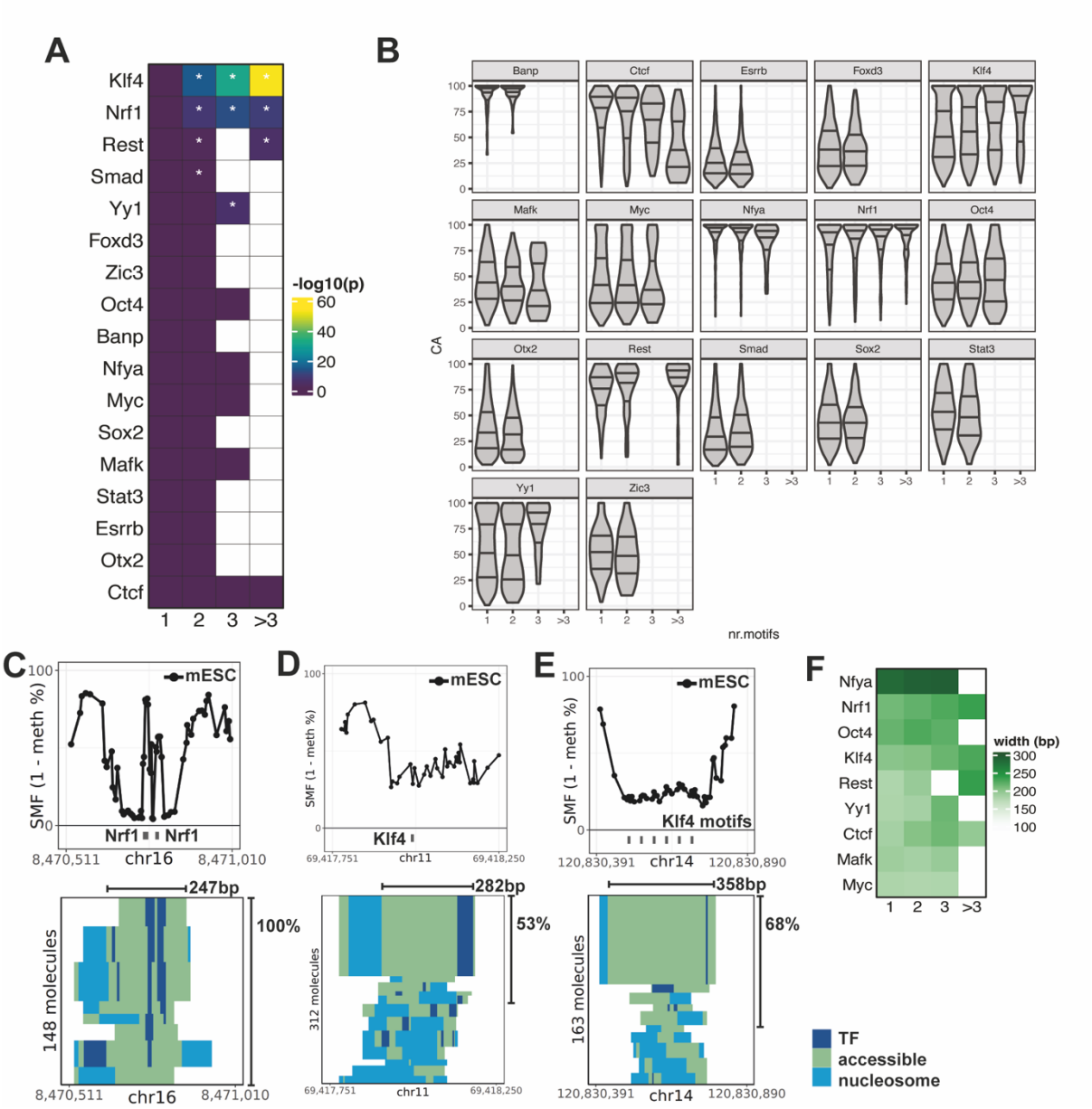

Figure 2.9. The combinatorial contribution of TFs to chromatin accessibility. A. -log10(p-value) resulting from testing for more frequent chromatin accessibility at CREs as a function of increasing number of bound TF motifs. The star indicates a statistically significant increase. **B.** Distributions of chromatin accessibility frequencies associated with regulatory activity for a CREs with an increasing number of TF-bound motifs. **C.** Single locus example of a CRE with two Nrf1 binding sites associated with 100% chromatin accessibility in the cell population. **D.** Single locus example of a CRE with one Klf4 binding site associated with 53% chromatin accessibility in the cell population. **E.** Single locus example of CRE with six Klf4 binding sites associated with 68% chromatin accessibility in the cell population. **F.** Median value of chromatin accessibility width for CREs containing an increasing number of bound TF motifs.

2.2.3 | Most TF instances are dispensable to maintain the frequency of CRE usage in the cell population

To validate the relative contribution of each TF instance to chromatin accessibility, I tested changes in chromatin accessibility upon controlled perturbation of individual or combination of TFs at CREs. I leveraged the natural genetic variation between different mice

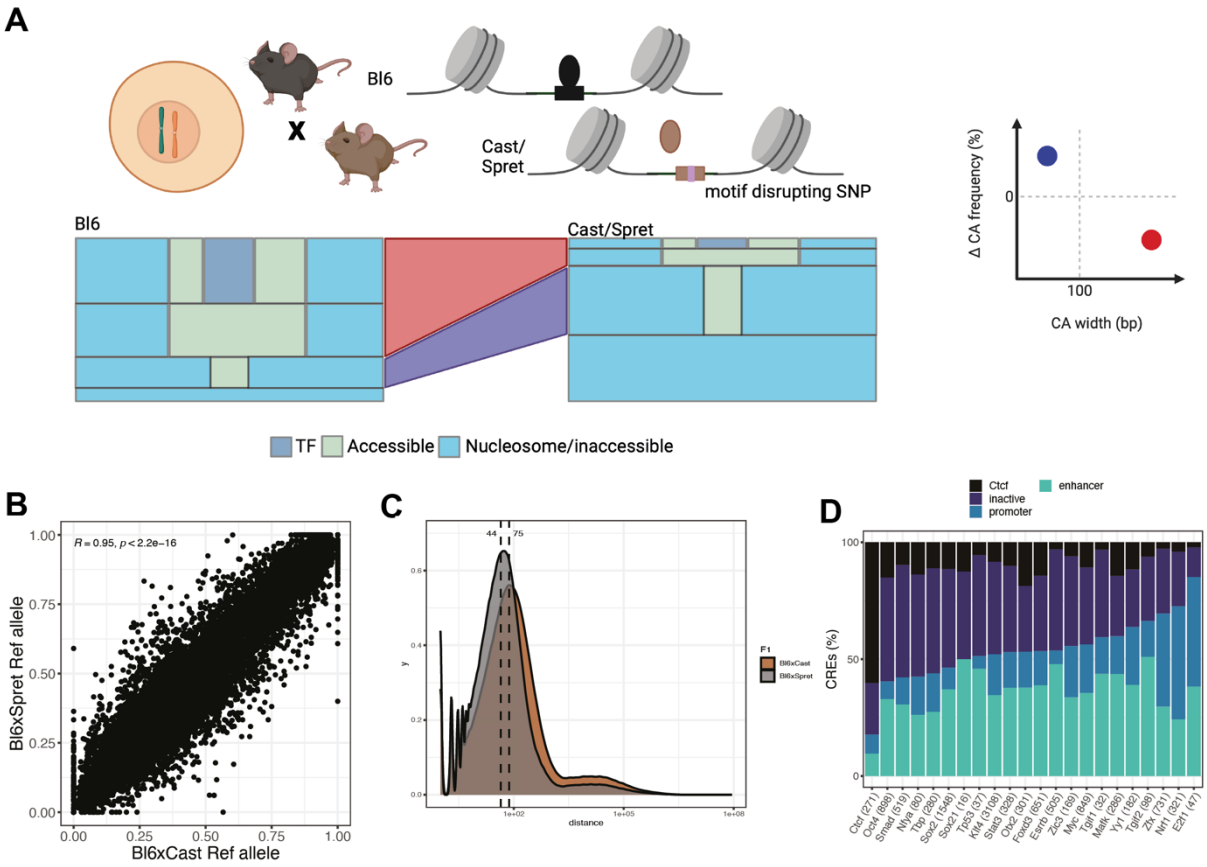

Figure 2.10. Systematic perturbation of TF binding at CREs. A. Systematic perturbation of TF binding across the mouse genome through natural genetic variation among mouse species crossed into F1 lines. Single molecules from both alleles in each F1 are run through *FootprintCharter* in a pool to systematically compare the allele-specific changes in frequency of each molecular state. The difference across alleles for each molecular state is summarized with two quantities: the width of its NDR and the delta in chromatin accessibility frequency. These quantities are plotted against each other as depicted on the right. **B.** Distribution of chromatin accessibility frequencies associated with regulatory activity (NDR width >100bp) for the Bl6 allele across F1s. The Pearson correlation coefficient (R) and p-value are indicated. **C.** Distribution of distances between consecutive SNPs in the Bl6xCast, brown, and Bl6xSpret, grey, F1s. The medians are indicated by vertical dashed lines. **D.** Distribution and number of TF motifs with a strong loss in affinity across alleles and types of CREs. The data for this figure was produced by Rozemarijn Kleinendorst.

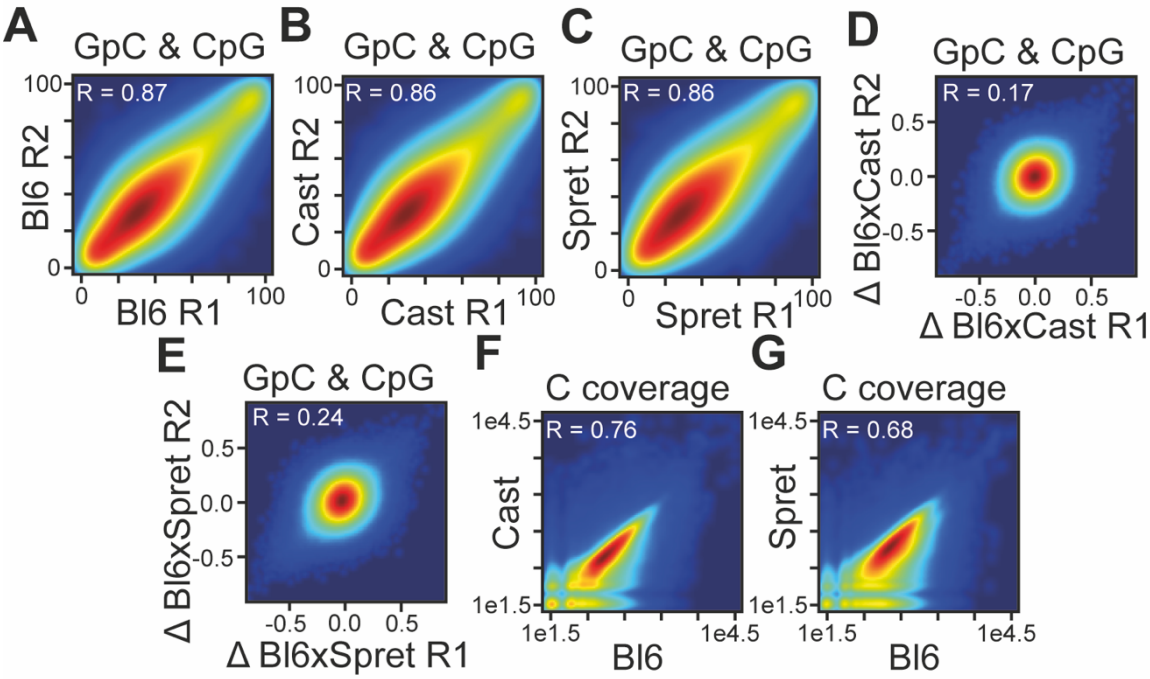

Figure 2.11. Depth and reproducibility of F1 SMF sequencing. Distributions of methylation rates (%) for GpCs and CpGs across biological replicates for the Bl6 (**A**), Cast (**B**) and Spret (**C**) alleles. The Pearson correlation coefficient (R) is indicated. Distributions of methylation rates deltas across biological replicates for the Bl6xCast (**D**) and Bl6xSpret (**E**) F1s. The Pearson correlation coefficient R is indicated. Distribution of cytosine coverage across alleles for the Bl6xCast (**F**) and Bl6xSpret (**G**) F1s. The Pearson correlation coefficient R is indicated. The data for this figure was produced by Rozemarijn Kleinendorst.

species to study the impact of nucleotide variants falling into TF binding motifs on the frequency of chromatin accessibility. Rozemarijn Kleinendorst performed SMF in two F1 lines derived from crosses between Mus musculus domesticus (Bl6) and Mus musculus castaneus (Cast), as well as Mus spretus (Spret), (Figure 2.10A). Both alleles are measured in the same nuclear environment, controlling for possible trans effects, such as changes in TF concentrations that may occur between species (Figure 2.10B) (Floc'hlay et al. 2021; Goncalves et al. 2012; Panten et al. 2024). These cell lines are characterised by a high density of Single Nucleotide Variants (SNPs) between the two alleles (1/44 bp and 1/75 bp for Bl6xCast and Bl6xSpret) (Figure 2.10C) (Keane et al. 2011; Lilue et al. 2018; Thybert et al. 2018). This high SNP density leads to 11,257 measurable perturbations within the motifs of the TFs expressed in my system, generating hundreds of perturbations of each TF in various genomic contexts (Figure 2.10D). Rozemarijn Kleinendorst sequenced the samples at high depth to enable reproducible quantification of chromatin accessibility differences between alleles (Supplementary Figure 2.11A-G). Figure 2.12A examplifies a Ctcf binding motif perturbed by a

single SNP for the Bl6xCast F1. This SNP imposes a strong reduction in the affinity for CTCF, leading to a 78% loss of TF binding as well as a 60% increase in nucleosome occupancy. This is in contrast with loci in which the

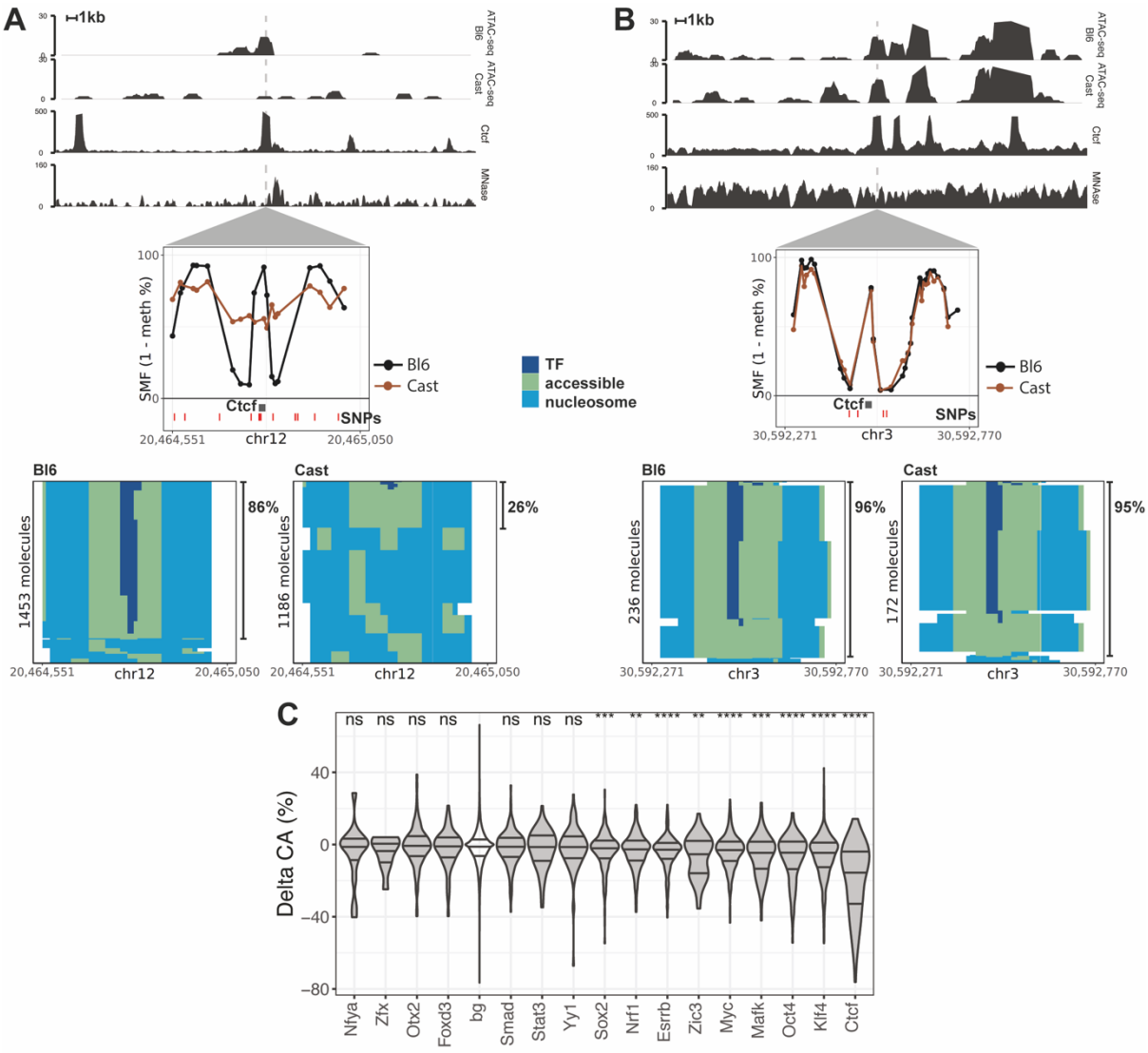

Figure 2.12. Single instances of pluripotency TFs are largely dispensable for maintaining accessibility at CREs. A. Single locus example of a bound Ctcf motif with a strong difference in affinity between the Bl6 and Castaneus alleles. This genetic variation is associated with a reduction in chromatin accessibility frequency of 60% in the cell population. **B.** Single locus example of a bound Ctcf motif with no difference in affinity between the Bl6 and Castaneus alleles and no change in chromatin accessibility frequency. **C.** Distributions of chromatin accessibility frequency deltas associated with the loss of TF binding affinity across alleles. The number of stars stands for the significance of a t-test performed for each TF against the distribution of frequencies at motifs that are genetically equivalent across alleles (ns: p>0.05; *: p<=0.05; **: p<=0.01; ***: p<=0.001; ****: p<=0.0001). The data for this figure was produced by Rozemarijn Kleinendorst.

CTCF motif is genetically identical between the two alleles and no difference in the distribution of molecular states is observed (Figure 2.12B).

Globally, the reduced motif affinity for Ctcf is associated with a frequent and strong loss of chromatin accessibility (t-test p-value < 0.01, Figure 2.12C) indicating that Ctcf is primarily responsible for the maintenance of chromatin accessibility at its motifs across the genome.

For 9 of the 16 TFs expressed in my system and with enough motif instances strongly perturbed by a SNP (see Methods and Materials section 4.2.16), I observed a significant enrichment for reduced chromatin accessibility frequency across alleles as a consequence of reduced motif affinity (t-test p-value < 0.01). Of these, Ctcf is associated with the largest reductions in chromatin accessibility frequencies, followed by TFs such as Oct4, Klf4, Mafk and Sox2, albeit with much lower effect sizes (Figure 2.12C).

Overall, the loss of TF-motif affinity leads to moderate loss of CRE activity, suggesting that the observed accessibility is the result of the cumulated function of multiple TFs and that the loss of binding of single TF instances rarely affects the frequency of CRE usage in the cell population.

2.2.4 | The width of NDRs at CREs is maintained in an all-or-nothing fashion

To further investigate the mechanistic contribution of TFs to chromatin accessibility, I focused on the subset of TF motifs whose loss of function results in a statistically significant reduction in chromatin accessibility across alleles (fisher's test p-value < 0.05). I excluded the CREs with more than one loss of function motif. I measured the width of the nucleosome depleted regions at the TF motifs for the molecular states that are either enriched or depleted across alleles. I observed that for TFs including the pluripotency regulators Esrrb, Oct4, Sox2 and Klf4, the enriched states are largely populated by accessibility associated with linker DNA (<100 bp) (Figure 2.13A, B). Notably, I did not observe any enrichment for molecular states with NDRs of intermediate widths, i.e., shorter than the unperturbed allele but longer than linker DNA. This implies that, at these genomic loci, the loss of a single TF binding event means the complete loss of function of the whole CRE (Fig. 2.13A, B, C). This is likely driven by the cooperative interactions within the clusters of TFs the pluripotency regulators are embedded in. TFs including Ctcf, Myc and Smad see the additional enrichment for larger accessible fractions (>100bp) upon motif loss of function. This suggests that the function of these CREs does not globally depend on the binding of such TFs.

These results indicate that for the pluripotency TFs, the combinatorial assembly of TFs increases the probability of usage of CREs in the cell population. However, at single molecule, CREs are maintained free of nucleosomes in an all-or-nothing fashion rather than modularly. This suggests that the width of NDRs is not directly determined by the physical displacement of nucleosomes by TFs but perhaps by the specific chromatin remodeler recruited at the CRE.

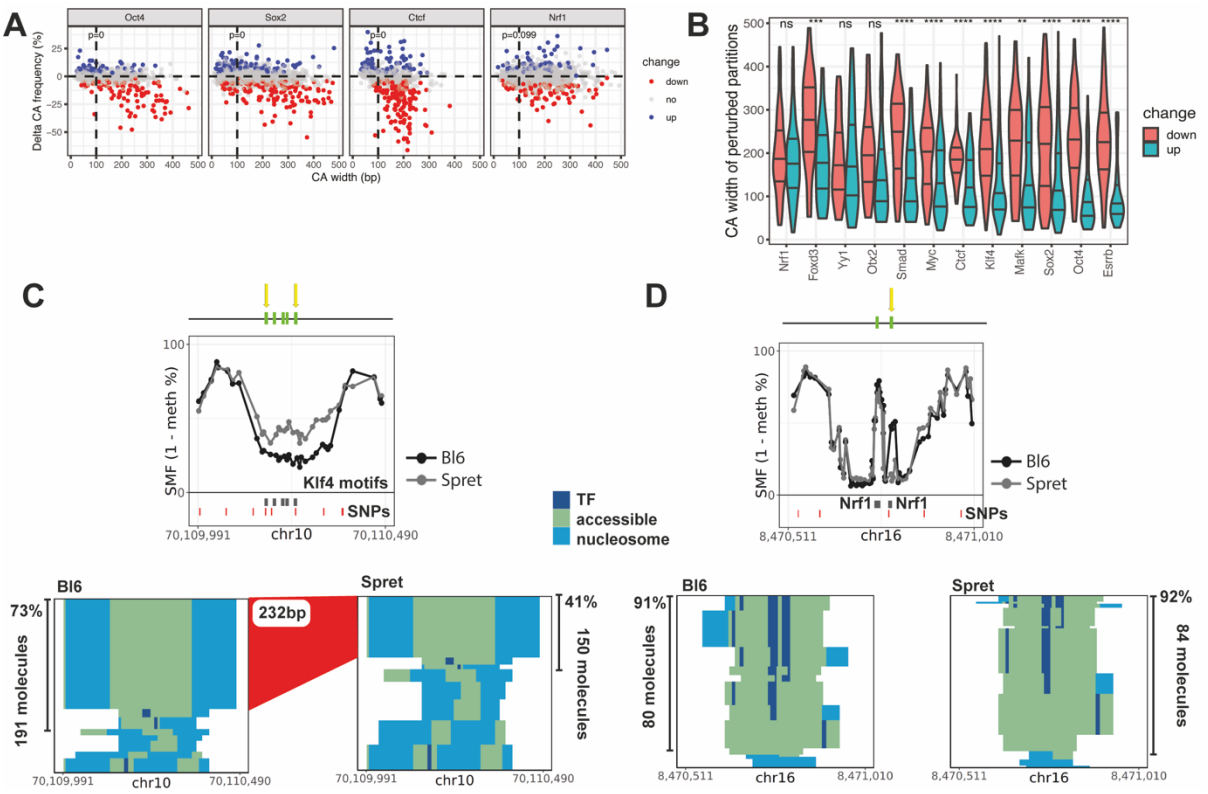

Figure 2.13. The width of NDR at pluripotency CREs is maintained in an all-or-nothing fashion. A. Allele-specific frequencies (y-axis) of molecular states with different NDR widths (x-axis). Oct4, Sox2 and Ctcf are largely associated with significant (Fisher's test p<0.05) enrichment of linker-long NDRs (upper left quadrant in each scatter-plot) but not of wider (>100bp) NDRs such as in the case of Nrf1. Each scatter plot is annotated with the Wilcoxon p-value testing the global difference in NDR width distribution between enriched and depleted molecular states. **B.** Distribution of NDR widths for molecular states which are either enriched (blue) or depleted (red) upon loss of TF binding affinity. The number of stars stands for the significance of a Wilcoxon test performed for each TF to compare the distributions of enriched and depleted widths (ns: p>0.05; *: p<=0.05; **: p<=0.01; ***: p<=0.001; ****: p<=0.0001). **C.** Single locus example of a CRE containing five Klf4 motifs of which the most upstream and the most downstream ones are subject to a strong reduction in binding affinity. Across alleles, the frequency of chromatin accessibility at these CREs is significantly reduced (p<0.05) however there is no enrichment for an accessible state with intermediate NDR width. **D.** Single locus example of CRE containing two Nrf1 motifs of which the downstream one is subject to a strong reduction in binding affinity. No allele-specific chromatin accessibility is detected at this locus. The data for this figure was produced by Rozemarijn Kleinendorst.

2.2.5 | Rapid and acute Sox2 protein depletion validates the all-or-nothing maintenance of NDRs at CREs

To validate that pluripotency TFs maintain chromatin accessibility cooperatively, as opposed to housekeeping TF such as Ctcf and Yy1, Laura Moniot-Perron induced the rapid and acute depletion of Sox2 at the protein level using a previously reported degradation tag (dTAG) system (Liu et al. 2021a; Maresca et al. 2023). Immunoblotting confirms the degradation of Sox2 upon 2h dTAG treatment (Figure 2.14A). Laura Moniot-Perron performed SMF on non-treated (NT) and 2h-treated (2h) samples and I used FootprintCharter to quantify chromatin accessibility at the single molecule level. I confirm a significant reduction in chromatin accessibility at Sox2 motifs upon dTAG treatment (Figure 2.14B, Wilcoxon test p-value < 2.2e-16), as exemplified in the single locus plot in Figure 2.14C. This is in line with previously reported bulk assays measurements (Maresca et al. 2023).

The perturbation of TF binding by natural genetic variation, as I reported it in this dissertation, is complex and requires careful interpretation due to the interactions that can occur between multiple SNPs within the same CRE. To control for such confounding effect, I sought to validate the contribution of Sox2 towards the maintenance of chromatin accessibility by leveraging the Sox2 dTAG dataset reported here. For the Sox2 sites associated with a reduction in binding affinity, I plotted the change in chromatin accessibility frequency as measured in the F1s against the change in frequency upon Sox2 protein depletion. This comparison validated 19 out of 42 instances of allele-specific frequency (precision=0.46, Figure 2.14D), indicating the genuine dependency of those CREs on Sox2 for the maintenance of the steady state frequency of chromatin accessibility. I want to highlight, that the Sox2 dTAG dataset has been so far only sequenced shallowly in comparison to the F1 or WT datasets. This implies that this dataset suffers from a reduced statistical power to call for significant changes

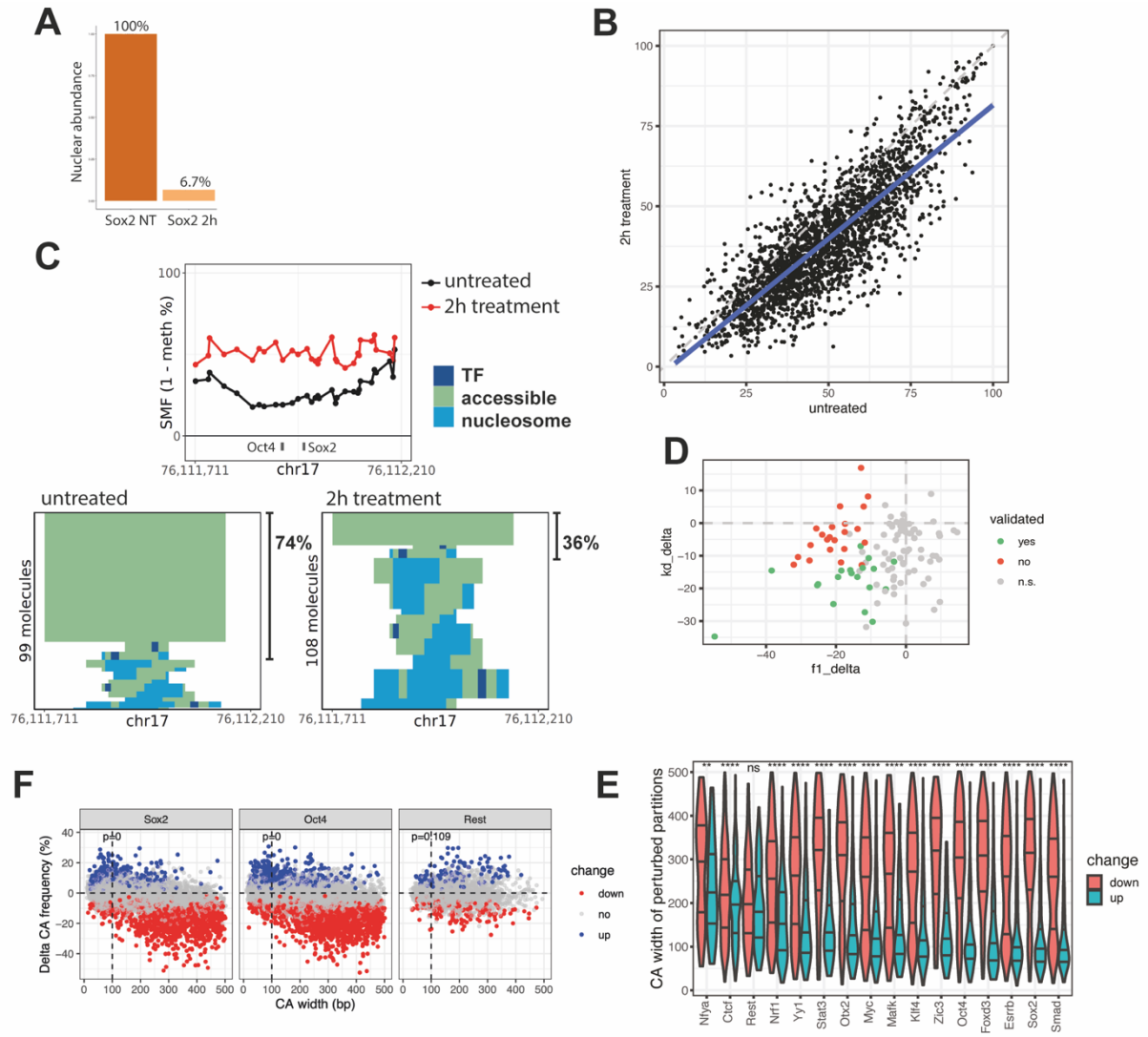

Figure 2.14. The width of NDR at pluripotency CREs is maintained in an all-or-nothing fashion. A. Nuclear abundance of Sox2 protein in the dTAG cell line for the untreated (NT) sample and the sample treated for 2 hours (2h) **B.** Chromatin accessibility frequency at Sox2 motifs, quantified by FootprintCharter in the untreated (x-axis) and treated (y-axis) samples **C.** Single locus example of a CRE containing a Sox2 and an Oct4 motif, displaying a significant (Fisher's test p<0.05) reduction in chromatin accessibility frequency upon Sox2 depletion. **D.** Delta in chromatin accessibility frequency at Sox2 motifs across F1 alleles (x-axis) or upon Sox2 protein depletion (y-axis). Green indicates motifs associated with a significant (Fisher's test p<0.05) reduction in frequency in both assays. Red indicates motifs associated with significant changes across F1 alleles but not upon protein depletion. Grey indicates motifs associated with no significant changes in both experiments. **E.** Distribution of NDR widths for molecular states which are either enriched (blue) or depleted (red) upon loss of TF binding affinity. Number of stars for the Wilcoxon test p-value (ns: p>0.05; *: p<=0.05; **: p<=0.01; ***: p<=0.001; ****: p<=0.0001). **F.** Allele-specific frequencies (y-axis) of molecular states with different NDR widths (x-axis). Annotated, the Wilcox p-value testing the global difference in NDR width distribution between enriched and depleted molecular states. The data for this figure was produced by Laura Moniot-Perron.

across conditions. Currently, additional libraries are currently being prepared by Laura Moniot-Perron to improve this metric.

Similarly, the rapid protein-level depletion of Sox2 confirms that the width of NDRs for CREs bound by pluripotency TFs is maintained in an all-or-nothing fashion (Figure 2.14E-F). More specifically, I observed a substantial depletion of states with large NDRs (>100bp) accompanied by an enrichment for molecular states with linker DNA (<100bp) at Sox2 motifs (Figure 2.14F, left panel). The same is true for the motifs of Oct4, which frequently occupies the same CREs as Sox2 and it reportedly engages in cooperative interactions with it (Figure 2.14F, central panel) (V. Malik et al. 2019; Soufi et al. 2015; Soufi, Donahue, and Zaret 2012). The same is not true for Rest, TF that is not known to engage with Sox2 (Figure 2.14F, right panel).

In addition to Oct4, I observed several other transcription factors, many of which are involved in the maintenance of pluripotency in ESCs (M. Li and Belmonte 2017; M. Li and Izpisua Belmonte 2018), which display statistically significant depletion in active CREs (>100bp) accompanied by the enrichment of linker DNA only (<100bp). These TFs include Smad, Stat3, Foxd3, Esrrb, Zic3, Klf4 and Myc (Figure 2.14E). These results not only highlight the dependency of these TFs on Sox2 for the maintenance of the steady state frequency of CRE usage, but also the fact that CREs they occupy maintain the width of NDRs in an all-or-nothing fashion. This implies that such CREs are activated as units and excludes their modularity, namely the possibility of usage of part of the CRE in a fraction of the cell population. These results do not exclude that such modularity is still applicable to CREs occupied by TFs such as Nrf1 and Nfya, for which I do observe enrichment of NDRs larger than 100bp (Figure 2.14E), or across cell types and differentiation stages where the definition of a CRE might change in response to the change in expression of cell type-specific TFs (Y. Hu et al. 2023).

3 | Discussion

3.1 | *FootprintCharter* conceptually advances how information is extracted from single molecule genomics data

3.1.1 | *FootprintCharter* annotates molecules in their full length

FootprintCharter is among the very first *ad-hoc* computational frameworks for the unsupervised detection and quantification of footprints from single molecule genomics data. Despite there being plenty of room for its further improvement, *FootprintCharter* represents a conceptual leap forward in the way information is extracted from single molecule footprinting data (SMF).

As detailed in the Results section 2.1.2, SMF has been analyzed until now using “single molecule sorting”. This is a supervised method that classifies molecules based on few key cytosines in and around previously annotated TF motifs. This implies that “single molecule sorting” inherits the TF-specific biases of sequence-based motif prediction. Therefore, TFs with more degenerate motifs can be erroneously quantified. Secondly, “single molecule sorting” does not explicitly detect footprints, this implies that it struggles to distinguish nucleosome occupancy from the binding of multiple adjacent TFs. Finally, “single molecule sorting” requires the enumeration of all the possible molecular states that can emerge from the analysis of a particular genomic locus. This severely limits the complexity of the loci that can be feasibly treated. In practice, genome-wide TF occupancy quantifications have so far been limited to two TF motifs at the time.

FootprintCharter explicitly detects and quantifies footprints in an unsupervised way and free of the biases of prior TF motif annotations. More generally, *FootprintCharter* annotates each molecule in its entirety with one of the following states: TF footprint, accessible, or nucleosome footprint. This latter feature implies that it can quantify molecular patterns of any complexity as long as they are represented in the data.

3.1.2 | Considerations on the applicability of *FootprintCharter* to other of single molecule genomics data modalities

FootprintCharter has been developed specifically to analyze single molecule footprinting data produced as described in (Kleinendorst et al. 2021). As exemplified in this study, *FootprintCharter* successfully captures the heterogeneity of SMF signal across genomic loci as well as the dynamic range of changes in frequencies of molecular states across conditions.

As exemplified through the analysis of F1 data in the section 2.2.3 and of Sox2 degradation data in the section 2.2.5, *FootprintCharter* successfully quantifies changes in molecular states both when applied to SMF data produced using single enzyme (GpC) and double enzyme (GpC and CpG) footprinting treatments. If provided with appropriate single molecule methylation matrixes, *FootprintCharter* should be applicable to single molecule genomics data modalities that include footprinting performed with any combination of genomic contexts, including adenosine methylation such as Fiber-seq or SMAC-seq (Shipony et al. 2020; Stergachis et al. 2020).

If provided with appropriate single molecule methylation matrixes, *FootprintCharter* should also be applicable to single molecule data modalities produced by different sequencing platforms including long read sequencing technologies such as Fiber-seq, SMAC-seq or nanoNOMe-seq (Battaglia et al. 2022; I. Lee et al. 2020). The running time of the workflow is largely independent of read length, except for the very first single molecule smoothing step which linearly grows with it. I advise experimenting without this step if working with long reads, also on account of the redundancy of its noise smoothing effect with the partitioning step.

3.1.3 | Consideration on future developments for *FootprintCharter* and computational approaches for single molecule genomics

As highlighted in multiple instances during this dissertation, *FootprintCharter* is meant to be a data-driven and unsupervised footprint detection algorithm. While this is true as far as footprint detection and quantification are concerned, the current workflow is still partially

dependent on previously predicted TF motifs for the aggregation of biologically equivalent footprints (Results section 2.1.3). When considering further developments of the method, besides adapting it to additional single molecule genomics data modalities, one potential area of experimentation could be the unsupervised annotation of the detected footprints. For example, one could attempt to discover *de novo* and then curate motifs from the collection of DNA sequences associated with TF footprints. This could give access to the great wealth of footprints which currently cannot be treated because of the lack of a strong enough but potentially influential TF motif instance (F. Lim et al. 2024), the lack of a sequence motif (Sandelin 2004) or of a good ChIP-seq dataset.

Additionally, it might be possible to computationally learn TF-specific sets of footprint "shapes". Different transcription factors tend to be associated with specific DNA sequences which often go a few base pairs beyond the core recognition motif, these are referred to as "flanking" sequences and have been demonstrated to be important for the affinity of TFs to DNA (De Almeida et al. 2022; Horton et al. 2023). Importantly for this discourse, TF footprints, when detected, very often span a genomic space that exceeds the length of the core motif, having flanking sequences as contributors to the footprint "shape". Additionally, the frequency of TF binding, as well as the positioning and frequencies of occupancy of the surrounding nucleosomes display a TF-specific connotation to some degree (Figure 2.8A, B). I should mention that, to this date, no deep learning architecture, nor in general machine learning method, is capable of dealing with single molecule genomics datasets, implying the need for developing an *ad hoc* deep learning architecture to attempt this task.

3.2 | Rethinking the contribution of transcription factors to chromatin accessibility at CREs

3.2.1 | Quantifying the regime of cis-regulatory elements usage across cell populations

Cis-regulatory elements (CREs) can be accessible, i.e., active, constitutively or in a cell-type specific manner. This reflects in the specificity of the genes they regulate, either housekeeping or cell-type specific (Ramalingam et al. 2023; Thurman et al. 2012). However, this distinction does not speak of the levels of usage of CREs within a homogeneous cell population. By regime of CRE activity here I mean how frequently a CRE is found active in a given cell population, i.e., what is the probability of observing at a given moment the CRE as accessible in an individual cell. Single cell measurements of chromatin accessibility would in principle be able to estimate this quantity. However, the sequence depth required to do so is a bottleneck for the quantitative estimation of this activity regime. Single molecule footprinting (SMF) is instead very suitable for estimating activity regimes. This is because SMF probes the chromatin accessibility at the single molecule level, at high sequencing depth and near-nucleotide resolution.

Through the lens of single molecule footprinting, I observed that within a cell population CREs are very dynamic and function under heterogeneous activity regimes. Promoters are very frequently active (>75%) while enhancers work at lower regimes (35-65%). Surprisingly, inactive elements show non-negligible levels of potential activity (up to 35%) in the form of TF motif exposure at linker DNA. Such CRE-specific regimes are reflected in the regimes of activity of the transcription factors occupying these CREs. Accordingly, TFs that occupy promoters more frequently (e.g., Banp, Nfy and Nrf1) are associated with higher levels of CRE activity than TFs associated primarily with enhancers. Interestingly, the activity regime of enhancers seems to scale with the number of TF motifs. This could represent a regulatory knob for the cell which can tune the nuclear abundance of TFs to modulate the activity regime of enhancers.

3.2.2 | TF pioneering: relaxing the rules and decoupling from accessibility maintenance

As discussed in the Introduction section 1.2.4.1, pioneer transcription factors are able to bind and activate inactive *cis*-regulatory elements *in-vivo* to initiate transcriptional programs that instruct cell differentiation. They are characterized by their ability to displace nucleosomes in an ATP-independent way and to access their DNA recognition motifs *in-vitro*. The pluripotency transcription factors Oct4, Sox2 and Klf4, are prominent examples of pioneers. It is not equally clear how pioneer TFs seed the activation of CREs *in-vivo,* where chromatin is considerably more complex and dynamic than what can be experimentally recapitulated *in-vitro*.

My results show that TF motifs at inactive CREs are accessible in a non-negligible fraction of the cell population (ca. 37%) due to the overlap with linker DNA. Such motif exposure might allow for opportunistic TF binding. Albeit probably infrequent and unstable, these TF binding events might be able to seed the activation of CREs through e.g., the recruitment of chromatin remodelers. If this was true, pioneer TFs would be able to opportunistically bind to such temporarily exposed motifs. The hypothesis is that this short-lived binding is sufficient to recruit chromatin remodelers in sufficient quantities to seed the activation of a CRE. Further research is required to test this specific mechanism; however, the community has been opening to the idea of ATP-dependent pioneering *in-vivo* (Barral and Zaret 2024; Bulyk et al. 2023; King and Klose 2017).

Furthermore, CREs bound by pioneer TFs are active in a fraction of the cell population that is much below 100%. This, and the fact that most pioneer TF instances at CREs are disposable for the maintenance of CRE activity, points towards the decoupling of the contribution that a TF can have towards chromatin accessibility. While pioneer TFs can seed chromatin accessibility at an inactive CRE, they do not need to be the primary responsible for its maintenance at the steady state.

3.2.3 | The pioneer TF Nrf1 is not a strong contributor to chromatin accessibility maintenance

Nrf1 has been included among pioneer TFs owing to its ability to establish chromatin accessibility *de novo* (Domcke et al. 2015; Mayran and Drouin 2018; Sherwood et al. 2014). Accordingly, I observed Nrf1 binding in association with some of the most frequently accessible CREs in the mESC population (Figure 2.8A).

However, when hindered from binding, the large majority of its CREs remain just as accessible (Figure 2.12C, Figure 2.13A, B), indicating Nrf1 as a disposable TF for the maintenance of chromatin accessibility. Even for those rarer CREs for which Nrf1 contributes to some degree to the steady state frequency of chromatin accessibility, its loss of binding is not associated with the loss of function of the CRE at the single molecule level. In fact, such CREs remain accessible albeit with a reduced width of nucleosome-depleted region (Figure 2.13A, B).

This latter observation marks a clear distinction between Nrf1 and the pluripotency TFs Oct4 and Sox2. This difference can be explained by the different genomic locations of these TFs. While Oct4 and Sox2 tend to mostly bind to enhancers, Nrf1 is more widespread at promoters (Figure 2.10D) which in turn tend to be more accessible more frequently (Figure 2.7B, C) and constitutively (Ramalingam et al. 2023; Thurman et al. 2012).

In conclusion, my results point towards a weak contribution of Nrf1 towards the maintenance of chromatin accessibility at its endogenous CREs. Such CREs, mostly promoters, rely on mechanisms other than the binding of Nrf1 to keep depleted of nucleosomes.

3.2.4 | CREs that are sensitive to sequence variation TF motif maintain their nucleosome depleted region width in an “all-or-nothing” fashion

Transcription factors assemble cooperatively at CREs through different mechanisms that reflect in the stringency of the syntax of the CRE itself (Deplancke, Alpern, and Gardeux 2016; Reiter, Wienerroither, and Stark 2017). By syntax, here, I mean the ensemble of sequence features such as motif sequences as well as the relative spacing and orientation of TF motifs

with respect to each other (Weingarten-Gabbay and Segal 2014). On one extreme, TFs can assemble modularly and outcompete nucleosomes by nucleosome-mediated cooperativity (Mirny 2010; Sönmezer et al. 2021). In this case, the syntax of the CRE is fairly loose and robust to sequence changes. On the other end of the spectrum, TFs can assemble through strict protein-protein interactions (Merika and Thanos 2001; Panne 2008). The DNA sequences at such CREs are highly constrained to mediate the clock-like assembly of TFs. Virtually any sequence variation disrupts the assembly at the entire enhancer. A prominent case example is the β-interferon enhanceosome (Panne, Maniatis, and Harrison 2007).

Accordingly, I observe that the CREs that respond to genetic variation at TF motifs lose their entire widths of nucleosome-depleted regions (NDRs) at the level of single molecules. In most cases, I do not observe that such CREs can function partially, i.e., no molecule shows accessibility for part of a CRE. This could be explained in two ways. One explanation is that these CREs are active in an enhanceosome-like fashion and they harbor TFs which mostly assemble by protein-protein interactions. In that case, the annotation of CREs by their sensitivity to sequence variation would return a conservative catalogue of CREs with enhanceosome-like proprieties. Alternatively, the width of the NDRs of these CREs might be governed indirectly through the recruitment of ATP-dependent chromatin remodelers. In that case, the loss of binding of one of the TFs in the CRE might reduce the frequency of recruitment of the remodeler in the cell population, resulting in a lower activity regime. The experimental perturbation of remodeler activity will be required to sort through these two possibilities.

On a final note, the fact that on single molecules such CREs shut down in an “all-or-nothing” fashion does not necessarily reflect in their complete loss of usage in the cell population. On the contrary, the vast majority of the CREs that respond in my system are subject to partial reduction in activity regime. It will be interesting to explore what are the transcriptional consequence of such subtle allele-specific CRE activity regimes.

3.2.5 | Regulating the regime of *cis*-regulatory element activity might influence transcriptional bursting and cellular decision making

Individual cells are confronted with a plethora of cues which guide their adaptive responses to their surrounding biochemical environment as well as the coordinated development of the organism. Such adaptive responses are crafted through the regulation of transcriptional programs by TFs at CREs.

Beyond simply switching on and off genes, cells finely modulate the dosage of gene expression. It has become widely accepted that single cells do not continuously transcribe genes, but alternate from moments in which genes are transcribed to moments with no transcriptional activity. This process is referred to as transcriptional bursting (Chubb et al. 2006; Fusco et al. 2003; Tunnacliffe and Chubb 2020). There, a cell can regulate two kinetic parameters to regulate genes in dosage and time: the duration of its bursts and its frequency. According to recent studies, promoters seem to be involved in the regulation of bursting duration (Pimmett et al. 2021), while distal enhancers seem to rather regulate frequency (Bartman et al. 2016; Fukaya, Lim, and Levine 2016). Accordingly, I observe that the CREs with the more variable, hence tunable, regimes of activity are distal enhancers.

Understanding how single cells modulate the enhancer activity through transcription factor binding can shed light on the regulation of bursting kinetics at genes. This in turn would offer a mechanistic overview on the decision-making processes of single cells.

4 | Materials and methods

4.1 | Experimental methods

4.1.1 | Cell culture

Rozemarijn Kleinendorst cultured F1 mouse embryonic stem cells Bl6xCast (Giorgetti et al. 2016) DNMT TKO and Bl6xSpret (Hochepied et al. 2004) DNMT TKO (Domcke et al. 2015) on 0.2% gelatin-coated plates in ESC medium (Dulbecco's Modified Eagle Medium (DMEM), supplemented with 15% Fetal Bovine Serum (FBS), Leukemia Inhibitory Factor (LIF), 2-Mercaptoethanol, 2 mM L-Glutamine and 1x non-essential amino acids) at 37°C and 5% CO2. She changed the medium daily and cells split cells every second day. Laura Moniot-Perron did the same with the Sox2-FKBP-2xHA cell line E14Tg2A (129/Ola) from the de Wit laboratory (Liu et al. 2021b).

Rozemarijn Kleinendorst worked with F1 hybrid cells, of which the Bl6XCast was originally generated in the Odom lab at the DKFZ in Heidelberg, Germany.

4.1.2 | Single molecule footprinting

Single Molecule Footprinting with targeted enrichment (SMF) was performed by Rozemarijn Kleinendorst as previously described (Kleinendorst et al. 2021; Sönmezer et al. 2021). In short, she harvested cultured cells using trypsin and washed twice with 1x PBS. She counted cells and used 250,000 cells per reaction. She resuspended cell pellets in ice-cold lysis buffer, incubated on ice for 10 min and spun at 1,000x g at 4°C for 5 min. She resuspended nuclei in 1x M.GpC buffer (NEB, #M0227L). For the GpC methyltransferase treatment, she added freshly made GpC methyltransferase mix (1x M.GpC buffer, 300 mM sucrose, 64 µM SAM (NEB, #B9003S)) and M.CviPI (NEB, #M0227L) and incubated at 37°C for 7.5 min. She added prewarmed stop solution and proteinase K and incubated overnight at 55°C. The next day, she extracted DNA using phenol-chloroform and treated with RNase A at 37°C for 30 min. For

making the Whole Genome Bisulfite Sequencing (WGBS) library with targeted enrichment, she fragmented DNA into 300 bp fragments via sonication using Covaris model S2. For library preparation and targeted enrichment of *cis*-regulatory elements, she used the SureSelect XT Mouse Methyl-Seq Kit Enrichment System for Illumina Multiplexed Sequencing Library protocol (Agilent Technologies, version E0, April 2018) as described in (Kleinendorst et al. 2021). She performed bisulfite conversion using the ZYMO EZ DNA Methylation-Gold Kit (Zymo, #D5005) according to the manufacturer's protocol. She PCR amplified the bisulfite-converted library and indexed using the SureSelect XT Mouse Methyl-Seq Kit (Agilent, # G9651A). She ran prepared libraries on an Illumina sequencing platform using a NextSeq High 150 bp paired-end mode.

Rozemarijn Kleinendorst produced respectively four and six biological replicates for the Bl6xCast and Bl6xSpret F1 lines.

4.1.3 | Sox2 degradation TAG

For conditional degradation, Laura Moniot-Perron treated cells during 2 hours with a final concentration of 500 nM of dTAG (Sigma, SML2601-5MG) or with DMSO for non-treated control.

4.1.4 | Western blotting

Laura Moniot-Perron harvested cells, washed twice in ice cold 1X PBS and resuspended in cold Lysis Buffer A (10 mM HEPES, 5mM MgCl2, 0.25M Sucrose, 0.1 % NP400). Laura Moniot-Perron lysed cytoplasmic membranes by incubating 20 min on ice and pipetting up and down several times. Laura Moniot-Perron spined down nuclei at 9800g during 10 min at 4ºC and then lysed during 10 min on ice in lysis Buffer B (2.5 mM HEPES, 1.5 mM MgCl2, 0.1 mM EDTA, 20 % Glycerol). Laura Moniot-Perron recovered nuclear extracts by vortexing samples at least 20 sec at max power and by spinning at 15000g 15 min at 4ºC. Laura Moniot-Perron determined protein concentration using Qubit protein quantification kit. Laura Moniot-Perron loaded 20µg of protein in a 10% SDS PAGE gel. Laura Moniot-Perron transferred proteins to a nitrocellulose membrane. Laura Moniot-Perron incubated membranes overnight at 4ºC with HA (1:1000, Biolegend, #901501) for Sox2 and H3 as a loading control (1:2000, Cell signaling, #4499T).

4.2 | Computational methods

4.2.1 | Single molecule footprinting data pre-processing

For all single molecule footprinting (SMF) datasets (WT, F1s, dTAG), I trimmed sequencing reads for Illumina adapters and low-quality bases using TrimGalore v0.6.7 (Krueger et al. 2023). I discarded reads shorter than 20bp after trimming. In the case of F1 data, to assign pre-processed reads to the correct allele of origin, I used SNPs annotations between mouse species I fetched from the Mouse Genome Project portal (Keane et al. 2011). I injected SNPs into the UCSC Mus musculus reference genome distributed through the Bioconductor (Gentleman et al. 2004; Huber et al. 2015) package BSgenome.Mmusculus.UCSC.mm10 (Team 2017) using the function qAlign provided with the Bioconductor package QuasR (Gaidatzis et al. 2015). This resulted in two separate genomes which I bisulfite converted using the QuasR function qAlign. I competitively aligned pre-processed reads to both genomes using the QuasR function qAlign with alignment parameters "-e 70-X 1000-k 2-best-strata". The best alignment determined the allelic assignment for each read. In the case of WT and dTAG data, I simply aligned pre-processed reads to the BSgenome.Mmusculus.UCSC.mm10 genome using the QuasR function qAlign with alignment parameters "-e 70-X 1000-k 2-best-strata". Finally, in all cases, I identified and removed duplicated using the Picard tool MarkDuplicates v2.15.0 (Picard Tools n.d.) as previously described (Kleinendorst et al. 2021).

4.2.2 | Definition of single molecule footprinting quantification windows

For the analysis of all SMF datasets, I designed ~3.3 million consecutive 80bp windows, overlapping each other by 40bp. To focus computational efforts, these windows cover exclusively DNase hypersensitivity regions in mESCs.

4.2.3 | SMF – single molecule methylation call

I performed single molecule methylation calls for cytosines covered by at least 20 reads separately for each sample. I used the function CallContextMethylation from the dev version of my previously reported *SingleMoleculeFootprinting* Bioconductor package and that is currently available at https://github.com/Krebslabrep/SingleMoleculeFootprinting (Kleinendorst et al. 2021). For F1 data, I further discarded from both alleles cytosines for which SNPs disrupted the sequence context recognized by the methyltransferase enzyme using custom R functions.

4.2.4 | SMF – quantification of footprints with *FootprintCharter*

I applied *FootprintCharter* to all SMF datasets with the maximum number of partitions k=12 and minimum number of reads per partition n=5. I ran WT SMF data individually, while I pooled reads from paired samples (i.e., Bl6 and Cast alleles, Bl6 and Spret alleles, untreated and 2h Sox2 dTAG treated samples). This allowed me to directly quantify the frequency changes of molecular states.

4.2.5 | SMF – calculation of chromatin accessibility frequency with *FootprintCharter*

For each TF motif, I calculated the frequency of chromatin accessibility across the cell population as the fraction of molecules for which a nucleosome depleted region longer than 100bp overlapped the TF motif.

4.2.6 | SMF – single locus plot, bulk plots

I produced the single molecule footprinting data bulk plots using a customization of the function PlotAvgSMF from the *SingleMoleculeFootprinting* package, dev version. I modified the function to include the annotation of SNPs.

4.2.7 | SMF – single locus plot, single molecule stacks

I produced the single molecule footprinting data single molecule stacks using the function PlotSM from the *SingleMoleculeFootprinting* package, dev version. I sorted molecules for plotting using the partitioning results from *FootprintCharter*.

4.2.8 | SMF – single locus plot, footprint detection heatmaps

I produced the single molecule footprinting data footprint detection heatmaps using custom functions. I sorted molecules for plotting using the partitioning results from *FootprintCharter*.

4.2.9 | TFBS annotation

I annotated transcription factor binding sites (TFBSs) by scanning the reference mouse genome with PWMs for vertebrate TFs I obtained from the JASPAR database (Sandelin 2004), through the JASPAR2016 Bioconductor package (Mathelier et al. 2016). I performed PWM scanning using the TFBStools Bioconductor package (Tan and Lenhard 2016). I retained for validation sequence matches with a score equal or greater than 10. I used the same publicly available ChIP-seq datasets as in (Sönmezer et al. 2021) to filter TFBSs predictions for the motif proven to be bound by TFs in vivo as measured by ChIP-seq. For the TF Banp I used ChIP-seq reported in (Grand et al. 2021). I treated the transcription factors Oct4, Sox2 and Klf4 differently. In short, I obtained TFBS predictions for Oct4, Sox2 and Klf4 from (Avsec, Weilert, et al. 2021). For comparability with the above-mentioned list, I scored these with the relevant PWMs from the JASPAR2016 collection as above and validated them using the relevant ChIP-nexus datasets reported in (Avsec, Weilert, et al. 2021).

4.2.10 | *Cis*-regulatory elements annotation with ChromHMM

I obtained ChromHMM *cis*-regulatory elements annotation obtained as described in (Kreibich et al. 2023).

4.2.11 | Genomic tracks plot

I plotted genomic tracks using custom R functions using publicly available datasets.

4.2.12 | Definition of the number of TF motifs at CREs

I defined the number of TF motifs at CREs as the number of motifs annotated (as described in the section 4.2.9) within the surrounding 300 base pairs centered around the motif for which the chromatin accessibility frequency or nucleosome-depleted region width is calculated.

4.2.13 | Statistical testing for increased chromatin accessibility

I tested for increased frequency of chromatin accessibility across categories (bound motifs against unbound motifs, loss of function motifs against, increasing number of TF motifs) using either a t-test or a Wilcoxon test. In both cases I did so using the R implementation of these tests and set the argument “alternative” to “greater”.

4.2.14 | Processing of publicly available ChIP-seq and ChIP-nexus datasets

I processed the publicly available ChIP-seq and ChIP-nexus datasets using the nf-core pipeline nf-core/chipseq v.2.0.0 with arguments “--narrow_peak--read_length 50” (P. Ewels et al. 2022).

4.2.15 | F1 Bl6xCast ATAC-seq data pre-processing

I fetched previously reported ATAC-seq data produced using an analogous Bl6xCast F1 ESC line as for this project (Xu et al. 2017) using the nf-core pipeline fetchngs (P. A. Ewels et al. 2020; Harshil Patel et al. 2024). I performed pre-processing similarly to what described for single molecule footprinting in the section 4.2.1. Briefly, I trimmed sequencing reads for Illumina adapters and low-quality bases using TrimGalore! v0.6.7 (Krueger et al. 2023). I discarded reads shorter than 20bp after trimming. To assign pre-processed reads to the correct allele of origin, I used SNPs annotations between mouse species I fetched from the Mouse Genome Project portal (Keane et al. 2011). I injected SNPs into the UCSC Mus musculus reference genome distributed through the Bioconductor (Gentleman et al. 2004; Huber et al. 2015) package BSgenome.Mmusculus.UCSC.mm10 (Team 2017) using the function qAlign provided with the Bioconductor package QuasR (Gaidatzis et al. 2015). This resulted in two separate genomes to which I competitively aligned pre-processed reads using the QuasR function qAlign with alignment parameters "-e 70-X 1000-k 2-best-strata". The best alignment determined the allelic assignment for each read. Finally, I identified and removed duplicated using the Picard tool MarkDuplicates v2.15.0 (Picard Tools n.d.).

4.2.16 | TFBS loss of function annotation and definition of the background

For each TFBS, I computed the relative change in PWM match score across alleles [(alt.-ref.) / abs(ref.)]. I considered a TFBS to be a loss of function on the alternative allele if its relative change was lower than-0.5. I discarded gain of function motifs (relative change greater than 1.5) due to the lack of ChIP-seq/-nexus data for validation. I considered all the remaining TF motifs unperturbed across alleles and included among the "background".

When I used multiple PWMs for the same TF and resulted in overlapping motif annotations, I retained only the motif with the most extreme relative change in PWM match score for further analyses. In some instances, the Cast and Spret genomes present the same genetic variation when compared to the reference Bl6. To avoid such redundancies, in these instances, I discarded from analyses the TFBS from the F1 cross with the lowest SMF coverage.

4.2.17 | Testing for statistically significant changes in frequency of molecular states

For each locus, I determined whether each molecular state (i.e., clustering partition) is associated with a statistically significant difference in frequency using a fisher's exact test. I used the R implementation of this test.

4.2.18 | Calculation of precision

I calculated the value for precision as validated / (validated + not validated). *Validated* is the number of TF motifs associated with a statistically significant difference in chromatin accessibility frequency both among F1 alleles and experimental conditions in the Sox2 dTAG experiment. *Not validated* is the number of TF motifs associated with a statistically significant difference in chromatin accessibility frequency among F1 alleles but not among experimental conditions in the Sox2 dTAG experiment.

4.2.19 | Data visualization and illustrations

I produced all the density plots, violin plots, scatter plots and barplots reported in this dissertation using the R package ggplot2, part of the tidyverse collection (Wickham et al. 2019). The only exceptions are the scatterplots in Figure 2.11, which I created using custom R functions.

I produced all the heatmaps reported in this dissertation using the Bioconductor package ComplexHeatmap (Gu, Eils, and Schlesner 2016).

I created with BioRender.com all the conceptual illustrations reported in this dissertation.

4.2.20 | Scripting, data analysis and high-performance computing

I performed all the and data analysis for this project using R-4.2.2 (R Core Team 2021) and RStudio Pro 2023.12.0+369.pro3 (RStudio Team 2020). I scripted the single molecule footprinting pre-processing pipeline using Nextflow-22.10.6 (Di Tommaso et al. 2017). I performed all high-performance computing using the SLURM workload manager (Yoo, Jette, and Grondona 2003) either through direct bash scripting or through the R package rslurm (Read et al. 2021).

4.3 | Materials

4.3.1 | Datasets

Dataset	Accession number	Reference
ChIP-seq	-	datasets reported in (Sönmezer et al. 2021)
DNase-seq	-	datasets reported in (Sönmezer et al. 2021)
MNase-seq	-	datasets reported in (Sönmezer et al. 2021)
Banp ChIP-seq	GSM4708468, GSM4708469, GSM4708470	(Grand et al. 2021)
Oct4 ChIP-nexus	GSM4072776	(Avsec, Weilert, et al. 2021)
Sox2 ChIP-nexus	GSM4072777	(Avsec, Weilert, et al. 2021)
Klf4 ChIP-nexus	GSM4072779	(Avsec, Weilert, et al. 2021)
Bl6xCast F1 ATAC-seq	GSM2247118, GSM2247119	(Xu et al. 2017)

Table 2. Publicly available datasets used in this study

4.3.2 | Software and databases

Name	Version	Reference
TrimGalore!	0.6.7	(Krueger et al. 2023)
BSgenome.Mmusculus.UCSC.mm10	1.4.3	(Team 2017)
QuasR	1.36.0	(Gaidatzis et al. 2015)
Picard	2.15.0	(Picard Tools n.d.)
SingleMoleculeFootprinting	dev	(Kleinendorst et al. 2021)

JASPAR2016	1.24.0	(Mathelier et al. 2016)
TFBSTools	1.36.0	(Tan and Lenhard 2016)
nf-core/fetchngs		(Harshil Patel et al. 2024)
tidyverse	2.0.0	(Wickham et al. 2019)
ComplexHeatmap	2.14.0	(Gu, Eils, and Schlesner 2016)
R	4.2.2	(R Core Team 2021)
RStudio Pro	2023.12.0+369.pro3	(RStudio Team 2020)
Nextflow	22.10.6	(Di Tommaso et al. 2017)
plyranges	1.16.0	(S. Lee, Cook, and Lawrence 2019)
rslurm	0.6.1	(Read et al. 2021)
parallelDist	0.2.6	-
cluster	2.1.4	-
qs	0.25.5	-
patchwork	1.1.2	(Thomas Lin Pedersen n.d.)
annotatr	1.22.0	(Cavalcante and Sartor 2017)

Table 3. External software used in this study

Name	Version	Reference
Mouse Genome Project	-	(Keane et al. 2011)
JASPAR	2016	(Mathelier et al. 2016)
chromHMM	1.25	(Ernst and Kellis 2012)

Table 4. External databases used in this study

Bibliography

Abdulhay, Nour J, Colin P McNally, Laura J Hsieh, Sivakanthan Kasinathan, Aidan Keith, Laurel S Estes, Mehran Karimzadeh, et al. 2020. 'Massively Multiplex Single-Molecule Oligonucleosome Footprinting'. *eLife* 9: e59404. doi:10.7554/eLife.59404.

Adams, C C, and J L Workman. 1995. 'Binding of Disparate Transcriptional Activators to Nucleosomal DNA Is Inherently Cooperative'. *Molecular and Cellular Biology* 15(3): 1405–21. doi:10.1128/MCB.15.3.1405.

Allshire, Robin C., and Hiten D. Madhani. 2018. 'Ten Principles of Heterochromatin Formation and Function'. *Nature Reviews Molecular Cell Biology* 19(4): 229–44. doi:10.1038/nrm.2017.119.

Amano, Takanori, Tomoko Sagai, Hideyuki Tanabe, Yoichi Mizushina, Hiromi Nakazawa, and Toshihiko Shiroishi. 2009. 'Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription'. *Developmental Cell* 16(1): 47–57. doi:10.1016/j.devcel.2008.11.011.

Andrews, Forest H, Brian D Strahl, and Tatiana G Kutateladze. 2016. 'Insights into Newly Discovered Marks and Readers of Epigenetic Information'. *Nature Chemical Biology* 12(9): 662–68. doi:10.1038/nchembio.2149.

Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. 2013. 'Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq'. *Science* 339(6123): 1074–77. doi:10.1126/science.1232542.

Avsec, Žiga, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, et al. 2021. 'Effective Gene Expression Prediction from Sequence by Integrating Long-Range Interactions'. *Nature Methods* 18(10): 1196–1203. doi:10.1038/s41592-021-01252-x.

Avsec, Žiga, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, et al. 2021. 'Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax'. *Nature Genetics* 53(3): 354–66. doi:10.1038/s41588-021-00782-6.

Balsalobre, Aurelio, and Jacques Drouin. 2022. 'Pioneer Factors as Master Regulators of the Epigenome and Cell Fate'. *Nature Reviews Molecular Cell Biology* 23(7): 449–64. doi:10.1038/s41580-022-00464-z.

Banerji, Julian, Sandro Rusconi, and Walter Schaffner. 1981. 'Expression of a β-Globin Gene Is Enhanced by Remote SV40 DNA Sequences'. *Cell* 27(2): 299–308. doi:10.1016/0092-8674(81)90413-X.

Bannister, Andrew J., Robert Schneider, Fiona A. Myers, Alan W. Thorne, Colyn Crane-Robinson, and Tony Kouzarides. 2005. 'Spatial Distribution of Di- and Tri-Methyl Lysine 36 of Histone H3 at Active Genes'. *Journal of Biological Chemistry* 280(18): 17732–36. doi:10.1074/jbc.M500796200.

Bannister, Andrew J., Philip Zegerman, Janet F. Partridge, Eric A. Miska, Jean O. Thomas, Robin C. Allshire, and Tony Kouzarides. 2001. 'Selective Recognition of Methylated Lysine 9 on Histone H3 by the HP1 Chromo Domain'. *Nature* 410(6824): 120–24. doi:10.1038/35065138.

Barisic, Darko, Michael B. Stadler, Mario Iurlaro, and Dirk Schübeler. 2019. 'Mammalian ISWI and SWI/SNF Selectively Mediate Binding of Distinct Transcription Factors'. *Nature* 569(7754): 136–40. doi:10.1038/s41586-019-1115-5.

Barral, Amandine, and Kenneth S. Zaret. 2024. 'Pioneer Factors: Roles and Their Regulation in Development'. *Trends in Genetics* 40(2): 134–48. doi:10.1016/j.tig.2023.10.007.

Bartman, Caroline R., Sarah C. Hsu, Chris C.-S. Hsiung, Arjun Raj, and Gerd A. Blobel. 2016. 'Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping'. *Molecular Cell* 62(2): 237–47. doi:10.1016/j.molcel.2016.03.007.

Battaglia, Sofia, Kevin Dong, Jingyi Wu, Zeyu Chen, Fadi J. Najm, Yuanyuan Zhang, Molly M. Moore, et al. 2022. 'Long-Range Phasing of Dynamic, Tissue-Specific and Allele-Specific Regulatory Elements'. *Nature Genetics* 54(10): 1504–13. doi:10.1038/s41588-022-01188-8.

Bell, Oliver, Vijay K. Tiwari, Nicolas H. Thomä, and Dirk Schübeler. 2011. 'Determinants and Dynamics of Genome Accessibility'. *Nature Reviews Genetics* 12(8): 554–64. doi:10.1038/nrg3017.

Benayoun, Bérénice A., Elizabeth A. Pollina, Duygu Ucar, Salah Mahmoudi, Kalpana Karra, Edith D. Wong, Keerthana Devarajan, et al. 2014. 'H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency'. *Cell* 158(3): 673–88. doi:10.1016/j.cell.2014.06.027.

Bleckwehl, Tore, Giuliano Crispatzu, Kaitlin Schaaf, Patricia Respuela, Michaela Bartusel, Laura Benson, Stephen J. Clark, et al. 2021. 'Enhancer-Associated H3K4 Methylation Safeguards in Vitro Germline Competence'. *Nature Communications* 12(1): 5771. doi:10.1038/s41467-021-26065-6.

Boltengagen, Mark, Daan Verhagen, Michael Roland Wolff, Elisa Oberbeckmann, Matthias Hanke, Ulrich Gerland, Philipp Korber, and Felix Mueller-Planitz. 2023. 'A Single Fiber View of the Nucleosome Organization in Eukaryotic Chromatin'. *Nucleic Acids Research*: gkad1098. doi:10.1093/nar/gkad1098.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. 'High-Resolution Mapping and Characterization of Open Chromatin across the Genome'. *Cell* 132(2): 311–22. doi:10.1016/j.cell.2007.12.014.

Brahma, Sandipan, and Steven Henikoff. 2024. 'The BAF Chromatin Remodeler Synergizes with RNA Polymerase II and Transcription Factors to Evict Nucleosomes'. *Nature Genetics* 56(1): 100–111. doi:10.1038/s41588-023-01603-8.

Buenrostro, Jason D, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. 2013. 'Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position'. *Nature Methods* 10(12): 1213–18. doi:10.1038/nmeth.2688.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. 'ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide'. *Current Protocols in Molecular Biology* 109(1). doi:10.1002/0471142727.mb2129s109.

Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. 'Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation'. *Nature* 523(7561): 486–90. doi:10.1038/nature14590.

Bulyk, Martha L., Jacques Drouin, Melissa M. Harrison, Jussi Taipale, and Kenneth S. Zaret. 2023. 'Pioneer Factors — Key Regulators of Chromatin and Gene Expression'. *Nature Reviews Genetics* 24(12): 809–15. doi:10.1038/s41576-023-00648-z.

Canver, Matthew C., Elenoe C. Smith, Falak Sher, Luca Pinello, Neville E. Sanjana, Ophir Shalem, Diane D. Chen, et al. 2015. 'BCL11A Enhancer Dissection by Cas9-Mediated in Situ Saturating Mutagenesis'. *Nature* 527(7577): 192–97. doi:10.1038/nature15521.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. 'Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells'. *Science* 361(6409): 1380–85. doi:10.1126/science.aau0730.

Cavalcante, Raymond G, and Maureen A Sartor. 2017. 'Annotatr: Genomic Regions in Context' ed. Alfonso Valencia. *Bioinformatics* 33(15): 2381–83. doi:10.1093/bioinformatics/btx183.

Chen, Kaifu, Zhong Chen, Dayong Wu, Lili Zhang, Xueqiu Lin, Jianzhong Su, Benjamin Rodriguez, et al. 2015. 'Broad H3K4me3 Is Associated with Increased Transcription Elongation and Enhancer Activity at Tumor-Suppressor Genes'. *Nature Genetics* 47(10): 1149–57. doi:10.1038/ng.3385.

Chen, Keshi, Qi Long, Guangsuo Xing, Tianyu Wang, Yi Wu, Linpeng Li, Juntao Qi, et al. 2020. 'Heterochromatin Loosening by the Oct4 Linker Region Facilitates Klf4 Binding and iPSC Reprogramming'. *The EMBO Journal* 39(1): e99165. doi:10.15252/embj.201899165.

Chen, Wei, Aaron McKenna, Jacob Schreiber, Maximilian Haeussler, Yi Yin, Vikram Agarwal, William Stafford Noble, and Jay Shendure. 2019. 'Massively Parallel Profiling and

Predictive Modeling of the Outcomes of CRISPR/Cas9-Mediated Double-Strand Break Repair'. *Nucleic Acids Research* 47(15): 7989–8003. doi:10.1093/nar/gkz487.

Chen, Zhou-Feng, Alice J. Paquette, and David J. Anderson. 1998. 'NRSF/REST Is Required in Vivo for Repression of Multiple Neuronal Target Genes during Embryogenesis'. *Nature Genetics* 20(2): 136–42. doi:10.1038/2431.

Cheung, Alan C.M., and Patrick Cramer. 2012. 'A Movie of RNA Polymerase II Transcription'. *Cell* 149(7): 1431–37. doi:10.1016/j.cell.2012.06.006.

Chew, Joon-Lin, Yuin-Han Loh, Wensheng Zhang, Xi Chen, Wai-Leong Tam, Leng-Siew Yeap, Pin Li, et al. 2005. 'Reciprocal Transcriptional Regulation of *Pou5f1* and *Sox2* via the Oct4/Sox2 Complex in Embryonic Stem Cells'. *Molecular and Cellular Biology* 25(14): 6031–46. doi:10.1128/MCB.25.14.6031-6046.2005.

Chi, Kelly Rae. 2016. 'The Dark Side of the Human Genome'. *Nature* 538(7624): 275–77. doi:10.1038/538275a.

Chronis, Constantinos, Petko Fiziev, Bernadett Papp, Stefan Butz, Giancarlo Bonora, Shan Sabri, Jason Ernst, and Kathrin Plath. 2017. 'Cooperative Binding of Transcription Factors Orchestrates Reprogramming'. *Cell* 168(3): 442-459.e20. doi:10.1016/j.cell.2016.12.016.

Chubb, Jonathan R., Tatjana Trcek, Shailesh M. Shenoy, and Robert H. Singer. 2006. 'Transcriptional Pulsing of a Developmental Gene'. *Current Biology* 16(10): 1018–25. doi:10.1016/j.cub.2006.03.092.

Cianfrocco, Michael A., George A. Kassavetis, Patricia Grob, Jie Fang, Tamar Juven-Gershon, James T. Kadonaga, and Eva Nogales. 2013. 'Human TFIID Binds to Core Promoter DNA in a Reorganized Structural State'. *Cell* 152(1–2): 120–31. doi:10.1016/j.cell.2012.12.005.

Cirillo, Lisa Ann, Frank Robert Lin, Isabel Cuesta, Dara Friedman, Michal Jarnik, and Kenneth S Zaret. 2002. 'Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4'. *Molecular Cell* 9(2): 279–89. doi:10.1016/S1097-2765(02)00459-8.

Clapier, Cedric R., Janet Iwasa, Bradley R. Cairns, and Craig L. Peterson. 2017. 'Mechanisms of Action and Regulation of ATP-Dependent Chromatin-Remodelling Complexes'. *Nature Reviews Molecular Cell Biology* 18(7): 407–22. doi:10.1038/nrm.2017.26.

Clark, Stephen J., Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, et al. 2018. 'scNMT-Seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and Transcription in Single Cells'. *Nature Communications* 9(1): 781. doi:10.1038/s41467-018-03149-4.

Cramer, P., K.-J. Armache, S. Baumli, S. Benkert, F. Brueckner, C. Buchen, G.E. Damsma, et al. 2008. 'Structure of Eukaryotic RNA Polymerases'. *Annual Review of Biophysics* 37(1): 337–52. doi:10.1146/annurev.biophys.37.032807.130008.

Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. 'Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State'. *Proceedings of the National Academy of Sciences* 107(50): 21931–36. doi:10.1073/pnas.1016071107.

Cusanovich, D. A., R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure. 2015. 'Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing'. *Science* 348(6237): 910–14. doi:10.1126/science.aab1601.

Dahl, John Arne, Inkyung Jung, Håvard Aanes, Gareth D. Greggains, Adeel Manaf, Mads Lerdrup, Guoqiang Li, et al. 2016. 'Broad Histone H3K4me3 Domains in Mouse Oocytes Modulate Maternal-to-Zygotic Transition'. *Nature* 537(7621): 548–52. doi:10.1038/nature19360.

Danino, Yehuda M., Dan Even, Diana Ideses, and Tamar Juven-Gershon. 2015. 'The Core Promoter: At the Heart of Gene Expression'. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1849(8): 1116–31. doi:10.1016/j.bbagrm.2015.04.003.

De Almeida, Bernardo P., Franziska Reiter, Michaela Pagani, and Alexander Stark. 2022. 'DeepSTARR Predicts Enhancer Activity from DNA Sequence and Enables the de Novo Design of Synthetic Enhancers'. *Nature Genetics* 54(5): 613–24. doi:10.1038/s41588-022-01048-5.

Deaton, Aimée M., and Adrian Bird. 2011. 'CpG Islands and the Regulation of Transcription'. *Genes & Development* 25(10): 1010–22. doi:10.1101/gad.2037511.

Deindl, Sebastian, William L. Hwang, Swetansu K. Hota, Timothy R. Blosser, Punit Prasad, Blaine Bartholomew, and Xiaowei Zhuang. 2013. 'ISWI Remodelers Slide Nucleosomes with Coordinated Multi-Base-Pair Entry Steps and Single-Base-Pair Exit Steps'. *Cell* 152(3): 442–52. doi:10.1016/j.cell.2012.12.040.

Deplancke, Bart, Daniel Alpern, and Vincent Gardeux. 2016. 'The Genetics of Transcription Factor DNA Binding Variation'. *Cell* 166(3): 538–54. doi:10.1016/j.cell.2016.07.012.

Di Giammartino, Dafne Campigli, Andreas Kloetgen, Alexander Polyzos, Yiyuan Liu, Daleum Kim, Dylan Murphy, Abderhman Abuhashem, et al. 2019. 'KLF4 Is Involved in the Organization and Regulation of Pluripotency-Associated Three-Dimensional Enhancer Networks'. *Nature Cell Biology* 21(10): 1179–90. doi:10.1038/s41556-019-0390-6.

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. 'Nextflow Enables Reproducible Computational Workflows'. *Nature Biotechnology* 35(4): 316–19. doi:10.1038/nbt.3820.

Dodonova, Svetlana O., Fangjie Zhu, Christian Dienemann, Jussi Taipale, and Patrick Cramer. 2020a. 'Nucleosome-Bound SOX2 and SOX11 Structures Elucidate Pioneer Factor Function'. *Nature* 580(7805): 669–72. doi:10.1038/s41586-020-2195-y.

Dodonova, Svetlana O., Fangjie Zhu, Christian Dienemann, Jussi Taipale, and Patrick Cramer. 2020b. 'Nucleosome-Bound SOX2 and SOX11 Structures Elucidate Pioneer Factor Function'. *Nature* 580(7805): 669–72. doi:10.1038/s41586-020-2195-y.

Dombrowski, Marco, Maik Engeholm, Christian Dienemann, Svetlana Dodonova, and Patrick Cramer. 2022. 'Histone H1 Binding to Nucleosome Arrays Depends on Linker DNA Length and Trajectory'. *Nature Structural & Molecular Biology* 29(5): 493–501. doi:10.1038/s41594-022-00768-w.

Domcke, Silvia, Anaïs Flore Bardet, Paul Adrian Ginno, Dominik Hartl, Lukas Burger, and Dirk Schübeler. 2015. 'Competition between DNA Methylation and Transcription Factors Determines Binding of NRF1'. *Nature* 528(7583): 575–79. doi:10.1038/nature16462.

Domingo, Júlia, Mariia Minaeva, John A Morris, Marcello Ziosi, Neville E Sanjana, and Tuuli Lappalainen. 2024. *Non-Linear Transcriptional Responses to Gradual Modulation of Transcription Factor Dosage*. Genomics. preprint. doi:10.1101/2024.03.01.582837.

Dorighi, Kristel M., Tomek Swigut, Telmo Henriques, Natarajan V. Bhanu, Benjamin S. Scruggs, Nataliya Nady, Christopher D. Still, et al. 2017. 'Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation'. *Molecular Cell* 66(4): 568-576.e4. doi:10.1016/j.molcel.2017.04.018.

Doughty, Benjamin R., Michaela M. Hinks, Julia M. Schaepe, Georgi K. Marinov, Abby R. Thurm, Carolina Rios-Martinez, Benjamin E. Parks, et al. 2024. *Single-Molecule Chromatin Configurations Link Transcription Factor Binding to Expression in Human Cells*. Molecular Biology. preprint. doi:10.1101/2024.02.02.578660.

Drozdz, Marek, Andrzej Piekarowicz, Janusz M. Bujnicki, and Monika Radlinska. 2012. 'Novel Non-Specific DNA Adenine Methyltransferases'. *Nucleic Acids Research* 40(5): 2119–30. doi:10.1093/nar/gkr1039.

Echigoya, Kenta, Masako Koyama, Lumi Negishi, Yoshimasa Takizawa, Yuka Mizukami, Hideki Shimabayashi, Akari Kuroda, and Hitoshi Kurumizaka. 2020. 'Nucleosome Binding by the Pioneer Transcription Factor OCT4'. *Scientific Reports* 10(1): 11832. doi:10.1038/s41598-020-68850-1.

El Khattabi, Laila, Haiyan Zhao, Jens Kalchschmidt, Natalie Young, Seolkyoung Jung, Peter Van Blerkom, Philippe Kieffer-Kwon, et al. 2019. 'A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers'. *Cell* 178(5): 1145-1158.e20. doi:10.1016/j.cell.2019.07.011.

Ernst, Jason, and Manolis Kellis. 2012. 'ChromHMM: Automating Chromatin-State Discovery and Characterization'. *Nature Methods* 9(3): 215–16. doi:10.1038/nmeth.1906.

Escobar, Thelma M., Ozgur Oksuz, Ricardo Saldaña-Meyer, Nicolas Descostes, Roberto Bonasio, and Danny Reinberg. 2019. 'Active and Repressed Chromatin Domains Exhibit Distinct Nucleosome Segregation during DNA Replication'. *Cell* 179(4): 953-963.e11. doi:10.1016/j.cell.2019.10.009.

Ewels, Philip A., Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2020. 'The Nf-Core Framework for Community-Curated Bioinformatics Pipelines'. *Nature Biotechnology* 38(3): 276–78. doi:10.1038/s41587-020-0439-x.

Ewels, Philip, Alexander Peltzer, Sven Fillinger, Harshil Patel, Johannes Alneberg, Andreas Wilm, Maxime Ulysse Garcia, Paolo Di Tommaso, and Sven Nahnsen. 2022. 'The Nf-Core Framework for Community-Curated Bioinformatics Pipelines.' doi:10.5281/ZENODO.3240506.

Fernandez Garcia, Meilin, Cedric D. Moore, Katharine N. Schulz, Oscar Alberto, Greg Donague, Melissa M. Harrison, Heng Zhu, and Kenneth S. Zaret. 2019. 'Structural Features of Transcription Factors Associating with Nucleosome Binding'. *Molecular Cell* 75(5): 921-932.e6. doi:10.1016/j.molcel.2019.06.009.

Floc'hlay, Swann, Emily S. Wong, Bingqing Zhao, Rebecca R. Viales, Morgane Thomas-Chollier, Denis Thieffry, David A. Garfield, and Eileen E.M. Furlong. 2021. '*Cis* -Acting Variation Is Common across Regulatory Layers but Is Often Buffered during Embryonic Development'. *Genome Research* 31(2): 211–24. doi:10.1101/gr.266338.120.

Friman, Elias T, Cédric Deluz, Antonio CA Meireles-Filho, Subashika Govindan, Vincent Gardeux, Bart Deplancke, and David M Suter. 2019. 'Dynamic Regulation of Chromatin Accessibility by Pluripotency Transcription Factors across the Cell Cycle'. *eLife* 8: e50087. doi:10.7554/eLife.50087.

Frommer, M, L E McDonald, D S Millar, C M Collis, F Watt, G W Grigg, P L Molloy, and C L Paul. 1992. 'A Genomic Sequencing Protocol That Yields a Positive Display of 5-Methylcytosine Residues in Individual DNA Strands.' *Proceedings of the National Academy of Sciences* 89(5): 1827–31. doi:10.1073/pnas.89.5.1827.

Fukaya, Takashi, Bomyi Lim, and Michael Levine. 2016. 'Enhancer Control of Transcriptional Bursting'. *Cell* 166(2): 358–68. doi:10.1016/j.cell.2016.05.025.

Fulco, Charles P., Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R. Grossman, Elizabeth M. Perez, Michael Kane, et al. 2016. 'Systematic Mapping of Functional Enhancer–Promoter Connections with CRISPR Interference'. *Science* 354(6313): 769–73. doi:10.1126/science.aag2445.

Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. 'Activity-by-Contact Model of Enhancer–Promoter Regulation from Thousands of CRISPR Perturbations'. *Nature Genetics* 51(12): 1664–69. doi:10.1038/s41588-019-0538-0.

Fusco, Dahlene, Nathalie Accornero, Brigitte Lavoie, Shailesh M. Shenoy, Jean-Marie Blanchard, Robert H. Singer, and Edouard Bertrand. 2003. 'Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells'. *Current Biology* 13(2): 161–67. doi:10.1016/S0960-9822(02)01436-7.

Gadea, Fabiana C. Malaga, and Evgenia N. Nikolova. 2022. *Nucleosome Topology and DNA Sequence Modulate the Engagement of Pioneer Factors SOX2 and OCT4*. Biochemistry. preprint. doi:10.1101/2022.01.18.476780.

Gaidatzis, Dimos, Anita Lerch, Florian Hahne, and Michael B. Stadler. 2015. 'QuasR: Quantification and Annotation of Short Reads in R'. *Bioinformatics* 31(7): 1130–32. doi:10.1093/bioinformatics/btu781.

Galupa, Rafael, and Edith Heard. 2018. 'X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation'. *Annual Review of Genetics* 52(1): 535–66. doi:10.1146/annurev-genet-120116-024611.

Garcia-Saez, Isabel, Hervé Menoni, Ramachandran Boopathi, Manu S. Shukla, Lama Soueidan, Marjolaine Noirclerc-Savoye, Aline Le Roy, et al. 2018. 'Structure of an H1-Bound 6-Nucleosome Array Reveals an Untwisted Two-Start Chromatin Fiber Conformation'. *Molecular Cell* 72(5): 902-915.e7. doi:10.1016/j.molcel.2018.09.027.

Gasperini, Molly, Gregory M. Findlay, Aaron McKenna, Jennifer H. Milbank, Choli Lee, Melissa D. Zhang, Darren A. Cusanovich, and Jay Shendure. 2017. 'CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions'. *The American Journal of Human Genetics* 101(2): 192–205. doi:10.1016/j.ajhg.2017.06.010.

Gasperini, Molly, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, et al. 2019. 'A Genome-Wide Framework for Mapping Gene Regulation via Cellular Genetic Screens'. *Cell* 176(1–2): 377-390.e19. doi:10.1016/j.cell.2018.11.029.

Gasperini, Molly, Jacob M. Tome, and Jay Shendure. 2020. 'Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers'. *Nature Reviews Genetics* 21(5): 292–310. doi:10.1038/s41576-019-0209-0.

Gehring, Walter J., Markus Affolter, and Thomas Bürglin. 1994. 'HOMEODOMAIN PROTEINS'. *Annual Review of Biochemistry* 63(1): 487–526. doi:10.1146/annurev.bi.63.070194.002415.

Gentleman, Robert C, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. 2004. 'Bioconductor: Open Software Development for Computational Biology and Bioinformatics'. *Genome Biology* 5(10): R80. doi:10.1186/gb-2004-5-10-r80.

Giorgetti, Luca, Bryan R. Lajoie, Ava C. Carter, Mikael Attia, Ye Zhan, Jin Xu, Chong Jian Chen, et al. 2016. 'Structural Organization of the Inactive X Chromosome in the Mouse'. *Nature* 535(7613): 575–79. doi:10.1038/nature18589.

Giresi, Paul G., Jonghwan Kim, Ryan M. McDaniell, Vishwanath R. Iyer, and Jason D. Lieb. 2007. 'FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Isolates Active Regulatory Elements from Human Chromatin'. *Genome Research* 17(6): 877–85. doi:10.1101/gr.5533506.

Goncalves, Angela, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro, Paul Flicek, Alvis Brazma, Duncan T. Odom, and John C. Marioni. 2012. 'Extensive Compensatory *Cis-Trans* Regulation in the Evolution of Mouse Gene Expression'. *Genome Research* 22(12): 2376–84. doi:10.1101/gr.142281.112.

Gong, Wuming, Satyabrata Das, Javier E. Sierra-Pagan, Erik Skie, Nikita Dsouza, Thijs A. Larson, Mary G. Garry, et al. 2022. 'ETV2 Functions as a Pioneer Factor to Regulate and Reprogram the Endothelial Lineage'. *Nature Cell Biology* 24(5): 672–84. doi:10.1038/s41556-022-00901-3.

Grand, Ralph S., Lukas Burger, Cathrin Gräwe, Alicia K. Michael, Luke Isbel, Daniel Hess, Leslie Hoerner, et al. 2021. 'BANP Opens Chromatin and Activates CpG-Island-Regulated Genes'. *Nature* 596(7870): 133–37. doi:10.1038/s41586-021-03689-8.

Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. 'Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data'. *Bioinformatics* 32(18): 2847–49. doi:10.1093/bioinformatics/btw313.

Gurard-Levin, Zachary A., Jean-Pierre Quivy, and Geneviève Almouzni. 2014. 'Histone Chaperones: Assisting Histone Traffic and Nucleosome Dynamics'. *Annual Review of Biochemistry* 83(1): 487–517. doi:10.1146/annurev-biochem-060713-035536.

Haberle, Vanja, and Alexander Stark. 2018. 'Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation'. *Nature Reviews Molecular Cell Biology* 19(10): 621–37. doi:10.1038/s41580-018-0028-8.

Hanna, Courtney W., Aaron Taudt, Jiahao Huang, Lenka Gahurova, Andrea Kranz, Simon Andrews, Wendy Dean, et al. 2018. 'MLL2 Conveys Transcription-Independent H3K4 Trimethylation in Oocytes'. *Nature Structural & Molecular Biology* 25(1): 73–82. doi:10.1038/s41594-017-0013-5.

Harada, Bryan T, William L Hwang, Sebastian Deindl, Nilanjana Chatterjee, Blaine Bartholomew, and Xiaowei Zhuang. 2016. 'Stepwise Nucleosome Translocation by RSC Remodeling Complexes'. *eLife* 5: e10051. doi:10.7554/eLife.10051.

Harrison, Melissa M., Xiao-Yong Li, Tommy Kaplan, Michael R. Botchan, and Michael B. Eisen. 2011. 'Zelda Binding in the Early Drosophila Melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition' ed. Gregory P. Copenhaver. *PLoS Genetics* 7(10): e1002266. doi:10.1371/journal.pgen.1002266.

Harshil Patel, Maxime U Garcia, Adam Talbot, Sateesh_Peri, Moritz E. Beber, Esha Joshi, Daisy Wenyan Han, et al. 2024. 'Nf-Core/Fetchngs: Nf-Core/Fetchngs v1.12.0 - Titanium Platypus'. doi:10.5281/ZENODO.5070524.

He, Qiye, Jeff Johnston, and Julia Zeitlinger. 2015. 'ChIP-Nexus Enables Improved Detection of in Vivo Transcription Factor Binding Footprints'. *Nature Biotechnology* 33(4): 395–401. doi:10.1038/nbt.3121.

He, Yuan, Jie Fang, Dylan J. Taatjes, and Eva Nogales. 2013. 'Structural Visualization of Key Steps in Human Transcription Initiation'. *Nature* 495(7442): 481–86. doi:10.1038/nature11991.

He, Yuan, Chunli Yan, Jie Fang, Carla Inouye, Robert Tjian, Ivaylo Ivanov, and Eva Nogales. 2016. 'Near-Atomic Resolution Visualization of Human Transcription Promoter Opening'. *Nature* 533(7603): 359–65. doi:10.1038/nature17970.

Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. 'Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression'. *Nature* 459(7243): 108–12. doi:10.1038/nature07829.

Heinz, S., C. E. Romanoski, C. Benner, K. A. Allison, M. U. Kaikkonen, L. D. Orozco, and C. K. Glass. 2013. 'Effect of Natural Genetic Variation on Enhancer Selection and Function'. *Nature* 503(7477): 487–92. doi:10.1038/nature12615.

Hochepied, Tino, Luc Schoonjans, Jan Staelens, Veerle Kreemers, Sophie Danloy, Leen Puimège, Désiré Collen, Frans Van Roy, and Claude Libert. 2004. 'Breaking the Species Barrier: Derivation of Germline-Competent Embryonic Stem Cells from *Mus Spretus* × C57BL/6 Hybrids'. *STEM CELLS* 22(4): 441–47. doi:10.1634/stemcells.22-4-441.

Hook, Paul W., and Winston Timp. 2023. 'Beyond Assembly: The Increasing Flexibility of Single-Molecule Sequencing Technology'. *Nature Reviews Genetics* 24(9): 627–41. doi:10.1038/s41576-023-00600-1.

Horton, Connor A., Amr M. Alexandari, Michael G. B. Hayes, Emil Marklund, Julia M. Schaepe, Arjun K. Aditham, Nilay Shah, et al. 2023. 'Short Tandem Repeats Bind Transcription Factors to Tune Eukaryotic Gene Expression'. *Science* 381(6664): eadd1250. doi:10.1126/science.add1250.

Hota, Swetansu K., and Benoit G. Bruneau. 2016. 'ATP-Dependent Chromatin Remodeling during Mammalian Development'. *Development* 143(16): 2882–97. doi:10.1242/dev.128892.

Hu, Gangqing, Dustin E. Schones, Kairong Cui, River Ybarra, Daniel Northrup, Qingsong Tang, Luca Gattinoni, et al. 2011. 'Regulation of Nucleosome Landscape and Transcription Factor Targeting at Tissue-Specific Enhancers by BRG1'. *Genome Research* 21(10): 1650–58. doi:10.1101/gr.121145.111.

Hu, Shaohui, Zhi Xie, Akishi Onishi, Xueping Yu, Lizhi Jiang, Jimmy Lin, Hee-sool Rho, et al. 2009. 'Profiling the Human Protein-DNA Interactome Reveals ERK2 as a Transcriptional Repressor of Interferon Signaling'. *Cell* 139(3): 610–22. doi:10.1016/j.cell.2009.08.037.

Hu, Yan, Sai Ma, Vinay K. Kartha, Fabiana M. Duarte, Max Horlbeck, Ruochi Zhang, Rojesh Shrestha, et al. 2023. *Single-Cell Multi-Scale Footprinting Reveals the Modular*

Organization of DNA Regulatory Elements. Genomics. preprint. doi:10.1101/2023.03.28.533945.

Huang, Huilin, Hengyou Weng, Keren Zhou, Tong Wu, Boxuan Simen Zhao, Mingli Sun, Zhenhua Chen, et al. 2019. 'Histone H3 Trimethylation at Lysine 36 Guides m6A RNA Modification Co-Transcriptionally'. *Nature* 567(7748): 414–19. doi:10.1038/s41586-019-1016-7.

Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, et al. 2015. 'Orchestrating High-Throughput Genomic Analysis with Bioconductor'. *Nature Methods* 12(2): 115–21. doi:10.1038/nmeth.3252.

Ibarra, Ignacio L., Nele M. Hollmann, Bernd Klaus, Sandra Augsten, Britta Velten, Janosch Hennig, and Judith B. Zaugg. 2020. 'Mechanistic Insights into Transcription Factor Cooperativity and Its Impact on Protein-Phenotype Interactions'. *Nature Communications* 11(1): 124. doi:10.1038/s41467-019-13888-7.

Illingworth, Robert S., and Adrian P. Bird. 2009. 'CpG Islands – "A Rough Guide"'. *FEBS Letters* 583(11): 1713–20. doi:10.1016/j.febslet.2009.04.012.

International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research:, Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, et al. 2001. 'Initial Sequencing and Analysis of the Human Genome'. *Nature* 409(6822): 860–921. doi:10.1038/35057062.

Isaac, R. Stefan, Thomas W. Tullius, Katja G. Hansen, Danilo Dubocanin, Mary Couvillion, Andrew B. Stergachis, and L. Stirling Churchman. 2024. 'Single-Nucleoid Architecture Reveals Heterogeneous Packaging of Mitochondrial DNA'. *Nature Structural & Molecular Biology*. doi:10.1038/s41594-024-01225-6.

Isbel, Luke, Ralph S. Grand, and Dirk Schübeler. 2022. 'Generating Specificity in Genome Regulation through Transcription Factor Sensitivity to Chromatin'. *Nature Reviews Genetics* 23(12): 728–40. doi:10.1038/s41576-022-00512-6.

Iurlaro, Mario, Michael B. Stadler, Francesca Masoni, Zainab Jagani, Giorgio G. Galli, and Dirk Schübeler. 2021. 'Mammalian SWI/SNF Continuously Restores Local Accessibility to Chromatin'. *Nature Genetics* 53(3): 279–87. doi:10.1038/s41588-020-00768-w.

Iwafuchi, Makiko, Isabel Cuesta, Greg Donahue, Naomi Takenaka, Anna B. Osipovich, Mark A. Magnuson, Heinrich Roder, et al. 2020. 'Gene Network Transitions in Embryos Depend upon Interactions between a Pioneer Transcription Factor and Core Histones'. *Nature Genetics* 52(4): 418–27. doi:10.1038/s41588-020-0591-8.

Jacob, François, and Jacques Monod. 1961. 'Genetic Regulatory Mechanisms in the Synthesis of Proteins'. *Journal of Molecular Biology* 3(3): 318–56. doi:10.1016/S0022-2836(61)80072-7.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. 'Genome-Wide Mapping of in Vivo Protein-DNA Interactions'. *Science* 316(5830): 1497–1502. doi:10.1126/science.1141319.

Jolma, Arttu, Jian Yan, Tom Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. 'DNA-Binding Specificities of Human Transcription Factors'. *Cell* 152(1): 327–39. doi:10.1016/j.cell.2012.12.009.

Jolma, Arttu, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, et al. 2015. 'DNA-Dependent Formation of Transcription Factor Pairs Alters Their Binding Specificity'. *Nature* 527(7578): 384–88. doi:10.1038/nature15518.

Jonkers, Iris, and John T. Lis. 2015. 'Getting up to Speed with Transcription Elongation by RNA Polymerase II'. *Nature Reviews Molecular Cell Biology* 16(3): 167–77. doi:10.1038/nrm3953.

Juven-Gershon, Tamar, Susan Cheng, and James T Kadonaga. 2006. 'Rational Design of a Super Core Promoter That Enhances Gene Expression'. *Nature Methods* 3(11): 917–22. doi:10.1038/nmeth937.

Juven-Gershon, Tamar, Jer-Yuan Hsu, Joshua Wm Theisen, and James T Kadonaga. 2008. 'The RNA Polymerase II Core Promoter — the Gateway to Transcription'. *Current Opinion in Cell Biology* 20(3): 253–59. doi:10.1016/j.ceb.2008.03.003.

Kasianowicz, John J., Eric Brandin, Daniel Branton, and David W. Deamer. 1996. 'Characterization of Individual Polynucleotide Molecules Using a Membrane Channel'. *Proceedings of the National Academy of Sciences* 93(24): 13770–73. doi:10.1073/pnas.93.24.13770.

Keane, Thomas M., Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, et al. 2011. 'Mouse Genomic Variation and Its Effect on Phenotypes and Gene Regulation'. *Nature* 477(7364): 289–94. doi:10.1038/nature10413.

Kelly, T. K., Y. Liu, F. D. Lay, G. Liang, B. P. Berman, and P. A. Jones. 2012. 'Genome-Wide Mapping of Nucleosome Positioning and DNA Methylation within Individual DNA Molecules'. *Genome Research* 22(12): 2497–2506. doi:10.1101/gr.143008.112.

Kerppola, Tom K., and Tom Curran. 1991. 'Fos-Jun Heterodimers and Jun Homodimers Bend DNA in Opposite Orientations: Implications for Transcription Factor Cooperativity'. *Cell* 66(2): 317–26. doi:10.1016/0092-8674(91)90621-5.

Khoueiry, Pierre, Charles Girardot, Lucia Ciglar, Pei-Chen Peng, E Hilary Gustafson, Saurabh Sinha, and Eileen Em Furlong. 2017. 'Uncoupling Evolutionary Changes in DNA Sequence, Transcription Factor Occupancy and Enhancer Activity'. *eLife* 6: e28440. doi:10.7554/eLife.28440.

Kilpinen, Helena, Sebastian M. Waszak, Andreas R. Gschwind, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, Eugenia Migliavacca, et al. 2013. 'Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription'. *Science* 342(6159): 744–47. doi:10.1126/science.1242463.

Kim, Seungsoo, Ekaterina Morgunova, Sahin Naqvi, Seppe Goovaerts, Maram Bader, Mervenaz Koska, Alexander Popov, et al. 2024. 'DNA-Guided Transcription Factor Cooperativity Shapes Face and Limb Mesenchyme'. *Cell* 187(3): 692-711.e26. doi:10.1016/j.cell.2023.12.032.

Kim, Seungsoo, and Joanna Wysocka. 2023. 'Deciphering the Multi-Scale, Quantitative Cis-Regulatory Code'. *Molecular Cell* 83(3): 373–92. doi:10.1016/j.molcel.2022.12.032.

King, Hamish W, and Robert J Klose. 2017. 'The Pioneer Factor OCT4 Requires the Chromatin Remodeller BRG1 to Support Gene Regulatory Element Function in Mouse Embryonic Stem Cells'. *eLife* 6: e22631. doi:10.7554/eLife.22631.

Kizer, Kelby O., Hemali P. Phatnani, Yoichiro Shibata, Hana Hall, Arno L. Greenleaf, and Brian D. Strahl. 2005. 'A Novel Domain in Set2 Mediates RNA Polymerase II Interaction and Couples Histone H3 K36 Methylation with Transcript Elongation'. *Molecular and Cellular Biology* 25(8): 3305–16. doi:10.1128/MCB.25.8.3305-3316.2005.

Klann, Tyler S, Joshua B Black, Malathi Chellappan, Alexias Safi, Lingyun Song, Isaac B Hilton, Gregory E Crawford, Timothy E Reddy, and Charles A Gersbach. 2017. 'CRISPR–Cas9 Epigenome Editing Enables High-Throughput Screening for Functional Regulatory Elements in the Human Genome'. *Nature Biotechnology* 35(6): 561–68. doi:10.1038/nbt.3853.

Kleinendorst, Rozemarijn W. D., Guido Barzaghi, Mike Smith, Judith B. Zaugg, and Arnaud R. Krebs. 2021. 'Genome-Wide Quantification of Transcription Factor Binding at Single-DNA-Molecule Resolution Using Methyl-Transferase Footprinting'. *Nature Protocols* 16(12): 5673–5706. doi:10.1038/s41596-021-00630-1.

Kornberg, R D. 2007. 'The Molecular Basis of Eucaryotic Transcription'. *Cell Death & Differentiation* 14(12): 1989–97. doi:10.1038/sj.cdd.4402251.

Kornberg, Roger D. 2005. 'Mediator and the Mechanism of Transcriptional Activation'. *Trends in Biochemical Sciences* 30(5): 235–39. doi:10.1016/j.tibs.2005.03.011.

Krebs, Arnaud R., Dilek Imanci, Leslie Hoerner, Dimos Gaidatzis, Lukas Burger, and Dirk Schübeler. 2017a. 'Genome-Wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters'. *Molecular Cell* 67(3): 411-422.e4. doi:10.1016/j.molcel.2017.06.027.

Krebs, Arnaud R., Dilek Imanci, Leslie Hoerner, Dimos Gaidatzis, Lukas Burger, and Dirk Schübeler. 2017b. 'Genome-Wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters'. *Molecular Cell* 67(3): 411-422.e4. doi:10.1016/j.molcel.2017.06.027.

Kreibich, Elisa, Rozemarijn Kleinendorst, Guido Barzaghi, Sarah Kaspar, and Arnaud R. Krebs. 2023. 'Single-Molecule Footprinting Identifies Context-Dependent Regulation of Enhancers by DNA Methylation'. *Molecular Cell* 83(5): 787-802.e9. doi:10.1016/j.molcel.2023.01.017.

Krueger, Felix, Frankie James, Phil Ewels, Ebrahim Afyounian, Michael Weinstein, Benjamin Schuster-Boeckler, Gert Hulselmans, and Sclamons. 2023. 'FelixKrueger/TrimGalore: V0.6.10 - Add Default Decompression Path'. doi:10.5281/ZENODO.7598955.

Kubik, Slawomir, Maria Jessica Bruzzone, Drice Challal, René Dreos, Stefano Mattarocci, Philipp Bucher, Domenico Libri, and David Shore. 2019. 'Opposing Chromatin Remodelers Control Transcription Initiation Frequency and Start Site Selection'. *Nature Structural & Molecular Biology* 26(8): 744–54. doi:10.1038/s41594-019-0273-3.

Kuehner, Jason N., Erika L. Pearson, and Claire Moore. 2011. 'Unravelling the Means to an End: RNA Polymerase II Transcription Termination'. *Nature Reviews Molecular Cell Biology* 12(5): 283–94. doi:10.1038/nrm3098.

Kulakovskiy, Ivan V, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, et al. 2018. 'HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP-Seq Analysis'. *Nucleic Acids Research* 46(D1): D252–59. doi:10.1093/nar/gkx1106.

Lago, Sara, Matteo Nadai, Filippo M. Cernilogar, Maryam Kazerani, Helena Domíniguez Moreno, Gunnar Schotta, and Sara N. Richter. 2021. 'Promoter G-Quadruplexes and Transcription Factors Cooperate to Shape the Cell Type-Specific Transcriptome'. *Nature Communications* 12(1): 3885. doi:10.1038/s41467-021-24198-2.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, et al. 2018. 'The Human Transcription Factors'. *Cell* 172(4): 650–65. doi:10.1016/j.cell.2018.01.029.

Landan, Gilad, Netta Mendelson Cohen, Zohar Mukamel, Amir Bar, Alina Molchadsky, Ran Brosh, Shirley Horn-Saban, et al. 2012. 'Epigenetic Polymorphism and the Stochastic Formation of Differentially Methylated Regions in Normal and Cancerous Tissues'. *Nature Genetics* 44(11): 1207–14. doi:10.1038/ng.2442.

Lee, Catherine S., Joshua R. Friedman, James T. Fulmer, and Klaus H. Kaestner. 2005. 'The Initiation of Liver Development Is Dependent on Foxa Transcription Factors'. *Nature* 435(7044): 944–47. doi:10.1038/nature03649.

Lee, Chanhyo, Xiaoyong Li, Aaron Hechmer, Michael Eisen, Mark D. Biggin, Bryan J. Venters, Cizhong Jiang, et al. 2008. 'NELF and GAGA Factor Are Linked to Promoter-Proximal Pausing at Many Genes in *Drosophila*'. *Molecular and Cellular Biology* 28(10): 3290–3300. doi:10.1128/MCB.02224-07.

Lee, Isac, Roham Razaghi, Timothy Gilpatrick, Michael Molnar, Ariel Gershman, Norah Sadowski, Fritz J. Sedlazeck, et al. 2020. 'Simultaneous Profiling of Chromatin Accessibility and Methylation on Human Cell Lines with Nanopore Sequencing'. *Nature Methods* 17(12): 1191–99. doi:10.1038/s41592-020-01000-7.

Lee, Kihyun, Hyunwoo Cho, Robert W. Rickert, Qing V. Li, Julian Pulecio, Christina S. Leslie, and Danwei Huangfu. 2019. 'FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation'. *Cell Reports* 28(2): 382-393.e7. doi:10.1016/j.celrep.2019.06.034.

Lee, Stuart, Dianne Cook, and Michael Lawrence. 2019. 'Plyranges: A Grammar of Genomic Data Transformation'. *Genome Biology* 20(1): 4. doi:10.1186/s13059-018-1597-8.

Lerner, Jonathan, Andrew Katznelson, Jingchao Zhang, and Kenneth S. Zaret. 2023. 'Different Chromatin-Scanning Modes Lead to Targeting of Compacted Chromatin by Pioneer Factors FOXA1 and SOX2'. *Cell Reports* 42(7): 112748. doi:10.1016/j.celrep.2023.112748.

Li, Mo, and Juan Carlos Izpisua Belmonte. 2017. 'Ground Rules of the Pluripotency Gene Regulatory Network'. *Nature Reviews Genetics* 18(3): 180–91. doi:10.1038/nrg.2016.156.

Li, Mo, and Juan Carlos Izpisua Belmonte. 2018. 'Deconstructing the Pluripotency Gene Regulatory Network'. *Nature Cell Biology* 20(4): 382–92. doi:10.1038/s41556-018-0067-6.

Li, Qiliang, Kenneth R. Peterson, Xiangdong Fang, and George Stamatoyannopoulos. 2002. 'Locus Control Regions'. *Blood* 100(9): 3077–86. doi:10.1182/blood-2002-04-1104.

Liang, Hsiao-Lan, Chung-Yi Nien, Hsiao-Yun Liu, Mark M. Metzstein, Nikolai Kirov, and Christine Rushlow. 2008. 'The Zinc-Finger Protein Zelda Is a Key Activator of the Early Zygotic Genome in Drosophila'. *Nature* 456(7220): 400–403. doi:10.1038/nature07388.

Lickwar, Colin R., Florian Mueller, Sean E. Hanlon, James G. McNally, and Jason D. Lieb. 2012. 'Genome-Wide Protein–DNA Binding Dynamics Suggest a Molecular Clutch for Transcription Factor Function'. *Nature* 484(7393): 251–55. doi:10.1038/nature10985.

Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. 'Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome'. *Science* 326(5950): 289–93. doi:10.1126/science.1181369.

Lilue, Jingtao, Anthony G. Doran, Ian T. Fiddes, Monica Abrudan, Joel Armstrong, Ruth Bennett, William Chow, et al. 2018. 'Sixteen Diverse Laboratory Mouse Reference Genomes Define Strain-Specific Haplotypes and Novel Functional Loci'. *Nature Genetics* 50(11): 1574–83. doi:10.1038/s41588-018-0223-8.

Lim, Chin Yan, Buyung Santoso, Thomas Boulay, Emily Dong, Uwe Ohler, and James T. Kadonaga. 2004. 'The MTE, a New Core Promoter Element for Transcription by RNA Polymerase II'. *Genes & Development* 18(13): 1606–17. doi:10.1101/gad.1193404.

Lim, Fabian, Joe J. Solvason, Genevieve E. Ryan, Sophia H. Le, Granton A. Jindal, Paige Steffen, Simran K. Jandu, and Emma K. Farley. 2024. 'Affinity-Optimizing Enhancer Variants Disrupt Development'. *Nature* 626(7997): 151–59. doi:10.1038/s41586-023-06922-8.

Liu, Ning Qing, Michela Maresca, Teun Van Den Brand, Luca Braccioli, Marijne M. G. A. Schijns, Hans Teunissen, Benoit G. Bruneau, Elphège P. Nora, and Elzo De Wit. 2021a. 'WAPL Maintains a Cohesin Loading Cycle to Preserve Cell-Type-Specific Distal Gene Regulation'. *Nature Genetics* 53(1): 100–109. doi:10.1038/s41588-020-00744-4.

Liu, Ning Qing, Michela Maresca, Teun Van Den Brand, Luca Braccioli, Marijne M. G. A. Schijns, Hans Teunissen, Benoit G. Bruneau, Elphège P. Nora, and Elzo De Wit. 2021b. 'WAPL Maintains a Cohesin Loading Cycle to Preserve Cell-Type-Specific Distal Gene Regulation'. *Nature Genetics* 53(1): 100–109. doi:10.1038/s41588-020-00744-4.

Lobato-Moreno, Sara, Umut Yildiz, Annique Claringbould, Nila H. Servaas, Evi P. Vlachou, Christian Arnold, Hanke Gwendolyn Bauersachs, et al. 2023. *Scalable Ultra-High-Throughput Single-Cell Chromatin and RNA Sequencing Reveals Gene Regulatory Dynamics Linking Macrophage Polarization to Autoimmune Disease*. Genomics. preprint. doi:10.1101/2023.12.26.573253.

Loh, Yuin-Han, Qiang Wu, Joon-Lin Chew, Vinsensius B Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, et al. 2006. 'The Oct4 and Nanog Transcription Network Regulates Pluripotency in Mouse Embryonic Stem Cells'. *Nature Genetics* 38(4): 431–40. doi:10.1038/ng1760.

Luger, Karolin, Armin W. Mäder, Robin K. Richmond, David F. Sargent, and Timothy J. Richmond. 1997. 'Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution'. *Nature* 389(6648): 251–60. doi:10.1038/38444.

Luger, Karolin, and Timothy J Richmond. 1998. 'The Histone Tails of the Nucleosome'. *Current Opinion in Genetics & Development* 8(2): 140–46. doi:10.1016/S0959-437X(98)80134-2.

Lukauskas, Saulius, Andrey Tvardovskiy, Nhuong V. Nguyen, Mara Stadler, Peter Faull, Tina Ravnsborg, Bihter Özdemir Aygenli, et al. 2024. 'Decoding Chromatin States by Proteomic Profiling of Nucleosome Readers'. *Nature*. doi:10.1038/s41586-024-07141-5.

Luscombe, Nicholas M, Susan E Austin, Helen M Berman, and Janet M Thornton. 2000. 'An Overview of the Structures of Protein-DNA Complexes'. *Genome Biology* 1(1): reviews001.1. doi:10.1186/gb-2000-1-1-reviews001.

Mahat, Dig Bijay, Hojoong Kwak, Gregory T Booth, Iris H Jonkers, Charles G Danko, Ravi K Patel, Colin T Waters, et al. 2016. 'Base-Pair-Resolution Genome-Wide Mapping of

Active RNA Polymerases Using Precision Nuclear Run-on (PRO-Seq)'. *Nature Protocols* 11(8): 1455–76. doi:10.1038/nprot.2016.086.

Malik, Harmit S, and Steven Henikoff. 2003. 'Phylogenomics of the Nucleosome'. *Nature Structural & Molecular Biology* 10(11): 882–91. doi:10.1038/nsb996.

Malik, Sohail, and Robert G. Roeder. 2005. 'Dynamic Regulation of Pol II Transcription by the Mammalian Mediator Complex'. *Trends in Biochemical Sciences* 30(5): 256–63. doi:10.1016/j.tibs.2005.03.009.

Malik, Sohail, and Robert G. Roeder. 2010. 'The Metazoan Mediator Co-Activator Complex as an Integrative Hub for Transcriptional Regulation'. *Nature Reviews Genetics* 11(11): 761–72. doi:10.1038/nrg2901.

Malik, Vikas, Laura V. Glaser, Dennis Zimmer, Sergiy Velychko, Mingxi Weng, Markus Holzner, Marius Arend, et al. 2019. 'Pluripotency Reprogramming by Competent and Incompetent POU Factors Uncovers Temporal Dependency for Oct4 and Sox2'. *Nature Communications* 10(1): 3477. doi:10.1038/s41467-019-11054-7.

Maresca, Michela, Teun Van Den Brand, Hangpeng Li, Hans Teunissen, James Davies, and Elzo De Wit. 2023. 'Pioneer Activity Distinguishes Activating from Non-activating SOX2 Binding Sites'. *The EMBO Journal* 42(20): e113150. doi:10.15252/embj.2022113150.

Mariani, Luca, Kathryn Weinand, Anastasia Vedenko, Luis A. Barrera, and Martha L. Bulyk. 2017. 'Identification of Human Lineage-Specific Transcriptional Coregulators Enabled by a Glossary of Binding Modules and Tunable Genomic Backgrounds'. *Cell Systems* 5(3): 187-201.e7. doi:10.1016/j.cels.2017.06.015.

Mathelier, Anthony, Oriol Fornes, David J. Arenillas, Chih-yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, et al. 2016. 'JASPAR 2016: A Major Expansion and Update of the Open-Access Database of Transcription Factor Binding Profiles'. *Nucleic Acids Research* 44(D1): D110–15. doi:10.1093/nar/gkv1176.

Maurano, Matthew T, Eric Haugen, Richard Sandstrom, Jeff Vierstra, Anthony Shafer, Rajinder Kaul, and John A Stamatoyannopoulos. 2015. 'Large-Scale Identification of Sequence Variants Influencing Human Transcription Factor Occupancy in Vivo'. *Nature Genetics* 47(12): 1393–1401. doi:10.1038/ng.3432.

Mayran, Alexandre, and Jacques Drouin. 2018. 'Pioneer Transcription Factors Shape the Epigenetic Landscape'. *Journal of Biological Chemistry* 293(36): 13795–804. doi:10.1074/jbc.R117.001232.

McKnight, Steven Lanier. 1991. 'Molecular Zippers in Gene Regulation'. *Scientific American* 264(4): 54–64. doi:10.1038/scientificamerican0491-54.

Meers, Michael P, Telmo Henriques, Christopher A Lavender, Daniel J McKay, Brian D Strahl, Robert J Duronio, Karen Adelman, and A Gregory Matera. 2017. 'Histone Gene

Replacement Reveals a Post-Transcriptional Role for H3K36 in Maintaining Metazoan Transcriptome Fidelity'. *eLife* 6: e23249. doi:10.7554/eLife.23249.

Melnikov, Alexandre, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, et al. 2012. 'Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay'. *Nature Biotechnology* 30(3): 271–77. doi:10.1038/nbt.2137.

Merika, Menie, and Dimitris Thanos. 2001. 'Enhanceosomes'. *Current Opinion in Genetics & Development* 11(2): 205–8. doi:10.1016/S0959-437X(00)00180-5.

Michael, Alicia K., Ralph S. Grand, Luke Isbel, Simone Cavadini, Zuzanna Kozicka, Georg Kempf, Richard D. Bunker, et al. 2020. 'Mechanisms of OCT4-SOX2 Motif Readout on Nucleosomes'. *Science* 368(6498): 1460–65. doi:10.1126/science.abb0074.

Michael, Alicia K., and Nicolas H. Thomä. 2021. 'Reading the Chromatinized Genome'. *Cell* 184(14): 3599–3611. doi:10.1016/j.cell.2021.05.029.

Millán-Zambrano, Gonzalo, Adam Burton, Andrew J. Bannister, and Robert Schneider. 2022. 'Histone Post-Translational Modifications — Cause and Consequence of Genome Function'. *Nature Reviews Genetics* 23(9): 563–80. doi:10.1038/s41576-022-00468-7.

Miller, Joanna A., and Jonathan Widom. 2003. 'Collaborative Competition Mechanism for Gene Activation In Vivo'. *Molecular and Cellular Biology* 23(5): 1623–32. doi:10.1128/MCB.23.5.1623-1632.2003.

Minderjahn, Julia, Andreas Schmidt, Andreas Fuchs, Rudolf Schill, Johanna Raithel, Magda Babina, Christian Schmidl, et al. 2020. 'Mechanisms Governing the Pioneering and Redistribution Capabilities of the Non-Classical Pioneer PU.1'. *Nature Communications* 11(1): 402. doi:10.1038/s41467-019-13960-2.

Mirny, L. A. 2010. 'Nucleosome-Mediated Cooperativity between Transcription Factors'. *Proceedings of the National Academy of Sciences* 107(52): 22534–39. doi:10.1073/pnas.0913805107.

Moreau, P., R. Hen, B. Wasylyk, R. Everett, M.P. Gaub, and P. Chambon. 1981. 'The SV40 72 Base Repair Repeat Has a Striking Effect on Gene Expression Both in SV40 and Other Chimeric Recombinants'. *Nucleic Acids Research* 9(22): 6047–68. doi:10.1093/nar/9.22.6047.

Morgunova, Ekaterina, and Jussi Taipale. 2017. 'Structural Perspective of Cooperative Transcription Factor Binding'. *Current Opinion in Structural Biology* 47: 1–8. doi:10.1016/j.sbi.2017.03.006.

Moyle-Heyrman, Georgette, Hannah S. Tims, and Jonathan Widom. 2011. 'Structural Constraints in Collaborative Competition of Transcription Factors against the Nucleosome'. *Journal of Molecular Biology* 412(4): 634–46. doi:10.1016/j.jmb.2011.07.032.

Murray, Iain A, Richard D Morgan, Yvette Luyten, Alexey Fomenkov, Ivan R. Corrêa, Nan Dai, Mohammed B Allaw, et al. 2018. 'The Non-Specific Adenine DNA Methyltransferase M.EcoGII'. *Nucleic Acids Research* 46(2): 840–48. doi:10.1093/nar/gkx1191.

Musselman, Catherine A, Marie-Eve Lalonde, Jacques Côté, and Tatiana G Kutateladze. 2012. 'Perceiving the Epigenetic Landscape through Histone Readers'. *Nature Structural & Molecular Biology* 19(12): 1218–27. doi:10.1038/nsmb.2436.

Nabilsi, N. H., L. P. Deleyrolle, R. P. Darst, A. Riva, B. A. Reynolds, and M. P. Kladde. 2014. 'Multiplex Mapping of Chromatin Accessibility and DNA Methylation within Targeted Single Molecules Identifies Epigenetic Heterogeneity in Neural Stem Cells and Glioblastoma'. *Genome Research* 24(2): 329–39. doi:10.1101/gr.161737.113.

Natoli, Gioacchino, and Jean-Christophe Andrau. 2012. 'Noncoding Transcription at Enhancers: General Principles and Functional Models'. *Annual Review of Genetics* 46(1): 1–19. doi:10.1146/annurev-genet-110711-155459.

Nechaev, Sergei, and Karen Adelman. 2011. 'Pol II Waiting in the Starting Gates: Regulating the Transition from Transcription Initiation into Productive Elongation'. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1809(1): 34–45. doi:10.1016/j.bbagrm.2010.11.001.

Neumayr, Christoph, Vanja Haberle, Leonid Serebreni, Katharina Karner, Oliver Hendy, Ann Boija, Jonathan E. Henninger, et al. 2022. 'Differential Cofactor Dependencies Define Distinct Types of Human Enhancers'. *Nature* 606(7913): 406–13. doi:10.1038/s41586-022-04779-x.

Nicetto, Dario, and Kenneth S. Zaret. 2019. 'Role of H3K9me3 Heterochromatin in Cell Identity Establishment and Maintenance'. *Current Opinion in Genetics & Development* 55: 1–10. doi:10.1016/j.gde.2019.04.013.

Nishimura, Masahiro, Yasuhiro Arimura, Kayo Nozawa, and Hitoshi Kurumizaka. 2020. 'Linker DNA and Histone Contributions in Nucleosome Binding by P53'. *The Journal of Biochemistry* 168(6): 669–75. doi:10.1093/jb/mvaa081.

Nogales, Eva, Robert K. Louder, and Yuan He. 2017. 'Structural Insights into the Eukaryotic Transcription Initiation Machinery'. *Annual Review of Biophysics* 46(1): 59–83. doi:10.1146/annurev-biophys-070816-033751.

Novakovsky, Gherman, Manu Saraswat, Oriol Fornes, Sara Mostafavi, and Wyeth W. Wasserman. 2021. 'Biologically Relevant Transfer Learning Improves Transcription Factor Binding Prediction'. *Genome Biology* 22(1): 280. doi:10.1186/s13059-021-02499-5.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. 'The Complete Sequence of a Human Genome'. *Science* 376(6588): 44–53. doi:10.1126/science.abj6987.

Oberbeckmann, Elisa, Kimberly Quililan, Patrick Cramer, and A. Marieke Oudelaar. 2024. 'In Vitro Reconstitution of Chromatin Domains Shows a Role for Nucleosome Positioning in 3D Genome Organization'. *Nature Genetics*. doi:10.1038/s41588-023-01649-8.

Oberbeckmann, Elisa, Michael Wolff, Nils Krietenstein, Mark Heron, Jessica L. Ellins, Andrea Schmid, Stefan Krebs, et al. 2019. 'Absolute Nucleosome Occupancy Map for the *Saccharomyces Cerevisiae* Genome'. *Genome Research* 29(12): 1996–2009. doi:10.1101/gr.253419.119.

Ohmori, Haruo, Jun-ichi Tomizawa, and Allan M. Maxam. 1978. 'Detection of 5-Methylcytosine in DNA Sequences'. *Nucleic Acids Research* 5(5): 1479–85. doi:10.1093/nar/5.5.1479.

Ohno, Masae, Tadashi Ando, David G. Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. 2019. 'Sub-Nucleosomal Genome Structure Reveals Distinct Nucleosome Folding Motifs'. *Cell* 176(3): 520-534.e25. doi:10.1016/j.cell.2018.12.014.

Ooi, Steen K. T., Chen Qiu, Emily Bernstein, Keqin Li, Da Jia, Zhe Yang, Hediye Erdjument-Bromage, et al. 2007. 'DNMT3L Connects Unmethylated Lysine 4 of Histone H3 to de Novo Methylation of DNA'. *Nature* 448(7154): 714–17. doi:10.1038/nature05987.

Orphanides, George, and Danny Reinberg. 2002. 'A Unified Theory of Gene Expression'. *Cell* 108(4): 439–51. doi:10.1016/S0092-8674(02)00655-4.

Pabo, Carl O., and Robert T. Sauer. 1992. 'TRANSCRIPTION FACTORS: Structural Families and Principles of DNA Recognition'. *Annual Review of Biochemistry* 61(1): 1053–95. doi:10.1146/annurev.bi.61.070192.005201.

Panne, Daniel. 2008. 'The Enhanceosome'. *Current Opinion in Structural Biology* 18(2): 236–42. doi:10.1016/j.sbi.2007.12.002.

Panne, Daniel, Tom Maniatis, and Stephen C. Harrison. 2007. 'An Atomic Model of the Interferon-β Enhanceosome'. *Cell* 129(6): 1111–23. doi:10.1016/j.cell.2007.05.019.

Panten, Jasper, Tobias Heinen, Christina Ernst, Nils Eling, Rebecca E. Wagner, Maja Satorius, John C. Marioni, Oliver Stegle, and Duncan T. Odom. 2024. 'The Dynamic Genetic Determinants of Increased Transcriptional Divergence in Spermatids'. *Nature Communications* 15(1): 1272. doi:10.1038/s41467-024-45133-1.

Park, Sungjoon, Yookyung Koh, Hwisang Jeon, Hyunjae Kim, Yoonsun Yeo, and Jaewoo Kang. 2020. 'Enhancing the Interpretability of Transcription Factor Binding Site Prediction Using Attention Mechanism'. *Scientific Reports* 10(1): 13413. doi:10.1038/s41598-020-70218-4.

Patwardhan, Rupali P, Joseph B Hiatt, Daniela M Witten, Mee J Kim, Robin P Smith, Dalit May, Choli Lee, et al. 2012. 'Massively Parallel Functional Dissection of Mammalian Enhancers in Vivo'. *Nature Biotechnology* 30(3): 265–70. doi:10.1038/nbt.2136.

Patwardhan, Rupali P, Choli Lee, Oren Litvin, David L Young, Dana Pe'er, and Jay Shendure. 2009. 'High-Resolution Analysis of DNA Regulatory Elements by Synthetic Saturation Mutagenesis'. *Nature Biotechnology* 27(12): 1173–75. doi:10.1038/nbt.1589.

Pengelly, Ana Raquel, Ömer Copur, Herbert Jäckle, Alf Herzig, and Jürg Müller. 2013. 'A Histone Mutant Reproduces the Phenotype Caused by Loss of Histone-Modifying Factor Polycomb'. *Science* 339(6120): 698–99. doi:10.1126/science.1231382.

Pennacchio, Len A., Wendy Bickmore, Ann Dean, Marcelo A. Nobrega, and Gill Bejerano. 2013. 'Enhancers: Five Essential Questions'. *Nature Reviews Genetics* 14(4): 288–95. doi:10.1038/nrg3458.

'Picard Tools'. http://broadinstitute.github.io/picard.

Pimmett, Virginia L., Matthieu Dejean, Carola Fernandez, Antonio Trullo, Edouard Bertrand, Ovidiu Radulescu, and Mounia Lagha. 2021. 'Quantitative Imaging of Transcription in Living Drosophila Embryos Reveals the Impact of Core Promoter Motifs on Promoter State Dynamics'. *Nature Communications* 12(1): 4504. doi:10.1038/s41467-021-24461-6.

Polach, K.J., and J. Widom. 1996. 'A Model for the Cooperative Binding of Eukaryotic Regulatory Proteins to Nucleosomal Target Sites'. *Journal of Molecular Biology* 258(5): 800–812. doi:10.1006/jmbi.1996.0288.

Porrua, Odil, Marc Boudvillain, and Domenico Libri. 2016. 'Transcription Termination: Variations on Common Themes'. *Trends in Genetics* 32(8): 508–22. doi:10.1016/j.tig.2016.05.007.

Pott, Sebastian. 2017. 'Simultaneous Measurement of Chromatin Accessibility, DNA Methylation, and Nucleosome Phasing in Single Cells'. *eLife* 6: e23203. doi:10.7554/eLife.23203.

R Core Team. 2021. 'R: A Language and Environment for Statistical Computing'. https://www.R-project.org/.

Rajagopal, Nisha, Sharanya Srinivasan, Kameron Kooshesh, Yuchun Guo, Matthew D Edwards, Budhaditya Banerjee, Tahin Syed, et al. 2016. 'High-Throughput Mapping of Regulatory DNA'. *Nature Biotechnology* 34(2): 167–74. doi:10.1038/nbt.3468.

Ramalingam, Vivekanandan, Xinyang Yu, Brian D. Slaughter, Jay R. Unruh, Kaelan J. Brennan, Anastasiia Onyshchenko, Jeffrey J. Lange, et al. 2023. 'Lola-I Is a Promoter Pioneer Factor That Establishes de Novo Pol II Pausing during Development'. *Nature Communications* 14(1): 5862. doi:10.1038/s41467-023-41408-1.

Ramani, Vijay, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. 2017. 'Massively Multiplex Single-Cell Hi-C'. *Nature Methods* 14(3): 263–66. doi:10.1038/nmeth.4155.

Read, Quentin, Philippe Marchand, Ian Carroll, Rachael E. Blake, Ben Fasoli, Rob Gilmore, Sebastian Schubert, ChristopherBarrington, and Dayne Filer. 2021. 'SESYNC-Ci/Rslurm: Rslurm Version 0.6.1'. doi:10.5281/ZENODO.5705429.

Reiter, Franziska, Sebastian Wienerroither, and Alexander Stark. 2017. 'Combinatorial Function of Transcription Factors and Cofactors'. *Current Opinion in Genetics & Development* 43: 73–81. doi:10.1016/j.gde.2016.12.007.

Renbaum, Paul, Dan Abrahamove, Abraham Fainsod, Geoffrey G. Wilson, Shlomo Rottem, and Aharon Razin. 1990. 'Cloning, Characterization, and Expression in *Escherichia Coli* of the Gene Coding for the CpG DNA Methylase from *Spiroplasma Sp.* Strain MQ1(M Sssl)'. *Nucleic Acids Research* 18(5): 1145–52. doi:10.1093/nar/18.5.1145.

Rhee, Ho Sung, and B. Franklin Pugh. 2012. 'Genome-Wide Structure and Organization of Eukaryotic Pre-Initiation Complexes'. *Nature* 483(7389): 295–301. doi:10.1038/nature10799.

Rhee, Ho Sung, and B. Franklin Pugh. 2011. 'Comprehensive Genome-Wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution'. *Cell* 147(6): 1408–19. doi:10.1016/j.cell.2011.11.013.

Rhoads, Anthony, and Kin Fai Au. 2015. 'PacBio Sequencing and Its Applications'. *Genomics, Proteomics & Bioinformatics* 13(5): 278–89. doi:10.1016/j.gpb.2015.08.002.

Rhodes, Daniela, and Aaron Klug. 1993. 'Zinc Fingers'. *Scientific American* 268(2): 56–65. doi:10.1038/scientificamerican0293-56.

Ricci, Maria Aurelia, Carlo Manzo, María Filomena García-Parajo, Melike Lakadamyali, and Maria Pia Cosma. 2015. 'Chromatin Fibers Are Formed by Heterogeneous Groups of Nucleosomes In Vivo'. *Cell* 160(6): 1145–58. doi:10.1016/j.cell.2015.01.054.

Rickels, Ryan, Hans-Martin Herz, Christie C Sze, Kaixiang Cao, Marc A Morgan, Clayton K Collings, Maria Gause, et al. 2017. 'Histone H3K4 Monomethylation Catalyzed by Trr and Mammalian COMPASS-like Proteins at Enhancers Is Dispensable for Development and Viability'. *Nature Genetics* 49(11): 1647–53. doi:10.1038/ng.3965.

Roberts, Gareth A., Burak Ozkan, Ivana Gachulincová, Michael R. O'Dwyer, Elisa Hall-Ponsele, Manoj Saxena, Philip J. Robinson, and Abdenour Soufi. 2021. 'Dissecting OCT4 Defines the Role of Nucleosome Binding in Pluripotency'. *Nature Cell Biology* 23(8): 834–45. doi:10.1038/s41556-021-00727-5.

Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. 'Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing'. *Nature Methods* 4(8): 651–57. doi:10.1038/nmeth1068.

Robinson, Philip Jj, and Daniela Rhodes. 2006. 'Structure of the "30nm" Chromatin Fibre: A Key Role for the Linker Histone'. *Current Opinion in Structural Biology* 16(3): 336–43. doi:10.1016/j.sbi.2006.05.007.

Rodda, David J., Joon-Lin Chew, Leng-Hiong Lim, Yuin-Han Loh, Bei Wang, Huck-Hui Ng, and Paul Robson. 2005. 'Transcriptional Regulation of Nanog by OCT4 and SOX2'. *Journal of Biological Chemistry* 280(26): 24731–37. doi:10.1074/jbc.M502573200.

Roeder, Robert G. 2019. '50+ Years of Eukaryotic Transcription: An Expanding Universe of Factors and Mechanisms'. *Nature Structural & Molecular Biology* 26(9): 783–91. doi:10.1038/s41594-019-0287-x.

RStudio Team. 2020. 'RStudio: Integrated Development Environment for R'. http://www.rstudio.com/.

Sandelin, A. 2004. 'JASPAR: An Open-Access Database for Eukaryotic Transcription Factor Binding Profiles'. *Nucleic Acids Research* 32(90001): 91D – 94. doi:10.1093/nar/gkh012.

Sanjana, Neville E., Jason Wright, Kaijie Zheng, Ophir Shalem, Pierre Fontanillas, Julia Joung, Christine Cheng, Aviv Regev, and Feng Zhang. 2016. 'High-Resolution Interrogation of Functional Elements in the Noncoding Genome'. *Science* 353(6307): 1545–49. doi:10.1126/science.aaf7613.

Santangelo, Thomas J., and Irina Artsimovitch. 2011. 'Termination and Antitermination: RNA Polymerase Runs a Stop Sign'. *Nature Reviews Microbiology* 9(5): 319–29. doi:10.1038/nrmicro2560.

Schick, Sandra, Sarah Grosche, Katharina Eva Kohl, Danica Drpic, Martin G. Jaeger, Nara C. Marella, Hana Imrichova, et al. 2021. 'Acute BAF Perturbation Causes Immediate Changes in Chromatin Accessibility'. *Nature Genetics* 53(3): 269–78. doi:10.1038/s41588-021-00777-3.

Schneider, Thomas D., and R.Michael Stephens. 1990. 'Sequence Logos: A New Way to Display Consensus Sequences'. *Nucleic Acids Research* 18(20): 6097–6100. doi:10.1093/nar/18.20.6097.

Schones, Dustin E., Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. 2008. 'Dynamic Regulation of Nucleosome Positioning in the Human Genome'. *Cell* 132(5): 887–98. doi:10.1016/j.cell.2008.02.022.

Schuettengruber, Bernd, and Giacomo Cavalli. 2009. 'Recruitment of Polycomb Group Complexes and Their Role in the Dynamic Regulation of Cell Fate Choice'. *Development* 136(21): 3531–42. doi:10.1242/dev.033902.

Segert, Julian A., Stephen S. Gisselbrecht, and Martha L. Bulyk. 2021. 'Transcriptional Silencers: Driving Gene Expression with the Brakes On'. *Trends in Genetics* 37(6): 514–27. doi:10.1016/j.tig.2021.02.002.

Selth, Luke A., Stefan Sigurdsson, and Jesper Q. Svejstrup. 2010. 'Transcript Elongation by RNA Polymerase II'. *Annual Review of Biochemistry* 79(1): 271–93. doi:10.1146/annurev.biochem.78.062807.091425.

Sherwood, Richard I, Tatsunori Hashimoto, Charles W O'Donnell, Sophia Lewis, Amira A Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. 2014. 'Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape'. *Nature Biotechnology* 32(2): 171–78. doi:10.1038/nbt.2798.

Shipony, Zohar, Georgi K. Marinov, Matthew P. Swaffer, Nicholas A. Sinnott-Armstrong, Jan M. Skotheim, Anshul Kundaje, and William J. Greenleaf. 2020. 'Long-Range Single-Molecule Mapping of Chromatin Accessibility in Eukaryotes'. *Nature Methods* 17(3): 319–27. doi:10.1038/s41592-019-0730-2.

Shipony, Zohar, Zohar Mukamel, Netta Mendelson Cohen, Gilad Landan, Elad Chomsky, Shlomit Reich Zeliger, Yael Chagit Fried, et al. 2014. 'Dynamic and Static Maintenance of Epigenetic Memory in Pluripotent and Somatic Cells'. *Nature* 513(7516): 115–19. doi:10.1038/nature13458.

Shlyueva, Daria, Gerald Stampfel, and Alexander Stark. 2014. 'Transcriptional Enhancers: From Properties to Genome-Wide Predictions'. *Nature Reviews Genetics* 15(4): 272–86. doi:10.1038/nrg3682.

Shoaib, Muhammad, Qinming Chen, Xiangyan Shi, Nidhi Nair, Chinmayi Prasanna, Renliang Yang, David Walter, et al. 2021. 'Histone H4 Lysine 20 Mono-Methylation Directly Facilitates Chromatin Openness and Promotes Transcription of Housekeeping Genes'. *Nature Communications* 12(1): 4800. doi:10.1038/s41467-021-25051-2.

Simeonov, Dimitre R., Benjamin G. Gowen, Mandy Boontanrart, Theodore L. Roth, John D. Gagnon, Maxwell R. Mumbach, Ansuman T. Satpathy, et al. 2017. 'Discovery of Stimulation-Responsive Immune Enhancers with CRISPR Activation'. *Nature* 549(7670): 111–15. doi:10.1038/nature23875.

Simon, Marek, Justin A. North, John C. Shimko, Robert A. Forties, Michelle B. Ferdinand, Mridula Manohar, Meng Zhang, et al. 2011. 'Histone Fold Modifications Control Nucleosome Unwrapping and Disassembly'. *Proceedings of the National Academy of Sciences* 108(31): 12711–16. doi:10.1073/pnas.1106264108.

Singleton, Martin R., Mark S. Dillingham, and Dale B. Wigley. 2007. 'Structure and Mechanism of Helicases and Nucleic Acid Translocases'. *Annual Review of Biochemistry* 76(1): 23–50. doi:10.1146/annurev.biochem.76.052305.115300.

Sinha, Kalyan K., Silvija Bilokapic, Yongming Du, Deepshikha Malik, and Mario Halic. 2023. 'Histone Modifications Regulate Pioneer Transcription Factor Cooperativity'. *Nature* 619(7969): 378–84. doi:10.1038/s41586-023-06112-6.

Sirinakis, George, Cedric R Clapier, Ying Gao, Ramya Viswanathan, Bradley R Cairns, and Yongli Zhang. 2011. 'The RSC Chromatin Remodelling ATPase Translocates DNA with High Force and Small Step Size: Remodeller Motor Translocates DNA with High Force'. *The EMBO Journal* 30(12): 2364–72. doi:10.1038/emboj.2011.141.

Skene, Peter J, and Steven Henikoff. 2017. 'An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites'. *eLife* 6: e21856. doi:10.7554/eLife.21856.

Solomon, Mark J., Pamela L. Larsen, and Alexander Varshavsky. 1988. 'Mapping proteinDNA Interactions in Vivo with Formaldehyde: Evidence That Histone H4 Is Retained on a Highly Transcribed Gene'. *Cell* 53(6): 937–47. doi:10.1016/S0092-8674(88)90469-2.

Song, Feng, Ping Chen, Dapeng Sun, Mingzhu Wang, Liping Dong, Dan Liang, Rui-Ming Xu, Ping Zhu, and Guohong Li. 2014. 'Cryo-EM Study of the Chromatin Fiber Reveals a Double Helix Twisted by Tetranucleosomal Units'. *Science* 344(6182): 376–80. doi:10.1126/science.1251413.

Sönmezer, Can, Rozemarijn Kleinendorst, Dilek Imanci, Guido Barzaghi, Laura Villacorta, Dirk Schübeler, Vladimir Benes, Nacho Molina, and Arnaud Regis Krebs. 2021. 'Molecular Co-Occupancy Identifies Transcription Factor Binding Cooperativity In Vivo'. *Molecular Cell* 81(2): 255-267.e6. doi:10.1016/j.molcel.2020.11.015.

Soufi, Abdenour, Greg Donahue, and Kenneth S. Zaret. 2012. 'Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome'. *Cell* 151(5): 994–1004. doi:10.1016/j.cell.2012.09.045.

Soufi, Abdenour, Meilin Fernandez Garcia, Artur Jaroszewicz, Nebiyu Osman, Matteo Pellegrini, and Kenneth S. Zaret. 2015. 'Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming'. *Cell* 161(3): 555–68. doi:10.1016/j.cell.2015.03.017.

Spiegel, Jochen, Sergio Martínez Cuesta, Santosh Adhikari, Robert Hänsel-Hertsch, David Tannahill, and Shankar Balasubramanian. 2021. 'G-Quadruplexes Are Transcription Factor Binding Hubs in Human Chromatin'. *Genome Biology* 22(1): 117. doi:10.1186/s13059-021-02324-z.

Spitz, François, and Eileen E. M. Furlong. 2012. 'Transcription Factors: From Enhancer Binding to Developmental Control'. *Nature Reviews Genetics* 13(9): 613–26. doi:10.1038/nrg3207.

Spivakov, Mikhail, Junaid Akhtar, Pouya Kheradpour, Kathryn Beal, Charles Girardot, Gautier Koscielny, Javier Herrero, et al. 2012. 'Analysis of Variation at Transcription Factor Binding Sites in Drosophila and Humans'. *Genome Biology* 13(9): R49. doi:10.1186/gb-2012-13-9-r49.

Stamatoyannopoulos, John A. 2012. 'What Does Our Genome Encode?' *Genome Research* 22(9): 1602–11. doi:10.1101/gr.146506.112.

Stefflova, Klara, David Thybert, Michael D. Wilson, Ian Streeter, Jelena Aleksic, Panagiota Karagianni, Alvis Brazma, et al. 2013. 'Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals'. *Cell* 154(3): 530–40. doi:10.1016/j.cell.2013.07.007.

Stergachis, Andrew B., Brian M. Debo, Eric Haugen, L. Stirling Churchman, and John A. Stamatoyannopoulos. 2020. 'Single-Molecule Regulatory Architectures Captured by Chromatin Fiber Sequencing'. *Science* 368(6498): 1449–54. doi:10.1126/science.aaz1646.

Stormo, Gary D., and Yue Zhao. 2010. 'Determining the Specificity of Protein–DNA Interactions'. *Nature Reviews Genetics* 11(11): 751–60. doi:10.1038/nrg2845.

Szczurek, Aleksander T., Emilia Dimitrova, Jessica R. Kelley, and Robert J. Klose. 2023. *Polycomb Sustains Promoters in a Deep OFF-State by Limiting PIC Formation to Counteract Transcription*. Cell Biology. preprint. doi:10.1101/2023.06.13.544762.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors'. *Cell* 126(4): 663–76. doi:10.1016/j.cell.2006.07.024.

Takaku, Motoki, Sara A Grimm, Bony De Kumar, Brian D Bennett, and Paul A Wade. 2020. 'Cancer-Specific Mutation of GATA3 Disrupts the Transcriptional Regulatory Network Governed by Estrogen Receptor Alpha, FOXA1 and GATA3'. *Nucleic Acids Research* 48(9): 4756–68. doi:10.1093/nar/gkaa179.

Talbert, Paul B., Michael P. Meers, and Steven Henikoff. 2019. 'Old Cogs, New Tricks: The Evolution of Gene Expression in a Chromatin Context'. *Nature Reviews Genetics* 20(5): 283–97. doi:10.1038/s41576-019-0105-7.

Tan, Ge, and Boris Lenhard. 2016. 'TFBSTools: An R/Bioconductor Package for Transcription Factor Binding Site Analysis'. *Bioinformatics* 32(10): 1555–56. doi:10.1093/bioinformatics/btw024.

Tanaka, Hiroki, Yoshimasa Takizawa, Motoki Takaku, Daiki Kato, Yusuke Kumagawa, Sara A. Grimm, Paul A. Wade, and Hitoshi Kurumizaka. 2020. 'Interaction of the Pioneer Transcription Factor GATA3 with Nucleosomes'. *Nature Communications* 11(1): 4136. doi:10.1038/s41467-020-17959-y.

Team, The Bioconductor Dev. 2017. 'BSgenome.Mmusculus.UCSC.Mm10'. doi:10.18129/B9.BIOC.BSGENOME.MMUSCULUS.UCSC.MM10.

Thakore, Pratiksha I, Anthony M D'Ippolito, Lingyun Song, Alexias Safi, Nishkala K Shivakumar, Ami M Kabadi, Timothy E Reddy, Gregory E Crawford, and Charles A Gersbach. 2015. 'Highly Specific Epigenome Editing by CRISPR-Cas9 Repressors for Silencing of Distal Regulatory Elements'. *Nature Methods* 12(12): 1143–49. doi:10.1038/nmeth.3630.

The ENCODE Project Consortium. 2012. 'An Integrated Encyclopedia of DNA Elements in the Human Genome'. *Nature* 489(7414): 57–74. doi:10.1038/nature11247.

The FANTOM Consortium, Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, et al. 2014. 'An Atlas of Active Enhancers across

Human Cell Types and Tissues'. *Nature* 507(7493): 455–61. doi:10.1038/nature12787.

The modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V. Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L. Eaton, et al. 2010. 'Identification of Functional Elements and Regulatory Circuits by *Drosophila* modENCODE'. *Science* 330(6012): 1787–97. doi:10.1126/science.1198374.

Thomas Lin Pedersen. 'Patchwork: The Composer of Plots'. https://patchwork.data-imaginist.com.

Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. 'The Accessible Chromatin Landscape of the Human Genome'. *Nature* 489(7414): 75–82. doi:10.1038/nature11232.

Thybert, David, Maša Roller, Fábio C.P. Navarro, Ian Fiddes, Ian Streeter, Christine Feig, David Martin-Galvez, et al. 2018. 'Repeat Associated Mechanisms of Genome Evolution and Function Revealed by the *Mus Caroli* and *Mus Pahari* Genomes'. *Genome Research* 28(4): 448–59. doi:10.1101/gr.234096.117.

Trojanowski, Jorge, and Karsten Rippe. 2022. 'Transcription Factor Binding and Activity on Chromatin'. *Current Opinion in Systems Biology* 31: 100438. doi:10.1016/j.coisb.2022.100438.

Tullius, Thomas W., R. Stefan Isaac, Jane Ranchalis, Danilo Dubocanin, L. Stirling Churchman, and Andrew B. Stergachis. 2023. *RNA Polymerases Reshape Chromatin and Coordinate Transcription on Individual Fibers*. Genomics. preprint. doi:10.1101/2023.12.22.573133.

Tunnacliffe, Edward, and Jonathan R. Chubb. 2020. 'What Is a Transcriptional Burst?' *Trends in Genetics* 36(4): 288–97. doi:10.1016/j.tig.2020.01.003.

Vakoc, Christopher R., Mira M. Sachdeva, Hongxin Wang, and Gerd A. Blobel. 2006. 'Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin'. *Molecular and Cellular Biology* 26(24): 9185–95. doi:10.1128/MCB.01529-06.

Vannini, Alessandro, and Patrick Cramer. 2012. 'Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries'. *Molecular Cell* 45(4): 439–46. doi:10.1016/j.molcel.2012.01.023.

van Overbeek, Megan, Daniel Capurso, Matthew M. Carter, Matthew S. Thompson, Elizabeth Frias, Carsten Russ, John S. Reece-Hoyes, et al. 2016. 'DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks'. *Molecular Cell* 63(4): 633–46. doi:10.1016/j.molcel.2016.06.037.

Vashee, Sanjay, Karsten Melcher, W.Vivianne Ding, Stephen Albert Johnston, and Thomas Kodadek. 1998. 'Evidence for Two Modes of Cooperative DNA Binding in Vivo That Do Not Involve Direct Protein–Protein Interactions'. *Current Biology* 8(8): 452–58. doi:10.1016/S0960-9822(98)70179-4.

Velankar, Sameer S, Panos Soultanas, Mark S Dillingham, Hosahalli S Subramanya, and Dale B Wigley. 1999. 'Crystal Structures of Complexes of PcrA DNA Helicase with a DNA Substrate Indicate an Inchworm Mechanism'. *Cell* 97(1): 75–84. doi:10.1016/S0092-8674(00)80716-3.

Venter, J. Craig, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, et al. 2001. 'The Sequence of the Human Genome'. *Science* 291(5507): 1304–51. doi:10.1126/science.1058040.

Vierstra, Jeff, John Lazar, Richard Sandstrom, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, et al. 2020. 'Global Reference Mapping of Human Transcription Factor Footprints'. *Nature* 583(7818): 729–36. doi:10.1038/s41586-020-2528-x.

Wang, Hua, Zheng Fan, Pavel V. Shliaha, Matthew Miele, Ronald C. Hendrickson, Xuejun Jiang, and Kristian Helin. 2023. 'H3K4me3 Regulates RNA Polymerase II Promoter-Proximal Pause-Release'. *Nature* 615(7951): 339–48. doi:10.1038/s41586-023-05780-8.

Wang, Yunhao, Anqi Wang, Zujun Liu, Andrew L. Thurman, Linda S. Powers, Meng Zou, Yue Zhao, et al. 2019. 'Single-Molecule Long-Read Sequencing Reveals the Chromatin Basis of Gene Expression'. *Genome Research* 29(8): 1329–42. doi:10.1101/gr.251116.119.

Wapinski, Orly L., Qian Yi Lee, Albert C. Chen, Rui Li, M. Ryan Corces, Cheen Euong Ang, Barbara Treutlein, et al. 2017. 'Rapid Chromatin Switch in the Direct Reprogramming of Fibroblasts to Neurons'. *Cell Reports* 20(13): 3236–47. doi:10.1016/j.celrep.2017.09.011.

Wasserman, Wyeth W., and Albin Sandelin. 2004. 'Applied Bioinformatics for the Identification of Regulatory Elements'. *Nature Reviews Genetics* 5(4): 276–87. doi:10.1038/nrg1315.

Weingarten-Gabbay, Shira, and Eran Segal. 2014. 'The Grammar of Transcriptional Regulation'. *Human Genetics* 133(6): 701–11. doi:10.1007/s00439-013-1413-1.

Weirather, Jason L, Mariateresa De Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. 2017. 'Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis'. *F1000Research* 6: 100. doi:10.12688/f1000research.10571.2.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. 'Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity'. *Cell* 158(6): 1431–43. doi:10.1016/j.cell.2014.08.009.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. 'Welcome to the Tidyverse'. *Journal of Open Source Software* 4(43): 1686. doi:10.21105/joss.01686.

Wong, Emily S., David Thybert, Bianca M. Schmitt, Klara Stefflova, Duncan T. Odom, and Paul Flicek. 2015. 'Decoupling of Evolutionary Changes in Transcription Factor Binding and Gene Expression in Mammals'. *Genome Research* 25(2): 167–78. doi:10.1101/gr.177840.114.

Worsley Hunt, Rebecca, and Wyeth W Wasserman. 2014. 'Non-Targeted Transcription Factors Motifs Are a Systemic Component of ChIP-Seq Datasets'. *Genome Biology* 15(7): 412. doi:10.1186/s13059-014-0412-4.

Wunderlich, Zeba, and Leonid A. Mirny. 2009. 'Different Gene Regulation Strategies Revealed by Analysis of Binding Motifs'. *Trends in Genetics* 25(10): 434–40. doi:10.1016/j.tig.2009.08.003.

Xie, Shiqi, Jialei Duan, Boxun Li, Pei Zhou, and Gary C. Hon. 2017. 'Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells'. *Molecular Cell* 66(2): 285-299.e5. doi:10.1016/j.molcel.2017.03.007.

Xu, Jin, Ava C Carter, Anne-Valerie Gendrel, Mikael Attia, Joshua Loftus, William J Greenleaf, Robert Tibshirani, Edith Heard, and Howard Y Chang. 2017. 'Landscape of Monoallelic DNA Accessibility in Mouse Embryonic Stem Cells and Neural Progenitor Cells'. *Nature Genetics* 49(3): 377–86. doi:10.1038/ng.3769.

Yang, Marty G, Emi Ling, Christopher J Cowley, Michael E Greenberg, and Thomas Vierbuchen. 2022. 'Characterization of Sequence Determinants of Enhancer Function Using Natural Genetic Variation'. *eLife* 11: e76500. doi:10.7554/eLife.76500.

Yoo, Andy B., Morris A. Jette, and Mark Grondona. 2003. 'SLURM: Simple Linux Utility for Resource Management'. In *Job Scheduling Strategies for Parallel Processing*, Lecture Notes in Computer Science, eds. Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn. Berlin, Heidelberg: Springer Berlin Heidelberg, 44–60. doi:10.1007/10968987_3.

Yu, Xinyang, and Michael J. Buck. 2019. 'Defining TP53 Pioneering Capabilities with Competitive Nucleosome Binding Assays'. *Genome Research* 29(1): 107–15. doi:10.1101/gr.234104.117.

Zaret, Kenneth S. 2020. 'Pioneer Transcription Factors Initiating Gene Network Changes'. *Annual Review of Genetics* 54(1): 367–85. doi:10.1146/annurev-genet-030220-015007.

Zaret, Kenneth S., and Jason S. Carroll. 2011. 'Pioneer Transcription Factors: Establishing Competence for Gene Expression'. *Genes & Development* 25(21): 2227–41. doi:10.1101/gad.176826.111.

Zhan, Yumeng, Frauke Grabbe, Elisa Oberbeckmann, Christian Dienemann, and Patrick Cramer. 2023. *Three-Step Mechanism of Promoter Escape by RNA Polymerase II*. Biochemistry. preprint. doi:10.1101/2023.12.22.572998.

Zhang, Tiantian, Zhuqiang Zhang, Qiang Dong, Jun Xiong, and Bing Zhu. 2020. 'Histone H3K27 Acetylation Is Dispensable for Enhancer Activity in Mouse Embryonic Stem Cells'. *Genome Biology* 21(1): 45. doi:10.1186/s13059-020-01957-w.

Zheng, An, Michael Lamkin, Hanqing Zhao, Cynthia Wu, Hao Su, and Melissa Gymrek. 2021. 'Deep Neural Networks Identify Sequence Context Features Predictive of Transcription Factor Binding'. *Nature Machine Intelligence* 3(2): 172–80. doi:10.1038/s42256-020-00282-y.

Zhou, Tianyin, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. 2013. 'DNAshape: A Method for the High-Throughput Prediction of DNA Structural Features on a Genomic Scale'. *Nucleic Acids Research* 41(W1): W56–62. doi:10.1093/nar/gkt437.

Zhu, Fangjie, Lucas Farnung, Eevi Kaasinen, Biswajyoti Sahu, Yimeng Yin, Bei Wei, Svetlana O. Dodonova, et al. 2018a. 'The Interaction Landscape between Transcription Factors and the Nucleosome'. *Nature* 562(7725): 76–81. doi:10.1038/s41586-018-0549-5.

Zhu, Fangjie, Lucas Farnung, Eevi Kaasinen, Biswajyoti Sahu, Yimeng Yin, Bei Wei, Svetlana O. Dodonova, et al. 2018b. 'The Interaction Landscape between Transcription Factors and the Nucleosome'. *Nature* 562(7725): 76–81. doi:10.1038/s41586-018-0549-5.