Development of bioinformatics methods to investigate characteristics of transcription factor binding during early embryonic development and cell differentiation

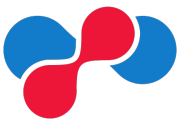
Cumulative inaugural dissertation
in partial fulfillment of the requirements for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

By
Mette Skou Bentsen

Submitted
May, 2023

Prepared in the:
Department of Cardiac Development and Remodelling
Max Planck Institute for Heart and Lung Research
Bad Nauheim, Germany

Submitted to the:
Faculty of Biology and Chemistry
Justus Liebig University Giessen
Giessen, Germany



Max Planck Institute
for Heart and Lung Research
W.G. Kerckhoff Institute

MAX-PLANCK-GESELLSCHAFT

JUSTUS-LIEBIG-
UNIVERSITÄT
GIESSEN

Preface

Thesis reviewers

- **First reviewer**
  Prof. Dr. Dr. habil. Thomas Braun
  Department of Biology and Chemistry, Justus Liebig University Giessen
  & Department of Cardiac Development and Remodelling, Max Planck Institute for Heart and Lung Research
- **Second reviewer**
  Prof. Dr. Mario Looso
  Department of Life Science Engineering, University of Applied Sciences Mittelhessen
  & Bioinformatics Core Unit, Max Planck Institute for Heart and Lung Research
- **Examiner**
  Prof. Dr. Stefan Janssen
  Algorithmic Bioinformatics, Justus Liebig University Giessen
- **Examiner**
  Prof. Dr. Michael Kracht
  Rudolf Buchheim Institute of Pharmacology, Justus Liebig University Giessen

Declaration

I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and cited all text passages that are derived verbatim from or are based on the content of published work of others, and all information relating to verbal communications. I consent to the use of an anti-plagiarism software to check my thesis. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ in carrying out the investigations described in the dissertation.

Bad Nauheim, May, 2023

Mette Skou Bentsen

Abstract

Epigenetic mechanisms drive determination of cell types from embryonic development to differentiated adult tissues. Transcription factors are particularly important for these processes due to their ability to bind specific DNA loci associated with the regulation of target gene expression, as well as their potential to modulate the global chromatin structure. In order to do this, transcription factors often interact with other factors and exhibit a certain binding grammar. However, due to both experimental and computational limitations, it is challenging to study the global characteristics of transcription factor binding. Especially at early developmental timepoints, where the input material is sparse, many aspects of transcription factor binding remain obscure. The objective of this thesis is to develop computational methods to study the effect of transcription factor binding during developmental processes.

This thesis presents the description of two bioinformatics tools suitable for identifying individual transcription factor binding sites and characterizing the grammar of co-occurring binding events. The first tool, named TOBIAS, is able to identify transcription factor binding using the method of ATAC-seq footprinting. Using this tool, it was possible to create a sequential map of transcription factor activity throughout the early cell divisions of human and mouse embryos. The second tool, TF-COMB, is a method to perform genome-wide transcription factor association analysis, which can utilize binding sites from TOBIAS or other methods. Investigations of both experimental and *in silico* data across multiple cell types showed that transcription factors bind DNA in the vicinity of other DNA-binding proteins, bind at sites labeled by specific chromatin marks and exhibit preferred binding conformations.

In conclusion, these two tools, published in two separate papers, provide a significant improvement to the existing bioinformatics methods for studying global transcription factor binding characteristics in the context of differentiation.

Zusammenfassung

Epigenetische Mechanismen steuern die Bestimmung von Zelltypen von der Embryonalentwicklung bis zum differenzierten Geweben im Erwachsenenalter. Transkriptionsfaktoren sind für diese Prozesse besonders wichtig, da sie in der Lage sind, an spezifische DNA-Loci zu binden, welche mit der Regulierung der Expression von Zielgenen verbunden sind. Zusätzlich können Transkriptionsfaktoren die globale Chromatinstruktur beeinflussen. Um dies zu erreichen, interagieren Transkriptionsfaktoren häufig mit anderen Faktoren und weisen eine bestimmte Bindungsgrammatik auf. Aufgrund Einschränkungen sowohl in der experimentellen als auch in der computergestützten Analyse ist es jedoch eine Herausforderung, die globalen Merkmale der Bindung von Transkriptionsfaktoren zu untersuchen. Vor allem zu frühen Entwicklungsstadien, wenn nur wenig Probenmaterial verfügbar ist, bleiben viele Aspekte der Bindung von Transkriptionsfaktoren unklar. Das Ziel dieser Arbeit ist es rechnerische Methoden zu entwickeln, um die Auswirkungen der Bindung von Transkriptionsfaktoren während der Entwicklungsprozesse zu untersuchen.

In diesem Zusammenhang werden zwei Bioinformatik-Tools beschrieben, die für die Identifizierung einzelner Transkriptionsfaktor-Bindungsstellen und die Charakterisierung der Grammatik von gemeinsam auftretenden Bindungsereignissen geeignet sind. Das erste Tool namens TOBIAS ist in der Lage, die Bindung von Transkriptionsfaktoren mit der Methode des ATAC-seq-Footprinting zu identifizieren. Hiermit konnte eine sequenzielle Abbildung der Aktivität von Transkriptionsfaktoren während der frühen Zellteilungen von Mensch- und Mausembryonen erstellt werden. Das zweite Tool, TF-COMB, ist eine Methode zur Durchführung genomweiter TF-Assoziationsanalysen, bei der Bindungsstellen aus TOBIAS oder anderen Methoden verwendet werden können. Untersuchungen von experimentellen und *in silico*-Daten über mehrere Zelltypen zeigten, dass Transkriptionsfaktoren DNA in der Nähe anderer DNA-bindender Proteine binden sowie an Stellen, die durch spezifische Chromatinmarkierungen gekennzeichnet sind. Außerdem wurde gezeigt, dass Transkriptionsfaktoren bevorzugte Bindungskonformationen aufweisen.

Zusammenfassend stellen diese beiden Tools, welche in zwei separaten Artikeln veröffentlicht wurden, eine signifikante Verbesserung der bestehenden bioinformatischen Methoden zur Untersuchung der globalen Bindungseigenschaften von Transkriptionsfaktoren im Kontext der Differenzierung bereit.

Contents

List of Figures

Figure statement

Figures 2-8 were created using the online scientific drawing tool BioRender.com.

Abbreviations

1/2/4/8C	1/2/4/8-cell embryonic stage
2CLC	2C-like cells
ATAC-seq	Assay for transposase-accessible chromatin with high-throughput sequencing
DWM	Di-nucleotide weight matrix
FAIR	Findability, accessibility, interoperability and reusability
FAIR4RS	FAIR principles for research software
FSHD	Facioscapulohumeral dystrophy
GTF	General transcription factor
GWAS	Genome-wide association study
HAT	Histone acetylase
HDAC	Histone deacetylase
hESC	Human embryonic stem cells
HP1	Heterochromatin protein 1
HT-SELEX	High throughput systematic evolution of ligands by exponential enrichment
ICM	Inner cell mass
IHEC	International Human Epigenome Consortium
iPSC	Induced pluripotent stem cells
LTR	Long terminal repeat
mESC	Mouse embryonic stem cells
OSKM	Oct4, Sox2, Klf4 & c-Myc
PIC	Transcriptional preinitiation complex
PolII	RNA polymerase II
PWM	Position weight matrix
RMS	Rhabdomyosarcoma
SCMC	Subcortical maternal complex
SNP	Single nucleotide polymorphism
TAD	Trans-activation domain (or) Topologically associating domain
TE	Transposable element
TF	Transcription factor
TF-COMB	Transcription factor co-occurrence using market basket analysis
TFBS	Transcription factor binding site
Tn5	Tn5 transposase
TOBIAS	Transcription factor occupancy prediction by investigation of ATAC-seq signal
TSS	Transcription start site
ZGA	Zygotic genome activation

1 Introduction

1.1 Motivation

The human body is estimated to consist of an astonishing ~30 trillion cells (Sender et al., 2016), but most remarkably, these cells all arose from one single fertilized egg cell, the zygote. Throughout the zygote's cell divisions to 2 cells, 4 cells, 8 cells, until 30 trillion cells, most cells of an organism contain the same genetic material, but still manage to modify their behavior to obtain unique properties. Consequently, the ability of cells to differentiate into individual cell lineages ensures the correct development of organs, limbs and highly specialized tissues, eventually achieving the staggering complexity of the human body. However, the question arises: How do cells control which cell type is to be chosen and maintained? And which biological processes provide the ability to make the choice?

The study of these mechanisms was first termed "epigenetics" by Conrad Waddington in 1942 (Waddington, 1942). He postulated that there must be an *epigenotype* which represents the developmental processes between the genotype and the observed phenotype of a cell. However, genome sequencing was not available at the time of Waddington's research. Consequently, when technologies to map the DNA sequence of genomes were developed, our understanding of the human genome was significantly extended. In this context, the Human Genome Project provided the first reference genome (Venter et al., 2001), which estimated that the human genome contains ~25,000 protein coding genes, but that the coding sequences comprise only ~1% of all bases. Building upon these fundamental findings, the 1000 Genomes Project extended our understanding of genetic variation between individuals by presenting 2,500 individual human genomes (The 1000 Genomes Project Consortium, 2015). Interestingly, this study showed that the majority of genetic variation occurs in the non-coding regions. Likewise, genome-wide association studies (GWAS), which link phenotype with genetic variation, have shown that most genetic diseases are associated with single nuclear polymorphisms (SNPs) in non-coding sequences (Maurano et al., 2012). In conclusion, these studies provided evidence that non-coding sequences are not just junk, but play a paramount role in the regulation of cellular processes.

To understand the influence of non-coding sequences, initiatives like the Roadmap Epigenomics Consortium went on to study epigenetic mechanisms such as histone modifications and DNA methylation (Roadmap Epigenomics Consortium et al., 2015). Through an integrated analysis of epigenetic marks for 111 epigenomes, the study found that epigenetic profiles are highly correlated with known cell types and underlying lineages. In particular, these epigenetic marks were shown to play an important role in establishing regions of open chromatin, which can act as *enhancers* of gene expression

through binding of the transcriptional machinery. In this context, it is noteworthy that 93% of the known GWAS SNPs in open chromatin were reported to overlap with binding sites for transcription factors (TFs) (Maurano et al., 2012). This emphasizes the role of TFs to read the non-coding genome code in order to influence transcription of target genes. Thus, through binding to DNA, TFs provide an interface between the static genome sequence and the phenotype as described by Waddington.

Despite having completed the human genome sequence, many aspects of epigenetics are still not fully understood. In particular, the mechanisms utilized by cells to switch between different epigenetic states, and how the genome sequence encodes the differentiation of tissues, are still under intense investigation. It is clear that the genome can encode different functionalities dependent on the circumstance, but we are still learning how to correctly read this information within the sequence (Figure 1). In this context, improving our knowledge of how TFs and other transcriptional regulators decode the genetic code to fulfill their functions is an integral part of understanding epigenetic regulation.

This thesis and its accompanying publications aim to contribute to the understanding of TF binding dynamics and how TFs shape decisions of cell fates throughout development. It is focused on the design, implementation and application of computational methods to analyze high-throughput next-generation sequencing data, and to integrate various levels of epigenetic information such as gene expression, chromatin accessibility and DNA-protein interactions to study TF binding.

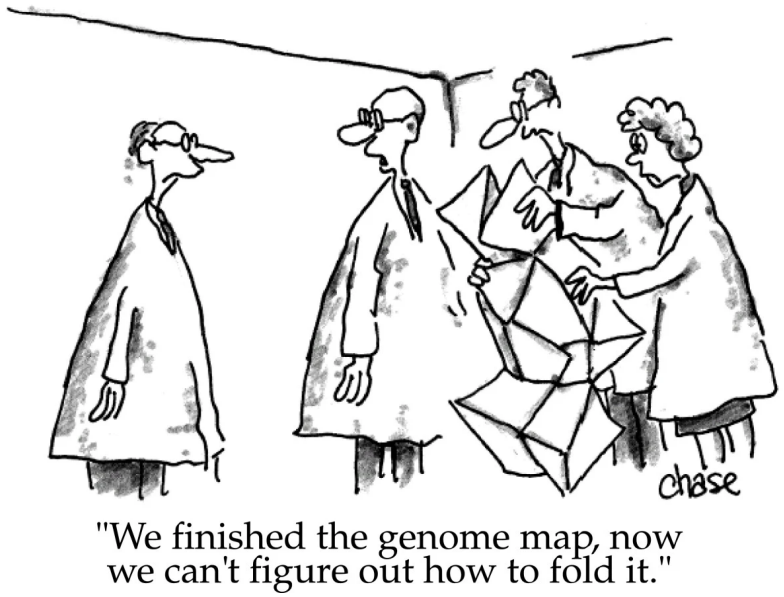


Figure 1: Interpretation of the genome map is a challenge.
While the sequence of the human genome has been uncovered, we still need to 'fold' the genome map to fully understand its meaning in different epigenetic contexts. Figure by Iyer et al., 2011 adapted from an illustration by cartoonist John Chase.

1.2 Background

In context of the concepts presented in the motivation, this section provides background on how epigenetic mechanisms shape cell fate decisions, and how the interplay of chromatin modifications and TF binding plays a role in embryonic development. Finally, the application of high-throughput sequencing and bioinformatics analysis as methods to study these mechanisms is discussed.

1.2.1 Epigenetic mechanisms of gene regulation

A major aim of epigenetic processes is to control gene expression and to ensure that the appropriate genes are transcribed at the right time. On a molecular level, this control is achieved using a variety of mechanisms which affect the level of chromatin accessibility and 3D organization, and thereby also influence binding of proteins to the regulatory regions. Some of the major mechanisms contributing to this system are introduced in Figure 2, and will be described in more detail here.

At the highest level, DNA is organized into chromatin, which is a complex of DNA and histone protein complexes, known as nucleosomes, ultimately forming a chromosome. The degree of chromatin condensation is controlled by the composition, positioning and organization of these nucleosomes, which each wrap 147 base pairs of DNA (Klemm et al., 2019). Depending on the positioning, nucleosomes can either pack tightly together (a state known as *heterochromatin*) or position themselves with nucleosome-free linker regions (known as *euchromatin*) (Figure 2A). Steric hindrance of heterochromatin generally prevents binding of transcriptional machinery to the DNA, whereas the loose conformation of euchromatin is associated with transcriptional activity. On average, open chromatin regions span ~2% of the genome (Maurano et al., 2012). By controlling chromatin accessibility, nucleosome organization is thereby indirectly influencing the potential for gene activation at a certain timepoint.

Nucleosomes consist of octamers of histones H2A, H2B, H3 and H4, as well as the linker histone H1, and different variants of these histones influence nucleosome positioning (Mariño-Ramírez et al., 2005) (Figure 2B). One example is H2A.Z, a variant of H2A, which is enriched in regulatory regions. During development, the integration of H2A.Z makes nucleosomes less stable and easier to evict from DNA, ultimately allowing dynamic changes in gene expression (Klemm et al., 2019). Nucleosome positioning is also influenced by post-translational modification of the histone amino acid chains. One such example is histone methylation, which is deposited by methyltransferases and, depending on the amino acid being methylated, has different effects on gene expression. For example, H3K4me3 is correlated with active transcription, whereas H3K9me3 is associated with repressed chromatin (Nicetto et al., 2019). Mechanistically, these marks can be recognized by protein machinery such as

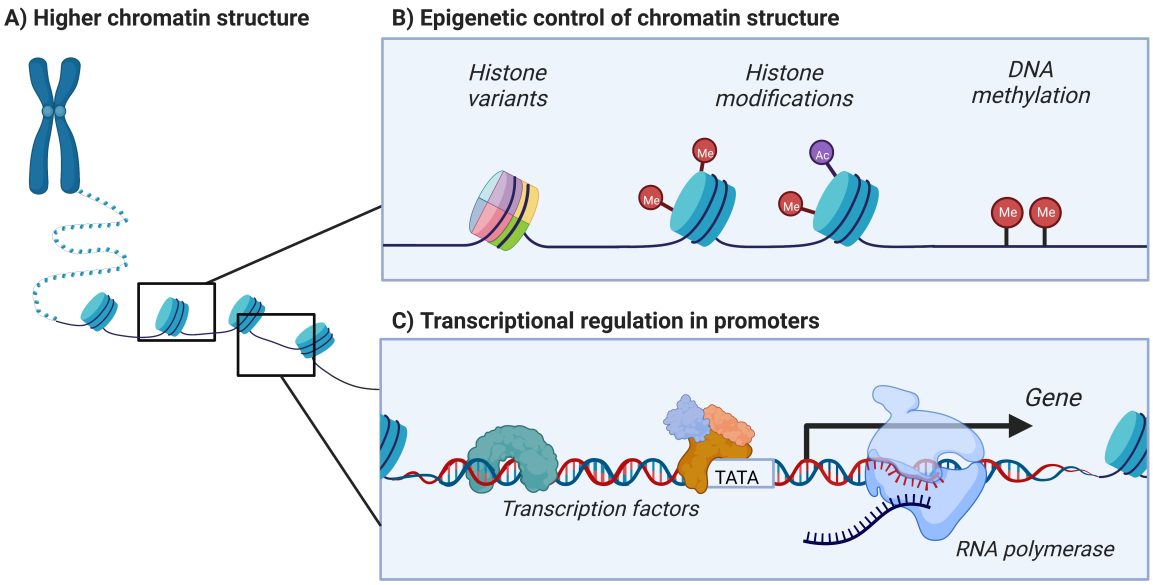


Figure 2: Mechanisms of epigenetic gene regulation.
A) DNA is wrapped around nucleosomes, which can be tightly bound or spaced further apart creating nucleosome free regions, also known as open chromatin. B) A variety of epigenetic marks influence the positioning of nucleosomes and thus the structure of chromatin. C) Within open chromatin, transcription factors and transcriptional machinery bind to DNA to activate expression of nearby genes by RNA polymerase.

heterochromatin protein 1 (HP1). HP1 interacts with H3K9me3 through its N-terminal domains, and forms a dimer with itself through the C-terminal domain, which strengthens the compaction of chromatin (Bannister et al., 2011). Additionally, HP1 interacts with the SUV39 methyltransferase, which reinforces methylation of H3K9 in nearby nucleosomes (Bannister et al., 2011).

Besides methylation of histones, methylation of cytosine residues in the CpG dinucleotides (known as DNA methylation) helps to establish gene repression (Cedar et al., 2009). There is also evidence that DNA methylation is not only a parallel mechanism, but is also linked with the deposition of histone modifications. For example, G9a, an H3K9 methyltransferase, is known to recruit HP1, but also recruits the DNA methyltransferase enzymes DNMT3A and DNMT3B, which finalize the silencing of genes (Cedar et al., 2009). To that effect, the type of histone methylation also helps create a distinction between facultative and constitutive heterochromatin regions. Facultative chromatin is enriched for H3K27me3 and contains genes that are differentially expressed depending on the developmental cue (Saksouk et al., 2015). In comparison, constitutive chromatin is characterized by enrichment of H3K9me3 and DNA methylation, and mainly locates at the pericentromeric and telomeric regions (Saksouk et al., 2015). Here, these marks help to maintain integrity of the chromatin, prevent spurious expression of satellite repeats, and facilitate successful cohesion of chromosomes during mitosis (Saksouk et al., 2015).

In contrast to repressive modifications, histone acetylation serves as a mechanism to enhance transcription. Mechanistically, histone acetylation is achieved by histone acetyltransferases (HATs) transferring an acetyl group to lysine amino acids (Bannister et al., 2011). In the case of H4K5ac, the modification causes a shift in charge which weakens the interaction between the histone and DNA, making the DNA more accessible to DNA binding proteins such as TFs (Bannister et al., 2011). As a consequence of this mechanism, histone acetylation is enriched in gene promoters and other regulatory regions. Opposite of HATs, histone deacetylases (HDACs) remove acetylation, thereby repressing chromatin accessibility. In fact, establishment of heterochromatin is dependent on histone deacetylases such as SIRT1-7 (Saksouk et al., 2015). In conclusion, histone and DNA modifications, as well as the removal of these, are mechanisms of controlling DNA accessibility.

As a result of dynamic chromatin structure, gene regulation is possible by the RNA polymerase II (polII) binding to DNA in the nucleosome free region upstream of the transcription start site (TSS) of genes. As polII is unable to recognize and bind to the DNA by itself, it utilizes the transcriptional preinitiation complex (PIC), which consists of multiple proteins including general transcription factors (GTFs) (Figure 2C). The GTFs are able to bind directly to specific promoter elements, such as seen for TATA binding protein (TBP), which recognizes the TATA-box commonly located 25-30bp upstream of the TSS (Thomas et al., 2006). By interacting directly with the PIC, the GTFs help to anchor polII at the correct position relative to the TSS, thereby enhancing transcription (Thomas et al., 2006). Besides the effects of GTFs, the stability of the PIC is also influenced by sequence-specific TFs (from now on referred to simply as TFs) both located within promoters and in distal enhancer elements (described in more detail in Section 1.2.4). These TFs are not part of the PIC directly, but are able to bind DNA to attract polII and support PIC assembly. The TFs therefore play a major role in linking sequence specificity to targeted gene expression of particular genes.

In conclusion, positioning of nucleosomes is an important mechanism for defining the regulatory state of regions in the genome. Through plastic chromatin structure, cells can control the access of transcriptional machinery and TFs to genes, ultimately regulating transcription itself. How these mechanisms govern the decisions of cell fates will be covered in the following section.

1.2.2 Differentiation and lineage specification

Specialization of cells is a prerequisite for obtaining the individual tissues needed for a complex organism. The process in which cells diverge into distinct cell types and lineages is called *differentiation*. Equipped with identical genetic backgrounds, each cell must translate the genetic code into blueprints for cell progression and final cell fate decision using the epigenetic mechanisms mentioned in Section 1.2.1. It has been estimated that the adult human body contains ~200 differentiated cell

types (Hatano et al., 2011). However, recent advances in technologies enabling the characterization of single cells have uncovered large-scale heterogeneity within cell populations, which complicates the quantification of terminally differentiated cell types (Roy et al., 2018; Goldman et al., 2019). This section will review the dynamics by which cell fate is decided for each progenitor cell.

The classic view of differentiation as an epigenetic landscape was outlined by Waddington (Waddington, 1957), building on his initial work on epigenetic mechanisms. His concept can be seen as a gravity-driven model in which cells are rolling through a landscape consisting of valleys and paths, ultimately dividing cells into individual subtypes at the bottom of the landscape (Figure 3A). In this model, the ridges and valleys represent underlying epigenetic mechanisms such as deposition of histone marks and binding of TFs to drive gene expression networks. At branching-points in the network, the relative height of nearby ridges decides the path of a cell. As cells progress through the landscape from top to bottom, their developmental potential decreases from totipotent (in the early embryo) to fully differentiated cells with limited potential for further progression (Melcer et al., 2010) (Figure 3B).

In order to secure final lineage commitment, dynamics of the epigenetic landscape generally ensure that the valleys beyond the branching-points are lower than the originating cell population, thus making it difficult for the cells to revert back to previous developmental stages (Moris et al.,

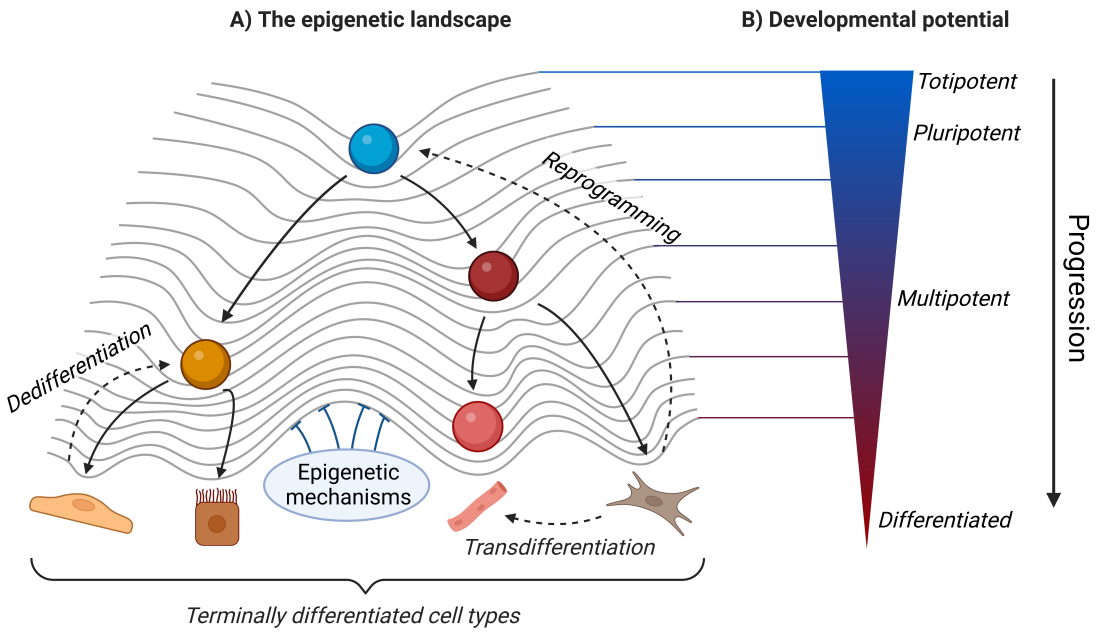


Figure 3: Dynamics of cell differentiation and reprogramming.
A) Waddington's epigenetic landscape. Progenitor cells are depicted as spheres, which roll through the epigenetic landscape towards a terminally differentiated cell type. Cells can convert their fate by methods of transdifferentiation, dedifferentiation or reprogramming as highlighted with dashed arrows. B) The developmental potential of cells decreases as they progress through the epigenetic landscape towards differentiated cell types.

2016). However, there are scenarios where cells can reverse back up the epigenetic hill to regain a larger potential for differentiation - a process known as *dedifferentiation* (Jopling et al., 2011). This phenomenon is particularly observed in amphibians, such as salamanders, which are well known for their unique ability to regenerate a wide variety of complex tissues, organs, and even whole limbs following amputation (Godwin et al., 2014). Using a combination of lineage tracing and single cell sequencing methods in the axolotl (*Ambystoma mexicanum*), Gerber et al., 2018 have shown that connective tissue cells near the amputation plane undergo dedifferentiation into a homogeneous population known as the blastema. During regeneration, these blastema cells lose their original adult phenotypes, and are converted into a multipotent progenitor state, similar to embryonic limb bud development. A marker for this state is the TF Prrx1, which is expressed in connective tissue of both the developing limb bud and the blastema after injury (Gerber et al., 2018). Finally, after dedifferentiation, these cells start re-expressing tissue markers to give rise to cell types including non-skeletal, cartilage and bone cells in the regenerating limb (Gerber et al., 2018). A similar process is observed during heart regeneration in zebrafish (*Danio rerio*), which involves partial dedifferentiation of cardiomyocytes and their subsequent reentry into the cell-cycle (Jopling et al., 2010). Using these mechanisms, zebrafish are able to fully regenerate a ventricle after amputation of up to 20% of the tissue (Jopling et al., 2010). Remarkably, in both axolotl and zebrafish, the blastema cells undergo a transient loss of phenotype, but still retain epigenetic memory of the original tissue and location of the amputation plane, ultimately preventing a tail from growing where a limb was amputated (Jopling et al., 2011).

In contrast, mammals show limited ability to regenerate tissue without scar formation, although the initial response to injury seems to promote regeneration. For example, damage to the adult human heart enables cardiomyocytes to partially dedifferentiate by shifting their transcriptional program to an immature state similar to that of fetal hearts. This shift includes re-expression of fetal genes such as α-smooth muscle actin, early cardiac TFs Gata4 and Nkx2.5, as well as the stem cell marker *Runx1* (Kubin et al., 2011; Zhang et al., 2010). However, while the initial process of dedifferentiation protects the heart by increasing resistance to stress and hypoxia, prolonged dedifferentiation leads to loss of sarcomeric structures and contractile force, which makes the human heart unable to regenerate after injury (Kubin et al., 2011). Interestingly, studies of porcine heart and mouse hearts show that neonatal hearts can regenerate following injury, but that this ability is lost just days after birth (Ye et al., 2018; Porrello et al., 2011). The reason for this loss of regenerative capacity, and the differences to organisms such as axolotl, is under heavy investigation and debate. For example, it has been shown that inactivation of *Rb* (retinoblastoma protein) and *Arf*, a known tumor suppressor which is not present in regenerating vertebrates, enables muscle cells to re-enter the cell cycle (Pajcini et al., 2010). Some studies also suggest an involvement of immune reactions, as the maturity an organism's immune system is inversely correlated with the capacity to

regenerate (Godwin et al., 2014). The continual research into the barriers of regeneration through dedifferentiation holds great promise for regenerative medicine (Yao, 2020).

While there is limited evidence of successful mammalian dedifferentiation, tissue damage can trigger other mechanisms for repair. One of these is the transformation of cells between terminally differentiated cell types, which is known as *transdifferentiation*. In the mouse liver, transdifferentiation is observed after toxin-mediated injury, as remaining hepatocytes switch fate to become biliary epithelial cells (Yanger et al., 2013). Throughout the slow conversion over several weeks, the cells co-express both the hepatocyte marker *HNF4α* and the biliary marker *Sox9*, suggesting that the cells pass through an intermediate state (Yanger et al., 2013). However, this state lacks markers for hepatocyte progenitor cells, which suggests that the conversion is horizontal in the epigenetic landscape, and does not require a dedifferentiation step. Another example of transdifferentiation is observed in the pancreas after loss of the β-cell population. A study in mice showed that diphtheria toxin-induced ablation of >99% of adult β-cells can induce proliferation of new β-cells, which resulted from changes in gene expression within α-cells, and not from proliferation of residual β-cells (Thorel et al., 2010). The authors propose that the loss of β-cells induces upregulation of TFs such as Pdx1 and Nkx6.1 within α-cells, which in turn activate expression of β-cell specific markers including insulin. Interestingly, this process differs in juvenile mice, where β-cell recovery is achieved by dedifferentiation of somatostatin-producing δ-cells (Chera et al., 2014). These observations are comparable to the loss of dedifferentiation potential of cardiomyocytes immediately following birth. Thus, there is evidence to suggest that mammalian dedifferentiation is restricted in adult cells, and that direct transdifferentiation might be the method of choice for adult tissue regeneration (Merrell et al., 2016). However, these ideas are under heavy debate, and are convoluted by the discovery of a persistent niche of immature β-cells in the pancreas, which produce insulin but lack expression of mature marker genes such as *Ucn3* (van der Meulen et al., 2017). It is unclear whether this cell population serves as a progenitor for β-cell differentiation after injury, and how remaining cells sense loss of tissue to trigger transdifferentiation or proliferation of existing cells. More work is needed to uncover how these processes might be leveraged to recover insulin production for the treatment of diabetes (Spears et al., 2021).

In the context of using the mechanisms of regeneration for treatment of lost cell populations, *reprogramming* is arising as a method for artificially harnessing the effects of dedifferentiation to push cells back to pluripotency (Aydin et al., 2019). A well-known example of this mechanisms is the usage of four TFs, Oct4, Sox2, Klf4 and c-Myc (OSKM), to reprogram mouse embryonic stem cells (mESCs) and adult human fibroblasts to induced pluripotent stem cells (iPSCs) (Takahashi et al., 2006). In the context of heart regeneration, it was recently shown that transient expression of OSKM in cardiomyocytes induces dedifferentiation and improves the outcome of myocardial damage by proliferation of existing cardiomyocytes (Chen et al., 2021). By using tissue specific TFs, it

is also possible to reprogram cells into other cell types. For example, overexpression of the TFs Gata4, Hand2, Mef2c, and Tbx5 can reprogram fibroblasts into pacemaker-like myocytes in mice (Fernandez-Perez et al., 2019). By using a different set of TFs, namely SOX10, OLIG2 and NKX6.2, Chanoumidou et al., 2021 were able to convert human fibroblasts into oligodendrocyte-like cells. While these TF-cocktails show immense potential for use in human medicine, there is also evidence that the molecular signals driving cellular reprogramming might make cells more prone to cancer, as for example seen by increased NOTCH signaling in human liver cancers (Sekiya et al., 2012). Thus, more research is needed in order to balance the effects of reprogramming with the risk of oncogenic transformation (Merrell et al., 2016)

In many of these examples, TFs play a major role in lineage specification, both for innate differentiation (downhill in the Waddington landscape) but also for efforts to dedifferentiate (uphill towards more immature cells) or reprogram (sideways across ridges and valleys) cells. Therefore, the next section will be focused on the mechanisms and abilities of TFs to control gene expression.

1.2.3 How do transcription factors target specific genes?

TFs are proteins capable of regulating the expression of genes. However, these proteins differ from other epigenetic mechanisms, such as histone modifiers, being able to recognize DNA in a sequence-specific manner (Lambert et al., 2018). To understand the influence of TFs in epigenetic regulation, this section will cover the mechanisms of how TFs recognize and bind to their target regions.

Recognition of target sites is in part elicited by the DNA-binding domains of TFs. While there are ~1600 known TFs in humans, these are comprised of only ~100 binding domain families, some of the most common being C2H2 zinc fingers, homeodomain and helix–loop–helix factors (Lambert et al., 2018; Vaquerizas et al., 2009). Some of these DNA-binding domains encode for a specific sequence preference, which is known as the TF binding motif (Figure 4A). Because the target DNA sequence is restricted by the particular fold of the DNA-binding domain, many family-members bind to similar motifs (Sandelin et al., 2004). Such redundancy within TF families can provide functional robustness in case mutations arise in essential TFs. In fact, a study in yeast showed that knockouts of individual TFs were largely compensated by the presence of other TFs with similar functional annotations (Wu et al., 2015). Such overlapping functionality has also been observed in mouse for the Krüppel-like TFs Klf2, Klf4 and Klf5, as loss of self-renewal properties within mESCs is evident only after *Klf2/4/5* triple-knockout, but not after individual single-knockout events (Yamane et al., 2018).

However, the presence of a TF target sequence is not always sufficient to enable TF binding, as the majority of sequence-predicted transcription factor binding sites (TFBS) are not functional

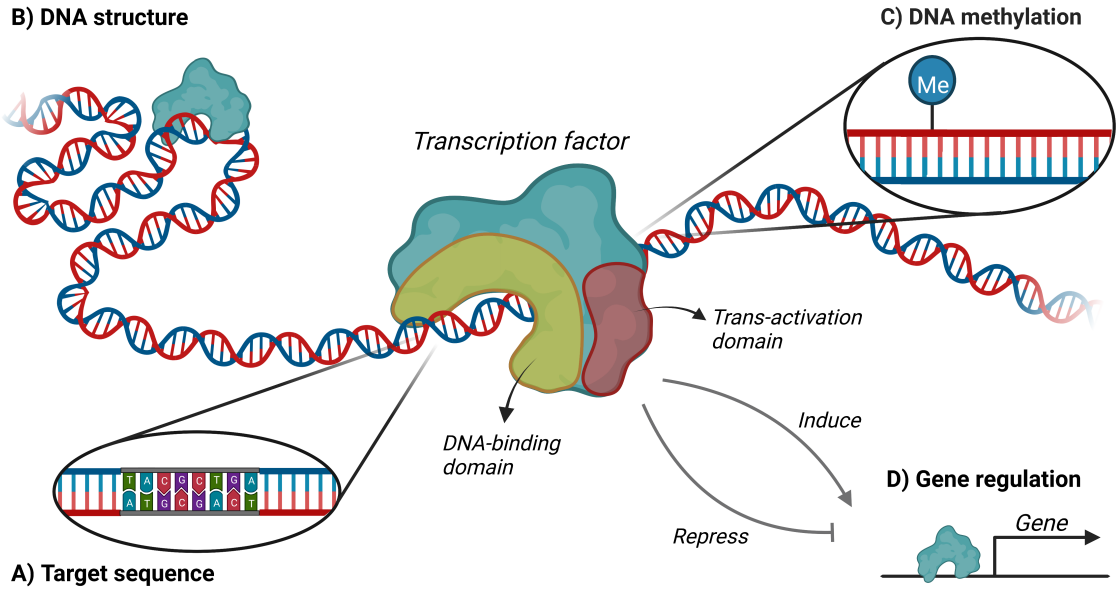


Figure 4: Influences and effects of transcription factor binding.
A) TFs bind to target sequences as defined by the DNA-binding domain. B) DNA shape (represented by looping of DNA) also affects binding affinity. C) Methylation of cytosine in target DNA is targeted differentially by individual TFs. D) The trans-activation domain of TFs can act to either induce or repress target gene expression.

in vivo (Wasserman et al., 2004). This indicates that there are additional factors influencing the binding of TFs to their target sequences. A study by Dror et al., 2015 highlighted that TFs recognize sequence features extending beyond the consensus motif, and that these help to distinguish occupied TFBS from unoccupied TFBS. Such additional sequence features include the preference of GC-rich sequences by C2H2 and ETS families, whereas homodomain factors prefer AT-rich environments. These sequence features also correlate with certain DNA shape features such as major groove width, helix twist and propeller twist (Figure 4B). For example, Mathelier et al., 2016 showed that MADS-box TFs are influenced by propeller twist within their motifs. Investigation of protein crystal structures of these TFs showed that this feature helps to enhance the DNA-protein contacts. Interestingly, for E2F TFs, DNA shape features outside the motif are also important for predicting occupied TFBS (Mathelier et al., 2016).

Another mechanism for specifying TF binding beyond sequence composition is methylation of the target DNA (Figure 4C). Using HT-SELEX (High Throughput Systematic Evolution of Ligands by EXponential enrichment) with methylation sensitivity (known as methyl-SELEX), Yin et al., 2017 investigated the responsiveness of ~500 TFs to methylation of their target sequences. Although methylation is generally believed to inhibit TF binding, 34% of the investigated TFs experienced increased affinity when their target sequences were methylated. In particular, TFs of the homeodomain family, which are involved in many developmental processes, were seen to have this property. This is particularly interesting, as the methylation of DNA in the early embryo is known to be highly

dynamic, and might thereby also guide homeodomain TFs through their susceptibility for binding to methylated sites (Guo et al., 2014). In contrast, 23% of TFs, mainly those of the bHLH-, bZIP and ETS-families, which are enriched for gene ontologies related to cell differentiation, were found to be inhibited by methylation of their target sequences. DNA methylation thereby serves as an additional regulation of TF binding, particularly during embryonic development.

When TFs have recognized their target, they act to influence transcription via a number of effector domains including trans-activation domains (TADs), which can interact with and stabilize the PIC (Figure 4D). For example, the TF Sp1, known as a *transcriptional activator*, contains a glutamine rich TAD which interacts directly with TFIID, a subunit of the PIC, to enhance transcriptional activation (Frietze et al., 2011). In contrast, other TFs function to inhibit expression of genes, and are known as *repressors*. For example, the Sp1-like repressors compete for access with Sp1, thereby indirectly suppressing expression of Sp1 target genes (Thiel et al., 2004). Other repressors function through interactions with histone modifiers, such as the TF REST, which recruits HDACs, and thus causes the chromatin to obtain a more compact state (Thiel et al., 2004). Another example of interactions with histone modifiers is seen for the members of the E2F TF family. Whereas E2F1/2/3 are known to active transcription, E2F4/5 are known to be repressors (Taubert et al., 2004). In quiescent cells, E2F4 associates with the pocket proteins p107 and p130, which in turn recruit HDAC1, resulting in repression of expression (Ferreira et al., 1998). However, upon entry into the cell cycle, the E2F4 repressive complexes dissociate from chromatin, allowing binding of E2F1, which recruits the HAT Tip60, resulting in histone acetylation and subsequent activation of target genes (Taubert et al., 2004). In that capacity, TFs such as the E2Fs link location specificity with sequence-unaware histone-modifications.

In conclusion, the epigenetic context of TF binding is just as important as the presence of a suitable target sequence. In this context, TFs do not only interact with epigenetic machinery, but can also bind with other TFs through their effector domains. Such combinatorial binding of TFs will be discussed next.

1.2.4 Combinatorial TF binding regulates target gene expression

As shown in the examples of TF cocktails used for reprogramming, TFs do not only act alone, but also in cooperation with one or multiple other TFs. In fact, analysis of TFBS in eukaryotes have shown that 10-15 TFBS are required to reach the level of specificity needed to uniquely identify a target region (Wunderlich et al., 2009). Thus, combinatorial binding of TFs is a crucial factor for cells to select the correct transcriptional program.

One way in which TFs can cooperate is through physical interactions, such as TF dimerization, either with a similar molecule (homodimer) or with another TF (heterodimer) (Figure 5A). In particular, heterodimerization can act as a switch between different abilities by the use of distinct binding partners. A well-known example of this is the Myc-Mad-Max network of TFs, which controls the balance between proliferation and apoptosis (Grinberg et al., 2004). All three TFs contain a helix-loop-helix leucine zipper domain, which enables dimerization in different combinations. In this network, Myc and Mad cannot homo- or heterodimerize with each other, thereby making Max essential for their function. In fact, Max has a higher efficiency in creating heterodimers than homodimers, thus promoting interactions with other partners (Grinberg et al., 2004). While it has been found that Myc-Max and Mad-Max dimers target many of the same promoters, the functional outcome is different, as Myc-Max dimers have the ability to recruit HATs, whereas Mad-Max dimers recruit HDACs (Grinberg et al., 2004). Thereby, the balance between Mad and Myc heterodimerization controls gene expression through epigenetic regulation.

The heterodimer networks are partially achieved by regulating the expression of the TFs themselves. Investigations of TF expression patterns have uncovered that TFs fall into two subsets - those which are ubiquitously expressed across many tissues and those which exhibit a tissue specific pattern (Vaquerizas et al., 2009). This promotes a model where constitutively expressed TFs are awaiting a signal or the expression of a potential binding partner. Indeed, in the case of Max-Mad-Myc, Max is ubiquitously expressed, whereas Myc and Mad are expressed in response to signaling (Grinberg et al., 2004). As such, dimerization is also a way of buffering spurious expression of individual TFs (Spitz et al., 2012). Other methods of regulating TF binding partners include post-translational protein modifications such seen for the ISGF3 complex. In resting-state macrophages, basal expression of interferon-stimulated genes is controlled by the binding of STAT2-IRF9 complexes (Platanitis et al., 2019). However, following activation of interferon receptors, STAT1-STAT2 heterodimers are phosphorylated by the JAK kinase, which enables their translocation to the nucleus, where they bind to promoters with IRF9 in the full ISGF3 complex (STAT-STAT2-IRF9) (Platanitis et al., 2019). Thus, intracellular signaling cascades can control the usage of different protein complex subunits.

Commonly, TFs bind together in cis-regulatory regions known as *enhancers*, which regulate transcription of individual target genes dependent on the epigenetic context (Pennacchio et al., 2013). Like rules for structuring a sentence in a language, the composition, location, orientation and affinity of binding sites within enhancers is known as *enhancer grammar* (Jindal et al., 2021). When TF binding sites adopt a very strict enhancer grammar in order to bind as a complex, this is known as an *enhanceosome* (Merika et al., 2001). A well-studied case of this is the expression of the eukaryotic interferon beta (IFN-β) gene, which requires assembly of an enhanceosome containing eight TFs in a ~50 bp interval upstream of the TSS (Panne, 2008) (Figure 5B). Molecules of the TFs ATF-2, c-Jun, IRF-3, IRF-7 and NFϰB bind on overlapping binding sites as they interrogate DNA on opposite

sides of the helix, creating a continuous binding interface. Without protein-protein interactions, the binding of IRF3 and IRF7 stabilizes the DNA in a bend of 25-30°, which allows ATF-2/c-Jun heterodimer binding to a low-affinity binding site (Panne, 2008). Thus, it is the correct spacing and order of the TFBSs which conveys function to the IFN-β enhancer.

Contrary to the enhanceosome model, enhancers might also adopt a *billboard structure*, which shows little constraint on the grammar of binding (Spitz et al., 2012) (Figure 5C). Parallel to this idea, the term *collaborative competition* describes a way for TFs to increase chromatin accessibility by competing for the same site (Spitz et al., 2012). With multiple TFs sharing the same TFBS, the temporal occupancy of each site is increased, which prevents nucleosome assembly. Thus, enhancer function can also be defined by the need for certain sequences, but not necessarily require a strict order or orientation of binding sites. However, it is also likely that enhancers display a combination of the enhanceosome and billboard models. For example, Farley et al., 2016 investigated notochord activity of synthetic enhancers in the embryos of the ascidian *Ciona intestinalis*, and found that strong affinity binding sites can compensate for a lack of motif grammar. In particular, the authors found that the optimal distance between ETS and ZicL binding sites in notochord enhancers is 11bp, but that this can be increased without disruption of enhancer activity if the affinity of the ETS site is simultaneously improved. However, when optimizing both binding affinity and distance,

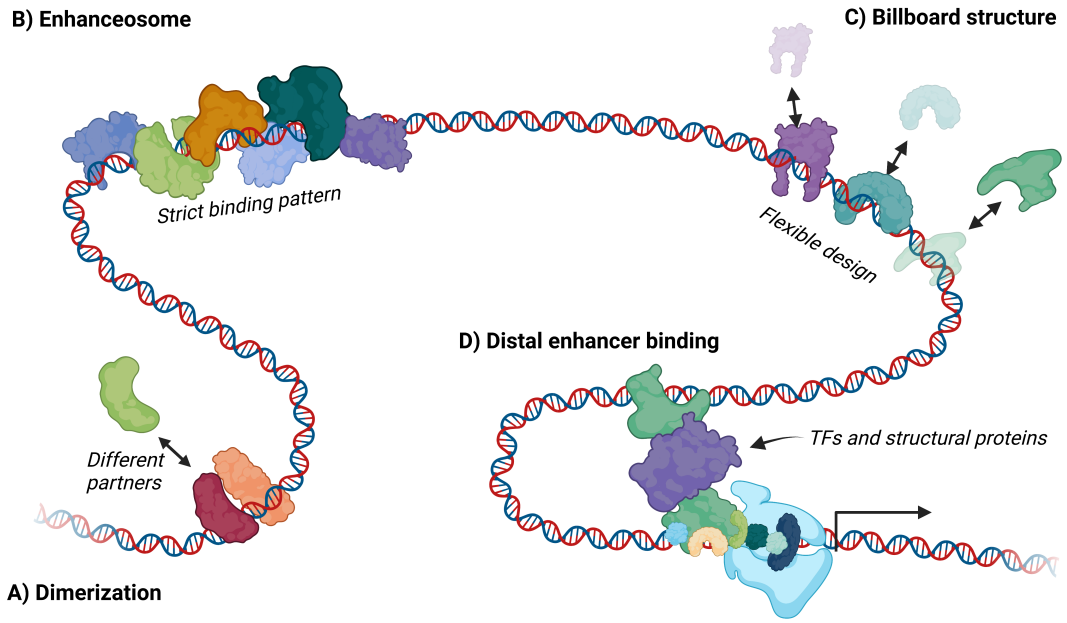


Figure 5: Combinatorial binding of TFs within and between enhancer elements.
A) Switching between dimerization partners can confer different functionalities. B) The enhanceosome requires strict organization of TFs in a continuous interface. C) The billboard structure is flexible in terms of TFs, which might compete for accessibility (inspired by Spitz et al., 2012, figure 3). D) Distal enhancers can interact with target promoters through DNA looping mediated by TFs and structural proteins.

the enhancer loses its tissue specificity. Thus, the trade-off of individual syntax elements serve as a way to carefully titrate gene regulation within individual tissues.

Besides describing the positioning of TFs within enhancers, *enhancer grammar* also describes the interactions of TFs and proteins between individual enhancer elements (Figure 5D). By DNA looping in 3D space, distal enhancers can make contact with promoter regions and thereby drive expression of genes from remote locations of the genome. On average, these enhancers are located ~20 to ~50 kilobases from their target genes, but there are also examples of distances of more than 100 kilobases (Furlong et al., 2018). These promoter-enhancer contacts are controlled by the existence of topologically associating domains (TADs; not to be confused with trans-activation domains of TFs), which are loop structures with a high frequency of interactions within each TAD, but not crossing the TAD borders (Szabo et al., 2019). In addition, the organization of TADs specifies two separate compartments of chromatin, known as the A and B compartments, which are activating and repressing gene expression, respectively (Szabo et al., 2019).

The formation of TADs is largely mediated by cohesin complexes, which form rings around chromatin fibers to enable active extrusion of chromatin loops (Furlong et al., 2018) and insulator proteins like CTCF, which act as barriers to limit extension of loops (Ong et al., 2014). However, other mechanisms exist as exemplified by the loop between the locus control region enhancer and the β-globin promoter in erythroid progenitor cells (Deng et al., 2012). This loop is independent of cohesin, and is instead driven by the interactions of GATA1 and the transcriptional cofactor Ldb1, through binding sites for GATA1 in both the enhancer and the target promoter. Interestingly, GATA1-mediated loop formation seems to be a general mechanism, as the majority (70%) of enhancer loops in erythrocyte precursors are mediated by GATA1/Ldb1 in the absence of cohesin (Krivega et al., 2017). Thus, collaboration of TFs and other DNA binding proteins, not only within enhancers, but also between enhancers and promoters, help to regulate gene expression in a cell type specific manner.

In conclusion, TFs collaborate in a myriad of ways to regulate target gene expression. However, there are also cases where TFs not only affect gene regulation, but also chromatin structure as a whole. These examples will be covered in the next section.

1.2.5 Mechanisms of pioneer transcription factors

In the previous sections, TF binding has largely been described as being limited to recognition of targets within already accessible chromatin. In this scenario, we can imagine regulatory regions as rooms for TFs to enter, where nucleosome deposition and repressive histone marks represent locked doors (Figure 6A). Indeed, nucleosome binding prevent most TFs from recognizing their target motifs, which helps cells to retain a particular transcriptional program. However, a subset of TFs known

as *pioneer factors* are able to overcome these restrictions to target DNA even in the presence of nucleosomes (Iwafuchi-Doi et al., 2014). Analogously, these TFs have a key to the locked doors, and are therefore important for initiating cell fate changes throughout differentiation (Figure 6B).

A well-known group of pioneer factors are Oct4, Sox2 and Klf4, which are part of the OSKM factors as mentioned in the context of reprogramming in Section 1.2.2. In fact, it is the pioneering abilities of these factors which make them successful at driving the massive chromatin remodeling required for reprogramming. This ability lies in the capacity of these factors to recognize their target motifs on DNA even when partially concealed by nucleosome binding. On nucleosome-free DNA, Oct4 recognizes its canonical 8bp motif by binding of its C- and N-terminal domains on both faces of the DNA (Soufi et al., 2015). However, this type of binding is impeded by nucleosome binding, which renders half of the DNA face inaccessible, but Oct4 overcomes this challenge by being able to recognize a partial motif using each binding domain separately (Soufi et al., 2015). Similarly, Klf4 uses three zinc-finger domains to associate with nucleosome free DNA, but only uses two in order to recognize a shortened motif on nucleosome-associated DNA (Soufi et al., 2015). Investigation of different pioneer factors uncovered that they preferentially bind DNA through scissor-like domains of short α-helixes (Fernandez Garcia et al., 2019). The shorter α-helix enables interaction on one face of the DNA, while the other face is being bound to the nucleosome. In contrast, weak nucleosome binders contain longer α-helixes, or lack these helixes completely, instead using β-sheets, short helical twists and unstructured regions to recognize DNA targets (Fernandez Garcia et al., 2019). Thus, specific properties allow certain TFs to be nucleosome binders.

Interestingly, while the pioneers are the first to gain access to the closed chromatin, they are not necessarily responsible for the subsequent increase in chromatin accessibility. An example of this is the case of the Pax7 and Tpit TFs, which distinguish the split between melanotropes and corticotropes in the pituitary gland. In an elegant study using mating of $Pax7^{-/-}$ and $Tpit^{-/-}$ mice, Mayran et al., 2019 showed that Pax7 is able to bind to nucleosomes in closed chromatin, but that its ability to open chromatin is dependent on the subsequent binding of Tpit. Pax7 is also a player in muscle cell specification, but the cooperation with Tpit ensures that these pituitary sites are only active when Tpit is expressed. In the analogy of the locked door, Pax7 can unlock the door, but can only open the door with the help of Tpit. Pioneer factors can thereby specify a primed state prior to lineage commitment (Iwafuchi-Doi et al., 2014).

Besides adjusting the expression of partner TFs, mechanisms such as histone modifications also help to control the effects of pioneer factors. In the case of fibroblast reprogramming to iPSCs using OSKM, 70% of the initial OSKM binding positions are found in closed chromatin regions, which subsequently become accessible (Soufi et al., 2012). However, comparison to known sites of OSKM in hES cells resulted in the identification of 264 regions with an average size of 2.2 megabases, which were obstructing OSKM binding in fibroblasts due to enrichment of H3K9me3 (Soufi et al.,

2012). Thus, this shows that H3K9me3 is impeding reprogramming by hindering the initial binding of OSKM. This mechanism is a great example of the cell fate restrictions imposed by the epigenetic landscape as discussed in Section 1.2.2. Ultimately, such restrictions commit the cell to a stable fate by protecting against detrimental chromatin opening by spurious activation of pioneers. In the analogy of the locked door, repressive histone marks act to block the door even to the pioneer factors which have keys (Figure 6C). Efforts have been made in trying to enhance reprogramming by removing these restrictions, such as with the use of HDAC inhibitors, H3K4 demethylases and DNA methyltransferases, as well as using demethylases against repressive chromatin marks H3K27 and H3K36 (Soufi, 2014).

These examples revealed that because pioneer factors are very powerful, they must also be tightly controlled. It is therefore not surprising that misregulation of pioneer TF binding can lead to disease. One example is the pioneer factor FOXA1/2, which bind to nucleosomes at liver-specific regulatory regions to enable binding of other TFs such as HNF4α and C/EBPβ (Iwafuchi-Doi et al., 2016). However, in the context of prostate cancer, Adams et al., 2019 found that mutations of the FOXA1 forkhead DNA-binding domain induce changes in chromatin accessibility in comparison to wildtype FOXA1, and that these mutations are associated with faster progression of metastasis. Mechanistically, the mutation of R219S altered the FOXA1 binding motif from GTAAA(C/T) to GTAAA(G/A), which allowed it to bind to novel sites, which were annotated to genes associated

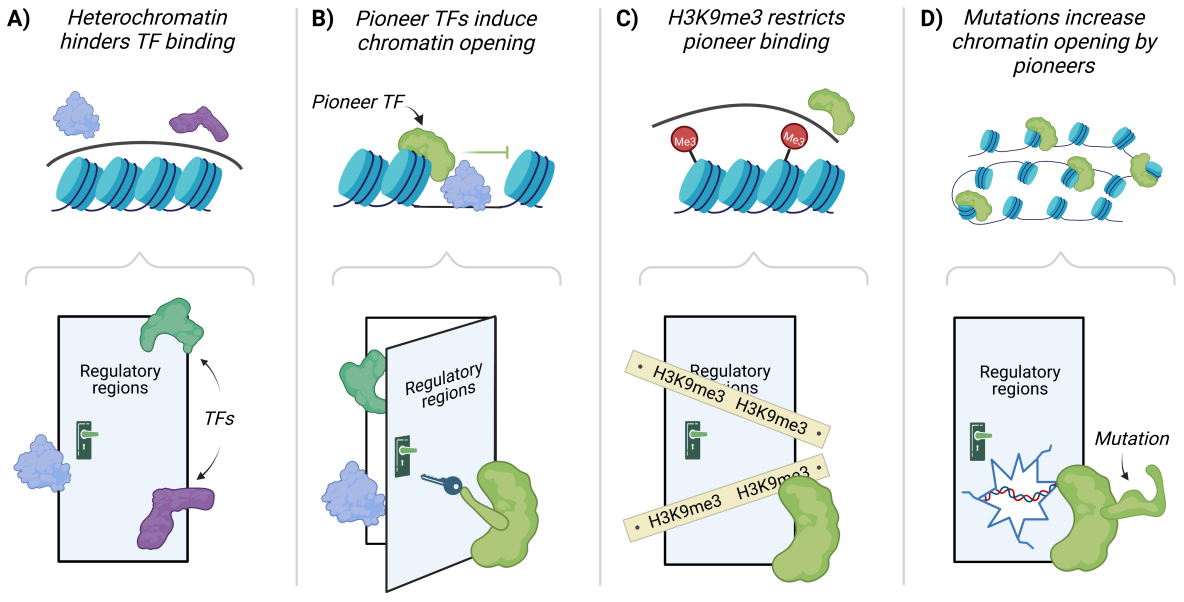


Figure 6: Pioneering factors influence epigenetic regulation by opening chromatin. A) TF binding is generally hindered by tightly-packed nucleosomes (blue cylinders), and the TFs can therefore not enter the door of regulatory regions. B) Pioneers can unlock heterochromatin and allow binding of additional TFs. C) Pioneer binding is restricted by H3K9me3. D) Mutations in the pioneer factor allows it to bind additional target sites, which were previously restricted.

with epithelial–mesenchymal-transition (Adams et al., 2019). In the analogy of the locked door, mutations in FOXA1 granted it powers to break through previously inaccessible doors, which lead to activation of inappropriate target genes (Figure 6D).

In conclusion, pioneer TFs are powerful tools for cells to switch between different transcriptional programs. In particular, the ability of pioneer TFs to massively change the chromatin landscape make them especially important for the initial lineage commitment of cells in the embryo. Thus, the next section will introduce the influence of pioneer factors, and many other epigenetic mechanisms, on the processes of early embryonic development.

1.2.6 Epigenetic control throughout early embryonic development

Preimplantation development of the embryo is a fascinating process, as it is a prerequisite for subsequent growth of the entire multi-cellular organism. During fertilization, two terminally differentiated cells, the oocyte and the spermatozoon, fuse to provide a totipotent zygote, which is capable of differentiating into both embryonic and extraembryonic cell lineages (Figure 7A) (Eckersley-Maslin et al., 2018). The successful progression of the zygote requires massive chromatin reorganization as well as activation of the transcriptionally silent genome - a process known as *zygotic genome activation* (ZGA) (Eckersley-Maslin et al., 2018). During this process, the zygotic genes are activated in two waves, known as the minor and major waves, which occur at the late 1-cell (1C) and late 2C stages in mouse, and at the 4C and 8C stages in humans (Schulz et al., 2019). This section will provide an overview of the ways in which ZGA, and embryonic development in general, are controlled by the types of epigenetic mechanisms previously described.

Directly following fertilization, the zygote is transcriptionally inactive, and is thus controlled exclusively by proteins provided by the oocyte. These so-called *maternal factors* include proteins such as ribosomes and spliceosomes, which are needed for the initial control of protein synthesis (Heyn et al., 2014). A complex of maternally provided proteins, namely Floped, Mater, Tle6 and Filia, known as the *Subcortical maternal complex* (SCMC), have also been shown to be required for normal development (Li et al., 2008). In fact, Li et al., 2008 showed that $Floped^{-/-}$ and $Mater^{-/-}$ females are infertile, and a closer look at any of the fertilized embryos from homozygous knockout females showed that they never progress beyond the 2C stage. This suggests that maternally provided proteins are important for proper development of the zygote. In addition to proteins, the oocyte also provides mRNAs, which encode for proteins needed during ZGA. However, as these require translation before becoming active, there is a delay in their activity, and these mRNA products are therefore also believed to partially explain the initial lag of ZGA following fertilization (Schulz et al., 2019).

While the maternally provided transcripts and proteins are important for initiation of ZGA, their timely removal is key to developmental success (Figure 7B). In fact, both maternal factors and genes expressed early in ZGA are known to mediate clearance of maternal mRNAs (Sha et al., 2020). In mice, maternally encoded pathways including BTG4-mediated deadenylation by CCR4-NOT, and terminal uridylation of mRNAs by uridylyltransferases TUT4 and TUT7, are responsible for degrading maternal mRNAs (Sha et al., 2020). However, final clearance of maternal factors requires additional factors transcribed during ZGA. For example, the TFs Nanog, Pou5f1 and SoxB1 have been shown to drive the first wave of ZGA in zebrafish, and in turn induce the expression of *miR-230* (Lee et al., 2013). *mirR-203*, a microRNA, has known functions in clearing maternal mRNAs by associating with its target mRNA molecules and accelerating their deadenylation and decay (Giraldez et al., 2006). Likewise, the maternally provided TF YAP1 is required for proper degradation of maternal transcripts through re-activation of *Tut4/7* expression from the zygotic genome in mice (Sha et al., 2020; Yu et al., 2016). In summary, maternally provided proteins and transcripts are needed for early control of the zygote, but also act via a negative feedback loop to remove remnants of the oocyte's transcriptional program, and control the timely initiation of ZGA.

Besides proteins for maternal factor clearance, the earliest ZGA genes include DNA binding proteins and histone modifiers (Gao et al., 2017; Heyn et al., 2014). In particular, the latter represents an important method for regulation of chromatin function in the early embryo. Although the zygote is transcriptionally inactive, the zygotic genome contains broad regions marked by H3K4me3 and the removal of these have been shown to be important for the progression of ZGA (Dahl et al., 2016). In the zygote and during the transition to 2C, the maternally provided demethylase KDM1A removes parts of the broad H3K4me3 marks (Ancelin et al., 2016). Subsequently, demethylases KDM5A and KDM5B are expressed from the zygotic genome during ZGA, and their expression correlates with the final removal of broad H3K4me3 domains and the establishment of canonical H3K4me3 and H3K27ac signals at promoters (Eckersley-Maslin et al., 2018). Whereas loss of KDM1A results in elevation of H3K4 methylation and failure to develop past the 2C stage in mice (Ancelin et al., 2016), embryos depleted for KDM5A and KDM5B exhibit developmental delays during the 4C to 8C transition (Dahl et al., 2016). This indicates that ZGA and histone modifications are highly intertwined, as regulation of histone modifications is necessary for ZGA to initiate, but transcripts from the zygotic genome are also necessary to regulate histone modifications during ZGA.

As described in Section 1.2.1, regulation of histone modifications is frequently coupled with changes in chromatin accessibility, which is also the case during ZGA. In fact, the existence of broad H3K4me3 regions coincide with regions of open chromatin prior to ZGA (Wu et al., 2018). Before ZGA, the chromatin is largely unstructured and does not exhibit any TADs (Ke et al., 2017). However, higher order TADs are progressively established during 4C-8C and become more defined throughout development (Ke et al., 2017). In parallel with the organization of higher chromatin

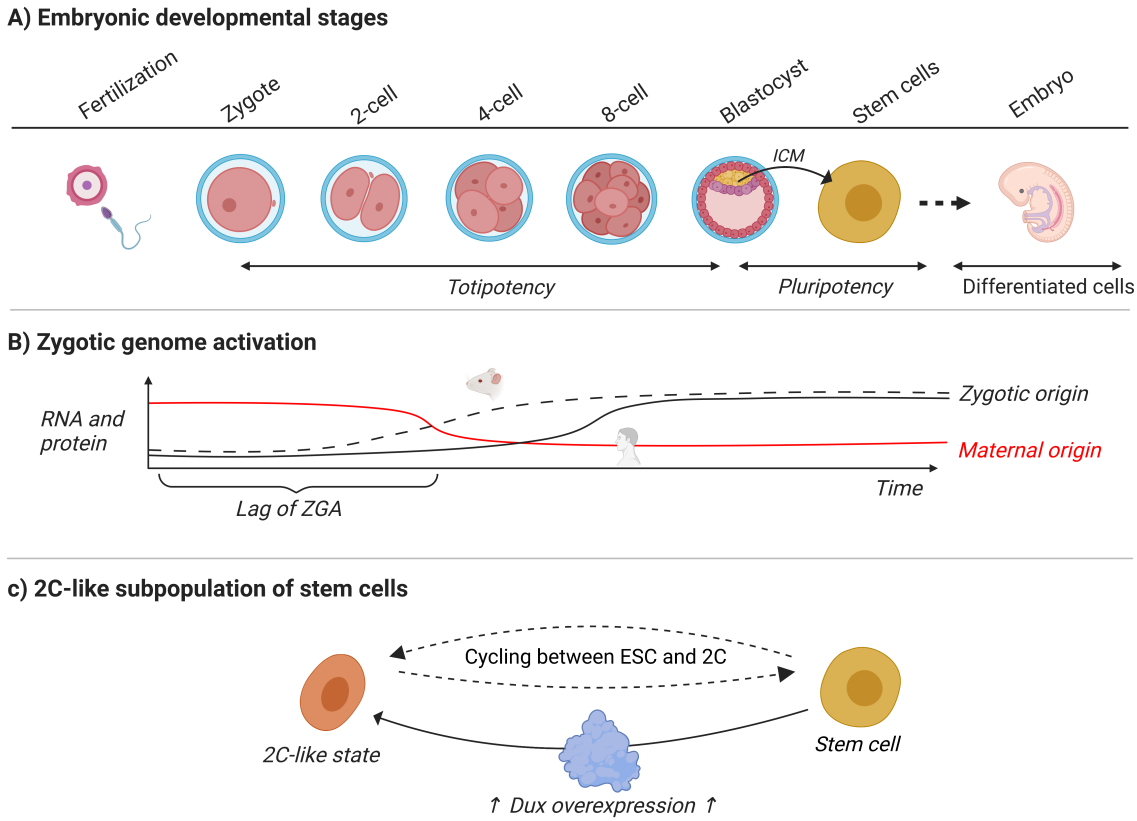


Figure 7: Epigenetic regulation through early embryonic development.
A) The developmental stages from fertilization to the expansion of the inner cell mass (ICM), which gives rise to the cells of the embryo. B) Zygotic genome activation is shifted between human and mouse. In parallel, the maternally provided proteins and transcripts are actively removed. C) Embryonic stem cells cycle in and out of a 2C-like state, which can also be induced artificially by Dux overexpression.

structure, the number of accessible regions also increases and their genomic distribution changes. At the 1C stage, 87% of open chromatin regions in mouse embryos are found in promoter regions, but this decreases to 77% in 2C and 50% from 8C and onward (Lu et al., 2016). Thus, the large increase of enhancers at 8C, as well as the establishment of higher order chromatin structure, indicates that distal regulatory elements play a role in the early lineage specification in the embryo.

In correlation with the lack of structured chromatin, the early zygote exhibits massive expression of repeats, including transposable elements (TEs) of the long terminal repeat (LTR) family (Wu et al., 2018; Macfarlan et al., 2012). Throughout evolution, TEs have arisen from germline introduction of retroviral DNA, and have obtained additional copies across the genome by means of active retrotransposition (Gifford et al., 2013). In that way, TEs have provided new regulatory networks through integration of TFBS, and are also thought to have played a role in the development of placental mammals by introduction of viral envelope proteins such as SyncytinA (Gifford et al., 2013; Dupressoir et al., 2009). However, retrotransposition can be detrimental, as seen for active L1 transposition, which is associated with the formation of many cancers (Payer et al., 2019). For

that reason, TEs these are largely silenced in somatic cells (Macfarlan et al., 2012). So why are these elements expressed in zygotes? Interestingly, LTRs have been found to serve as alternative promoters for early ZGA genes (Peaston et al., 2004), and expression of LINE-1 elements helps to regulate chromatin accessibility (Jachowicz et al., 2017). Thus, some TEs are suggested to have been actively maintained in the genome due to their roles in the processes of embryonic development.

As already mentioned, several TFs have been implicated in ZGA, many of which are thought to be pioneer factors. For example, the maternally provided TF Zelda is a major regulator of ZGA in *Drosophila*, where it directly activates early ZGA genes as well as regulate chromatin accessibility prior to binding of additional TFs (Harrison et al., 2011). In vertebrates, other maternal proteins, such as Nfya, fulfill similar roles in establishing early chromatin accessibility. In fact, silencing Nfya with siRNAs in mouse oocytes resulted in loss of ~30% of 2C open regions, failure to activate expression for 15% of ZGA genes, and developmental arrest before reaching blastocyst stage (Lu et al., 2016). In contrast, the pioneer factor Oct4 plays a role in establishing open chromatin regions in 8C, as knockout of Oct4 in mouse embryos showed a failure to open 27.4% of wildtype-gained 8C regions (Lu et al., 2016). In summary, the massive changes in chromatin accessibility are driven partly by pioneer factors which reprogram chromatin structure throughout preimplantation development.

As seen for Zelda and Nfya, the early TFs implicated in ZGA seem to have evolved to fulfill similar functionalities despite not being evolutionary related. Indeed, it has been shown that genes expressed during ZGA are younger than genes expressed at other times during development (Heyn et al., 2014). For example, Madissoon et al., 2016 characterized a group of PAIRED-like TF homeodomain genes, *ARGFX*, *CPHX1/2*, *DUXA/B*, *TPRX1/2*, *DPRX* and *NOBOX*, which are mainly found in primates (Töhönen et al., 2015). These genes are expressed in 8C human embryos and they are silenced in somatic tissue due to methylation of their gene promoters, suggesting that they are important during ZGA (Töhönen et al., 2015). While most of these were lost in mice, two clusters of distant orthologs have been discovered, showing that *TPRX1* and *TPRX2* gave rise to the *Crxos* and *Obox* families respectively. However, whereas human TPRX1/2 are believed to act mainly as repressors, Crxos is found to upregulate known 2C genes (Madissoon et al., 2016; Royall et al., 2018). Interestingly, there is a significant overlap of human ARGFX targets with Crxos targets in mouse (Royall et al., 2018). This case represents an interesting case of evolution where murine Crxos seems to fulfill the missing functionalities of ARGFX, which is lost in mice, even though their genes are not orthologous.

Another member of the PAIRED-like family is the TF DUX4. Like the other genes in this family, the *DUX* locus has undergone extensive gene duplication and divergence between species (Leidenroth et al., 2010). However, both DUX4 and its mouse homolog Dux have conserved functions within ZGA (Hendrickson et al., 2017). In line with the shift of ZGA between human and mouse, *DUX4* is expressed during 4C stage in human, and *Dux* during early 2C stage in mouse. In addition, many of the Dux target genes are orthologous to the DUX4 target genes, including *Zscan4*, which is one of

the earliest expressed genes during ZGA (Hendrickson et al., 2017). Thus, DUX4/Dux is believed to be one of the master regulators of ZGA.

Interestingly, although the expression of *Dux* and its targets are restricted to the 2C stage in mouse, it is also found upregulated in a small subpopulation of mESCs. This subpopulation of mESCs show a spontaneous increase in expression of LTRs and 2C-genes including *Zscan4*, *Tcstv1/3* and *Tdpoz1-5*, and these cells likewise obtain broad H3K4 domains (Macfarlan et al., 2012). Thus, this subpopulation is known as *2C-like cells* (2CLCs) (Macfarlan et al., 2012) (Figure 7C). These cells comprise less than one percent of the mESC pool at any time, but if mESCs are cultured long enough, nearly every cell will have entered this subpopulation (Macfarlan et al., 2012). In mESCs, the conversion to the 2CLC state is driven in part by Dux, as depletion of Dux has been shown to decrease the spontaneous reprogramming of mESCs (Hendrickson et al., 2017). Likewise, overexpression of *Dux* reprograms mESCs to 2CLCs through massive chromatin changes, suggesting that Dux acts as a pioneer factor (Hendrickson et al., 2017; De Iaco et al., 2017). In addition, mutations of either demethylase Kdm1a, transcriptional repressor Kap1 or H3K9 histone methyltransferase G9a, as well as inhibition of histone deacetylases using trichostatin A, leads to significant upregulation of the 2CLC population (Macfarlan et al., 2011; Macfarlan et al., 2012). Interestingly, mESC cultures depleted of 2CLCs are still viable, but are shown to have increased differentiation into mesoderm and ectoderm lineages *in vitro* (Macfarlan et al., 2012). This observation suggests that cycling through the 2CLC state is important for maintaining pluripotency in mESCs.

In conclusion, early embryonic development is a highly specialized process employing multiple branches of epigenetic regulation including histone modifications, chromatin accessibility and binding of pioneer TFs. However, many aspects of how ZGA is initiated, and which TFs are playing a role, are still poorly understood. In that context, the next section will review a number of experimental and computational methods for studying TF binding.

1.2.7 High-throughput techniques and bioinformatics analysis for studying TF binding

With the emergence of high-throughput sequencing techniques, it has become possible to investigate epigenetic mechanisms in great detail. As the amount of data increases, the requirements of interpretation increase as well. The task of extracting information from high-throughput assays calls for customized bioinformatics analysis (Gauthier et al., 2019). Some of the most relevant technologies, as well as the discussion of how to utilize the data generated, will be covered in this section.

One of the primary steps of epigenetics research has been to identify active regulatory regions, and this has been fulfilled using chromatin-accessibility assays such as DNase-seq and later ATAC-seq

(Figure 8A). These technologies take advantage of the susceptibility of accessible DNA to be cut by enzymes. In the case of ATAC-seq, a modified Tn5 transposase (referred to as Tn5) predominantly inserts sequencing adapters into open chromatin (Buenrostro et al., 2013). After sequencing of these fragments, computational mapping and peak-calling yields a list of open chromatin regions (Figure 8B). Ultimately, these peaks highlight regions where TFs might bind to regulate gene expression. In this context, RNA-seq can be used as a read-out of transcription (Figure 8A). Thus, the combination of ATAC-seq and RNA-seq is a powerful tool to identify whether changes in chromatin accessibility have consequences for gene expression, and can help to assign regulatory enhancers to individual genes.

To investigate the direct association of TFs to DNA, ChIP-based methods such as ChIP-seq and ChIP-exo have commonly been utilized (Rossi et al., 2018). Recently, CUT&RUN and CUT&Tag are gaining popularity due to their ability to work with a low amount of cells and without the need for cross-linking protein to DNA (Skene et al., 2017; Kaya-Okur et al., 2019). These methodologies all work by applying antibodies for a target protein (e.g. TF or histone modification), cleaving DNA at the target locations by either sonication (ChIP-seq), MNase (CUT&RUN) or Tn5 (CUT&Tag), and subsequently sequencing the resulting fragments. Using peak-calling analysis similar to chromatin accessibility assays, these methods can identify the exact locations of TF binding throughout the genome. However, while these experimental methods provide accurate identification of binding sites, they are limited by the need for high quality target antibodies and individual experiments per TF (Park, 2009). The use of experimental assays to study the effect of multiple TFs in large-scale transcriptional networks is therefore challenging.

As a supplement to experimental methods, the use of bioinformatics analysis for studying TF binding is intensifying. Based on ChIP-seq and HT-SELEX assays, the sequence preferences for many TFs have been established and represented as position weight matrices (PWM) and logo plots (Dror et al., 2015; Schneider et al., 1990). These motifs are typically 6-20 base pairs wide and are collected in databases such as JASPAR (Fornes et al., 2020) and HOCOMOCO (Kulakovskiy et al., 2016). These PWMs can be utilized to scan the genome for possible TFBS, but as discussed in Section 1.2.3, there is more to TF binding than just the presence of a target sequence. TFs can act differently depending on the cell type, tissue and cellular condition, or even differ in functionality across organisms. Thus, searching for binding sites without being aware of the epigenetic landscape does not properly reflect TF binding.

An analysis known as *genomic footprinting* aims to bridge the gap between experimental and *in silico* prediction of TFBS. The concept of this technique was shown for the first time in 1978 using DNase on a gel (Galas et al., 1978). Since DNA is accessible to cleavage by DNase-seq, binding of a TF leaves a region of limited cuts - known as a *footprint*. While the original assay was limited to very small regions, DNase-seq and ATAC-seq have allowed for massive genome-wide

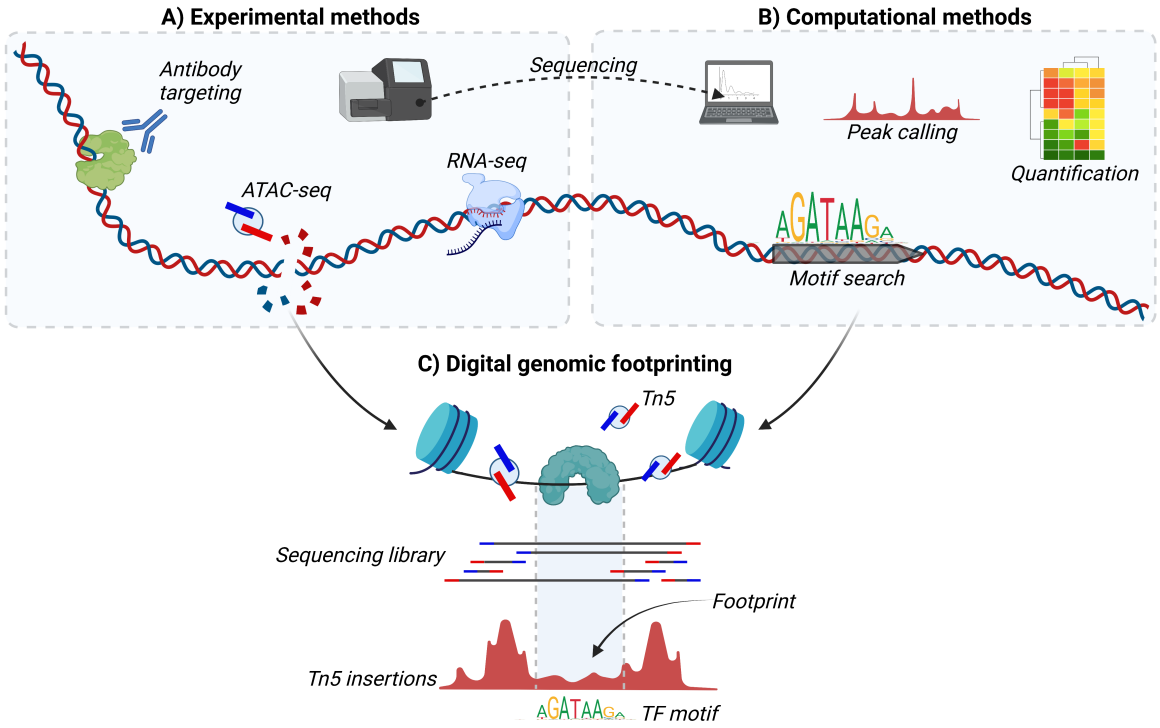


Figure 8: High-throughput sequencing technologies.
A) Experimental sequencing methods, including antibody targeting for protein binding, ATAC-seq for chromatin accessibility and RNA-seq for quantifying gene expression, are used to study epigenetic mechanisms. B) Interpretation of sequencing data requires computational methods, including peak calling and quantification between samples. Likewise, purely computational analysis such as motif search are used. C) Digital genomic footprinting is a computational method for identifying TF binding from chromatin accessibility assays such as ATAC-seq.

analysis of insertion patterns (Hesselberth et al., 2009; Neph et al., 2012), updating the term to *digital* genomic footprinting (Figure 8C). Footprinting analysis holds great potential to uncover the dynamics of parallel TF binding from a single experimental assay. However, with more than 700 known transcription factor motifs, and more than a million human enhancer elements (Gasperini et al., 2020), the computational task is not trivial.

In general, there are considerable challenges in software development for bioinformatics tasks. Besides producing accurate results, bioinformatics tools must be well-maintained and provide good documentation in order to be adopted by the scientific community. In the context of data, the FAIR principles, which stand for *F*indability, *A*ccessibility, *I*nteroperability and *R*eusability, describe a standard for data handling in scientific research (Wilkinson et al., 2016). Building on FAIR, the *FAIR for research software* (FAIR4RS) principles have recently been established, which provide a guideline for best practices when developing and sharing software (Barker et al., 2022). Comparable to the FAIR principles, these guidelines state that software must be easily retrievable, should read and write data in community standard formats, and be interoperable with other software (Barker et al., 2022). In this context, open-source repositories such as GitHub enable version control, commu-

nity driven issues and continuous maintenance of the code. While this seems trivial, these needs are a huge challenge, as 1/3 of bioinformatics tools have never been updated after publication (Russell et al., 2018). Likewise, for web-services, the availability of tools decreases linearly to the time after publication (Kern et al., 2020). While web-services are important for bridging the gap between bioinformaticians and biologists, this comes at the cost of reproducibility over time. Another challenge with web-services is that they are difficult to integrate into existing workflows, which is a requirement for FAIR4RS. In recent years, multiple workflow languages such as Nextflow (Di Tommaso et al., 2017) and Snakemake (Koster et al., 2012) have been developed to solve the problem of organizing complex analysis pipelines. These tools simplify the usage of multiple types of software, but also raises the necessity of common file formats for individual tools to pass information to the subsequent steps of the analysis. The quality of bioinformatics software should therefore not only be measured on the results, but also on the documentation, usage and provided interfaces to other tools.

In conclusion, bioinformatics analysis is an integral part of epigenetic research. Experimental and bioinformatics approaches are heavily intertwined, and in the best-case scenario, bioinformatics analysis on experimental data yields a discovery, which brings the investigators back to the lab, enabling close interactions between experimental and computational efforts. As such, advances within the field of epigenetics are driven by the interplay of establishing new exciting techniques and the development of high quality bioinformatics solutions to analyze large amounts of data.

1.3 Objectives

Within the introduction of this thesis it has been argued that epigenetic mechanisms, and TFs in particular, act to regulate gene expression throughout development. It is therefore of great interest to study the mechanisms of TF binding in different cell types. However, especially for scarce samples such as early embryos, some experimental assays are difficult to perform. As a result, the knowledge of the TFs initiating ZGA, as well as their target genes, is incomplete. Likewise, the lack of a global map of TF binding has hindered the investigation of TF cooperation and binding grammar in the context of transcriptional networks.

While some of these challenges can be solved by bioinformatics approaches, purely *in silico* methods, like scanning for genome-wide occurrences of TF motifs, are poor predictors of TF binding. Thus, there is a need to include information of the epigenetic landscape into these models. Footprinting analysis was presented in Section 1.2.7 as the combination of these experimental and computational research efforts. However, while some tools exist for footprinting (Li et al., 2019; Ouyang et al., 2020; Gusmao et al., 2014; Raj et al., 2015; Piper et al., 2013; Kähärä et al., 2015), the majority

of these were developed for DNase-seq and can therefore not be used for ATAC-seq due to assay-specific biases. In addition, many of these methods require input ChIP-seq for supervised learning, which limit their application to well-studied TFs for which ChIP-seq antibodies are available. There is a need for an unbiased ATAC-seq footprinting tool able to predict TF binding independently of ChIP-seq data.

The ability to predict TF binding positions adds a new dimension of potential investigations relating to TF binding mechanisms. In the context of combinatorial TF binding as described in Section 1.2.4, genome-wide TFBS can for example be utilized to study the interaction networks of individual TFs. However, similarly to existing tools for footprinting analysis, many of the current tools for studying TF co-occurrence and binding grammar rely on ChIP-seq as input, and do not support input from other sources such as footprinting (Levitsky et al., 2019; Whitington et al., 2011). This currently limits co-occurrence analysis to those cell types where ChIP-seq quantification has been performed on multiple factors. In addition, many tools within this area of research are provided as web-services, which are no longer maintained (Perna et al., 2018; Kazemian et al., 2013; Zhang et al., 2011). There is a need for a flexible tool for co-occurrence analysis and binding grammar, which also complies with the FAIR4RS principles.

As a result of these challenges, the objective of this thesis was to develop and apply bioinformatics tools in order to:

- Utilize ATAC-seq footprinting for the prediction of TF binding
- Investigate TF binding in the context of early embryonic development
- Characterize TF co-occurrence and binding grammar within enhancers

2 Results

2.1 Publication 1: ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation

This publication introduces TOBIAS (Transcription factor Occupancy prediction By Investigation of ATAC-seq Signal), a bioinformatics tool for predicting binding locations of TFs through ATAC-seq footprinting analysis. By investigating Tn5 insertions in relation to its known background bias, and quantifying whether footprints are observed, TOBIAS creates a genome-wide network of TF binding. Additionally, TOBIAS compares TF binding between biological conditions to provide a differential footprint score per TF. Validation on a set of experimentally verified TF binding sites showed that TOBIAS outperforms previously published tools for ATAC-seq footprinting. Particularly in terms of correction of Tn5 bias, TOBIAS was shown to correctly uncover footprints previously hidden by background signal. Thereby, this method serves to solve the problem of surveying binding of multiple TFs from one ATAC-seq assay.

Because of the challenges of investigating TF binding in early embryonic development, TOBIAS was applied to previously published ATAC-seq data from preimplantation embryos of human and mouse. The analysis showed an intriguing timeline, in which specific TF binding is visible as ATAC-seq footprints for certain timepoints throughout development. By integration of RNA-seq, it was also possible to comment on the relationship between the transcriptome and the binding of TFs. Using an additional dataset on overexpression of *Dux*, TOBIAS correctly identified Dux as being one of the top differentially bound TFs between the control and perturbed experiments. In addition, the ability of TOBIAS to identify local TF binding sites allowed for the collection of a list of predicted target genes of Dux, which correlated with changes in the transcriptome as well as repeat elements of both LTR and LINE-1 families.

In conclusion, TOBIAS enables researchers to predict changes in TF binding across different stages of differentiation, thereby mapping the influence of TFs in driving distinct cell lineages. An overview of the individual contributions of the thesis author to the publication is found in Table 1.

Table 1: Contributions by the thesis author to publication 1.

Area	**Contributions**
Conceptualization	Contributed to the definition of the project goals.
Software	Developed and implemented the TOBIAS software and the TOBIAS Snakemake pipeline. Supervised the creation of the Nextflow pipeline.
Data	Contributed to obtaining and processing datasets from public databases.
Analysis	Performed validation of TOBIAS using ChIP-seq data. Performed extensive comparison to existing footprinting tools. Application of TOBIAS to datasets in the context of early embryonic development.
Visualization	Created Figure 1, 2b-c, 3, 4a-d, 5 and all supplementary figures.
Manuscript	Wrote the manuscript draft and contributed to the review and editing of the final manuscript.

The full article is found in the following pages and the supplementary figures are found in Appendix A1.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-18035-1 OPEN

ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation

Mette Bentsen[1], Philipp Goymann[1], Hendrik Schultheis[1], Kathrin Klee[1], Anastasiia Petrova[1], René Wiegandt[1], Annika Fust[1], Jens Preussner[1,2], Carsten Kuenne[1], Thomas Braun[2,3], Johnny Kim[2,3] & Mario Looso[1,2]✉

While footprinting analysis of ATAC-seq data can theoretically enable investigation of transcription factor (TF) binding, the lack of a computational tool able to conduct different levels of footprinting analysis has so-far hindered the widespread application of this method. Here we present TOBIAS, a comprehensive, accurate, and fast footprinting framework enabling genome-wide investigation of TF binding dynamics for hundreds of TFs simultaneously. We validate TOBIAS using paired ATAC-seq and ChIP-seq data, and find that TOBIAS outperforms existing methods for bias correction and footprinting. As a proof-of-concept, we illustrate how TOBIAS can unveil complex TF dynamics during zygotic genome activation in both humans and mice, and propose how zygotic Dux activates cascades of TFs, binds to repeat elements and induces expression of novel genetic elements.

[1] Bioinformatics Core Unit (BCU), Max Planck Institute for Heart and Lung Research, 61231 Bad Nauheim, Germany. [2] German Centre for Cardiovascular Research (DZHK), Partner Site Rhine-Main, 60596 Frankfurt am Main, Germany. [3] Department of Cardiac Development and Remodeling, Max Planck Institute for Heart and Lung Research, 61231 Bad Nauheim, Germany. ✉email: mario.looso@mpi-bn.mpg.de

Epigenetic mechanisms governing chromatin organization and transcription factor (TF) binding are critical components of transcriptional regulation and cellular transitions. In recent years, rapid improvements of pioneering sequencing methods such as ATAC-seq (Assay of Transposase Accessible Chromatin)[1], have allowed for systematic, global scale investigation of epigenetic mechanisms controlling gene expression. While ATAC-seq can uncover accessible regions where TFs might bind, reliable identification of specific TF binding sites (TFBS) still relies on chromatin immunoprecipitation methods such as ChIP-seq. However, ChIP-seq methods require high input cell numbers, are limited to one TF per assay, and are further restricted to TFs for which antibodies are readily available. Therefore, it remains costly, or even impossible, to study the binding of multiple TFs in parallel.

Current limits to the investigation of TF binding become particularly apparent when investigating processes involving a very limited number of cells, such as preimplantation development (PD) and zygotic genome activation (ZGA) of early zygotes. Integration of multiple omics-based profiling methods have revealed a set of key TFs that are expressed at the onset of and during ZGA including Dux[2], Zscan4[3], and other homeobox-containing TFs[4]. However, due to the limitations of ChIP-seq, the exact genetic elements bound and regulated by different TFs during PD remain to be fully discovered. Consequently, the global network of TF binding dynamics throughout PD remains mostly obscure.

A computational method known as digital genomic footprinting (DGF)[5] has emerged as an alternative means, which can overcome some of the limitations of ChIP-based methods. DGF is a computational analysis of chromatin accessibility assays such as ATAC-seq, which employs DNA effector enzymes that only cut accessible DNA regions. Similarly to nucleosomes, bound TFs hinder cleavage of DNA, resulting in defined regions of decreased signal strength within larger regions of high signal—known as footprints[6] (Fig. 1a).

Surprisingly, although this concept shows considerable potential to survey genome-wide binding of multiple TFs in parallel from a single experiment, DGF analysis is rarely applied when investigating TF binding mechanisms. The skepticism towards DGF has been driven by the discovery that enzymes used in chromatin accessibility assays (e.g., DNase-I) are biased towards certain sequence compositions, an effect which has been well characterized for DNase-seq[7,8]. The influence of Tn5 transposase bias in the context of ATAC-seq footprinting has, however, only been described very recently[9,10] and still represents an uncertainty during discovery of true footprints. Besides the identification of footprints, comparing footprints across biological conditions remains challenging as well. While there have been efforts to estimate differential TF binding on a genome-wide scale[11,12], investigation of epigenetic processes often requires more in-depth information on the individual differentially bound TFBS and genes targeted by these TFs. Furthermore, many footprinting methods suffer from performance issues due to missing support for multiprocessing, inflexible software architecture, and the use of non-standard file-formats. These obstacles complicate the assembly of different tools for advanced analysis workflows. Consequently, despite its compelling potential, these issues have rendered footprinting on ATAC-seq cumbersome to apply to biological questions. Essentially, a comprehensive framework enabling large-scale ATAC-seq footprinting is missing.

Here, we describe TOBIAS (Transcription factor Occupancy prediction By Investigation of ATAC-seq Signal), a comprehensive computational framework that we created for footprinting analysis (Fig. 1b–f). TOBIAS is a collection of command-line tools utilizing a minimal input of ATAC-seq reads, TF motifs and genome information (Fig. 1b) to perform all levels of footprinting analysis including bias correction (Fig. 1c), footprinting (Fig. 1d), and comparison between conditions (Fig. 1e). Furthermore, TOBIAS includes a variety of auxiliary tools such as TF network inference and visualization of footprints, which allow for various downstream analysis (Fig. 1f, Supplementary Fig. 1). In this investigation, we apply TOBIAS to ATAC-seq data from both human and mouse PD and show how visible TF footprints correlate with the timings of TF activity throughout development. We additionally focus on the TF Dux, an important TF during ZGA, and use TOBIAS to unravel its target genes and influence on the global transcriptional network throughput PD.

Results

Impact of bias correction on footprint visibility. To validate the results of the TOBIAS method, we utilized 217 paired ChIP-seq/ATAC-seq datasets across four different cell types (GM12878, A549, HepG2, and K562). Here, the ChIP-seq peaks represent the true binding sites for each TF, which we used for validating the accuracy of the binding sites predicted by footprinting (see Supplementary Methods part 3).

As it has been shown that the Tn5 transposase has a large effect on footprinting[10], the first step of the TOBIAS footprinting pipeline is Tn5 bias correction. The TOBIAS bias correction module (named ATACorrect) utilizes a dinucleotide weight matrix (DWM)[13] to estimate the background bias of the Tn5 transposase (Fig. 1c). This DWM is used to calculate an expected Tn5 signal for each genomic region, representing the influence of the Tn5 bias (Fig. 1c; expected cutsites). Subtracting these expected cuts from the uncorrected signals yields a corrected track, highlighting the effect of protein binding (details are available in Supplementary Methods part 1). In order to evaluate the performance of TOBIAS in comparison to existing bias correction tools, we utilized the paired ChIP-seq/ATAC-seq data mentioned above to visualize aggregated footprints across bound and unbound subsets of TFBS. We found TOBIAS to outperform other bias correction tools in uncovering footprints and thereby distinguishing between bound/unbound sites (Supplementary Fig. 2a, Supplementary Data 1). Next, we wanted to quantify the depths of the aggregated footprints and utilized a footprint depth (FPD) metric as described by Baek et al.[12] (Supplementary Fig. 2b). In line with the visual impression, TOBIAS has the most significant difference in FPD between bound and unbound subsets of TFBS (Supplementary Fig. 2c). Importantly, the FPD's of unbound sites are minimally affected by bias correction, indicating that bias correction only uncovers footprints for truly bound sites.

Of note, the TOBIAS 'ATACorrect' method relies on the calculation of the expected Tn5 cuts based on the influence of Tn5 bias. Interestingly, besides identifying cases where the footprint was hidden by Tn5 bias (Supplementary Fig. 2d; JDP2), the track of expected signal also identifies TFs for which the motif itself disfavors Tn5 integration, thereby creating a false-positive footprint in uncorrected signals (Supplementary Fig. 2d; FOXD3). We wanted to investigate this effect in more detail and found that there is a high correlation between the footprint depths of uncorrected and expected Tn5 signals across all TFs, which vanishes after TOBIAS correction (Supplementary Fig. 2e). This observation demonstrates that bias correction effectively uncovers TF footprints, which were otherwise superimposed by Tn5 bias. It has previously been suggested that only 20% of all TFs leave measurable footprints[12], and we were able to confirm this observation using the uncorrected footprint depths and the same metric (Supplementary Fig. 2f; uncorrected). However, in contrast, we observed a measurable footprint for 59% of the TFs

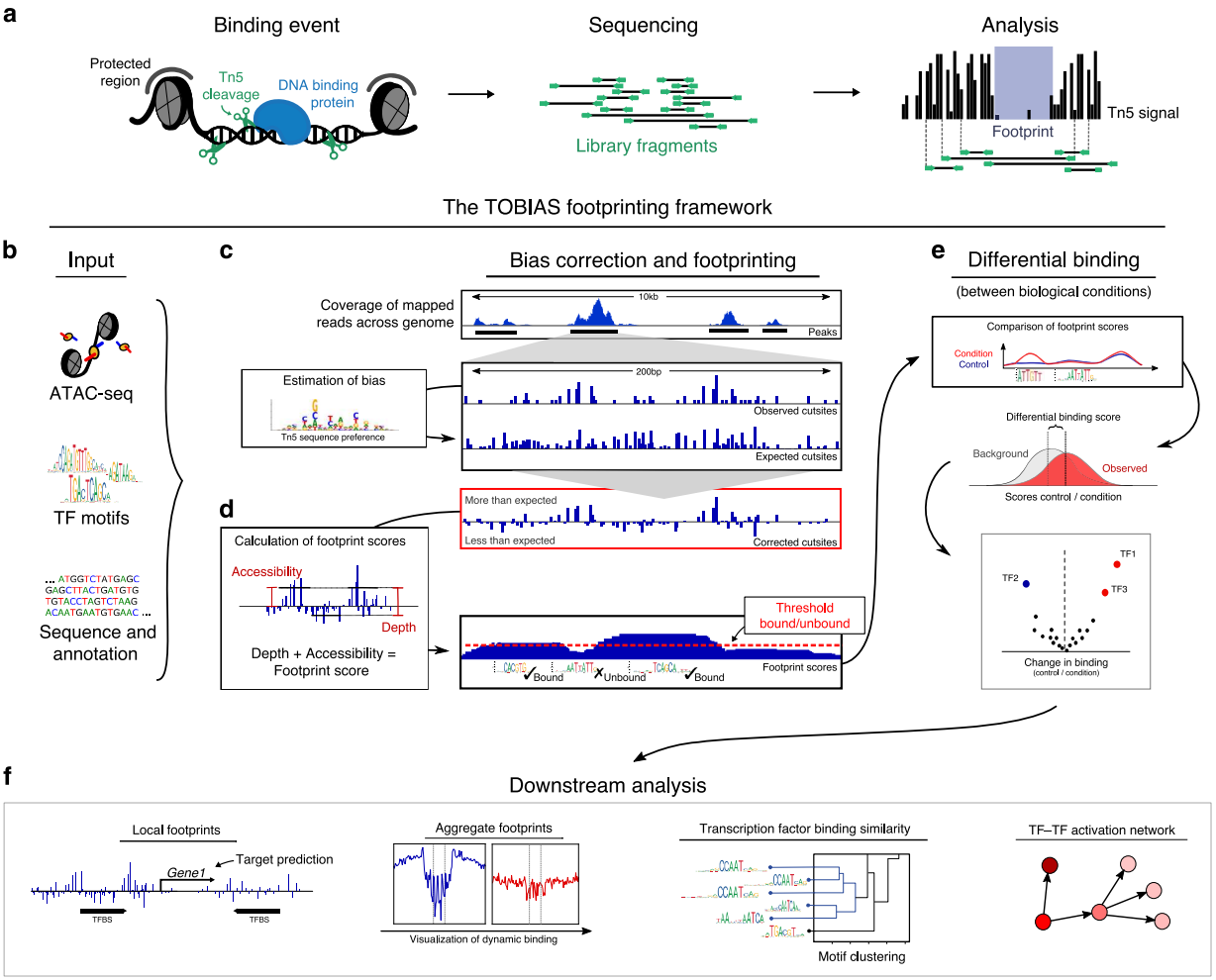


Fig. 1 The TOBIAS digital genomic footprinting framework. a The concept of footprinting using ATAC-seq. Tn5 transposase cleaves DNA and inserts sequencing adapters, but is unable to cut chromatin occupied by proteins such as nucleosomes (gray) and other DNA binding proteins e.g., transcription factors (blue). Sequencing libraries of DNA fragments are sequenced to yield reads (green). During analysis, each read is mapped to the genome and used to create a signal of single Tn5 insertion events (black bars), in which binding of protein is visible as depletion of the signal (defined as the footprint). **b** TOBIAS uses reads from ATAC-seq, transcription factor motifs and sequence annotation in standard formats as input. **c** Bias correction of Tn5 signal. In the first step, TOBIAS reads the observed Tn5 cutsites and estimates the underlying Tn5 sequence preference. TOBIAS then calculates the expected Tn5 cutsites per region, which represent the background probability of Tn5 insertion. Using the expected signal track, the Tn5 bias corrected cutsites are obtained (red box). **d** Footprinting to estimate transcription factor binding. The corrected cutsites enable calculation of footprint scores with a scoring function taking into account both the accessibility and depth of the local footprint (as depicted in the box labeled "Calculation of footprint scores"). This continuous footprint score is correlated with the presence of transcription factor binding sites in the genome, and a threshold is set to distinguish between bound and unbound sites. **e** Differential footprinting. If multiple conditions are investigated, the differential binding module summarizes individual site scores (upper black box) for each TF, and compares them between conditions (gray/red curve center) in order to define differentially bound TFs. Performed on all TFs under investigation, a volcano plot illustrates the global changes in transcription factor binding. **f** Additional analysis modules. After the main TOBIAS analysis, a variety of downstream analysis can be applied including visualization of local and aggregated footprints across conditions, comparison of binding specificity between individual transcription factors and TF network prediction.

when using the TOBIAS corrected signals (Supplementary Fig. 2f; corrected). As the fitted two-component model is a limited estimator to classify bound/unbound sites, we additionally calculated null distributions of randomized corrected footprints. By this approach, we similarly found the number of measurable footprints to be consistent at ~65% across all four cell types investigated (Supplementary Fig. 2g). This demonstrates that failure to correct for Tn5 bias can lead to false negative footprints, while bias correction uncovers the true amount of measurable footprints to be above 50%.

Validation of TOBIAS footprinting. For the task of protein binding prediction (i.e., footprinting), we collected four popular tools for ATAC-seq footprinting (HINT-ATAC, PIQ, Wellington, and msCentipede) and compared these to the individual TOBIAS framework features where applicable. While we found that some functionalities are overlapping between tools, we found a substantial set of features, such as differential footprinting for more than two conditions, to be exclusively covered by TOBIAS (Supplementary Table 1). Evaluating the results of each tool, we found that TOBIAS significantly outperformed the other de novo

tools HINT-ATAC, PIQ, and Wellington (Supplementary Fig. 3a) and performed equally well as msCentipede overall (Supplementary Fig. 3b). Notably, TOBIAS also showed robust performance across individual cell types (Supplementary Fig. 3c). Looking at individual TFs, TOBIAS outperforms msCentipede for factors such as CEBPB, which has a notable gain of footprints after Tn5 bias correction (Supplementary Fig. 3d), once again highlighting the advantage of taking Tn5 bias into account. Although msCentipede implements a motif centric learning approach, which can take TF specific binding patterns into account, it did not yield overall higher accuracy in comparison to TOBIAS. Additionally, the approach of building individual TF models took 300 times longer to compute than performing footprinting using TOBIAS (Supplementary Fig. 3e). Such learning approaches are therefore greatly limited in the number of TFs and conditions, which can realistically be analyzed.

Although we have shown that more than half of TFs create visible aggregated footprints, the footprints at individual loci are much more difficult to detect due to the sparsity of the ATAC-seq signal (as seen in Fig. 1f; Local footprints). In order to take this sparsity into account, we have designed the TOBIAS footprinting score as a combined score taking into account both depletion and accessibility (Supplementary Fig. 3f). In comparison, previous scoring methods such as the Footprint Occupancy Score (FOS)[14], calculate the difference in signal level between the background and the footprint (Supplementary Fig. 3g). To test the impact of this novel scoring approach, we compared the results of the TOBIAS (depletion + accessibility) score (as calculated from corrected cutsites) with the FOS score (pure depletion) (Supplementary Fig. 3h). While there is a limited improvement in the FOS footprinting score by using TOBIAS corrected cutsites, we found that there is a significant increase in predictive ability by using the TOBIAS score. This shows that, although bias correction is highly important for visualizing aggregated footprints, the influence of accessibility in the calculation of footprint scores is of considerable importance as well. Along this line, these findings illustrate the relationship between aggregated footprints and individual TFBS footprints. While the number of TFs with footprints in aggregated signals is above 50%, the proportion of individual TFBS supported by footprints might be considerably lower. Consequently, a score like FOS, which requires a footprint depletion for prediction, is inherently limited when predicting protein binding. In conclusion, we found that TOBIAS exceeded other tools in terms of uncovering footprints hidden by bias and correctly identifying bound TF binding sites. The improvement in accuracy is achieved by the alternative approaches for bias correction as well as by the novel footprinting score.

TF binding dynamics in mammalian ZGA. To demonstrate the potential of TOBIAS to predict differential TF binding across multiple conditions, in particular in the investigation of processes involving only few cells, we analyzed a series of ATAC-seq datasets derived from both human and murine preimplantation embryos at different developmental stages ranging from 2C, 4C, 8C to ICM in addition to embryonic stem cells of their respective species[15,16] (Fig. 2a). Altogether, TOBIAS was used to calculate footprint scores for a list of 590 and 464 individual TFs across the entire process of PD of human and mouse embryos, respectively. After clustering TFs into co-active groups within one or multiple developmental timepoints, we first asked whether the predicted timing of TF activation reflects known processes in human PD. Intriguingly, we found 10 defined clusters of specific binding patterns, the majority of which peaked between 4C and 8C, fully concordant with the transcriptional burst and termination of ZGA (Fig. 2b).

Two clusters of TFs (Cluster 1 + 2; $n = 83$) displayed highest activity at the 2–4C stage and strongly decreased thereafter, suggesting that factors within these clusters are likely involved in ZGA initiation. We set out to classify these TFs, and observed a high overlap with known maternally transferred transcripts[17] (LHX8, BACH1, EBF1, LHX2, EMX1, MIXL1, HIC2, FIGLA, SALL4, and ZNF449), explaining their activity before ZGA onset. Importantly, DUX4 and DUXA, which are amongst the earliest expressed TFs during ZGA[2,18], were also contained in these clusters. Additional TFs included HOXD1, which is known to be expressed in human unfertilized oocytes and preimplantation embryos[19] and ZBTB17, a TF mandatory to generate viable embryos[20]. Cluster 6 ($n = 67$) displayed a particularly prominent 8C specific signature, that harbored well-known TFs involved in lineage specification such as PITX1, PITX3, SOX8, MEF2A, MEF2D, OTX2, PAX5, and NKX3.2. Furthermore, overlapping TFs within Cluster 6 with RNA expression datasets ranging from the germinal vesicle to cleavage stage[2], 12 additional TFs (FOXJ3, HNF1A, ARID5A, RARB, HOXD8, TBP, ZFP28, ARID3B, ZNF136, IRF6, ARGFX, MYC, and ZSCAN4) were confirmed to be exclusively expressed within this time frame. Taken together, these data show that TOBIAS reliably uncovers massively parallel TF binding dynamics at specific timepoints during early embryonic development.

TF binding correlates with visible footprints. To confirm that TOBIAS-based footprinting scores are indeed associated with leaving bona fide footprints, we utilized the ability to visualize aggregated footprint plots as implemented within the framework. Indeed, bias corrected footprint scores were highly congruent with explicitly defined footprints (Fig. 2c) of prime ZGA regulators at developmental stages in which these have been shown to be active[3]. For example, footprints associated with DUX4, a master inducer of ZGA, were clearly visible from 2C–4C, decreased from 8C onwards and were completely lost in later stages, consistent with known expression levels[15] and ZGA onset in humans. Footprints for ZSCAN4, a primary DUX4 target[2], were exclusively visible at the 8C stage. Interestingly, GATA2 footprints were exhibited from 8C to ICM stages which is in line with its known function in regulating trophoblast differentiation[21]. As expected, CTCF creates footprints across all timepoints. Strikingly, we observed that these defined footprints were not detectable without TOBIAS-mediated Tn5 bias correction (Supplementary Fig. 4a). These data show that footprint scores can be reliably confirmed by footprint visualizations, which further allow to infer TF binding dynamics.

To test if the global footprinting scores of individual TFs correlate with the incidence and level of their RNA expression, we matched them to RNA expression datasets derived from individual timepoints throughout zygotic development, taking TF motif similarity into account. Indeed, we found that TOBIAS scores for the majority of TFs either correlated well with the timing of their expression profiles or displayed a slightly delayed activity after expression peaked (Supplementary Fig. 4b). This is important because it shows that in conjunction with expression data, TOBIAS can indicate the kinetics between TF expression (mRNA) and the actual binding activity of their translated proteins. The value of this added information becomes particularly apparent when analyzing activities of TFs that did not correlate with the timing of their RNA expression (Supplementary Fig. 4b; not correlated).

For example, within the non-correlated cluster 13 TFs were identified, which are of putative maternal origin[17] including SALL4. In mice, Sall4 protein is maternally contributed to the zygote, subsequently degraded at 2C and then re-expressed after

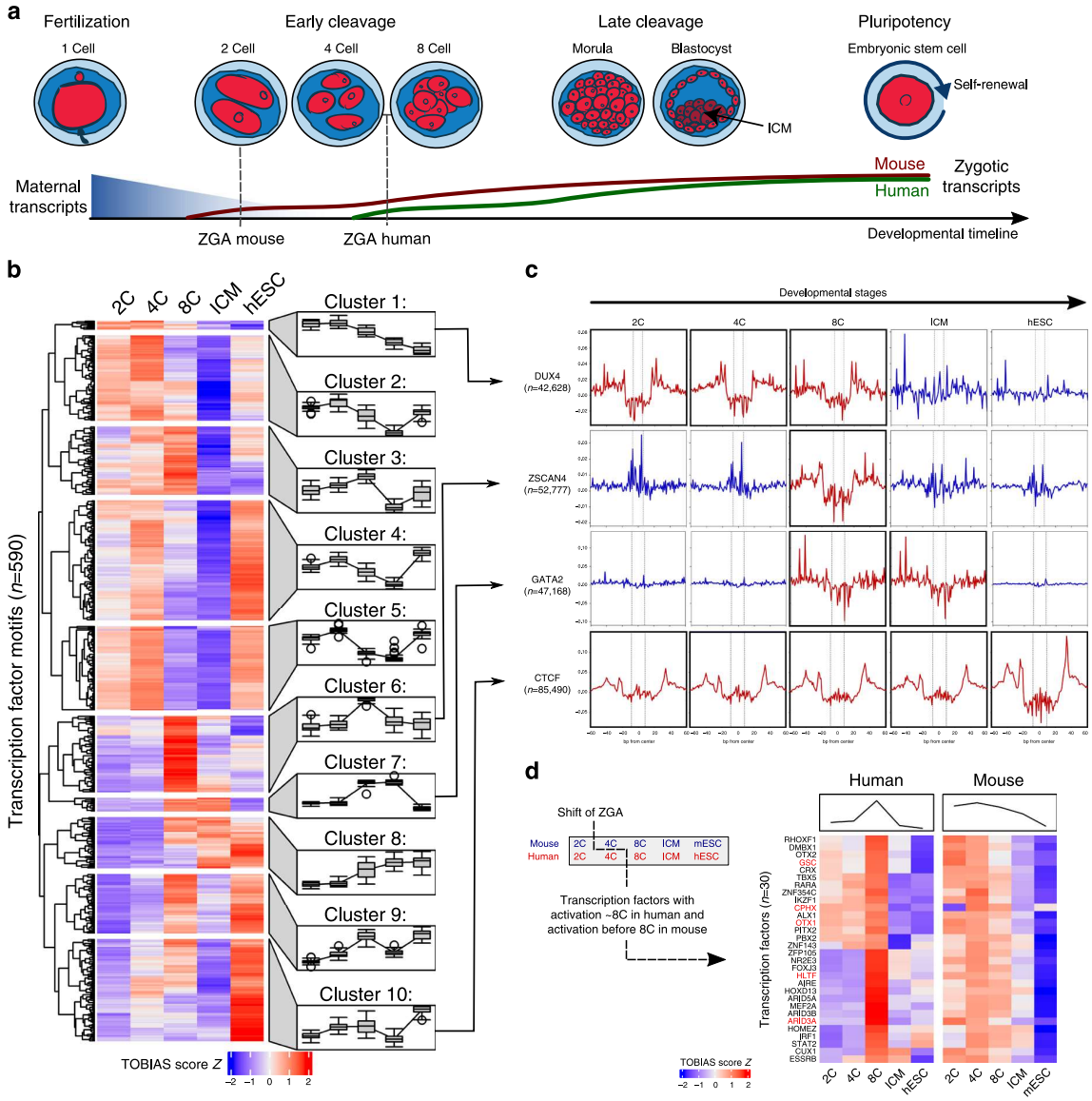


Fig. 2 Global changes in transcription factor binding throughout embryonic development. a Early embryonic development in human and mouse. While the fertilized egg undergoes a series of divisions, the maternal transcripts (blue curve) provided by the egg are depleted over time, and the zygotic genome is activated in waves (ZGA, red and green lines). ZGA initiates at the 2-cell stage in mouse and at the 4–8-cell stage in human. **b** Clustering of transcription factor activities throughout development. Each row represents one TF, each column a human developmental stage. TF activity scores from TOBIAS are *Z*-score transformed across rows. Blue color indicates low activity, red color indicates high activity. In order to visualize cluster trends, each cluster is associated with a mean trend line (left to right) and timepoint specific boxplots respectively. Source data are available in the Source Data file. **c** Bias corrected ATAC-seq footprints. For selected TFs with known roles in early development originating from four clusters (arrows from **b**), an aggregated footprinting plot matrix for all associated transcription factor binding sites is shown. Individual plots are centered around binding motifs (n = asterisk (*) relates to the number of binding sites). Rows indicate TFs DUX4, ZSCAN4, GATA2, and CTCF; columns illustrate developmental stages from left to right. Active binding of the individual TFs is visible as depletion in the signal around the binding site (highlighted in red). See Supplementary Figure 4a for corresponding uncorrected footprints. **d** TF activity onset in human and mouse. Heatmaps show activity of known ZGA-related TFs for human (left) and mouse (right) across matched timepoints 2C/8C/ICM/hESC (mESC). Transcription factors with known roles in ZGA are highlighted in bold red.

zygotic transcription has initiated[22]. Consistent with this, *SALL4* expression increases dramatically from 8C onwards (see Source Data file). In contrast to the expression values, TOBIAS predicted SALL4 to have the highest activity in 2C, with decreasing activity in 8C, which is in line with the presence of maternal SALL4 in the zygote. Comparing this change to all TF changes between 2C and 8C (log2 fold-changes estimated from TOBIAS activity scores), we find that SALL4 is at the 7th percentile of all changes ranked from decreasing to increasing, which is consistent with the degradation of the protein after the 2C stage. These data show that TOBIAS can provide significant insight into TF activities, in particular for those where determining their expression patterns alone does not suffice to explain when they exert their biological function.

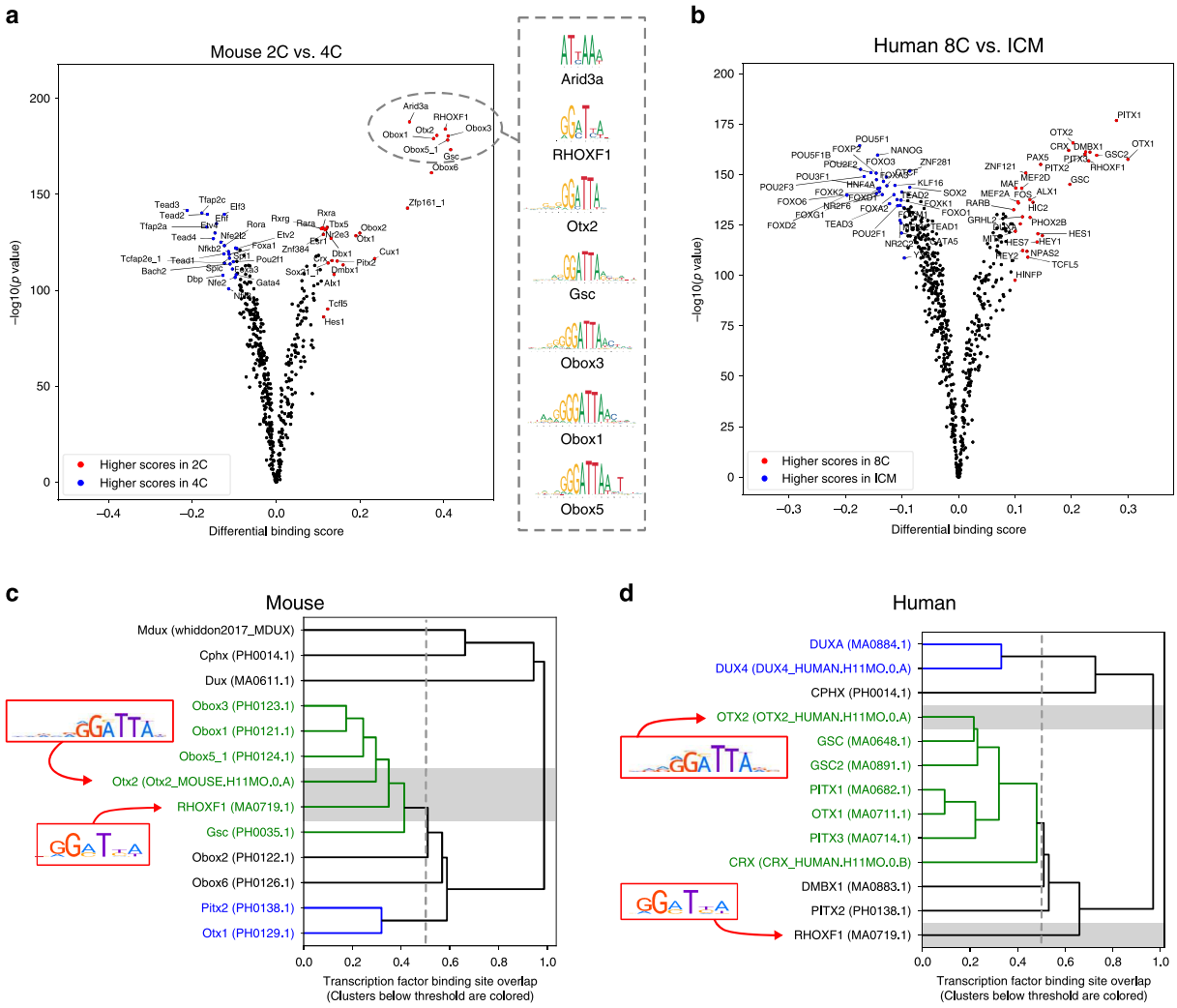


Fig. 3 Specification of ZGA functions between mouse and human. a, **b** Pairwise comparison of TF activity between developmental stages. The volcano plots show the differential binding activity against the −log10(*p* value) (both provided by TOBIAS) of all investigated TF motifs; each dot represents one motif. For **a** 2C stage specific TFs are labeled in red, 4C specific factors in blue. From the 2C specific TFs, seven prominent examples are chosen and illustrated by their motif. For **b** 8C stage specific TFs are labeled in red, ICM specific factors in blue. **c**, **d** Clustering of TF motifs based on binding-site overlap. Excerpt of the global TF clustering based on TF binding location, illustrating individual TFs as rows. The trees indicate genomic positional overlap of individual TFBS. A tree depth of 0.2 represents an overlap of 80% of the motifs. Each TF is indicated by name and unique ID in brackets. Clusters of TFs with more than 50% overlap (below 0.5 tree distance) are colored in green/blue. The position of TF motifs RHOXF1 and Otx2/OTX2 are highlighted. **c** shows overlap of motifs included in the mouse analysis. **d** shows clustering of human motifs. Complete TF trees are provided in Supplementary Notes 1 and 2.

Comparison of TF binding between human and mouse ZGA. The timing of ZGA varies between mice (2C) and humans (4C–8C) (reviewed in ref. [23]). By integrating the TOBIAS scores from human and mouse (Supplementary Fig. 4c), and instrumentalizing the capability of TOBIAS to generate differential TF binding plots for all timepoints automatically, we investigated similarities and differences of PD between these species. Firstly, reflecting the shift of ZGA onset, we identified 30 TFs, which appeared to be ZGA specific in both human and mouse (Fig. 2d), including OTX1, GSC, CPHX, and HLTF, which already have described functions within ZGA[4,24]. Moreover, this list also includes ARID3A, which has been shown to play a role in cell fate decisions in creating trophectoderm[25].

Next, we wanted to investigate specific differentially bound TFs, not only across the whole timeline, but also between individual conditions. We therefore utilized the differential TF binding plots created by TOBIAS, and chose to focus on the cellular transition initiated at and following ZGA, which corresponds to the transition between 2C and 4C in mice (Fig. 3a, Supplementary Note 1 for all pairwise comparisons), and between 8C and ICM in humans (Fig. 3b, Supplementary Note 2 for all pairwise comparisons). In mice, we observed a shift of Obox-factor activity in 2C to an activation of Tead (Tead1-4) and AP-2 (Tfap2a/c/e) motifs in 4C. Notably, AP-2/Tfap2c is required for normal embryogenesis in mice[26] and was also recently shown to act as a chromatin modifier that opens enhancers proximal to pluripotency factors in human[27]. We observed a similar shift of TF activity for homeobox factors such as PITX1-3, RHOXF1, CRX, and DMBX1 at the human 8C stage towards higher scores in ICM for known pluripotency factors, such as POU5F1 (OCT4) and other POU-factors.

Throughout the pairwise comparisons, we observed that TFs from the same families often display similar binding kinetics within species, which is not surprising since they often possess highly similar binding motifs (Fig. 3a; right). To characterize TF similarity, TOBIAS clusters TFs based on the overlap of TFBS within investigated samples (Fig. 3c, d). This enables quantification of the similarity and clustering of individual TFs that appear to be active at the same time. Thereby, we observed a group of homeobox motifs, which cluster together with more than 50% overlap of their respective binding sites in mouse (Fig. 3c). In contrast, other TFs such as Tead and AP-2 cluster separately, indicating that these factors utilize independent motifs (the full tree is found in Supplementary Note 1). While this might appear trivial, this clustering of TFs in fact also highlights differences in motif usage between human and mouse. One prominent example is the RHOXF1 motif, which shows high binding-site overlap with Obox 1/3/5 and Otx2 binding sites in mouse (Fig. 3c; ~60% overlap), but does not cluster with OTX2 in human (Fig. 3d; ~35% overlap). This observation could suggest important functional differences of RHOX/Rhox TFs between mice and humans. In support of this hypothesis, *RHOXF1*, *RHOXF2*, and *RHOXF2B* are exclusively expressed at 8C and ICM in humans, whereas Rhox factors are not expressed in corresponding developmental stages of preimplantation in mice (expression values are given in the Source Data file). Conceivably, this observation, together with the finding that murine Obox factors share the same motif as RHOX-factors in humans, suggests that Obox TFs might function similarly to RHOX-factors during ZGA. Altogether, the TOBIAS-mediated TF clustering based on TFBS overlap allows for quantification of target-similarity and divergence of TF function between motif families.

Dux expression induces massive changes in TF networks. Throughout the investigations of human and mouse development we became particularly interested in the Dux/DUX4 TF, which TOBIAS predicted to be one of the earliest factors to be active in both organisms (Fig. 2b and Supplementary Fig. 4c). Interestingly, despite the fact that Dux has already been proved to play a prominent role in ZGA[2], there is still a poor understanding of how Dux regulates its primary downstream targets, and consequently its secondary targets, during this process. We therefore applied TOBIAS to identify Dux binding sites utilizing an ATAC-seq dataset of *Dux* overexpression (*Dux*OE) in mESC[2].

As expected, the differential TF activity predicted by TOBIAS showed an increase in activity of Dux and Obox TFs, as well as Hltf, which was already highlighted to be common between mouse and human ZGA (Fig. 4a). Interestingly, this was accompanied by a massive loss of TF binding for pluripotency markers, such as Nanog, Pou5f1 (OCT4), and Sox2 upon *Dux*OE, indicating that Dux renders previously accessible chromatin sites associated with pluripotency inaccessible.

Consistently, Dux footprints (Fig. 4b; left) were clearly evident upon *Dux*OE. In comparison to existing bias correction methods, we found TOBIAS to be better at uncovering this footprint between Control and *Dux*OE conditions (Supplementary Fig. 5a). Importantly, TOBIAS additionally discriminated ~30% of all potential binding sites within open chromatin regions to be bound in the *Dux*OE condition (Fig. 4b; right). To rank the biological relevance of the individually changed binding sites between control and *Dux*OE conditions, we linked all annotated gene loci to RNA expression. A striking correlation between the gain-of-footprint and gain-of-expression of corresponding loci was clearly observed and mirrored by the TOBIAS predicted bound/unbound state (Fig. 4c). Among the genes within the list of bound Dux binding sites were well-known Dux targets including *Zscan4c* and *Pramef25*[28], for which local footprints for Dux were clearly visible (Fig. 4d). The high resolution of footprints is particularly pronounced for *Tdpoz1*, which harbors two potential Dux binding sites of which one is clearly footprinted in the score track, while the other is predicted to be unoccupied (Fig. 4d; bottom). In line with this, *Tdpoz1* expression is significantly upregulated upon *Dux*OE as revealed by RNA-seq (log2FC: 6,95). Consistently, *Tdpoz1* expression levels are highest at 2C and decrease thereafter, indicating that *Tdpoz1* is likely a direct target of Dux during PD both in vitro and in vivo. Footprinting scores also directly correlated with ChIP-seq peaks for Dux in the *Tdpoz1* promoter (Supplementary Fig. 5b), an observation which we also found at other positions (Supplementary Fig. 5c, d).

Many of the TOBIAS-predicted Dux targets encode TFs themselves. Therefore, we applied the TOBIAS network module to subset and match all activated binding sites to TF target genes with the aim of inferring how these TF activities might connect. Thereby, we could model an intriguing pseudo timed TF-activation network. This directed network predicted a TF-activation cascade initiated by Dux, resulting in the activation of 7 primary TFs which appear to subsequently activate 32 further TFs (first three layers depicted in Fig. 4e). As Dux is a regulator of ZGA, we asked how the in vitro activated Dux network compared to gene expression throughout PD in vivo. Strikingly, the in vivo RNA-seq data of the developmental stages[16] confirmed an early 2C specific expression of *Dux*, followed by a slightly shifted activation pattern for all direct Dux targets except for *Rxrg* (Fig. 4f). However, it is of note that *Rxrg* is significantly upregulated in the in vitro *Dux*OE from which the network is inferred (see Source Data for Fig. 4c), pointing to both the similarities and differences between the in vivo 2C and in vitro 2C-like stages induced by Dux. In conclusion, these data suggest that beyond identifying specific target genes of individual TFs, TOBIAS can promote biological insight by predicting entire TF-activation networks.

Dux targets repeat elements. Notably, many of the predicted Dux binding sites (40%) are not annotated to genes (Fig. 4g), raising the question what role these sites play in ZGA. Dux is known to induce expression of repeat regions such as long terminal repeats (LTRs)[2] and consistently, we found that more than half of the DUX-bound sites without annotation to genes are indeed located within known LTR sequences (Fig. 5a), which were transcribed both in vitro and in vivo (Fig. 5b; LTR). Interestingly, we additionally found that 28% of all non-annotated Dux binding sites overlap with genomic loci encoding LINE1 elements. Although LINE1 expression does not appear to be altered in mESC, there is a striking pattern of increasing LINE1 transcription from 4C–8C (Fig. 5b; LINE1) in vivo, pointing to a possible role of LINE1 regulation throughout PD. Finally, we found a portion of the Dux binding sites, which do not overlap with any annotated gene nor with putative regulatory repeat sequences, even though transcription clearly occurs at these sites (Fig. 5b; no overlap). One example is a predicted Dux binding site on chromosome 13, which coincides with a spliced region of increased expression between control mESC/*Dux*OE and comparable high expression in 2C, 4C, and 8C (Supplementary Fig. 6). These data suggest the existence of novel transcribed genetic elements, the function of which remains unknown, but which are likely controlled by Dux and may play a role during PD.

In conclusion, TOBIAS predicted the locations of Dux binding in promoters of target genes, and could propose how Dux initiates TF-activation networks and induces expression of repeat regions. Importantly, these data further show that TOBIAS

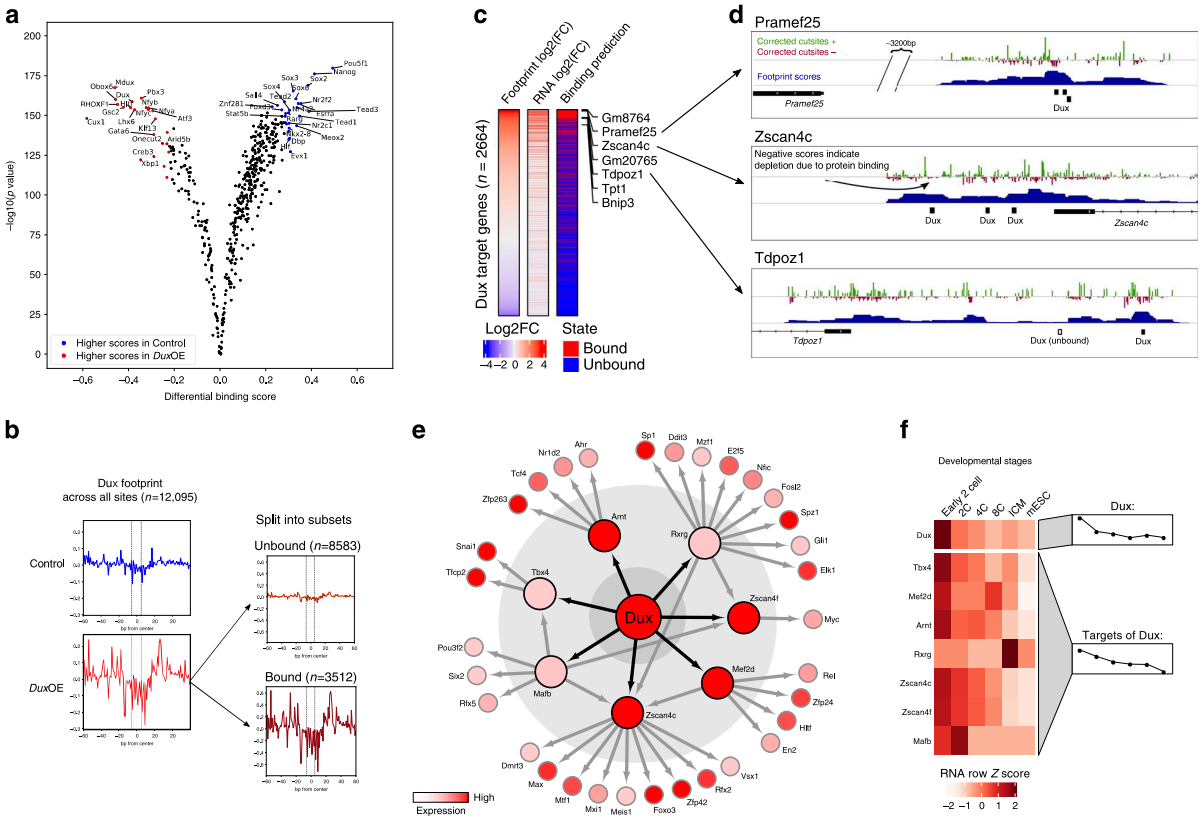


Fig. 4 Dux binding induces transcription at gene promoters. a Comparison of TF activities between mDux GFP- (Control; labeled in blue) and mDux GFP + (*Dux*OE; labeled in red). Volcano plot showing the TOBIAS differential binding score on the x-axis and -log10 (p value) on the y-axis; each dot represents one TF. **b** Aggregated footprint plots for Dux. The aggregated plots are centered on the predicted binding sites for Dux between Control and *Dux*OE conditions (left: all genomic sites). The total possible binding sites for *Dux*OE ($n = 12{,}095$) are separated into bound and unbound sites (right). The dashed lines represent the edges of the Dux motif. **c** Change in expression of genes near Dux binding sites. The heatmap shows $n = 2664$ Dux binding sites found in gene promoters. Footprint log2(FC) and RNA log2(FC) represent the matched changes between Control and *Dux*OE for footprints and gene expression, respectively. Log2(FC) is calculated as log2(*Dux*OE/Control). The column Binding prediction depicts whether the binding site was predicted by TOBIAS to be bound/unbound in the *Dux*OE condition. **d** Genomic tracks indicating three exemplary Dux binding sites and their target gene promoters and respective tracks for corrected cutsite signals (red/blue), TOBIAS footprint scores (blue), detected motifs (black boxes), and gene locations (solid black boxes with arrows indicating gene strand). **e** Dux transcription factor network. The TF-TF network is built of all TFBS with binding in TF promoters with increasing strength in *Dux*OE (log2(FC) > 0). Sizes of nodes represent the level of the network starting with Dux (Large: Dux, Medium: 1st level, Small: 2nd level). Nodes are colored based on corresponding RNA level in the *Dux*OE condition. Directed edges indicate binding sites in the respective gene promoter found by the TOBIAS CreateNetwork module. **f** Correlation of the Dux transcription factor network to expression during development. The heatmap depicts the in vivo gene expression during developmental stages. The right-hand group annotation highlights the difference in mean expression for each timepoint. The heatmap is split into Dux and target genes of Dux. Source data are available in the Source Data file.

predicts any TFBS with increased binding, not only those limited to annotated genes, which aids in uncovering novel regulatory genetic elements.

Discussion

To the best of our knowledge, this is the first application of a DGF approach to visualize gain and loss of individual TF footprints in the context of time series, TF overexpression, and TF-DNA binding for a wide-range of TFs in parallel. Importantly, we found that these advances could in large part be attributed to the framework approach we took in developing TOBIAS, which enabled us to simultaneously compare global TF binding across samples and quantify changes in TF binding at specific loci. The modularity of the framework also allowed us to apply a multitude of downstream analysis tools to easily visualize footprints and gain even more information about TF binding dynamics as exemplified by the prediction of the Dux TF-activation network.

The power of this framework to handle time-series data becomes especially apparent when integrating the TOBIAS-based prediction of TF binding with RNA-seq data from the same timepoints. For instance, TOBIAS predicted that the maternally transferred TF SALL4 is active in 2C, while its gene expression pattern alone suggests later activation. While SALL4 was one of the TFs with the largest decrease in binding from 2C to 8C, it is, however, also worth noting that since TFs have different baseline activities, large changes between timepoints can also arise from very low activity scores. Although the scores are normalized towards global TF activity, differences in the quality of footprinting (due to sample-specific biases) can also influence the prediction of differential TF binding between conditions, and this should be considered as a limitation of this method. In this context, it is tempting to speculate that TFs for which footprinting scores are low, even though their RNA expression is high, might act as transcriptional repressors, because footprinting relies on the premise that TFs will increase chromatin

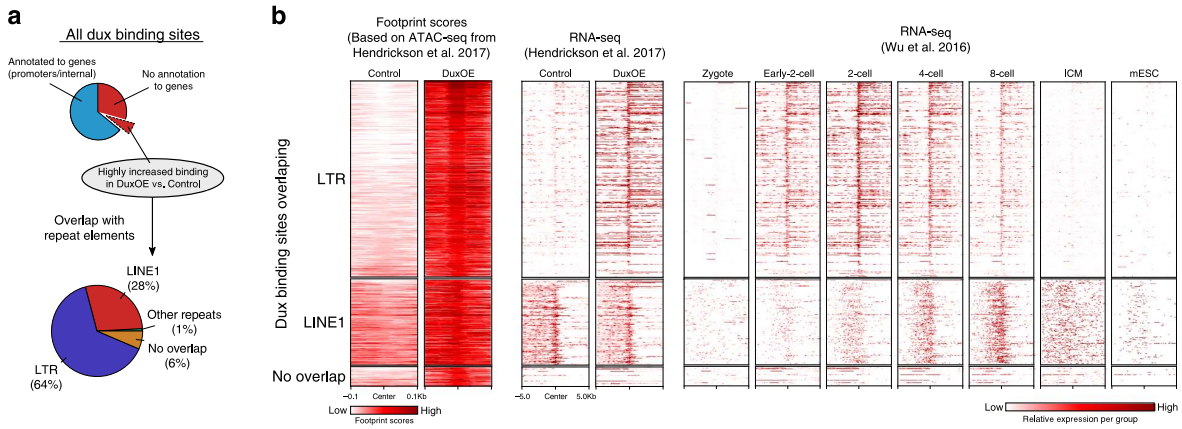


Fig. 5 Dux binding influences expression of repeat elements. a Dux binding sites overlap with repeat elements. All potential Dux binding sites are split into sites either overlapping promoters/genes or without annotation to any known genes (upper circle, blue/red). The bottom pie chart shows a subset of the latter, additionally having highly increased binding (log2(FC) > 1), annotated to repeat elements including LTR/LINE1 elements. **b** Dux induces expression of transcripts specific for preimplantation. Genomic signals for the Dux binding sites which are bound in *Dux*OE with log2(FC) footprint score >1 (i.e., upregulated in *Dux*OE) are split into overlapping either LTR, LINE1 or no known genetic elements (top to bottom); each row indicates one binding-site/associated gene loci. Footprint scores (±100 bp from Dux binding sites, left column) indicate the differential Dux binding between control and *Dux*OE (in vitro). RNA-seq shows the normalized read-counts from matched RNA-seq samples (center columns, in vitro) and throughout development (right columns, in vivo) within ±5 kb of the respective Dux binding sites. Dark red color indicates high expression.

accessibility around the binding site. In support of this hypothesis, recent investigations have suggested that repressors display a decreased footprinting effect in comparison to activators[29]. Therefore, the integration of ATAC-seq footprinting and RNA-seq is an important step in revealing additional information such as classification of TFs into repressors and activators, as well as the kinetics between expression and binding.

In the context of TF target prediction, we showed that TOBIAS could identify almost all known Dux targets. In addition to coding genes, our analysis disclosed novel Dux binding sites and significant footprint scores at LINE1 encoding genomic loci, which appear to be activated at the 4C/8C stage. This finding is especially interesting because a recent study has shown that LINE1 RNA can interact with Nucleolin and Kap1 to repress *Dux* expression[30]. Therefore, our findings suggest a kinetics driven model in which Dux not only initiates ZGA but also regulates its own termination by a temporally delayed negative feedback loop. How exactly this feedback loop could be controlled remains to be determined.

Despite the striking capability of DGF analysis, some limitations and dependencies of this method still remain. Among these is the need of high-quality TF motifs for matching footprint scores to individual TFs with high confidence. In other words, while the binding of a TF might create an effect that can be interpreted as a footprint, without a known motif, this effect cannot be matched to the corresponding TF. It also needs to be noted that footprinting analysis cannot take effects into account that arise from heterogeneous mixtures of cells wherein TFs are bound in some cells and in others not. Therefore, if not separated, the classification of differential binding will be an observation averaged across many cells, possibly masking subpopulation effects. Recent advances have enabled the application of ATAC-seq in single cells, but this generates sparse matrices, rendering footprinting approaches on single cells elusive. However, we speculate that by creating aggregated pseudo-bulk signals from large clustered single-cell ATAC datasets, DGF analysis might also become possible in single cells.

In conclusion, we present TOBIAS as the first comprehensive software that performs all steps of DGF analysis, natively supports multiple experimental conditions and performs visualization within one single framework. Although we utilized the process of PD as a proof of principle, the modularity and universal nature of the TOBIAS framework enables investigations of various biological conditions beyond PD. We believe that continued work in the field of DGF, including advances in both software and wet-lab methods, will validate this method as a resourceful tool to extend our understanding of a variety of epigenetic processes involving TF binding.

Methods

Processing of ATAC-seq data. Raw sequencing fastq files were assessed for quality, adapter content and duplication rates with FastQC v0.11.7, trimmed using cutadapt[31] and aligned with STAR v2.6.0c[32] (parameters: --alignEndsType EndToEnd --outFilterMismatchNoverLmax 0.1 --outFilterScoreMinOverLread 0.66 --outFilterMatchNminOverLread 0.66 --outFilterMatchNmin 20 --alignIntronMax 1 --alignSJDBoverhangMin 999 --alignEndsProtrude 10 ConcordantPair --alignMatesGapMax 2000 --outMultimapperOrder Random --outFilterMultimapNmax 999 --outSAMmultNmax 1) to either the mouse or human genome using Mus_musculus.GRCm38 or Homo_sapiens.GRCh38 versions from Ensembl[33]. Accessible regions were identified by peak calling for each sample separately using MACS2 (parameters: --nomodel --shift -100 --extsize 200 --broad)[34]. Peaks from each sample were merged to a set of union peaks across all conditions using bedtools merge. Each union peak was annotated to the transcriptional start site of genes (GENCODE[35]) in a distance of −10000/+1000 from the gene start using UROPA[36].

Processing of RNA-seq data. Raw reads were assessed for quality, adapter content and duplication rates with FastQC v0.11.7, trimmed using cutadapt[31] and aligned with STAR v2.6.0c[32] (parameters: --outFilterMismatchNoverLmax 0.1 --outFilterScoreMinOverLread 0.9 --outFilterMatchNminOverLread 0.9 --outFilterMatchNmin 20 --alignIntronMax 200000 --alignMatesGapMax 2000 --alignEndsProtrude 10 ConcordantPair --outMultimapperOrder Random --outFilterMultimapNmax 999) to either the mouse or human genome using Mus_musculus.GRCm38 or Homo_sapiens.GRCh38 versions from Ensembl[33].

Processing of ChIP-seq data. Raw sequencing files in fastq format were quality assessed by Trimmomatic by trimming reads after a quality drop below a mean of Q15 in a window of five nucleotides[37]. All reads longer than 15 nucleotides were aligned versus the mouse genome version mm10, keeping just unique alignments (parameters: --outFilterMismatchNoverLmax 0.2 --outFilterScoreMinOverLread 0.66 --outFilterMatchNminOverLread 0.66 --outFilterMatchNmin 20 --alignIntronMax 1 --alignSJDBoverhangMin 999

--outFilterMultimapNmax 1 --alignEndsProtrude 10 ConcordantPair) by using the STAR mapper[32]. Read deduplication was done by Picard (http://broadinstitute.github.io/picard/).

Processing of TF motifs. TF motifs were downloaded from JASPAR CORE 2018[38], the JASPAR PBM HOMEO collection and Hocomoco V11[39] databases. We further included the human ARGFX_3 motif from footprintDB[40] which originates from a HT-SELEX assay[41]. In addition to the Dux/Dux4 motifs of JASPAR and Hocomoco, we also included two TF motifs for Dux/DUX4 created using MEME-ChIP[42] with standard parameters on the ChIP-seq peaks of[28] (GSE87279).

JASPAR motifs were linked to Ensembl gene ids by mapping the provided Uniprot id to the Ensembl gene id through biomaRt[43]. Hocomoco motifs were likewise linked to genes through the provided HGNC/MGI annotation. Due to the redundancy of motifs between JASPAR and Hocomoco, we further filtered the TF motifs to one motif per gene, preferentially choosing motifs originating from mouse/human, respectively. For each TOBIAS run, we created sets of expressed TFs as estimated from RNA-seq in the respective conditions. This amounted to 590 motifs for the dataset on human preimplantation stages, 464 motifs for the dataset on mouse preimplantation, and 459 for the *Dux*OE dataset.

Maternal genes. Maternal genes for human and mouse were downloaded from the REGULATOR database[17]. Entrez gene ids were converted to Ensembl gene ids using biomaRt[43] and subsequently matched to available TF motifs as previously explained.

Overlap of Dux binding sites with repeat elements. Repeat elements for mm10 were downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath/mm10/database/rmsk.txt.gz). Overlap of Dux sites to individual repeat elements was performed using bedtools intersect. The sum of overlaps were counted per repeat class (LINE1/LTR).

Visualization. All TF-score heatmaps were generated by R Version 3.5.3 and ComplexHeatmap package version 3.6[44]. Individual gene views were generated by loading TOBIAS output tracks into IGV version 2.6.2[45] or using the TOBIAS PlotTracks module, which is a wrapper for the svist4get visualization tool[46]. TF networks were drawn with Cytoscape version 3.7.1[47]. Heatmaps of genomic signal density were generated using Deeptools version 3.3.0[48]. All other figures, such as footprint plots, volcano plots and motif clustering dendrograms were generated by the TOBIAS visualization modules.

The TOBIAS framework. Details on the TOBIAS algorithms and framework setup are found in the Supplementary Methods part 1 and 2.

Comparison of TOBIAS to existing methods. Details on the validation and comparison of TOBIAS to existing methods for bias correction and footprinting are found in the Supplementary Methods part 3.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The source data for Figs. 2b, 4c, f, Supplementary Figs. 2c, e, f, g, 3, 4b, c, as well as expression values for *Rhox* and *Obox* genes throughout human and mouse development are available in the Source Data file. Raw ATAC-seq and RNA-seq data for human and mouse embryonic development are available from GEO under the accessions GSE66390 (mouse) and GSE101571 (human). Raw ATAC-seq, RNA-seq, and ChIP-seq data from *Dux* overexpression experiments are available from GEO under the accession GSE85632. Data for validation are available from ENCODE as explained in Supplementary Methods. Excerpts of the TOBIAS analysis results are accessible for dynamic visualization at: http://loosolab.mpi-bn.mpg.de/tobias-meets-wilson. UCSC track hubs (for viewing in the UCSC genome browser) of corrected Tn5 and footprint signals are available at: https://genome.ucsc.edu/cgi-bin/hgTracks?hubUrl=https://s3.mpi-bn.mpg.de/data-tobias-ucsc/hub.txt&genome=mm10 and https://genome.ucsc.edu/cgi-bin/hgTracks?hubUrl=https://s3.mpi-bn.mpg.de/data-tobias-ucsc/hub.txt&genome=hg38 for mouse and human respectively. All data are available from the authors upon reasonable request.

Code availability

The TOBIAS software is publicly available at GitHub (https://github.com/loosolab/TOBIAS) and can additionally be obtained through PyPI and Bioconda.

Received: 17 February 2020; Accepted: 23 July 2020;
Published online: 26 August 2020

References

1. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
2. Hendrickson, P. G. et al. Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nat. Genet.* **49**, 925–934 (2017).
3. Eckersley-Maslin, M. A. et al. MERVL/Zscan4 network activation results in transient genome-wide DNA demethylation of mESCs. *Cell Rep.* **17**, 179–192 (2016).
4. Madissoon, E. et al. Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Sci. Rep.* **6**, 28995 (2016).
5. Hesselberth, J. R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
6. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
7. Sung, M. H., Baek, S. & Hager, G. L. Genome-wide footprinting: ready for prime time? *Nat. Methods* **13**, 222–228 (2016).
8. Vierstra, J. & Stamatoyannopoulos, J. A. Genomic footprinting. *Nat. Methods* **13**, 213–221 (2016).
9. Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol.* **20**, 42 (2019).
10. Li, Z. et al. Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45–45 (2019).
11. Tripodi, I. J., Allen, M. A. & Dowell, R. D. Detecting differential transcription factor activity from ATAC-seq data. *Molecules* **23**, 1136 (2018).
12. Baek, S., Goldstein, I. & Hager, G. L. Bivariate genomic footprinting detects changes in transcription factor activity. *Cell Rep.* **19**, 1710–1722 (2017).
13. Siddharthan, R. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PLoS ONE* **5**, e9722 (2010).
14. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
15. Wu, J. et al. Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature* **557**, 256–260 (2018).
16. Wu, J. et al. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652–657 (2016).
17. Wang, K. & Nishida, H. REGULATOR: a database of metazoan transcription factors and maternal factors for developmental studies. *BMC Bioinforma.* **16**, 114 (2015).
18. De Iaco, A. et al. DUX-family transcription factors regulate zygotic genome activation in placental mammals. *Nat. Genet.* **49**, 941–945 (2017).
19. Adjaye, J. & Monk, M. Transcription of homeobox-containing genes detected in cDNA libraries derived from human unfertilized oocytes and preimplantation embryos. *Mol. Hum. Reprod.* **6**, 707–711 (2000).
20. Adhikary, S. et al. Miz1 is required for early embryonic development during gastrulation. *Mol. Cell Biol.* **23**, 7648–7657 (2003).
21. Home, P. et al. Genetic redundancy of GATA factors in the extraembryonic trophoblast lineage ensures the progression of preimplantation and postimplantation mammalian development. *Development* **144**, 876–888 (2017).
22. Xu, K. et al. Maternal Sall4 is indispensable for epigenetic maturation of mouse oocytes. *J. Biol. Chem.* **292**, 1798–1807 (2017).
23. Svoboda, P. Mammalian zygotic genome activation. *Semin. Cell Dev. Biol.* **84**, 118–126 (2018).
24. Tohonen, V. et al. Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development. *Nat. Commun.* **6**, 8207 (2015).
25. Rhee, C. et al. ARID3A is required for mammalian placenta development. *Dev. Biol.* **422**, 83–91 (2017).
26. Winger, Q., Huang, J., Auman, H. J., Lewandoski, M. & Williams, T. Analysis of transcription factor AP-2 expression and function during mouse preimplantation development. *Biol. Reprod.* **75**, 324–333 (2006).
27. Pastor, W. A. et al. TFAP2C regulates transcription in human naive pluripotency by opening enhancers. *Nat. Cell Biol.* **20**, 553–564 (2018).
28. Whiddon, J. L., Langford, A. T., Wong, C. J., Zhong, J. W. & Tapscott, S. J. Conservation and innovation in the DUX4-family gene network. *Nat. Genet.* **49**, 935–940 (2017).
29. Berest, I. et al. Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF. *Cell Rep.* **29**, 3147–3159 (2019).
30. Percharde, M. et al. A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell* **174**, 391–405 (2018).

31. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 3 (2011).
32. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
33. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
34. Feng, J. X., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
35. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2018).
36. Kondili, M. et al. UROPA: a tool for Universal RObust Peak Annotation. *Sci. Rep.* **7**, 2593 (2017).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
38. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D1284 (2018).
39. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2017).
40. Sebastian, A. & Contreras-Moreira, B. footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* **30**, 258–265 (2013).
41. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
42. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
43. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184 (2009).
44. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
45. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
46. Egorov, A. A. et al. svist4get: a simple visualization tool for genomic tracks from sequencing experiments. *BMC Bioinforma.* **20**, 113 (2019).
47. Shannon, P. et al. Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
48. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

Acknowledgements

We would like to thank the IT-group at MPI-BN for continued support with IT-infrastructure. We would also like to thank Marius Dieckmann, the administrator of the Kubernetes cluster in the deNBi project at JLU (https://cloud.denbi.de/giessen/), for his support and help in implementing the TOBIAS-Nextflow Cloud version. This work was funded by the Max Planck Society, the German Research Foundation (DFG), grant KFO309 (project number 284237345, epigenetics core unit) to M.L., DZHK Rhine-Main Site, and by the Cardio-Pulmonary Institute (CPI), EXC 2026, Project ID: 390649896 to M.L. Open access funding provided by Projekt DEAL.

Author contributions

M.B., C.K., J.K., and M.L. wrote the manuscript. M.B. developed the TOBIAS software. M.B., P.G., H.S., A.P., K.K., R.W., A.F., and J.P. performed additional bioinformatics analysis. T.B., J.K., and M.L. directed, coordinated, and supervised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-18035-1.

Correspondence and requests for materials should be addressed to M.L.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

2.2 Publication 2: TF-COMB - discovering grammar of transcription factor binding sites

This publication presents the TF-COMB (Transcription Factor Co-Occurrence using Market Basket analysis) framework for investigating co-occurrence of TFBS and other epigenetic features. As implied in the name, TF-COMB relies on the market basket analysis, which is primarily utilized to investigate shopping habits of customers, e.g. if a customer buys cereal, they are also likely to buy milk. Similarly, TF-COMB uses a modified market basket analysis to investigate the co-occurrence relationships of TFs, e.g. if TF1 binds, TF2 is also likely to bind. In contrast to earlier methods, the strength of TF-COMB is its ability to work with any type of input data in the form of genomic regions, which enables the use of TFBS predicted by footprinting methods, such as TOBIAS.

Using 1663 TF ChIP-seq experiments across 8 cell lines, TF-COMB was able to predict a number of commonly co-occurring TF pairs including MAX-MYC, NFYA-NFYB and CTCF-ZNF143. Extending this data with locations of open chromatin from ATAC-seq revealed that individual TFs have locational preferences within open chromatin. Data from 3D chromatin maps also revealed that a smaller group of individual TFs, including CTCF, co-occur within interacting regions. Likewise, co-occurrence analysis of TFs with histone modifications uncovered an interesting set of TFs which are devoid of binding in the promoter-associated H3K9ac, and instead bind preferentially in regions marked by H3K27ac and H3K4me3.

While ChIP-seq data provide binding positions with high precision, one of its limitations is the lack of resolution and orientation of each TFBS. TF-COMB allows to read TFBS from TOBIAS footprinting output, and with the increased resolution from motif-based sites, TF-COMB identified characteristics of grammar in terms of distance and orientation for individual TF pairs. This analysis identified TF pairs with preferred distance, of which some exhibit more than one peak. Interestingly, a number of these preferred distances could be verified by ChIP-seq data as highlighted by the pair THAP11-ZNF143.

Lastly, TF-COMB provides functionality to analyze TF co-occurrences in a network context, as TFs and co-occurrences can be treated as nodes and edges, respectively. This network analysis uncovered clusters of TFs with high connectivity, which could be mapped to known protein complexes and gene ontology terms. Finally, the analysis demonstrates that the majority of networks exhibit a powerlaw-distribution of node degree, meaning that the networks are robust against perturbations.

TF-COMB is a highly flexible tool for studying TF co-occurrence, binding grammar and cooperativity networks from a variety of input sources. An overview of the individual contributions of the thesis author to the publication is found in Table 2.

Table 2: Contributions by the thesis author to publication 2.

Area	**Contributions**
Conceptualization	Contributed to the definition of the project goals.
Software	Developed and implemented the main classes of the TF-COMB software. Supervised the creation of individual analysis modules.
Data	Downloaded and processed ChIP-seq and ATAC-seq datasets from ENCODE. Obtained protein-protein interactions from STRINGDB and protein sequences from Uniprot.
Analysis	Performed validation and network analysis using TF-COMB with ChIP-seq data in integration with known protein-protein interactions and protein similarity. Calculated co-occurrence between TFs and histone modifications, open chromatin and HIC-peaks. Compared ChIP-seq with motif-derived co-occurrences.
Visualization	Created figures 2, 3, 4a-b, 4e-g, 5a-b, 6 and supplementary figures S2, S3, S4 and S6. Supervised the creation of the remaining figures.
Manuscript	Contributed to the writing of the manuscript draft and editing of the final manuscript.

The full article is found in the following pages and the supplementary figures are found in Appendix A2.

Computational and Structural Biotechnology Journal 20 (2022) 4040–4051

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

TF-COMB – Discovering grammar of transcription factor binding sites

Mette Bentsen [a], Vanessa Heger [a], Hendrik Schultheis [a], Carsten Kuenne [a], Mario Looso [a,b,*]

[a] Bioinformatics Core Unit (BCU), Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany
[b] Cardio-Pulmonary Institute (CPI), Bad Nauheim, Germany

ARTICLE INFO

Article history:
Received 14 June 2022
Accepted 12 July 2022
Available online 21 July 2022

Keywords:
Transcription factors
Co-occurrence
Grammar
Networks
Gene regulation

ABSTRACT

Cooperativity between transcription factors is important to regulate target gene expression. In particular, the binding grammar of TFs in relation to each other, as well as in the context of other genomic elements, is crucial for TF functionality. However, tools to easily uncover co-occurrence between DNA-binding proteins, and investigate the regulatory modules of TFs, are limited. Here we present TF-COMB (Transcription Factor Co-Occurrence using Market Basket analysis) - a tool to investigate co-occurring TFs and binding grammar within regulatory regions. We found that TF-COMB can accurately identify known co-occurring TFs from ChIP-seq data, as well as uncover preferential localization to other genomic elements. With the use of ATAC-seq footprinting and TF motif locations, we found that TFs exhibit both preferred orientation and distance in relation to each other, and that these are biologically significant. Finally, we extended the analysis to not only investigate individual TF pairs, but also TF pairs in the context of networks, which enabled the investigation of TF complexes and TF hubs. In conclusion, TF-COMB is a flexible tool to investigate various aspects of TF binding grammar.

1. Introduction

Epigenetic regulation of gene expression is crucial for all processes occurring in cell biology, including cell differentiation and cell development among others. In this context, transcription factors (TFs) are a well-known class of DNA-binding proteins regulating gene expression, operating in direct proximity to target genes, as well as through regulatory enhancer elements further up- or downstream. While some TFs are known to be specific for a certain cell type or a dedicated cellular process, in many cases, combinations of multiple TFs are needed to elicit a specific response - a concept known as TF co-occurrence (Fig. 1a). In fact, the amount and types of involved TFs varies, and in higher eukaryotes, it has been shown that 10–15 TF binding sites (TFBS) are required to target a specific gene [1], which can differ in terms of the interaction mode. In the context of enhancers, different models of TF-enhancer interactions have been described [2], which include the enhanceosome, the billboard and the TF collective models (Fig. 1b). Of these, the enhanceosome is defined by a certain arrangement, order, distance, affinity and/or relative orientation of the given TFBS. Such rules for binding are known as *TF binding grammar* (please see [3] for a review). In contrast, the billboard and the TF collective models allow for a more flexible grammar where TFBS arrangement and TF binding states are not restricted. In these models, protein–protein interactions (PPI) can enable co-occurrence of TFs through dimerization and binding of additional TF co-factors, but also allow for TFs to co-occur without direct physical interactions. Additionally, the binding of TFs is influenced by other factors including the presence of histone modifications, chromatin accessibility and 3D chromatin organization. A smaller subset of TFs, known as pioneers, are even able to influence chromatin accessibility by opening previously closed chromatin regions, thereby initiating and enabling subsequent binding of other factors [4]. Thus, a higher level of enhancer organization is a prominent factor in gene regulation.

The detection and analysis of TF binding grammar is complex, as the number of known human TFs is reported to be ~1600 [5], genes are reported in the range of ~22000 protein-coding genes [6], and enhancers are reported in the range of millions [7]. Moreover, the collection of active TFs, as well as the chromatin organization, differs for each single cell type. To detect TF co-occurrence, one can broadly distinguish between experimental and computational methods. In the case of experimental methods, the de-facto gold standard in the field are chromatin immunoprecipitation based assays followed by deep sequencing (e.g. ChIP-seq and CUT&RUN-seq). These assays utilize TF specific antibodies and

* Corresponding author at: Bioinformatics Core Unit (BCU), Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany.
E-mail address: mario.looso@mpi-bn.mpg.de (M. Looso).

https://doi.org/10.1016/j.csbj.2022.07.025

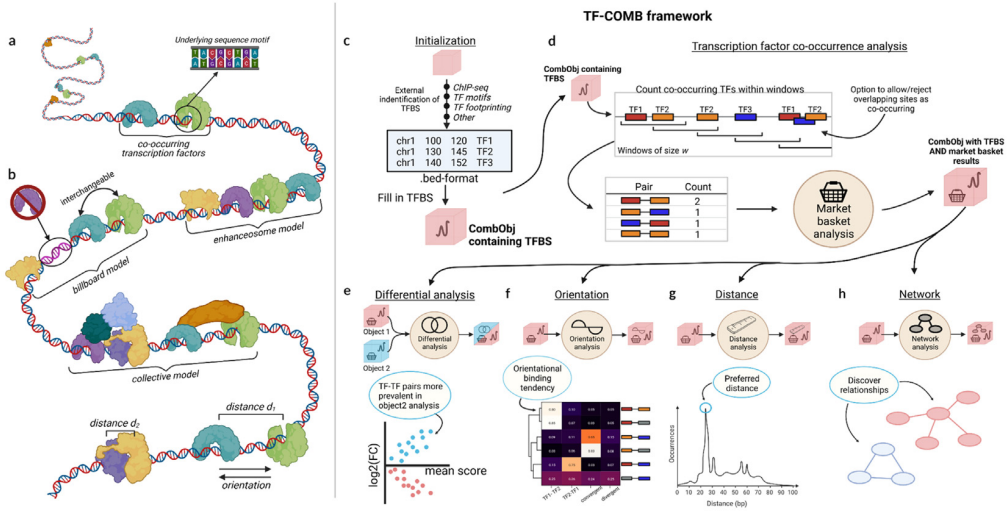


Fig. 1. Grammar of regulatory elements and the TF-COMB framework. a) Concept of two co-occurring transcription factors (green + blue) bound in immediate proximity directly to the DNA. b) Models of TF-enhancer interactions and TF binding characteristics. The enhanceosome is defined by strict positioning of TFs, whereas the billboard allows for interchanged positions (green + blue) and absent TF factors (purple). The collective model allows TFs to bind on top of other factors (dark orange, light blue, dark green). TF pairs also exhibit additional characteristics, such as preferred binding distance (d_1 and d_2), as well as binding in different orientations on the DNA. Drawn with inspiration from [2]. c-h) The TF-COMB framework: c) Initialization of a TF-COMB object (red square) by providing TFBS and regions of interest from any data origin (e.g. ChIP-Seq, footprinting, ATAC-peaks). d) Co-occurrences are identified, counted and analyzed with an adapted market basket analysis, and are stored in the object for further analyses. e) Differential analysis module allows for the comparison of two independent TF-COMB objects. The module visualizes data to indicate TF pairs more frequent in Object1 (red dots) and Object2 (blue dots) respectively. f) The orientational binding module calculates strand specificity of TF pairs and visualizes preferences via heatmaps. g) TF pairs are analyzed in context of their binding distance, and pairs with prefered binding distance are classified and visualized as histograms. h) Network analysis and visualization module allows to identify higher order relationships between TFs and/or other features. All subfigures were created with BioRender.com. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are performed in one cell type at a time. While it is technically possible to run these assays multiple times on different factors of interest, the cost and the dependency on highly specific antibodies generally render these assays unsuitable to investigate global TF co-occurrence. Due to these limitations, the topic of co-occurring TFs have preferably been investigated by *in silico* methods, which utilize a variety of statistical methods [8], linguistic models [9] and enrichment-based algorithms [10]. Most of these methods are based on association analysis of two TFs, and thus simplify the complexity of co-occurring TFs to one specific pair at a time. Additionally, while most of the available tools perform TF motif searches to screen the genome for potential co-occurring events, they are mostly restricted to a single TF anchor point, derived from e.g. ChIP-seq. These limitations motivated us to design a tool that would enable global TF co-occurring analysis independent of data origin. In this context, chromatin accessibility data such as ATAC-seq is of high interest, as we and other groups have recently shown that ATAC-seq can be utilized to find TF binding sites via genome-wide TF footprinting [11,12]. In addition, co-occurrence of TFs with histone modifications, locations of genes, and other genomic elements are likewise important to accommodate when analyzing TF binding as an epigenetic mechanism. Here we introduce the TF-COMB (Transcription Factor Co-Occurrence using Market Basket analysis) framework, which uses an enhanced market basket analysis to identify and investigate the grammar of co-occurring TFs from a variety of data sources.

2. Approach of TF-COMB

In order to detect co-occurring TF binding sites, TF-COMB utilizes an association analysis known as market basket analysis (MBA) [13]. This method has classically been applied to identify shopping habits, also known as association rules, such as "if the customer buys cereal, they are likely to buy milk". The same approach can also be applied to TF co-occurrence analysis as "if TF1 binds, it is also likely that TF2 binds", for which the association rule is called a TF co-occurring pair. However, the classical MBA carries a number of shortcomings in the context of TF binding data. For example, the classical MBA reduces items occurring multiple times per transaction to one, and would thus mask the effects of robustness by TFBS replicates in biological networks. In addition, previous applications of MBAs to TF data have likewise excluded all overlapping TFs [14], which influences the discovery of TFs binding in dimers and complexes, where the binding sites might overlap. Moreover, in the context of binding grammar, the order and orientation of TF binding is also of high importance, which has not been taken into account in previous implementations of MBAs. Thus, the current state of publicly available software lacks support for certain aspects needed to characterize the effects of TF co-occurrence. The TF-COMB framework is intended to overcome these limitations.

Firstly, TF-COMB counts all co-occurring TFs in a predefined genomic window. As the source of binding data is flexible, the analysis can be performed for both ChIP-seq peaks, motif positions, and footprints, as well as other input regions. Rather than setting up non-overlapping genomic regions, TF-COMB utilizes sliding windows beginning at each given TFBS, allowing to count all co-occurring TFs relative to that position. A default window size of 100 bp was used throughout this paper, however, the parameters of the framework can be changed in order to control the size of the window (minimum and maximum distance), whether to count TFs more than once, as well as basic grammar parameters of TFs such as the inclusion of directionality and strandedness of TF pairs. Thus, the framework is very flexible in terms of investigating a variety of aspects of co-occurring TFs.

Depending on the input data, the total number of TF binding sites and the resulting number of TF combinations can be immense, which gives rise to the need for a measure to rank biologically or statistically important TF combinations. Using the information of pairwise counted TF1/TF2 pairs, different association metrics are supported. The classical MBA utilizes the support, confidence and lift measures to filter and rank interesting associations, however, additional scoring schemes exist, such as the cosine association score [15]. Of note, there is no established method of choosing the correct threshold for association scores of MBA [16]. To overcome this limitation, TF-COMB calculates a Z-score of significance, which helps to reduce false-positives from TFs which are ubiquitously present across the whole genome. This is done through a null-model of random co-occurrences, which is calculated by shuffling the TF labels, rather than randomly shuffling the positions across the genome, as TFBS positions naturally appear in clusters [17]. In summary, the TF-COMB supported metrics allow to rate and select a subset of TF co-occurrences of interest.

3. Materials and methods

3.1. Availability and implementation

TF-COMB is a Python package intended to be used as a toolbox within Jupyter notebooks or within custom analysis scripts. Due to the high computational needs, it is supported by C-code through Cython [18] integration in Python. Additionally, given functions support multiprocessing when applicable.

TF-COMB is open source and freely available on github at: https://github.com/loosolab/TF-COMB. Details on the individual TF-COMB modules are given at: https://tf-comb.readthedocs.io.

3.2. Sequencing data

We obtained TF ChIP-seq peaks, histone ChIP-seq peaks, RNA-seq and ATAC-seq for cell lines HepG2, K562, HEK293, GM12878, MCF-7, H1, A549 and HeLa-S3 from ENCODE [19]. The cell lines were chosen based on a requirement of at least 50 unique TF ChIP-seq experiments available. In case of more than one available experiment per cell line and/or ChIP target, the most recent experiment was used. For ChIP-seq, all peaks were centered at the peak summit and reduced to 1 bp regions, while peaks overlapping blacklisted regions were excluded. For ATAC-seq, individual replicates were merged per experimental condition. Accession numbers for all ENCODE datasets used are given in Supplementary Table 1.

Additionally, we obtained genomic coordinates for HiC anchor regions for cell line GM12878 (GEO accession GSE63525; file "GSE63525_GM12878_primary + replicate_HiCCUPS_looplist.txt") [20].

3.3. Transcription factor motifs

Motifs were obtained from the JASPAR database (JASPAR 2022 CORE vertebrates) [21]. Annotated dimers (e.g. "Ahr::Arnt") and any additional motif variations for the same TF (e.g. "var.2") were excluded.

3.4. Metadata for TF interactions

For validation of TF pairs, we obtained a variety of metadata. Known PPIs were obtained from Biogrid [22]. Protein sequences for individual TFs were obtained from UniProt [23] and pairwise protein similarity was calculated using EMBOSS Stretcher [24]. Literature association of TFs was calculated by querying PubMed abstracts and titles for the common presence of each TF pair. The global counts of publications containing either TF1, TF2, and/or TF1 + TF2 were used to calculate the cosine similarity measure representing the PubMed association score. Transcription factor families were obtained from AnimalTFDB (v3.0) [25] and using this classification, TF pairs were estimated to be either same-family or different-family pairs. GO-term analysis of TF clusters were performed using the goatools python package [26].

3.5. TOBIAS footprinting analysis

Footprinting analysis was performed with the TOBIAS pipeline [11] with default parameters. The input peaks, ATAC-seq reads and motifs were obtained from ENCODE and JASPAR respectively, as explained above. In order to reduce the effect of repetitive elements on the co-occurrence analysis, we used the masked hg38 genome [27].

3.6. Comparison to existing tools

In order to compare TF-COMB to existing computational tools, we performed a literature search and obtained 12 *in silico* tools for investigation of co-occurring TFs, taking webservices and command line usage into account (Supplementary Table 2). Namely, these 12 tools were CENTDIST [28], iTFs [29], INSECT 2.0 [30], TICA [31], NAUTICA [32], PC-Traff [9], SpaMo [10], COPS [33], TACO [8], MCOT [34], CisMiner [35] and coTRaCTE [36]. Of these, most tools were discarded from comparison due to different reasons including unreachable weblinks. Briefly for all accessible methods, PC-Traff is limited to predefined motifs, CisMiner does not provide a functional example of the expected input, COPS is limited to *Drosophila melanogaster* and *Mus musculus* genomes, TACO needs at least two replicate experiments to run and coTRaCTE needs differentially regulated chromatin regions from multiple cell types. The remaining two tools, SpaMo and MCOT, were used for an exemplary analysis based on ChIP data from ENCODE (cell line GM12878; Supplementary Table 1) on 74 TFs and hg38 genome version. We recorded the total runtime on a VM with 64 GB RAM and 8 cores at 2.6 GHz CPU for each tested tool individually. Identified TF pairs were ranked for each tool independently. Resulting lists of co-occurring TFs were aligned where applicable, and top ten exclusively found pairs per tool were manually evaluated via literature search in the PubMed database [37].

4. Results

4.1. TF-COMB: A universal tool to investigate grammar of enhancers

The typical workflow of a TF-COMB-based analysis is presented in Fig. 1c-h. Briefly, the analysis starts with the initialization of an TF-COMB object with regions of interest. As genomic positions in standardized BED file format are supported, TF-COMB can handle (but is not limited to) binding sites from ChIP-seq, pre-calculated motif positions, histone modifications, locations of genes, enhancers, and open chromatin peaks (Fig. 1c). In the next step, the genomic positions of the TF-COMB object are internally processed by a sliding window approach and an adjusted MBA is calculated in order to identify TF combinations (Fig. 1d). At this stage, the framework provides a variety of analysis and visualization methods, as well as the functionality to compare different conditions with each other (Fig. 1e). In order to further examine the TF combination data, the TF-COMB tools provide functionality to investigate TF binding grammar, which includes binding orientation (Fig. 1f) and binding distance (Fig. 1g), as well as the opportunity to investigate TF pair networks (Fig. 1h).

In order to rate the features and performance of TF-COMB, we reviewed 12 existing implementations for TF co-occurrence prediction (Supplementary Table 2; methods segment 3.6). Only tools classified as comparable in functionality and able to run the test data were accepted for assessment. We identified two command-line tools, MCOT and SpaMo, as suitable for comparison with TF-COMB. Where applicable, we aligned individual tool functionality, and we identified a substantial set of features to be exclusively covered by TF-COMB (Supplementary Table 3). Briefly, TF network functionality, binding entity classification (orientation, distance) and quantitative support between conditions of TF-COMB render our software to be a significant extension of existing tools. In order to compare the quality of the results, we ran an exemplary analysis on public ChIP-data and validated the result to known interacting TFs from the BioGrid database [22]. Using a receiver operating characteristic (ROC) curve, we found that TF-COMB has the best predictive ability, with MCOT ranking second best (Supplementary Fig. 1a). As expected, we found that SpaMo has a low ability to predict co-occurring TFs, as this tool is particularly focused on identifying motif spacing and not necessarily general motif co-occurrence. In addition to the ROC analysis, we manually rated the top ten candidates per tool via literature search, and found considerably more TF-COMB specific TF pairs verified by literature than for the other tools (Supplementary Table 4). In terms of runtime, we found TF-COMB to outperform the other tools, even though it generates an all-against-all analysis instead of using a single anchor TF (Supplementary Fig. 1b). In conclusion, we present TF-COMB as a novel tool for the investigation of TF co-occurrence and TF binding grammar.

4.2. TF-COMB detects co-occurring TFs from ChIP-seq data

In order to illustrate the basic functionality of TF-COMB to detect co-occurring TFs, we have utilized the collection of high quality ChIP-seq datasets deposited by the ENCODE project [19,38]. We collected a total of 1663 ChIP-seq experiments across 8 human cell lines (HepG2, K562, HEK293, GM12878, MCF-7, H1, A549, HeLa-S3) (Supplementary Table 1), and used TF-COMB to find TF associations.

By subsetting pairs based on cosine score and significance (Z-score) across all 8 cell lines, we were able to specify a total of 1938 (1877 unique) TF-TF co-occurring pairs, which correspond to 1–3% of all pairs per cell line (Fig. 2a). Within the individual cell lines, TF-COMB predicted the top co-occurring TF pairs to be the well-known pairs MAX-MYC [39] in MCF-7, AP-1 (FOSL2-JUNB) [40] in A549, and CTCF-ZNF143 [41] in H1 cells (Fig. 2b-c; Supplementary Table 5). Interestingly, besides highlighting significantly co-occurring pairs, the analysis is also informative in terms of establishing seemingly anti-*co*-occurring sites, as a negative Z-score represents TFs with less co-occurrences than expected. In MCF-7, such pairs included CTCF-FOS, CTCF-GATA3 and CTCF-CEBPB, indicating that CTCF possibly avoids binding to certain partners (Fig. 2b). The reason for this might be related to CTCF's involvement in chromatin looping, as other TFs carry out separate functionalities, which should not interfere with chromatin organization. Thus, the TF-COMB co-occurrence analysis highlights both preferred and unpreferred TF pairs.

Because the list of available ChIP-seq experiments differs for each cell line (Supplementary Fig. S2a), with the only TFs available in all cell lines being CTCF and REST, it is not possible to directly compare the top co-occurring pairs across cell lines. For co-occurring TF pairs present in at least 4 cell lines, we found NFYA-NFYB, USF1-USF2, JUN-JUND, E2F6-MAX and CTCF-ZNF143 pairs among others to have the highest median association scores across multiple cell lines (Fig. 2d), indicating TF pairs of general importance. In confirmation of this result, we found all these pairs to either form a complex or to interact with each other [41–44]. In addition, by comparing association scores between cell lines, we found that the median spearman correlation is 0.49, suggesting the presence of both common, as well as cell-line specific co-occurring TF-pairs (Supplementary Fig. S2b-c). Thus, we conclude TF-COMB to reliably identify known co-occurring TFs across different biological conditions.

In order to further validate the identified ChIP-based co-occurring TF-pairs within individual cell lines, we asked whether the found pairs recapitulate known PPIs from the BioGrid database [22]. Even though physical interactions are not a necessity for co-occurrence, we found that the majority of cell lines exhibit significantly higher association scores for the TF pairs with a known PPI in comparison to other pairs (Fig. 2e; Supplementary Fig. S2d). However, all cell lines also contain cases of high-association scores without known PPI. To ensure that the default window size of 100 bp is sufficient to catch potential PPIs, we ran TF-COMB iteratively with increasing window sizes and correlated the results of each distance to the known PPIs. We found that while most physically interacting TF-TF pairs are collected at distances between 10 and 50 bp (Supplementary Fig. S2e), associations are still found for larger windows (Supplementary Fig. S2f), meaning that there are other types of TF-TF co-occurrences than those explained by PPI. In this context, we observed that many of the top TF-COMB predictions without a known PPI are TFs from the same families such as FOXA1-FOXA2, MAFF-MAFK and SOX5-SOX13. With the purpose of ruling out cross-reactive ChIP-seq antibodies as the source of this effect, we correlated the association scores with the protein similarity within each TF pair. While we did find some examples of simultaneous high-similarity and high association score, there is no global effect of protein similarity on the co-occurrence analysis (Supplementary Fig. S2g). Finally, to explain the association of low similarity TF pairs without known PPI, we also investigated the association of PubMed terms from literature. This analysis could confirm that some pairs, such as BATF-IRF4 were previously described, despite not being annotated with a known PPI in BioGrid (Fig. 2f). However, this analysis still leaves a number of pairs including ARID5B-FOXA3, which has high co-occurrence association, but low PubMed association (Fig. 2f; lower right). For these, more in-depth analysis will be needed in order to confirm the biological mechanisms of their observed association. In conclusion, we regard TF-COMB as a powerful tool to identify co-occurring TFs, both those physically interacting and those applying other modes of co-occurrence.

4.3. Integration of epigenetic marks reveals positional identity of TFs

Besides the expression of certain TFs in individual cell types, multiple epigenetic processes play a role in TF binding, such as the accessibility of chromatin, the presence of histone marks and the 3D chromatin organization. Thus, we sought to use TF-COMB to investigate the co-occurrence of TFs with other epigenetic signals by extending our ChIP-seq co-occurrence analysis with transcriptional co-factors and other DNA-binding proteins (e.g. DNA polymerase II), positions of known histone marks, positional information of genes, chromatin loop anchors from HiC data and open chromatin regions from ATAC-seq. While not all DNA-binding factors in this analysis are strictly identified as TFs, we will still use this term for simplicity.

First, we characterized TFBS in the context of chromatin accessibility and gene promoters. Although gene promoters make up only ~3% of the entire genome, we found that the majority of factors have promoter association in the range of 10–20%, supporting the enrichment of TF binding to directly regulate gene expression (Fig. 3a). In line with TF binding in enhancer regions, we find that the majority of TF binding sites are located in open chromatin

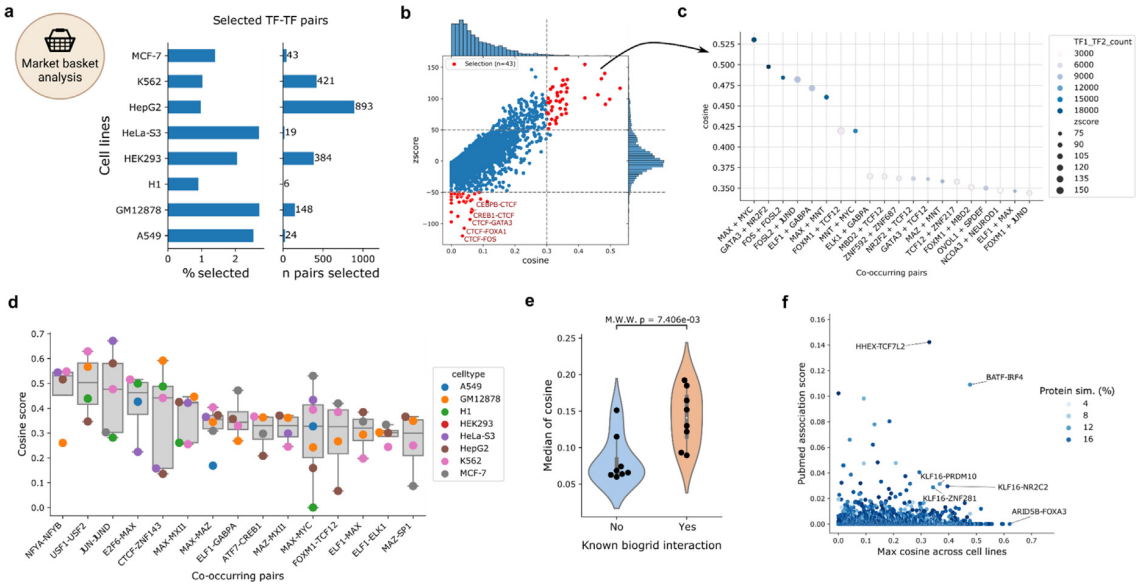


Fig. 2. Co-occurrence of TF ChIP-seq peaks in ENCODE cell lines. a) Cell lines and their highly co-occurring TF pairs in percentage and numbers. b) Relationship between cosine and Z-score for each TF pair within cell line MCF-7. The upper right section (red) indicates pairs selected as highly co-occurring. The lower left section contains pairs with less-than-expected co-occurrence. c) Selected co-occurring TF pairs sorted by cosine score. Color indicates the number of TF1-TF2 occurrences, size illustrates the Z-score of the pair. d) The TF pairs identified to have the highest cosine across different cell lines. Only TF pairs present in at least 4 cell lines were included. Each point indicates the cousine score of the given TF pair (column), distribution of scores across cell lines is indicated by a boxplot. e) Violin plot of median cosine score for all TF interactions with (right) or without (left) known protein–protein interactions across the 8 cell lines. The significance of the difference in distributions is calculated using the Mann-Whitney *U* test. f) Correlation of maximum cosine (across cell lines) and known PubMed association score for TF pairs without known protein–protein interactions. Pairs with high cosine or high PubMed association are highlighted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

outside of promoter regions. However, we also observed that the percentage of sites found in open chromatin is ranging from ~95% for POU5F1 to below 10% for CBX8, which highlights the differences in functionality of individual TFs and co-factors. Interestingly, we found a number of known pioneer factors including PBX1, GATA3 and FOXA3 to be less associated with open chromatin than other factors, which reflects their ability to also bind to closed chromatin.

By performing a distance analysis between the TFBS and annotated genes, we found a strong enrichment of binding sites close to the TSS, with 80% of these binding sites occurring within 2000 bp (Fig. 3b). In contrast, when we investigated the more global TF binding patterns in the context of open chromatin, we found that the TF binding sites show locational preferences. As TFs are known to bind to open chromatin, it was not surprising that the vast majority of TFs are addressing the center region (+/- 25%) of ATAC-seq peaks, such as shown for ATF2, FOS, ELK1 and the histone deacetylase HDAC2 (Fig. 3c). However, we also found TFs with their binding sites located without preference across the whole open region (EHF; Fig. 3d), as well as some TF candidates with a preference to the outer bounds of open chromatin regions (CHD4; Fig. 3e). This localization of CHD4 is well explained by the fact that CHD4 has been shown to slide nucleosomes, which are found at the borders of open chromatin [45]. Overall, the relative locations of TFBS in open chromatin peaks showed a significant correlation between cell lines, which confirms that there are groups of TFs and co-factors with locational preferences regardless of cell type (Supplementary Fig. S3a).

Within the group of TFs located at the center of ATAC-seq regions, we also observed CTCF, which has a highly centered peak around the +/- 15% core of the peak (Fig. 3f). As CTCF is known to mediate chromatin looping [46], we used TF-COMB to investigate the co-occurrence of ChIP-seq defined TF sites with chromatin loop anchor regions as defined by HiC. As expected, we found that CTCF had high co-occurrence with loop regions, and we additionally found SMC3, TRIM22, RAD21, ZNF143, YY1 and RUNX3 (Fig. 3g). These results are perfectly in line with previous investigations showing that CTCF, RAD21, ZNF143, TRIM22 and RUNX3 can accurately predict the position of chromatin loops [47]. Additionally, SMC3 is an essential component of cohesin [48], and YY1 has been shown to mediate 3D chromatin interactions in collaboration with CTCF [49]. As seen for CTCF, these proteins likewise show strong positional specificities within open chromatin (Supplementary Fig. S3b). As such, the relative positioning of these factors might be an important mechanism for higher chromatin organization.

Finally, we investigated the higher order binding patterns of TFs in the context of activating and repressing histone modifications (Fig. 3h). Since the association of histone modifications with TFs is not necessarily symmetrical, we used the association ‘confidence’ score, which represents the fractional overlap between sites. Firstly, we confirmed the association of respective histone marks to chromatin, and not surprisingly, we found the active histone marks H3K4me1/2 and H3K27ac to have the highest association with open chromatin, and in contrast, the repressive histone marks H3K36me3 and H3K9me1/2/3 with the lowest association (Supplementary Fig. S3c). Correspondingly, we found that H3K4me2 and H3K27ac have the highest overall association with ChIP-seq defined TF targets (Fig. 3i), which has also been described previously [50]. In contrast, the repressing marks H3K27me3, H3K9me1/2/3, H3K36me3, H4K20me1 and histone variant H3F3A had a low overall association with TF binding. However, despite the minimal association with open chromatin and TF binding, we identified factors such as EZH2, CBX2/3/8, ZNF184, MCM3/5, XRCC3, ZNF280A, SRSF9 and PLRG1 to be prominently overlapping with H3K9me3 and H3K27me3, while simultaneously being depleted for association with active histone marks (Fig. 3i). Thus, these proteins have an ability to

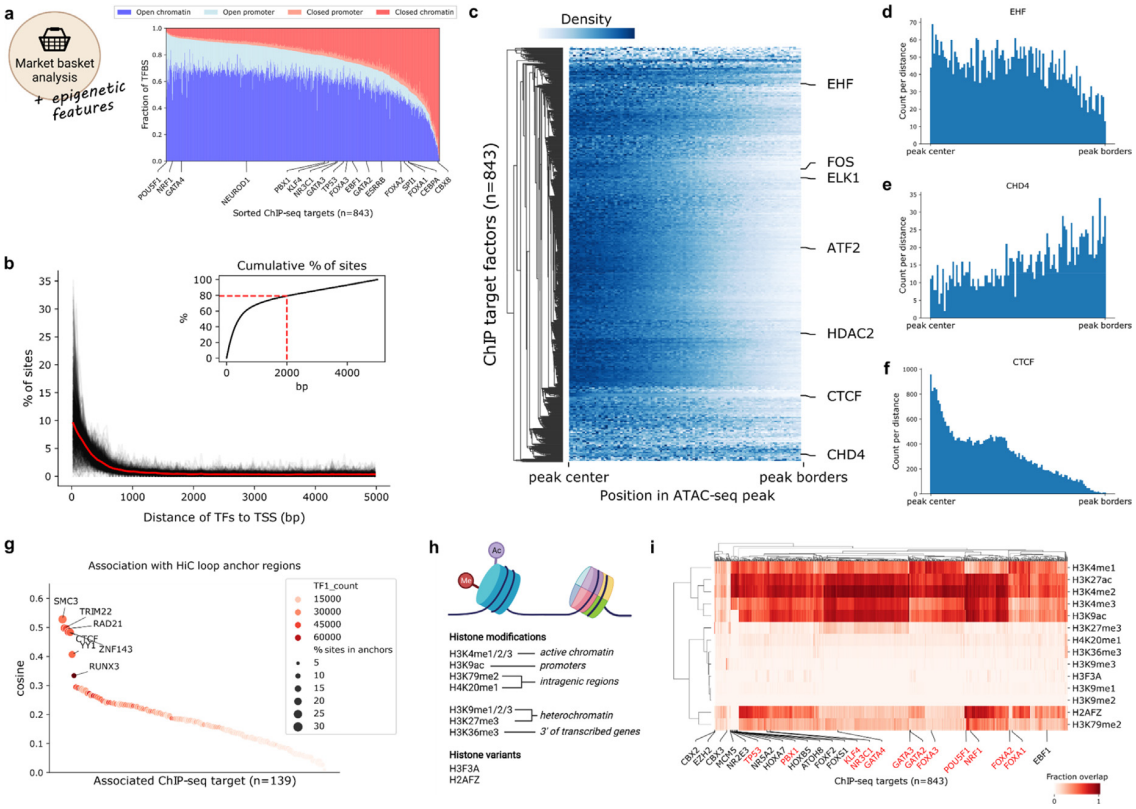


Fig. 3. Integration of epigenetic marks reveals positional identity of TFs and co-factors. a) Percentage of ChIP-seq peaks (y-axis) in open/closed (defined by overlap with ATAC-seq peaks) promoter regions (defined as 5000 bp upstream of the TSS) and chromatin regions (all regions not defined as promoter) for individual factors (x-axis). b) Distance of ChIP-seq peak summits to transcription start sites (TSS) of genes. Distributions for individual factors are shown in black, and the mean distance is shown in bold red. Upper right corner shows the cumulative distribution of sites with 80% of sites marked with a dashed line. c) Relative location of ChIP-seq peak summits in open chromatin regions. Counts for the same TF in different cell lines were merged by taking the mean at each distance. The colorbar represents the scaled number of positions found to co-localize at different percentages of the peak length. Counts to the left/right of the peak center are aggregated to a range from center to border. d-f) Relative TF binding positions in ATAC-seq peaks from peak-center (left) to outer peak-borders (right). g) Co-occurrence of TFs and HiC loop regions in the cell line GM12878. TF1_count represents the total count of ChIP-seq peaks per factor. h) Scheme of histone modifications and histone variants investigated. Created with BioRender.com. i) Co-occurrence of TFs and histone marks as calculated by fractional overlap of regions. A selection of TFs from interesting clusters are annotated on the x-axis. Known pioneer factors are marked in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

bind in otherwise inactive chromatin, which the majority of other proteins do not. Indeed, a number of these TFs are members of the Polycomb Group of proteins (namely EZH2 and CBX family proteins), which assemble in multi-protein complexes to repress genes.

Interestingly, this analysis also highlighted another prominent cluster of TFs, which is defined by an overall strong association with active histone marks, but with an exclusive depletion of association with H3K9ac and H3K79me2, which are markers for active promoters and intragenic regions respectively. This cluster contains several HOX and FOX factors, nuclear receptors NR2E3 and NR5A2, as well as PBX1, ATOH8 and TP53 among others. Many of these factors are known to be pioneer factors (TP53 and PBX1) or part of families with many known pioneers (nuclear receptors and the FOX family) [51]. However, while some other known pioneers also show a decrease in association with H3K9ac (e.g. FOXA1/2), it is not an universal rule (e.g. NRF1), and the pioneer hypothesis is therefore not the only explanation for the depletion of H3K9ac for this cluster. Alternatively, the effect could be explained by the role of these factors in controlling lineage specification, as is well described for the HOX factors [52]. Thus, the discovery of TFs specifically co-occurring or restricting binding to certain histone modifications, can uncover hallmarks of TF binding to enhancers.

In conclusion, we find that TF-COMB analysis integrating epigenetic signatures uncovers DNA-binding proteins with locational specificity corresponding to individual biological functions.

4.4. Co-occurrence analysis utilizing TF footprinting

While we were able to use gold standard ChIP-seq data to identify positional locations of TFs within larger regulatory regions, this data comes with some fundamental challenges in the context of investigating local binding grammar. Mainly, TF ChIP-seq peaks are several hundred base pairs wide, and thus do not clearly indicate the exact location of TF binding sites. As a result, ChIP-seq will generally fail to find multiple TF sites from one factor in close proximity, and will lose the information of TF binding orientation, which impedes the investigation of a higher order of TF binding grammar from ChIP-seq data. In contrast, the identification of TFBS through methods such as motif prediction or digital genomic footprinting requires only one chromatin accessibility assay per cell type to estimate binding events for hundreds of TFs in parallel, while preserving location and orientation of the TFBS (Fig. 4a). Thus, we obtained ATAC-seq experiments for cell lines A549, GM12878, HepG2, K562, MCF-7 from ENCODE, and ran our previously published ATAC-seq footprinting pipeline TOBIAS on the data [11]. The pipeline identifies bound TFBS on the basis of Tn5 inser-

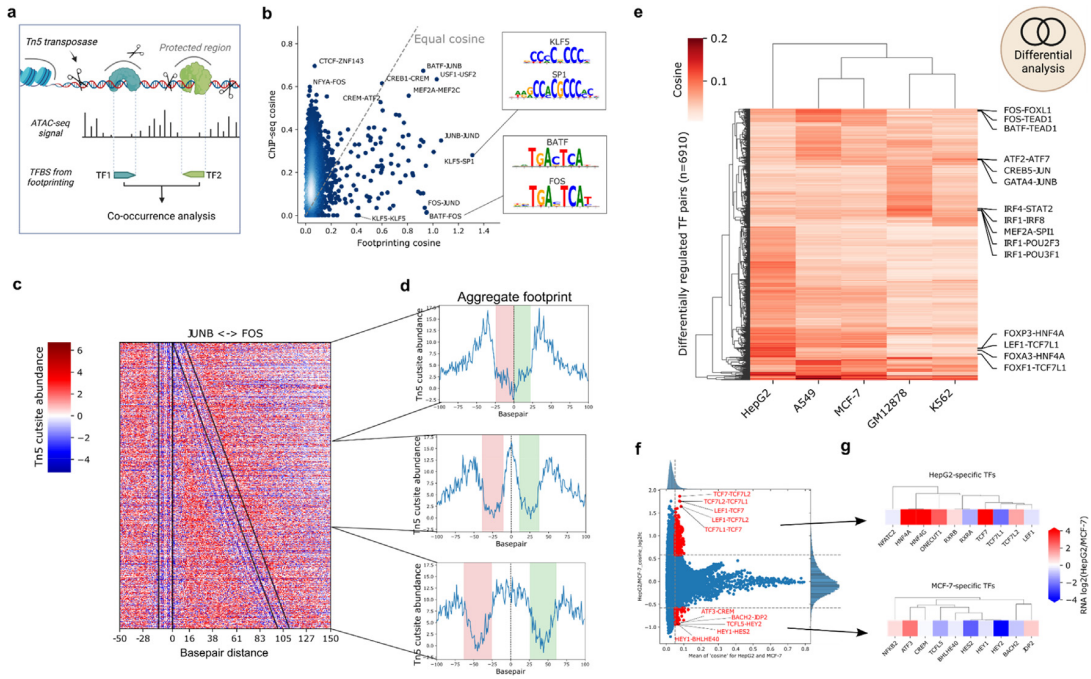


Fig. 4. Footprinting data uncovers cell line specific TF co-occurrence. a) Scheme of TF co-occurrence and Tn5 mediated digital genomic footprinting. Prepared using BioRender.com. b) Direct comparison of scores derived for TF-pairs via ChIP-seq and ATAC-seq (footprinting) analysis. Two pairs with high footprinting scores are highlighted and corresponding motifs are illustrated. c) Footprinting heatmap of all co-occurring JUNB-FOS sites. Colored for Tn5 cutsites appearing more than expected (red) or less than expected (blue) if DNA is inaccessible. Black lines represent the edges of JUNB (left) and FOS (right) motifs. Edges show the binding strand of the respective TF. d) Aggregated views of the scores shown in e). Increasing distance (top to bottom) causes the combined TFs footprint to split into two distinct ones. e) Heatmap showing cosine scores for differentially co-occurring TF pairs across five cell lines. A subset of prominent cell line specific pairs are labeled on the right side. f) Activity of TF-pairs in direct comparison between HepG2 and MCF-7 cell lines. Significantly changed TF pairs are marked in red. g) Differential RNA expressions of the top 10 TFs selected in f) for each group. TFs are clustered by motif similarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

tion patterns, which were used to subsequently find co-occurring TFs with TF-COMB.

Overall, we find both coincident and differing TF pairs when comparing the co-occurrence of ChIP-seq based and footprinting-derived TFBS (Fig. 4b). For example, we observed that well-known factor pairs BATF-JUNB, USF1-USF2 and CREB1-CREM were found to have high cosine scores in both footprinting and ChIP-seq analysis. However, we also observed some pairs specifically found in ChIP-seq data including CTCF-ZNF143 and NFYA-FOS. As the footprinting analysis is dependent on the presence of sequence motifs, we performed a global analysis of the presence of motifs within ChIP-seq peaks, and found the match between ChIP and motifs to be very different across factors (Supplementary Fig. S4a). For example, while a high percentage of CTCF sites contain at least one motif, less than 10% of ZNF143 ChIP-seq sites contain known motifs, thus making the CTCF-ZNF143 difficult to discover from motif-based data. As such, the lack of co-occurrence between ChIP-seq and footprinting-derived sites can be partially explained by a lack of identifiable motifs within these ChIP-seq peaks.

In contrast, there are also a number of co-occurring pairs which are more commonly found in footprinting data, including FOS-JUND, ATF3-ATF7 and KLF5-SP1, all of which have very similar motifs. We also observed self-pairs such as SP1-SP1, which we could not observe in ChIP-seq due to the lack of resolution of ChIP-seq peaks. When analyzing the underlying genomic sequence, we indeed found multiple SP1 motifs in the vicinity of one SP1 ChIP-seq peak, which naturally increases the cosine scores of this pair for footprinting data (Supplementary Fig. S4b). In general, we found that the co-occurrence scores for footprinting data are correlated with motif similarity to a much higher degree than the ChIP-seq derived data (Supplementary Fig. S4c). While this is expected due to motif overlap, the correlation to motif similarity persists, even when overlapping between sites is disallowed, which suggests that the effect is not only due to direct motif overlap. As is the case for SP1, this effect can arise due to multiple motifs within peaks, as exemplified by NFYA-NFYC (Supplementary Fig. S4d). Whereas the association of two ChIP-seq peaks is only counted once, the co-occurrence of similar motifs will lead to high scoring pairs for NFYA-NFYA, NFYC-NFYC and NFYA-NFYC, as these are counted multiple times within the same window. Thus, TFs do in fact co-occur with similar motifs on two levels; firstly by direct motif overlap, and secondly by multiple copies of a motif in close proximity. The latter case suggests a certain importance of a genomic loci for an individual factor, as multiple binding sites provide an increased probability of binding - even in the event of mutations. In conclusion, co-occurrence of footprinting data reflects ChIP-seq derived data analysis, and additionally unravels genomic sequence compositions that utilize motif redundancy at target regions.

The gain of resolution by utilizing footprinting data additionally allows for a TF distance analysis as exemplarily shown for the TF TBP, which has a preferred distance of 16 bp to the TSS (Supplementary Fig. S4e), while the ChIP-based analysis did not show any preferred binding distances to TSS (Supplementary Fig. S4f). We asked whether this increased resolution also enables the visualization of paired TF footprints and therefore utilized the TF-COMB plotting module for paired TF sites (Fig. 4c). As exemplarily shown for the well characterized TF pair JUN-FOS, sites with close motif distances create one common footprint, and by increasing

the motif distance of bound sites, the individual footprints appear distinguishable starting from distances above 20 bp (Fig. 4d). As footprints are generated by Tn5 transposase cutting patterns, we conclude that the Tn5 transposase is unable to insert adapters between closely bound TFs. Thus, in the context of footprinting analysis, individual footprints might not be directly mappable to single TFs, but might be the result of several closely bound TFs. This is an important factor to take into account for future footprinting algorithms.

As ChIP-seq data is limited to certain factors in each cell type, it can be tricky to compare co-occurrences between datasets. However, footprinting analysis contains the same TF motifs across all cell types, and just differs at the respective footprinting score levels. Thus, we used our previously calculated co-occurring TFs based on TOBIAS footprints and added a differential analysis to the TF-COMB object in order to quantify co-occurring TF pairs between cell lines in a global manner. As expected, we found the majority of TF pairs commonly active across cell lines (spearman correlations 0.8–0.9) (Supplementary Fig. S4g), however, by selecting the enriched TF pairs from each contrast, we identified 3.2% (n = 6910) of the potential TF pairs as differentially active between cell types (Fig. 4e). Not surprisingly, the overall clustering of the cell lines reflected the respective cell origin, by grouping epithelial cells (A549 and MCF7), as well as the lymphocyte cell lines (GM12878 and K562). The cell line specific TF pairs nicely mapped to the biological background, which included FOXA3-HNF4A for HepG2 cells, which are well known liver TFs able to program fibroblasts into hepatocyte-like cells [53], and multiple pairs containing IRF for GM12878-cells, which supports the importance of these factors in lymphocyte differentiation [54].

Focusing on the prominent changes between HepG2 and MCF-7, we used TF-COMB to highlight ~1% differential TF pairs specific for this contrast (Fig. 4f). The changes in expression (log2FC RNA-seq) of the top 10 TFs between these cell types likewise showed the majority of the TFs to be upregulated in the cell type, for which they participate in co-occurring pairs (Fig. 4g). In contrast, we observed TCF7L1 in many co-occurring pairs in HepG2, while it is actually downregulated on RNA level in comparison to MCF-7. This effect might be driven by motif similarity between TFs, as we see that both TCF7 and TCF7L2, which have highly similar motifs to TCF7L1, are upregulated in HepG2. Thus, the use of motifs makes it difficult to directly link motif activity with a certain TF, but integration of TF expression data might help to uncover which TF is most likely to be the participating partner in a co-occurring pair.

In summary, we conclude the application of co-occurrence analysis to digital genomic footprinting data to be a valuable approach for uncovering global changes of TF co-occurrence and TF binding grammar between biological conditions. In addition, the association of motifs allows to untangle TF relationships driven by motif similarity and motif redundancy.

4.5. TF binding grammar encodes biological relevance

Considering the higher resolution and completeness of TF binding activity for motif derived data, we asked if we are able to infer detailed information on TF binding grammar in the context of local binding site arrangement. Literature has several examples of highly regulated enhancers with specific distances and strandedness of TFs, such as the NFYB-USF1 pair, found at a preferred binding distance of 17–18 bp in a converged orientation [55] or a preferred distance of 37 bp for the CTCF-ZNF143 pair [41]. Therefore, we sought to use TF-COMB to investigate whether these exemplary binding characteristics are rare, or constitute a more global property of TF pairs that allow for a classification of TF pairs and enhancer organization. To be able to uncover the global presence of grammar in the genome, and not only for the sites which are predicted to be bound in the footprinting analysis, we used the full set of motif positions within the HepG2 ATAC-peak regions as input for TF-COMB.

Firstly we analyzed the global directionality of TF pairs, and found that a subset of ~2% of TF pairs exhibit preferential directionality with more than 60% of sites in either the same or opposite orientations (Fig. 5a). Of these, the orientations can be further divided into groups of TF1-TF2/TF2-TF1 and convergent/divergent for same and opposite groups respectively. While some TF pairs show equal distribution of the subdivided groups, the split highlights an additional preference for the exact order of binding (Fig. 5b).

Next, we analyzed the preferred distance between TFs using TF-COMB, and found that 36.6% of all TF pairs exhibit at least one preferred binding distance (Supplementary Fig. S5a). For the majority of pairs with a preferred binding distance, exactly-one distance was predicted (35.5%), while the remaining pairs (1.1%) exhibited multiple distances (Supplementary Fig. S5b). Among the candidates with predicted preferred binding distances, we found well established pairs, like THAP11-ZNF143 [56], NFY-USF members [57] (Fig. 5c) and BATF-JUN [58,79] (Supplementary Fig. S5c). In addition, among others, we found not yet described pairs like ETV3-ERG. In contrast, E2F2-FOXC2 is an example for a TF pair with no predicted preferred distance (Fig. 5c). This highlights that preferred distances between TFs is a global property applicable when investigating binding grammar. Furthermore, this might also hint to additional characteristics of grammar, such as changes between biological conditions. Exemplary, the FOXA2-GATA4 pair, which was recently described as liver specific [4], differs between HepG2 and GM12878 cell lines (Fig. 5d). In contrast, the ubiquitous pair NFYB-NFYC, which is known to form the trimeric NFY complex in collaboration with NFYA [42], remains similar between different cell lines (Supplementary Fig. S5d). Thus, parallel to the general co-occurrence analysis, the individual TF binding distances are also indicative of cell line specific co-occurrence. In addition, by analyzing all distances per TF pair, we detected that many pairs have distributions of distances which seem to occur with certain periodicity. One such example is the FOXJ2-SP8 pair, which displays an apparent structure in the distribution of binding sites, as it translates to a period of ~20 bp between two peaks (Fig. 5e). In contrast, we find other pairs with differing periods (Supplementary Fig. S5e), indicating that individual pairs exhibit different binding preferences.

In order to evaluate the biological relevance of the preferred distance sites from a TF pair compared to the non preferred distance sites from the same pair, we split the data into three groups. The first group contains all sites corresponding to TF pairs classified to have no preferred binding distance, which we call “no preference sites”. The second group covers the sites for TF pairs classified to have a preferred binding distance, filtered for the “distance peak” sites, called “preferred distance sites”. Finally, the third group contains the remaining sites of the TF pairs from the prior group outside of the preferred binding distance, named “no preferred distance sites”. Firstly, we hypothesized that the preferred distance sites represent important functional units, and are therefore more likely to occur within regulatory features, such as gene promoters. Indeed, after annotating all paired sites with UROPA [59], we found a significant increase of the gene annotation rate for sites with a preferred distance compared to both groups without preferred distances (Supplementary Fig. S5f). Next, we asked whether these motif-derived preferred distance sites can be used as a classifier to discriminate between real binding sites and potential binding sites. To this end, we overlapped the motif-derived sites with corresponding ChIP-seq peaks in HepG2 cells (Fig. 5f). As suggested by our prior findings on the gene annotation level, we detected a significant increase of overlapping “true” ChIP-seq

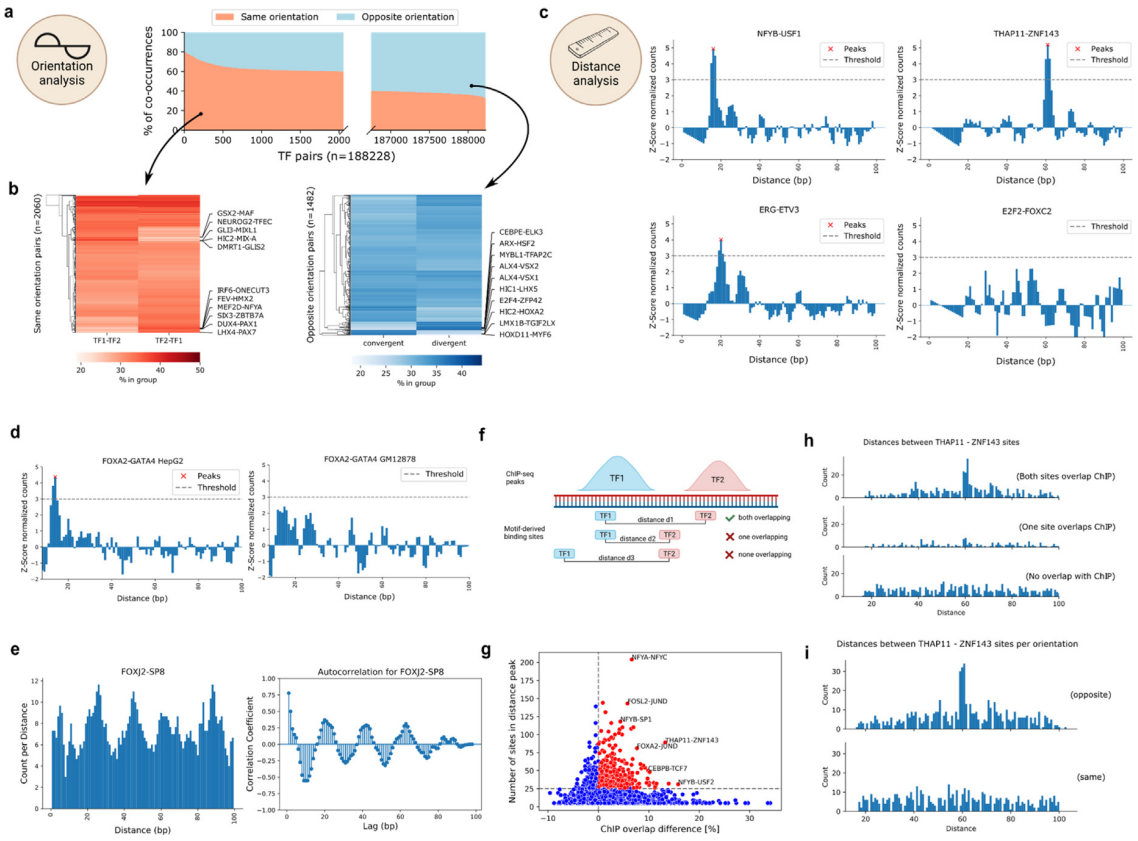


Fig. 5. TF pairs exhibit local binding grammar. a) Percentage of TF-pair locations (y-axis) with both TFs on the same or opposite strand. X-axis gives the ranking of TF pairs with regards to orientation. Only pairs with more than 60% for either group are shown (axis is not continuous). b) Percentage of TF-pair locations derived from a), splitted by TFs orientation. Red heatmap highlights top pairs with orientation on the same strand, blue heatmap highlights top pairs with orientation on opposite strands. c) Z-score normalized TF-pair binding counts sorted by distance. Peaks above threshold are called by TF-COMB and considered preferred binding distance. d) Difference in binding distance distribution for FOXA2-GATA4 in HepG2 (left) and GM12878 (right) cells. e) Binding distance periodicity of the FOXJ2-SP8 pair. Left plot shows the distribution of binding site distances. Right plot shows the calculated autocorrelation for the signal, indicating a lag of 20 bp. f) Scheme of motif-derived binding sites overlapping with ChIP peaks. Prepared using BioRender.com. g) Difference in ChIP-overlap fraction (x-axis) between preferred distance sites and no preferred distance sites per pair and power (y-axis). The most prominent pairs are annotated. h) Number of sites by distance between THAP11 and ZNF143 binding sites. Subplots indicate split of loci into groups as indicated in f). i) Number of sites by distance between THAP11 and ZNF143 binding sites. Upper and lower plots indicate strand orientation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

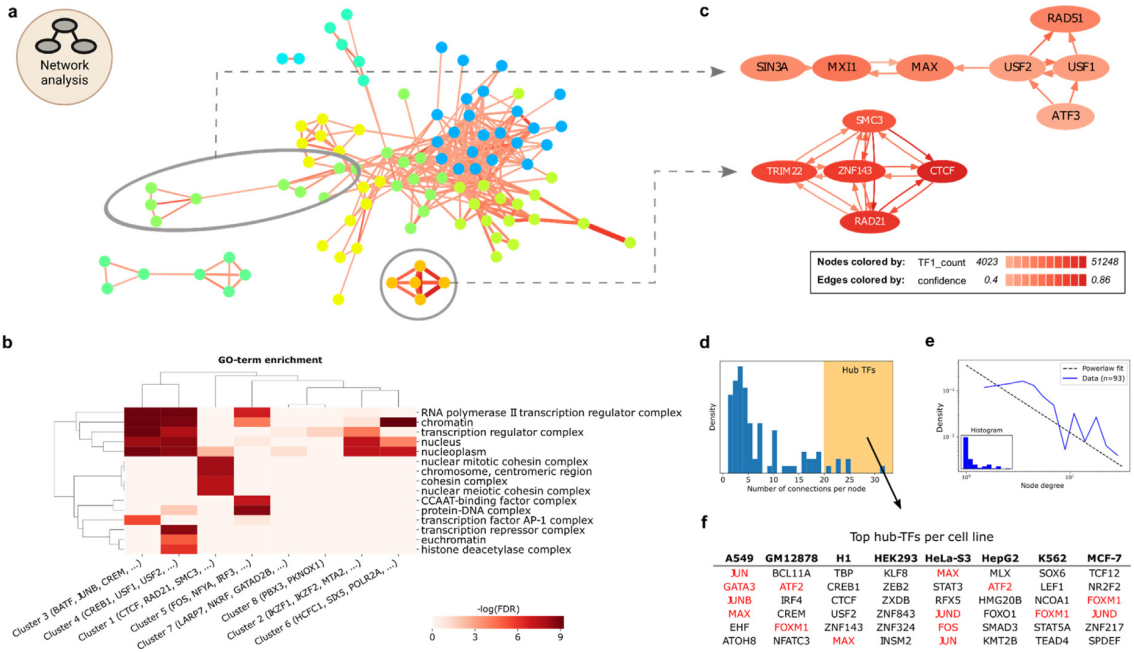
sites for the preferred distance sites group (Supplementary Fig. S5g). Of note, the TF pairs THAP11-ZNF143 and NFY-USF, both well described in literature and already found in earlier sections (Fig. 5c), were among the pairs showing the strongest differences (Fig. 5g). This finding holds true when visualizing the distance plots for the ChIP-seq overlap of both TF sites, only one TF site, or no overlapping sites, respectively (Fig. 5h). Finally, we combined distance and orientation analysis, and exemplarily found that the majority of preferred distance sites for THAP11-ZNF143 are located in opposite directions, exhibiting a preferred distance around 61 bp (Fig. 5i). This confirms that the preferred distance and orientation encodes for true co-occurrence of both factors.

In conclusion, we find that TF-COMB is able to globally uncover TF pairs that exhibit local binding grammar characteristics such as TF pair distance, and relative TF orientation. The overlap with ChIP-seq derived binding sites suggest biological relevance for the preferred binding distances, a finding that might contribute to future methods utilizing motifs as an approximation for DNA binding.

4.6. TF network analysis uncovers regulatory complexes

While we have investigated TF binding in pairs using TF-COMB, TFs often participate in multi-TF modules, which regulate complex biological networks through additive or synergistic binding [60]. Therefore, the discovery of these relationships is crucial for understanding gene regulation. As a result, TF-COMB can utilize the full set of TF rules to deduce a network in terms of nodes and edges. In order to identify potential protein complexes with the network module, we utilized the ChIP-seq data of TFs, co-factors and other DNA-binding proteins for the GM12878 cell line from ENCODE.

After initial filtering, TF-COMB generated a core network consisting of 329 edges (TF co-occurrences) and 93 nodes (TFs). The network view (Fig. 6a) uncovers noticeable substructures, including isolated TF clusters not connected to the main network, some barely interconnected substructures with tight internal links, as well as dense subgroups, driven by highly interconnected nodes. In order to quantify this structure, TF-COMB uses the louvain method for community detection [61], which partitioned the exemplary network into 8 clusters (Fig. 6a; Supplementary Fig. S6a left). Not surprisingly when analyzing transcriptional regulators, GO-term analysis of the individual clusters revealed enrichment of terms such as “chromatin”, “transcription regulator complex” and “nucleus” among others (Fig. 6b). However, besides GO-terms indicative of positive regulation, we also identified clusters, such as cluster 4, enriched for terms related to transcriptional repression. Further, we annotated cluster 1 to the cohesin complex,



A549	GM12878	H1	HEK293	HeLa-S3	HepG2	K562	MCF-7
JUN	BCL11A	TBP	KLF8	MAX	MLX	SOX6	TCF12
GATA3	ATF2	CREB1	ZEB2	STAT3	ATF2	LEF1	NR2F2
JUNB	IRF4	CTCF	ZXDB	RFX5	HMG20B	NCOA1	FOXM1
MAX	CREM	USF2	ZNF843	JUND	FOXO1	FOXM1	JUND
EHF	FOXM1	ZNF143	ZNF324	FOS	SMAD3	STAT5A	ZNF217
ATOH8	NFATC3	MAX	INSM2	JUN	KMT2B	TEAD4	SPDEF

Fig. 6. Network approach uncovers TF complexes and hubs. a) The GM12878 co-occurrence network. Nodes represent TFs, edges represent associated TF pairs. Edges are colored by cosine score and sized by the number of associated TF1-TF2 co-occurring sites. Coloring of nodes illustrates Louvain community clustering. Two sub-clusters are indicated for zooming in c). b) GO-term analysis on TF groups extracted from network clusters (columns). Coloring indicates enrichment for GO terms (rows). c) Directional sub-networks from a). Each arrow represents a dependency defined by the confidence score, e.g. ATF3 is dependent on USF2. Node coloring indicates the number of sites assigned to the respective TF. Only edges with confidence scores above 0.4 are shown. d) Distribution of node degrees. High node degree is shaded with yellow in the plot. e) Number of connections per TF (blue), approximates power-law distribution (grey dashed line). f) Table of TFs per cell line with the highest number of interactions (hub creators) from d). Marked TFs (red) are discussed in the main text. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and cluster 3, containing BATF and JUN, to be enriched for "transcription factor AP-1 complex". Of note, we observed many same-family TFs for cluster 3 (Supplementary Fig. S6a right). Thus, we wanted to test whether same-family TF pairs are significantly enriched within individual clusters. Indeed, by randomly selecting pairs within and between clusters, we observed a significant increase in percentage of same-family pairs within network clusters (Supplementary Fig. S6b). To summarize, the network analysis enabled us to identify protein clusters and complexes with particular biological functionality.

Within the individual clusters, the TF-COMB network analysis can additionally uncover dependency relationships. For this, a directional network using the 'confidence' score, which represents the probability that TF2 is found, if TF1 is present, is used. For cluster 1, which includes ZNF143/CTCF/RAD21, all TFs are dependent on each other, which suggests that these factors bind in a protein complex (Fig. 6c; upper). However, in the cluster of USF1/USF2/RAD51/ATF3, we found USF1 and USF2 to have directional relationships with ATF3 and RAD51 (Fig. 6c; lower). The association of USF and RAD51 is supported by a recent study on the location of USF motifs at RAD51-bound elements [62], however, our analysis additionally indicates a significant number of RAD51 sites to have a completely independent role of USF. Interestingly, the network analysis also reveals ATF3 to be highly connected to USF factors, but not vice versa, with more than 50% of its binding sites in the vicinity of USF binding sites. Moreover, the lack of a link between ATF3 and RAD51 suggests that the interactions with USF are contained in independent regulatory circuits.

Finally, the network representation of TF co-occurrence also allows to draw conclusions on potential TF hubs, defined as TFs with many partners in the network. The distribution of node degrees has an apparent tail, indicating most TFs to have few partners, yet some TFs exhibit many co-occurrences with other factors (Fig. 6d). Noteworthy, we found the networks of most cell types to follow a power-law distribution, which is a characteristic of biological networks [63] (Fig. 6e, Supplementary Fig. S6c). Taking all cell lines into account, the TFs with the highest node-degree include MAX, JUN, GATA3 and FOXM1 (Fig. 6f). MAX is known to orchestrate a large network [64]. Likewise, TFs JUN, JUNB, JUND, FOS and ATF2 are all part of the AP-1 family of TFs, which can dimerize, thus explaining the hub characteristics of these TFs. Interestingly, we also find GATA3, which is a known pioneer factor, as well as FOXM1. While FOXM1 is not known to have pioneer activity, previous publication showing an overlap of 71% between FOXM1 and FOXA1 binding events [65], thus allows us to speculate that FOXM1 might indeed be able to function as a pioneer in line with FOXA1.

In conclusion, we found network analysis on TF co-occurrence as a highly flexible tool to explore relationships between transcriptional regulators in a wider perspective and to extract subgroups of factors as a valuable source for the hypothesis of potential TF complexes.

5. Discussion & conclusion

This study was performed to demonstrate the functionality and usability of a new software framework, named TF-COMB, intended to gather, analyze, visualize and explore data in the field of co-occurrence and grammar of TF binding. Due to its generalized setup, TF-COMB is able to help to unravel epigenetic related aspects of TF binding grammar, such as the interplay of chromatin accessibility, histone modifications, and gene locations, as well as local grammar in terms of distance and orientation of TFs. In addition, TF-COMB allows a broader look at the interdependencies of TF

co-occurrences in a global network context, which allows to draw conclusions on the binding of TFs not only in pairs, but also in larger compositions, including e.g. protein complexes.

The exemplary investigations on widely accepted gold-standard datasets from ENCODE are intended to illustrate potential analysis workflows provided via TF-COMB. To this end, we used TF binding data of both ChIP-seq peaks and motif sites derived from ATAC-seq based TF footprinting to identify commonalities, as well as unique aspects that can be derived for each data source. A major strength of ChIP-seq data was found to be the independence of TF sequence motifs [66] enabling to calculate co-occurrence of TFs with co-factors and other DNA-associating factors without known sequence motifs (e.g. RAD21, SMC3, TRIM22). In contrast, we demonstrated the advantage of TF footprinting and motif positions in general, which allow for global and more detailed analysis on TF binding grammar, as shown in the analysis of distances and orientations between individual TFs. Using the footprinting data, we found that TFs often co-occur with similar motifs at the same loci, a characteristic which has already been described to be important in both promoters and developmental enhancers [67]. These results hint towards a certain level of redundancy in TF binding, with similar transcription factors such as FOXA1 and FOXA2 substituting each other, as described previously [68]. In summary, all our findings illustrated that the selection of the input data plays a crucial role in the identification of resulting associations. Software design should ideally permit integration of data from various sources including newly emerging assays. For example, a recent method has improved on existing footprinting methods by applying methylation and bisulfite sequencing to interrogate single molecule footprints in single cells [69], which might therefore help to reduce the noise of footprinting in bulk samples. TF-COMB is ready to use such data, and with the advent of more technologies and available collections of TF binding positions, TF-COMB will help to further improve the investigation of co-occurrences and TF binding grammar.

By utilizing a pure motif analysis, we gained high resolution TFBS data, and could thereby isolate TF pairs which exhibit both preferred orientation and distance to each other. This, together with the observation of binding site periodicity, suggests the existence of a "Goldilocks distance" for many TF pairs, which we define by a set of locational parameters that probably optimizes complex building, binding duration or binding itself. The fact that these locations have significantly higher overlap with both regulatory features and "real" ChIP-seq derived binding sites, strongly supports this hypothesis. Of note, as only a minority of TFs are proven by wet lab based methods to physically interact, the preferred distance of TF pairs might also represent the right distance for comfortably fitting two proteins on the DNA without being sterically hindered by each other. In other cases, such as seen for the collective model, binding on the correct side of the DNA might also be essential in binding of co-factors. However, the majority of TFs do not show any particular binding grammar. This observation is not necessarily a rejection of their co-occurring status, but rather a sign of flexible TF binding. In fact, it has been shown that transient binding of multiple TFs is a mechanism to compete with nucleosome binding and keep DNA accessible - a mechanism known as 'assisted loading' [70]. In such cases, the exact location of TFs might be disregarded, as also observed in the billboard enhancer model.

Many areas of TF co-occurrence are still open for investigation, including the correlation of the size of open chromatin in relation to the number of TFs binding, and how TFs became hub proteins throughout evolution. For this study, we have focused on the co-occurrence of TFs within the same regulatory region, but our results have also identified a number of TFs, including CTCF and ZNF143, which are highly co-occurring at chromatin loop anchors, and are involved in connecting distant regulatory elements. There is thus an additional layer of 3D co-occurrences built between regulatory regions, as well as within looped enhancers, which we currently cannot track with our software. However, we were able to provide evidence for these structures in our co-occurrence analysis of histone modifications and TFs, where we found a number of TFs restricted to enhancer binding, avoiding cis regulatory regions. This suggests a model of e.g. differentiation, in which cell type specific TFs primarily control gene expression from regulatory enhancer elements *in trans*. This increases regulatory complexity, while simultaneously preventing spurious activation of target genes by TFs such as pioneers. Indeed, such regulation of enhancer activity by formation of topologically associated domains (TADs) is well-known for the HoxD cluster of genes, which are important for limb development [71]. Thus, the influence of chromatin organization should not be disregarded when discussing co-occurrence of transcription factors.

In conclusion, we have used TF-COMB to investigate a variety of aspects of TF binding grammar. Understanding the effect of TF co-occurrence is important for uncovering the direct targets of TFs, as multiple TFs create complicated AND/OR/XOR logic, as known from studies on systems biology. In particular, TFs such as pioneer factors can act as primers to subsequent binding of other TFs. It is therefore of great interest to discover potential sets of co-occurring TFs for individual cell lineages, and we believe that TF-COMB represents a valuable resource to identify, study and understand such TF co-occurrences in the context of gene regulation.

6. Figure attribution statement

Plots were produced using TF-COMB framework functionalities, and using matplotlib and seaborn packages in Python. The graphical abstract as well as explanatory (sub-)figures were created using BioRender.com as stated in the individual figure descriptions. Module icons as included in Figures 2-6 were taken from Figure 1.

Data availability

The TF-COMB tool is freely available under MIT License at Github: https://github.com/loosolab/TF-COMB. Full documentation and examples are available through ReadTheDocs at: https://tf-comb.readthedocs.io.

Author contributions

M.B. and M.L conceived the study. M.B. designed TF-COMB. M.B., V.H. and H.S. implemented the framework. M.B, C.K. and V.H. preprocessed data. M.B., V.H. and H.S. performed the analysis. M.B., V.H., H.S., C.K. and M.L. wrote the manuscript. M.L. supervised the project.

Funding

This study was funded by the German Research Foundation (DFG):EXC2026/1 and KFO309 Project Z1, 284237345 to M.L, as well as the Max Planck Society.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.025.

References

[1] Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet 2009;25(10):434–40.
[2] Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 2012;13(9):613–26.
[3] Jindal GA, Farley EK. Enhancer grammar in development, evolution, and disease: dependencies and interplay. Dev Cell 2021;56(5):575–87.
[4] Balsalobre A, Drouin J. Pioneer factors as master regulators of the epigenome and cell fate. Nat Rev Mol Cell Biol 2022.
[5] Lambert SA et al. The Human Transcription Factors. Cell 2018;172(4):650–65.
[6] Salzberg SL. Open questions: How many genes do we have? BMC Biol 2018;16 (1):94.
[7] Meuleman W et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature 2020;584(7820):244–51.
[8] Jankowski A, Prabhakar S, Tiuryn J. TACO: a general-purpose tool for predicting cell-type–specific transcription factor dimers. BMC Genom 2014;15(208).
[9] Meckbach C et al. PC-TraFF: identification of potentially collaborating transcription factors using pointwise mutual information. BMC Bioinf 2015;16:400.
[10] Whitington T et al. Inferring transcription factor complexes from ChIP-seq data. Nucleic Acids Res 2011;39(15):e98.
[11] Bentsen M et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun 2020;11(1):4267.
[12] Li Z et al. Identification of transcription factor binding sites using ATAC-seq. Genome Biol 2019;20(1):45.
[13] Kotu, V. and B. Deshpande, *Chapter 6 - Association Analysis*, in *Data Science (Second Edition)*, V. Kotu and B. Deshpande, Editors. 2019, Morgan Kaufmann. p. 199-220.
[14] Anandhavalli M, Ghose M, Gauthaman M. Association Rule Mining in Genomics. Int J Comput Theor Eng 2010;2.
[15] Tan P-N, Kumar V, Srivastava J. Selecting the right objective measure for association analysis. Inf Syst 2004;29(4):293–313.
[16] Raeder T, Chawla NV. Market basket analysis with networks. Social Network Anal Mining 2011;1(2):97–113.
[17] Haiminen N, Mannila H, Terzi E. Determining significance of pairwise co-occurrences of events in bursty sequences. BMC Bioinf 2008;9(1):336.
[18] Behnel S et al. Cython: The Best of Both Worlds. Comput Sci Eng 2011;13:31–9.
[19] Davis CA et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 2018;46(D1):D794–801.
[20] Rao Suhas SP et al. 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 2014;159(7):1665–80.
[21] Castro-Mondragon JA et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2022;50 (D1):D165–73.
[22] Oughtred R et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Prot Sci: Publ Protein Soc 2021;30(1):187–200.
[23] UniProt C. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480–9.
[24] Madeira F et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res 2022:p. gkac240.
[25] Hu H et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res 2019;47(D1): D33–8.
[26] Klopfenstein DV et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 2018;8(1):10872.
[27] Smit, A.H., R; Green, P. . *RepeatMasker Open-4.0*. 2013-2015; Available from: http://www.repeatmasker.org.
[28] Zhang, Z., et al., *CENTDIST: discovery of co-associated factors by motif distribution.* Nucleic Acids Res, 2011. **39** (Web Server issue): p. W391-9.
[29] Kazemian M et al. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res 2013;41 (17):8237–52.
[30] Parra RG et al. INSECT 2.0: a web-server for genome-wide cis-regulatory modules prediction. Bioinformatics 2016;32(8):1229–31.
[31] Perna S et al. TICA: Transcriptional Interaction and Coregulation Analyzer. Genom Proteom Bioinfo 2018;16(5):342–53.
[32] Perna S et al. NAUTICA: classifying transcription factor interactions by positional and protein-protein interaction information. Biol Direct 2020;15 (1):13.
[33] Ha N, Polychronidou M, Lohmann I. COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. PLoS One 2012;7(12):e52055.
[34] Levitsky V et al. A single ChIP-seq dataset is sufficient for comprehensive analysis of motifs co-occurrence with MCOT package. Nucleic Acids Res 2019;47(21):e139.
[35] Navarro C et al. CisMiner: genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining. PLoS One 2014;9(9):e108065.
[36] van Bommel A et al. coTRaCTE predicts co-occurring transcription factors within cell-type specific enhancers. PLoS Comput Biol 2018;14(8):e1006372.
[37] Sayers EW et al. Database resources of the national center for biotechnology information. Nucleic Acids Res 2022;50(D1):D20–6.
[38] Dunham I et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74.
[39] Amati B, Land H. Myc—Max—Mad: a transcription factor network controlling cell cycle progression, differentiation and death. Curr Opin Genet Dev 1994;4 (1):102–8.
[40] de Los G, Fayos Alonso I, et al. The Role of Activator Protein-1 (AP-1) Family Members in CD30-Positive Lymphomas. Cancers 2018;10(4):93.
[41] Zhou Q et al. ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function. Nat Commun 2021;12(1):43.
[42] Ly LL, Yoshida H, Yamaguchi M. Nuclear transcription factor Y and its roles in cellular processes related to human disease. Am J Cancer Res 2013;3 (4):339–46.
[43] Sirito M et al. Members of the USF family of helix-loop-helix proteins bind DNA as homo- as well as heterodimers. Gene Expr 1992;2(3):231–40.
[44] Ogawa H et al. A Complex with Chromatin Modifiers That Occupies E2F- and Myc-Responsive Genes in G0 Cells. Science 2002;296(5570):1132–6.
[45] Zhong Y et al. CHD4 slides nucleosomes by decoupling entry- and exit-side DNA translocation. Nat Commun 2020;11(1):1519.
[46] Pugacheva Elena M et al. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. Proc Natl Acad Sci 2020;117 (4):2020–31.
[47] Ibn-Salem J, Andrade-Navarro MA. 7C: Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs. BMC Genomics 2019;20(1):777.
[48] Sun M, Nishino T, Marko JF. The SMC1-SMC3 cohesin heterodimer structures DNA through supercoiling-dependent loop formation. Nucleic Acids Res 2013;41(12):6149–60.
[49] Beagan JA et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res 2017;27(7):1139–52.
[50] Wang Y, Li X, Hu H. H3K4me2 reliably defines transcription factor binding regions in different cells. Genomics 2014;103(2):222–8.
[51] Lai X et al. Pioneer Factors in Animals and Plants—Colonizing Chromatin for Gene Regulation. Molecules 2018;23(8).
[52] Pearson JC, Lemons D, McGinnis W. Modulating Hox gene functions during animal body patterning. Nat Rev Genet 2005;6(12):893–904.
[53] Nakamori D et al. Direct conversion of human fibroblasts into hepatocyte-like cells by ATF5, PROX1, FOXA2, FOXA3, and HNF4A transduction. Sci Rep 2017;7 (1):16675.
[54] Hagman J. Critical Functions of IRF4 in B and T Lymphocytes. J Immunol 2017;199(11):3715.
[55] Ronzio M et al. Integrating Peak Colocalization and Motif Enrichment Analysis for the Discovery of Genome-Wide Regulatory Modules and Transcription Factor Recruitment Rules. Front Genet 2020;11:72.
[56] Parker JB et al. Host Cell Factor-1 Recruitment to E2F-Bound and Cell-Cycle-Control Genes Is Mediated by THAP11 and ZNF143. Cell Reports 2014;9 (3):967–82.
[57] Zhu J et al. NF-Y cooperates with USF1/2 to induce the hematopoietic expression of HOXB4. Blood 2003;102(7):2420–7.
[58] Chinenov Y, Kerppola TK. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. Oncogene 2001;20:2438–52.
[59] Kondili M et al. UROPA: a tool for Universal RObust Peak Annotation. Sci Rep 2017;7(1):2593.
[60] Li X et al. Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. Mol Syst Biol 2015;11(1):775.
[61] Blondel VD et al. Fast unfolding of communities in large networks. J Stat Mech: Theory Exp 2008;2008(10):P10008.
[62] Kang K et al. Epigenomic Analysis of RAD51 ChIP-seq Data Reveals cis-regulatory Elements Associated with Autophagy in Cancer Cell Lines. Cancers 2021;13(11).
[63] Albert RK. Scale-free networks in cell biology. J Cell Sci 2005;118(21):4947–57.
[64] Wahlström T, Henriksson M. Mnt Takes Control as Key Regulator of the Myc/Max/Mxd Network. In: Advances in Cancer Research. Academic Press; 2007. p. 61–80.
[65] Sanders DA et al. Genome-wide mapping of FOXM1 binding reveals co-binding with estrogen receptor alpha in breast cancer cells. Genome Biol 2013;14(1): R6–R.
[66] Wang J et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 2012;22 (9):1798–812.
[67] Gotea V et al. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res 2010;20 (5):565–77.
[68] Wan H et al. Compensatory Roles of Foxa1 and Foxa2 during Lung Morphogenesis*. J Biol Chem 2005;280(14):13809–16.
[69] Sönmezer C et al. Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. Mol Cell 2021;81(2):255–267.e6.
[70] Voss TC et al. Dynamic Exchange at Regulatory Elements during Chromatin Remodeling Underlies Assisted Loading Mechanism. Cell 2011;146(4):544–54.
[71] Rodríguez-Carballo E et al. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. Genes Dev 2017;31(22):2264–81.

3 Discussion

3.1 Challenges of ATAC-seq footprinting

This thesis presents the development of the TOBIAS tool to satisfy the need for an ATAC-seq footprinting tool. In terms of software architecture, TOBIAS is a python package giving access to different analysis modules, such as *ScoreBigwig* for calculating footprint scores and *BINDetect* for comparing TF binding between conditions. On the front-end, users can query the individual modules on the command-line, which makes it possible to employ different combinations of tools and also enables easy integration into existing analysis pipelines. In addition to the Python package, the analyses for the publication were performed using the supplied pipelines in Snakemake and Nextflow, which enable a start-to-finish analysis with automatic processing of replicates across several experimental conditions. Finally, the code for TOBIAS is hosted on *github.com*, which enables version control as well as community support through the issues page. This combination of features is novel in comparison to existing footprinting tools.

An important feature of TOBIAS is the Tn5 bias correction, accomplished by the module named *ATACorrect*. It is well known that the enzymes used for chromatin accessibility assays have a preference for certain sequences (Green et al., 2012), and DNase bias has been successfully characterized and corrected using a simple 6-mer sequence (Martins et al., 2018). However, previous efforts at bias correction have stated that k-mer based methods are insufficient to fully correct the bias for ATAC-seq (Martins et al., 2018). This challenge arises because Tn5 binds as a dimer, and has even been proposed to create multimeric filament-like structures on DNA (Goryshin et al., 1998). It was also shown that its protein-DNA contacts are more important than protein-protein interactions for maintaining dimer binding (Steiniger-White et al., 2004), which indicates that both Tn5 binding sites must be taken into account for bias correction. In order to overcome this challenge, TOBIAS estimates Tn5 insertion bias using a di-nucleotide weight matrix (DWM), which does not assume independence of sites (Siddharthan, 2010). ATACorrect uses this bias to estimate the predicted background, and subtracts this from the observed signal. To this end, TOBIAS successfully uncovered footprints between unbound and bound subsets. TOBIAS includes implementation of both PWM and DWM as representations of Tn5 bias, and comparisons with existing tools showed that the DWM model was superior in uncovering ATAC-seq footprints. In conclusion, the application of a DWM in *TOBIAS ATACorrect* is an improvement for the capture and correction of Tn5 bias.

Besides being utilized for ATAC-seq, Tn5 is widely applied across other high-throughput sequencing methods including chromatin conformation capture assays to map 3D structure of chromatin (e.g. Dip-C), whole-genome bisulfite sequencing (e.g. Tn5mC-seq), as well as in the CUT&Tag method (Li et al., 2020). Thus, being able to estimate and correct Tn5 bias is not only relevant in the context of ATAC-seq, but also influences the use of Tn5 in other assays, where non-random insertion probabilities can influence interpretation of results.

While bias correction by TOBIAS improved the number of previously footprintable factors from 20% (Baek et al., 2017) to 65%, the remaining 35% prove that there are still TFs for which footprinting is intrinsically limited. For example, some TFs might not bind DNA long enough to block Tn5 adapter insertion, and therefore leave no measurable footprints (Sung et al., 2014). Transitive binding of these TFs might nevertheless increase overall chromatin accessibility by means of collaborative competition (as described in section 1.2.4). TOBIAS alleviates this issue by taking both accessibility and footprint depletion into account during scoring. However, the subsequent visualization of a global footprint is still impossible for many factors. Along these lines, there are also indications that repressors show less prominent footprints, as these TFs would act to preferentially close chromatin rather than opening it (Berest et al., 2019). The biological functions of TFs therefore limits the use of footprinting. More work is needed to unravel the influence of TF binding properties on footprinting results.

3.2 An overview of dynamic transcription factor binding in ZGA

In the context of early embryonic development and ZGA, limited sample material has previously hampered large-scale investigations of TF binding within these processes. However, two publications have provided ATAC-seq data from developing human and mouse embryos, which enabled an overview of dynamic chromatin changes throughout ZGA. By using the TOBIAS Snakemake pipeline, it was possible to process and automatically compare TF binding across multiple conditions, in this case for developmental timepoints 2C, 4C, 8C, inner cell mass (ICM) and mESC/hESC.

A major result of this analysis is the temporal map of TF binding dynamics, which shows distinct clusters of specific activation patterns. In parallel to this, it was possible to visualize stage-specific aggregate footprints for a number of TFs known to play a role during early development. For example, the results showed stage-specific footprints for DUX4 and ZSCAN4 during early ZGA and 8C human stages, respectively. The aggregate footprints also uncovered interesting binding dynamics of ubiquitous proteins such as CTCF, which showed footprints across all timepoints. CTCF is a maternally provided factor, explaining its activity prior to ZGA (Wan et al., 2008). However, keeping in mind that the chromatin of early embryo is highly unstructured, it was surprising to see

CTCF binding at the 2C and 4C stages. This observation suggests that CTCF binding precedes the establishment of TADs, perhaps priming chromatin for future 3D organization. It is also possible that CTCF acts directly as a TF in these cells, as has been shown within oocytes, which are also devoid of TADs (Wan et al., 2008). Interestingly, in the context of enhancer grammar, it has also been shown that the orientation of CTCF binding sites is important for the direction of loop formation, as loops are preferably created by dimerization of CTCF at convergent binding sites (Guo et al., 2015). Further investigation of the exact positions of CTCF footprints, and the orientation of these sites, may provide insights into the function of CTCF binding before ZGA.

Besides ATAC-seq, Wu et al., 2016 and Wu et al., 2018 also performed RNA-seq of the same developmental timepoints, which allowed integration of the changes in expression with the footprinting analysis. One of the most interesting observations was that while the majority of TFs are expressed during the time of strongest binding, a large subset is also expressed before. This observation suggests that the dynamics of expression and translation differ, and indeed, a study of proteomics and RNA-seq data within mouse embryos confirms this assumption (Gao et al., 2017). In the context of collaborative TF binding, individual TFs might exhibit a temporal shift in occupancy as they wait for binding partners to become expressed. The differences might also be driven by regulation of protein stability, as previously shown for KLF4, which gains stability through interaction with other proteins (Dhaliwal et al., 2019) and Pax8, which is controlled by sumoylation (de Cristofaro et al., 2009). In the early embryo, these dynamics are also highly affected by the extensive regulation of maternal mRNAs as described in Section 1.2.6. Technologies like Ribo-seq (Ingolia et al., 2009), which measures protein translation, might help bridge this knowledge gap. Continued research into the interplay between gene expression, mRNA stability and translation, and temporal TF binding is required to fully understand the dynamics between initial TF expression and final target activation.

While previous experimental studies used pooled embryos for bulk ATAC-seq and RNA-seq, advances in the field of single cell genomics have recently made it possible to quantify both expression and accessibility at a single cell resolution (Hwang et al., 2018; Cusanovich et al., 2015). Such methods may help to uncover the variability of individual cells as well as identify subpopulations of cells during differentiation processes. However, in the context of footprinting, scATAC-seq is limited by the sparsity of the input signal, as most regions will contain either 0, 1 or 2 counts per cell. Classical identification of depleted regions is therefore not applicable for single cell data. One way to solve the issue of sparsity is by joining groups of similar cells into pseudo-bulks during analysis, which was shown to be powerful in estimating differential expression in scRNA-seq (Squair et al., 2021). In recent years, new methodologies have been developed for quantification of TF binding in single cells, such as scCUT&Tag (Bartosovic et al., 2021). Another method to interrogate TF binding in single cells is known as *single-molecule footprinting* (Sönmezer et al., 2021). Instead of using Tn5 to query DNA accessibility, this assay applies recombinant methyltransferases, which cannot methylate

TF-bound DNA. Combined with bisulfite sequencing, this assay can thereby map footprints across the genome in single cells. However, as we rely on capturing the TF binding event, all of these methods suffer from the limitations of TF residence time as discussed in Section 3.1. It will be exciting to see whether the methods can be tweaked to increase the capture of transiently bound TFs in single cells, for example by applying cross-linking of TF-DNA interactions or adjusting Tn5 and methyltransferase concentrations, or whether other assays are developed for this purpose.

3.3 Dux is a driver of the 2-cell stage

The global map of TF binding during ZGA uncovered that Dux and DUX4 are active in the early stages of ZGA in both mouse and human, respectively. Clustering with known maternal factors, this observation recapitulates previous results, showing that Dux/DUX4 are master regulators of ZGA. To investigate the direct influence of Dux to drive expression of genes related to ZGA, TOBIAS was applied to an additional dataset of *Dux* overexpression in mESCs (Hendrickson et al., 2017).

The analysis presents a global overview of up- and downregulated TF binding between *Dux* overexpression and control (mESC). In agreement with the experimental setup, Dux is one of the most up-regulated TFs, whereas the downregulated TFs include known pluripotency factors, such as Pou5f1, Nanog and Sox2. Interestingly, although Pou5f1 and Sox2 are known to be pioneer factors, and can induce chromatin accessibility, their binding is limited by *Dux* overexpression. This is likely a result of the reversal of these cells into the 2-like state, which is incompatible with the pluripotency of mESCs. Thus, evidence for an internal hierarchy between pioneer factors was obtained, where reprogramming by Dux exceeds the effects of Pou5f1 and Sox2. This hierarchy also explains the ability of Dux to drive the spontaneous transition of mESCs to 2CLCs in culture. Likewise, it helps us to understand the detrimental effects of *Dux* expression in somatic tissues. For example, facioscapulohumeral dystrophy (FSHD) is a muscle disorder which causes weakness in the muscles of the face, back and upper arms, and is caused by failure to silence *DUX4* in skeletal muscle (Daxinger et al., 2015). The *Dux*-family has also been implicated in the development of a number of cancers including childhood rhabdomyosarcoma (RMS), for which a subset of tumors exhibit a ZGA-like signature driven by aberrant expression of *Dux*-family genes (Preussner et al., 2018). This suggests that Dux TFs are strong regulators of the ZGA program, even outside the bounds of the embryo, and must therefore be tightly regulated in all other circumstances.

In this context, it is surprising that *Dux* is expressed in mESCs, and even has a beneficial effect on maintaining the mESC population. This might be explained by the increased chromatin plasticity of mESCs, as these cells have yet to make any lineage-decisions which might restrict differentiation potential (Melcer et al., 2010). However, later in development, when the chromatin landscape is

locked in a certain direction, the expression of *Dux* family genes is poorly tolerated. In the case of FSHD, the expression of *Dux* triggers muscle degeneration and cell death, whereas RMS can be caused by *Dux*-expressing cells, which are no longer under control by apoptotic pathways.

While Dux is clearly able to initiate expression of ZGA genes, recent studies have indicated that Dux is not strictly essential for successful ZGA (Chen et al., 2019; De Iaco et al., 2020; Bosnakovski et al., 2021). This observation would suggest that there is a certain robustness of TF functions in early development and in support of this hypothesis, a recent preprint shows that Obox4 can rescue the effects of *Dux* knockout with only minor defects in blastocyst formation (Guo et al., 2022). As explained in Section 1.2.6, the *Obox* genes are homologous to human *TPRX2*, which might suggest a similar redundant role for TPRX2 in human embryos. A recent study showed that knockout of *TPRX1/2/L* led to severe developmental defects and failure to activate ZGA (Zou et al., 2022). However, if TPRX has a similar function as Obox, it is unclear why DUX4 does not compensate for the lack of *TPRX* during early embryogenesis. This might be related to the divergent functionality of TPRX as a repressor rather than a pioneer factor, but further studies are needed to unravel this mechanism.

In order to identify potential ZGA regulators, Alda-Catalinas et al., 2020 utilized a single cell CRISPR activation screen to measure the ability of 230 proteins to induce a 2CLC signature. Interestingly, overexpression of a number of genes, including *Yap1*, *Smarca5*, *Patz1* and *Dppa2*, were found to initiate expression of 2C genes, including LTRs. However, many of the identified factors are ubiquitously expressed in adult differentiated tissues. It is therefore not completely clear whether these proteins play a specific role in ZGA, or whether the plastic chromatin state of mESCs simply enables conversion into 2CLCs through a variety of pathways. In this context, one has to keep in mind that the 2CLC state lacks maternal factors otherwise present in the 2C embryo. Moreover, the 2CLC state is not totipotent but the 2C stage is. Thus, in order to simulate a truly totipotent state, we must first understand the structure of the epigenetic landscape within endogenous embryos. Full understanding of the epigenetic landscape is a desirable goal that might facilitate creation of totipotent cells from mESCs, which, on the other hand, warrants careful ethical considerations, in particular if such cells are used for the production of artificial embryos.

In summary, the identification of the early embryonic TFs is highly important for understanding the mechanisms responsible for the success of preimplantation development. Indeed, mutations in *FIGLA*, encoding a TF regulating expression of the SCMC proteins, has already been identified as a cause of premature ovarian failure (Zhao et al., 2008). In addition, early embryonic factors, such as Zscan4, have been shown to enhance reprogramming of fibroblasts to iPSCs (Hirata et al., 2012). Thus, the continued investigation of early embryonic development will not only have implication for fertility treatment, but also for the future of cell type reprogramming.

3.4 Transcription factor co-occurrence

The second paper of this thesis presents TF-COMB, a software package for uncovering co-occurring TFs using a modified *market basket* approach. Whereas the original market basket analysis collects all baskets (genomic regions) in a table, this is not feasible for whole-genome quantification of co-occurring TFs. Instead, TF-COMB applies a sliding-window approach which increments counts for co-occurring TFs as they are observed. This enables TF-COMB to handle large amounts of input data, as well as data from a variety of sources. Similar to TOBIAS, TF-COMB is hosted on *github.com*, where a number of tutorials and use-cases are also available.

To validate the method, TF-COMB was first applied to experimentally identified TFBS from ChIP-seq, as well as broad ChIP-seq peaks of activating and repressive histone modifications. Through these investigations, TF-COMB uncovered known TF pairs across multiple human cell types, which were enriched for known protein-protein interactions. Interestingly, the subsequent integration of histone ChIP-seq identified a number of TFs with preferential binding to H3K4me1/2/3 and H3K27ac, but completely abolished binding in regions marked by H3K9ac, H3K79me2 and the H2 variant H2A.Z. Thus, it seems that these factors are actively avoiding genes, and instead bind to distal enhancer elements. Indeed, the combination of H3K4me1 and H3K27ac is indicative of active enhancers, whereas H3K4me1 alone represents primed enhancers (Creyghton et al., 2010). The factors in this group include several *HOX* and *FOX* genes, which are important for limb and organ development (Hannenhalli et al., 2009; Pearson et al., 2005). In addition, a number of these, including FOXA1 as described in Section 1.2.5, have known pioneering abilities. Altogether, the TF-COMB analysis indicates that these factors are important for cell type specification by defining the grammar of individual enhancers, rather than directly binding in promoters to regulate gene expression. This raises the potential of computationally driven characterization of promoter- and enhancer specific TFs, which will improve our understanding of promoter-enhancer interactions during differentiation.

In the context of the 3D structure of chromatin, TF-COMB was also used to identify the co-occurrences of TF ChIP-seq with Hi-C chromatin capture maps of the lymphoblastoid cell line GM12878. The analysis highlighted a number of factors to be associated with loop anchors including CTCF, TRIM22 and RAD21. However, additional Hi-C experiments across multiple cell types will be required to understand the potential differences between cell types, such as described for GATA1/Ldb1 in erythrocyte progenitors in Section 1.2.4. New computational methods such as *Cicero* for predicting promoter-enhancer interactions from scATAC-seq might help to explain these variations (Pliner et al., 2018). In addition, a recent method could also predict chromatin loops on the basis of ChIP-seq of CTCF, RAD21 and ZNF143, which, as also shown by TF-COMB, are highly associated with loop anchors (Ibn-Salem et al., 2019). Combining predicted genome topology with

co-occurrence analysis will be extremely interesting to uncover TFs involved in long-range chromatin interactions within individual cell types.

Due to the flexible input format of TF-COMB, it was also possible to quantify the differences between ChIP-seq and motif-based TF co-occurrences. ChIP-seq is inherently limited by the resolution of peaks. Accordingly, TF-COMB identified the preferred binding distance of TBP to TSSs from motif sites, which was not found using ChIP-seq peaks. This lack of resolution also means that multiple binding sites might be perceived as one peak. The comparison of ChIP-seq peaks and motifs highlighted many examples of this, as self-TF pairs were highly enriched within motif-derived data. While these might be due to technical effects of motif similarity with homologous factors, homotypic clusters of TFs have been observed throughout the genome (Ezer et al., 2014). In addition, ChIP-seq differs from motif data by being able to capture proteins bound to DNA-binding factors, but not necessarily bound to DNA themselves - a type of binding known as piggy-backing (Kato et al., 2004). Thus, if two TFs are highly co-occurring in ChIP-seq data, but the pair is not confirmed by motif data, this might be indicative of piggy-backing behavior. Integration of ChIP-seq and motif analysis might be interesting for identifying such global characteristics of TF binding, and will be important for understanding TF binding from a mechanistic point of view.

Another aspect to consider in the correlation between ChIP-seq and motifs is the assumption that TF binding motifs accurately predict binding of TFs. As described in Section 1.2.3, there are many additional properties of TF binding, which are poorly covered by the current PWM models. In addition, without external information such as expression levels of individual TFs, it is difficult to confidently assign a motif to one TF if several TFs share the same motif. To complicate matters further, it has been shown that TFs can alter their recognition sites when binding cooperatively with other TFs (Jolma et al., 2015). To improve *in silico* prediction of TFBS, the current motif models need to be extended with information about DNA-binding domain restrictions (Sandelin et al., 2004), methylation status, DNA shape and the possibility for co-binding TFs. To pursue this goal, the recent database *TFBSshape* has been established, which collects data on the influence of both DNA shape and DNA methylation on the preference of TF binding (Chiu et al., 2020). However, we are still far from understanding the mechanisms by which TFs obtain target specificity. Furthermore, in the context of the FAIR and FAIR4RS principles, the establishment of a common file-format for these multi-modal motifs might turn out to be the greatest challenge of all.

In conclusion, TF-COMB is able to identify known co-occurring TFs and characteristics for binding. Although not discussed here, it is also able to visualize TFs in a network context. As seen in the case of Sp1 and the Sp1-like repressors, co-occurrence of factors might indicate competition for the same binding locations. Adding another layer of individual TF functionalities might help to improve the interpretation of TF co-occurrence networks.

3.5 The future of epigenetics research

With the cost of sequencing decreasing faster than predicted by Moore's law (November, 2018), application of assays such as ATAC-seq are becoming increasingly relevant to the study of epigenetic mechanisms. Since publication of TOBIAS in 2020, the software has been used on ATAC-seq data from a variety of contexts including the study of TF binding in human esophageal carcinoma (Rogerson et al., 2020), the study of AP-1 binding in injured zebrafish heart (Beisaw et al., 2020), for identifying mechanisms of color vision in *Heliconius* butterflies (McCulloch et al., 2022) and for comparing different cell types within maize (Dai et al., 2022). In conclusion, TF binding is clearly a focus of research throughout the entire tree of life.

Likewise, future method developments will help to investigate TF binding and epigenetics. In particular, improvements within the field of single cell sequencing have recently provided new possibilities for research through multi-omics approaches including parallel quantification of chromatin accessibility with methylation (Guerin et al., 2021) and transcriptomics (Ma et al., 2020). A recent method known as *Perturb-ATAC* combines the effects of scATAC-seq and CRISPR technology to identify effects of TF knockouts on the chromatin landscape (Rubin et al., 2019). One of the latest advances in the single cell field is the ability to map chromatin accessibility in correlation with the spatial location of individual cells (Deng et al., 2022). Thus, rather than studying epigenetic mechanisms in isolation, these combinatorial assays will be instrumental in understanding the interplay of features within epigenetic regulation. The International Human Epigenome Consortium (IHEC) collects complete reference epigenome maps for various cell types. By integrating data from a number of international consortia, the IHEC data portal enables analysis of multiple modalities of epigenetic data, which is paramount for understanding their interplay within each tissue (Bujold et al., 2016).

In parallel with advances in method development, there is a continuous evolution of bioinformatics solutions. Recently, a number of deep-learning methods utilizing neural networks have emerged, such as by Yang et al., 2022, which presented a new method for identifying TF binding from ATAC-seq. Interestingly, this algorithm does not explicitly rely on footprints, but on learning features of Tn5 insertions around occupied TFBS. In the context of motif grammar, Avsec et al., 2021 applied a deep-learning method to ChIP-nexus experiments and found that the model learned a soft binding syntax between pluripotency factors Oct4, Sox2, Nanog and Kl4. However, with the development of increasingly advanced tools, it might be worth to consider the impacts of data storage and heavy computations on energy and CO_2 consumption. In fact, a study recently highlighted that parameters such as optimizing algorithms, controlling RAM usage and running tools in multiprocessing lowers carbon emissions (Grealey et al., 2022). With the increasing focus on climate change, the future usage and development of bioinformatics tools might have to take the carbon footprint into account as well. Replacement of brute force computing by smart solutions will help to save important resources.

Finally, through combination of experimental and computational advances, epigenetic research is starting to be applicable in a clinical context. Drugs such as HDAC inhibitors were the first epigenetic drugs to be approved for treatment against T-cell lymphoma (Eckschlager et al., 2017). As discussed throughout this thesis, TFs hold great promise to be used for reprogramming of terminally differentiated cells (Ulasov et al., 2018). Several studies have already demonstrated the clinical potential to use TFs for promoting regeneration of damaged tissue (Chen et al., 2021) and for reprogramming of cancer cells (Gong et al., 2019). Of course, there are several challenges to the *in vivo* application of TFs and chromatin modifiers as drugs, including stable delivery of the proteins to the target cell nucleus, but several delivery systems are currently under investigation, which may solve such problems (Ulasov et al., 2018). In fact, TFs are already used in the clinic now, as seen by the drug *Gendicine*, a viral gene therapy approved by the Chinese food and drug administration for the treatment of cancer, which works by delivering a wildtype p53 directly to cancerous tumor cells, thus increasing apoptosis (Zhang et al., 2018). In conclusion, the future is now in terms of method development for clinical epigenetics using TFs.

4 Conclusion

This thesis has provided a general overview of epigenetic mechanisms as the drivers of cell type specification throughout development. In particular, TFs have proven to be especially important for changing expression of cell type specific genes, but also have great implications for the reprogramming of cells. It is therefore of great interest to study the ways in which TFs act to regulate gene expression.

The first project focused on the development of a comprehensive tool for identifying occupied TFBS using ATAC-seq data. The results showed that utilizing intrinsic features of ATAC-seq data can successfully uncover factors which are differentially active throughout embryonic development. The second project went into more detail with the collaboration of TFs, which is a necessary mechanism of cells to increase the complexity of gene regulatory networks. The analysis uncovered several characteristics of TF binding, including preferences of TF location in relation to open chromatin, histone marks and 3D genome structure, specific TF-TF binding grammar, and potential TF complexes through network analysis.

While this thesis has a strong focus on developmental mechanisms, such investigations are equally important for understanding the maintenance and repair of differentiated tissues. In fact, the study of early embryonic development has many parallels to reprogramming of cells through dedifferentiation, and understanding these mechanisms is therefore of high priority for clinical applications.

In the context of understanding the genome map, we are still working on learning how to fold, read and interpret it correctly (see Figure 1). In a literal way, enhancer-promoter interactions fold the genome into a 3D shape, which controls gene activation through TF binding, among other mechanisms. But figuratively, different options of folding the genome map can also represent the different ways the genome is able to employ different transcriptional programs depending on the current conformation of the epigenetic landscape. Thus, we still need to study the many different ways in which the genome map can be configured, and understand how this affects the outcome in terms of cell phenotype.

In conclusion, while the influence of TFs in epigenetic regulation is far from understood, the publications of this thesis have contributed to the investigation of these mechanisms through specialized bioinformatics software. Ultimately, continuation of such work, particularly in the understanding of combinatorial logic of TFs within regulatory networks, will further extend our knowledge of TF binding in shaping differentiation throughout development.

References

Adams, E. J., Karthaus, W. R., Hoover, E., (...), and Sawyers, C. L. (2019). "FOXA1 Mutations Alter Pioneering Activity, Differentiation and Prostate Cancer Phenotypes". *Nature* 571.7765, pp. 408–412.

Alda-Catalinas, C., Bredikhin, D., Hernando-Herraez, I., (...), and Reik, W. (2020). "A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program". *Cell Systems* 11.1, 25–41.e9.

Ancelin, K., Syx, L., Borensztein, M., (...), and Heard, E. (2016). "Maternal LSD1/KDM1A Is an Essential Regulator of Chromatin and Transcription Landscapes during Zygotic Genome Activation". *eLife* 5, e08851.

Avsec, Ž., Weilert, M., Shrikumar, A., (...), and Zeitlinger, J. (2021). "Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax". *Nature Genetics* 53.3, pp. 354–366.

Aydin, B. and Mazzoni, E. O. (2019). "Cell Reprogramming: The Many Roads to Success". *Annual Review of Cell and Developmental Biology* 35.1, pp. 433–452.

Baek, S., Goldstein, I., and Hager, G. L. (2017). "Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity". *Cell Reports* 19.8, pp. 1710–1722.

Bannister, A. J. and Kouzarides, T. (2011). "Regulation of Chromatin by Histone Modifications". *Cell Research* 21.3, pp. 381–395.

Barker, M., Chue Hong, N. P., Katz, D. S., (...), and Honeyman, T. (2022). "Introducing the FAIR Principles for Research Software". *Scientific Data* 9.1, p. 622.

Bartosovic, M., Kabbe, M., and Castelo-Branco, G. (2021). "Single-Cell CUT&Tag Profiles Histone Modifications and Transcription Factors in Complex Tissues". *Nature Biotechnology* 39.7, pp. 825–835.

Beisaw, A., Kuenne, C., Guenther, S., (...), and Stainier, D. Y. R. (2020). "AP-1 Contributes to Chromatin Accessibility to Promote Sarcomere Disassembly and Cardiomyocyte Protrusion During Zebrafish Heart Regeneration". *Circulation Research* 126.12, pp. 1760–1778.

Berest, I., Arnold, C., Reyes-Palomares, A., (...), and Zaugg, J. B. (2019). "Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF". *Cell Reports* 29.10, 3147–3159.e12.

Bosnakovski, D., Gearhart, M. D., Ho Choi, S., and Kyba, M. (2021). "Dux Facilitates Post-Implantation Development, but Is Not Essential for Zygotic Genome Activation". *Biology of Reproduction* 104.1, pp. 83–93.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., (...), and Greenleaf, W. J. (2013). "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position". *Nature Methods* 10.12, pp. 1213–1218.

Bujold, D., Morais, D. A. d. L., Gauthier, C., (...), and Bourque, G. (2016). "The International Human Epigenome Consortium Data Portal". *Cell Systems* 3.5, 496–499.e2.

Cedar, H. and Bergman, Y. (2009). "Linking DNA Methylation and Histone Modification: Patterns and Paradigms". *Nature Reviews Genetics* 10.5, pp. 295–304.

Chanoumidou, K., Hernández-Rodríguez, B., Windener, F., (...), and Kuhlmann, T. (2021). "One-Step Reprogramming of Human Fibroblasts into Oligodendrocyte-like Cells by SOX10, OLIG2, and NKX6.2". *Stem Cell Reports* 16.4, pp. 771–783.

Chen, Y., Lüttmann, F. F., Schoger, E., (...), and Braun, T. (2021). "Reversible Reprogramming of Cardiomyocytes to a Fetal State Drives Heart Regeneration in Mice". *Science* 373.6562, pp. 1537–1540.

Chen, Z. and Zhang, Y. (2019). "Loss of DUX Causes Minor Defects in Zygotic Genome Activation and Is Compatible with Mouse Development". *Nature Genetics* 51.6, pp. 947–951.

Chera, S., Baronnier, D., Ghila, L., (...), and Herrera, P. L. (2014). "Diabetes Recovery by Age-Dependent Conversion of Pancreatic δ-Cells into Insulin Producers". *Nature* 514.7523, pp. 503–507.

Chiu, T.-P., Xin, B., Markarian, N., (...), and Rohs, R. (2020). "TFBSshape: An Expanded Motif Database for DNA Shape Features of Transcription Factor Binding Sites". *Nucleic Acids Research* 48.D1, pp. D246–D255.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., (...), and Jaenisch, R. (2010). "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State". *Proceedings of the National Academy of Sciences* 107.50, pp. 21931–21936.

Cusanovich, D. A., Daza, R., Adey, A., (...), and Shendure, J. (2015). "Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing". *Science* 348.6237, pp. 910–914.

Dahl, J. A., Jung, I., Aanes, H., (...), and Klungland, A. (2016). "Broad Histone H3K4me3 Domains in Mouse Oocytes Modulate Maternal-to-Zygotic Transition". *Nature* 537.7621, pp. 548–552.

Dai, X., Tu, X., Du, B., (...), and Li, P. (2022). "Chromatin and Regulatory Differentiation between Bundle Sheath and Mesophyll Cells in Maize". *The Plant Journal* 109.3, pp. 675–692.

Daxinger, L., Tapscott, S. J., and van der Maarel, S. M. (2015). "Genetic and Epigenetic Contributors to FSHD". *Current Opinion in Genetics & Development* 33, pp. 56–61.

de Cristofaro, T., Mascia, A., Pappalardo, A., (...), and Zannini, M. (2009). "Pax8 Protein Stability Is Controlled by Sumoylation". *Journal of Molecular Endocrinology* 42.1, pp. 35–46.

De Iaco, A., Planet, E., Coluccio, A., (...), and Trono, D. (2017). "DUX-family Transcription Factors Regulate Zygotic Genome Activation in Placental Mammals". *Nature Genetics* 49.6, pp. 941–945.

De Iaco, A., Verp, S., Offner, S., (...), and Trono, D. (2020). "DUX Is a Non-Essential Synchronizer of Zygotic Genome Activation". *Development* 147.2, dev177725.

Deng, W., Lee, J., Wang, H., (...), and Blobel, G. A. (2012). "Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor". *Cell* 149.6, pp. 1233–1244.

Deng, Y., Bartosovic, M., Ma, S., (...), and Fan, R. (2022). "Spatial Profiling of Chromatin Accessibility in Mouse and Human Tissues". *Nature* 609.7926, pp. 375–383.

Dhaliwal, N. K., Abatti, L. E., and Mitchell, J. A. (2019). "KLF4 Protein Stability Regulated by Interaction with Pluripotency Transcription Factors Overrides Transcriptional Control." *Genes & development* 33.15-16, pp. 1069–1082.

Di Tommaso, P., Chatzou, M., Floden, E. W., (...), and Notredame, C. (2017). "Nextflow Enables Reproducible Computational Workflows". *Nature Biotechnology* 35.4, pp. 316–319.

Dror, I., Golan, T., Levy, C., (...), and Mandel-Gutfreund, Y. (2015). "A Widespread Role of the Motif Environment in Transcription Factor Binding across Diverse Protein Families". *Genome Research* 25.9, pp. 1268–1280.

Dupressoir, A., Vernochet, C., Bawa, O., (...), and Heidmann, T. (2009). "Syncytin-A Knockout Mice Demonstrate the Critical Role in Placentation of a Fusogenic, Endogenous Retrovirus-Derived, Envelope Gene". *Proceedings of the National Academy of Sciences* 106.29, pp. 12127–12132.

Eckersley-Maslin, M. A., Alda-Catalinas, C., and Reik, W. (2018). "Dynamics of the Epigenetic Landscape during the Maternal-to-Zygotic Transition". *Nature Reviews Molecular Cell Biology* 19.7, pp. 436–450.

Eckschlager, T., Plch, J., Stiborova, M., and Hrabeta, J. (2017). "Histone Deacetylase Inhibitors as Anticancer Drugs". *International Journal of Molecular Sciences* 18.7, p. 1414.

Ezer, D., Zabet, N. R., and Adryan, B. (2014). "Homotypic Clusters of Transcription Factor Binding Sites: A Model System for Understanding the Physical Mechanics of Gene Expression". *Computational and Structural Biotechnology Journal* 10.17, pp. 63–69.

Farley, E. K., Olson, K. M., Zhang, W., (...), and Levine, M. S. (2016). "Syntax Compensates for Poor Binding Sites to Encode Tissue Specificity of Developmental Enhancers". *Proceedings of the National Academy of Sciences* 113.23, pp. 6508–6513.

Fernandez Garcia, M., Moore, C. D., Schulz, K. N., (...), and Zaret, K. S. (2019). "Structural Features of Transcription Factors Associating with Nucleosome Binding". *Molecular Cell* 75.5, 921–932.e6.

Fernandez-Perez, A., Sathe, A. A., Bhakta, M., (...), and Munshi, N. V. (2019). "Hand2 Selectively Reorganizes Chromatin Accessibility to Induce Pacemaker-like Transcriptional Reprogramming". *Cell Reports* 27.8, 2354–2369.e7.

Ferreira, R., Magnaghi-Jaulin, L., Robin, P., (...), and Trouche, D. (1998). "The Three Members of the Pocket Proteins Family Share the Ability to Repress E2F Activity through Recruitment of a Histone Deacetylase". *Proceedings of the National Academy of Sciences* 95.18, pp. 10493–10498.

Fornes, O., Castro-Mondragon, J. A., Khan, A., (...), and Mathelier, A. (2020). "JASPAR 2020: Update of the Open-Access Database of Transcription Factor Binding Profiles". *Nucleic Acids Research* 48.D1, pp. D87–D92.

Frietze, S. and Farnham, P. J. (2011). "Transcription Factor Effector Domains". *A Handbook of Transcription Factors*. Springer Netherlands, pp. 261–277.

Furlong, E. E. M. and Levine, M. (2018). "Developmental Enhancers and Chromosome Topology". *Science* 361.6409, pp. 1341–1345.

Galas, D. J. and Schmitz, A. (1978). "DNAse Footprinting: A Simple Method for the Detection of Protein-DNA Binding Specificity." *Nucleic acids research* 5.9, pp. 3157–3170.

Gao, Y., Liu, X., Tang, B., (...), and Gao, S. (2017). "Protein Expression Landscape of Mouse Embryos during Pre-implantation Development". *Cell Reports* 21.13, pp. 3957–3969.

Gasperini, M., Tome, J. M., and Shendure, J. (2020). "Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers". *Nature Reviews Genetics* 21.5, pp. 292–310.

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). "A Brief History of Bioinformatics". *Briefings in Bioinformatics* 20.6, pp. 1981–1996.

Gerber, T., Murawala, P., Knapp, D., (...), and Treutlein, B. (2018). "Single-Cell Analysis Uncovers Convergence of Cell Identities during Axolotl Limb Regeneration". *Science* 362.6413, eaaq0681.

Gifford, W. D., Pfaff, S. L., and Macfarlan, T. S. (2013). "Transposable Elements as Genetic Regulatory Substrates in Early Development". *Trends in Cell Biology* 23.5, pp. 218–226.

Giraldez, A. J., Mishima, Y., Rihel, J., (...), and Schier, A. F. (2006). "Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal mRNAs". *Science* 312.5770, pp. 75–79.

Godwin, J. W. and Rosenthal, N. (2014). "Scar-Free Wound Healing and Regeneration in Amphibians: Immunological Influences on Regenerative Success". *Differentiation* 87.1-2, pp. 66–75.

Goldman, S. L., MacKay, M., Afshinnekoo, E., (...), and Mason, C. E. (2019). "The Impact of Heterogeneity on Single-Cell Sequencing". *Frontiers in Genetics* 10, p. 8.

Gong, L., Yan, Q., Zhang, Y., (...), and Guan, X. (2019). "Cancer Cell Reprogramming: A Promising Therapy Converting Malignancy to Benignity". *Cancer Communications* 39.1, p. 48.

Goryshin, I. Y., Miller, J. A., Kil, Y. V., (...), and Reznikoff, W. S. (1998). "Tn5/IS50 Target Recognition". *Proceedings of the National Academy of Sciences* 95.18, pp. 10716–10721.

Grealey, J., Lannelongue, L., Saw, W.-Y., (...), and Inouye, M. (2022). "The Carbon Footprint of Bioinformatics". *Molecular Biology and Evolution* 39.3, msac034.

Green, B., Bouchier, C., Fairhead, C., (...), and Cormack, B. P. (2012). "Insertion Site Preference of Mu, Tn5, and Tn7 Transposons". *Mobile DNA* 3.1, p. 3.

Grinberg, A. V., Hu, C.-D., and Kerppola, T. K. (2004). "Visualization of Myc/Max/Mad Family Dimers and the Competition for Dimerization in Living Cells". *Molecular and Cellular Biology* 24.10, pp. 4294–4308.

Guerin, L. N., Barnett, K. R., and Hodges, E. (2021). "Dual Detection of Chromatin Accessibility and DNA Methylation Using ATAC-Me". *Nature Protocols* 16.12, pp. 5377–5397.

Guo, H., Zhu, P., Yan, L., (...), and Qiao, J. (2014). "The DNA Methylation Landscape of Human Early Embryos". *Nature* 511.7511, pp. 606–610.

Guo, Y., Xu, Q., Canzio, D., (...), and Wu, Q. (2015). "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function". *Cell* 162.4, pp. 900–910.

Guo, Y., Kitano, T., Inoue, K., (...), and Siomi, H. (2022). "Obox4 Secures Zygotic Genome Activation upon Loss of Dux". *bioRxiv*, p. 2022.07.04.498763.

Gusmao, E. G., Dieterich, C., Zenke, M., and Costa, I. G. (2014). "Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications". *Bioinformatics* 30.22, pp. 3143–3151.

Hannenhalli, S. and Kaestner, K. H. (2009). "The Evolution of Fox Genes and Their Role in Development and Disease". *Nature Reviews Genetics* 10.4, pp. 233–240.

Harrison, M. M., Li, X.-Y., Kaplan, T., (...), and Eisen, M. B. (2011). "Zelda Binding in the Early Drosophila Melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition". *PLOS Genetics* 7.10, e1002266.

Hatano, A., Chiba, H., Moesa, H. A., (...), and Fujibuchi, W. (2011). "CELLPEDIA: A Repository for Human Cell Information for Cell Studies and Differentiation Analyses". *Database* 2011, bar046.

Hendrickson, P. G., Doráis, J. A., Grow, E. J., (...), and Cairns, B. R. (2017). "Conserved Roles of Mouse DUX and Human DUX4 in Activating Cleavage-Stage Genes and MERVL/HERVL Retrotransposons". *Nature Genetics* 49.6, pp. 925–934.

Hesselberth, J. R., Chen, X., Zhang, Z., (...), and Stamatoyannopoulos, J. A. (2009). "Global Mapping of Protein-DNA Interactions in Vivo by Digital Genomic Footprinting". *Nature Methods* 6.4, pp. 283–289.

Heyn, P., Kircher, M., Dahl, A., (...), and Neugebauer, K. M. (2014). "The Earliest Transcribed Zygotic Genes Are Short, Newly Evolved, and Different across Species". *Cell Reports* 6.2, pp. 285–292.

Hirata, T., Amano, T., Nakatake, Y., (...), and Ko, M. S. H. (2012). "Zscan4 Transiently Reactivates Early Embryonic Genes during the Generation of Induced Pluripotent Stem Cells". *Scientific Reports* 2.1, p. 208.

Hwang, B., Lee, J. H., and Bang, D. (2018). "Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines". *Experimental & Molecular Medicine* 50.8, pp. 1–14.

Ibn-Salem, J. and Andrade-Navarro, M. A. (2019). "7C: Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF Motifs". *BMC Genomics* 20.1, p. 777.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling". *Science* 324.5924, pp. 218–223.

Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., (...), and Zaret, K. S. (2016). "The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation". *Molecular Cell* 62.1, pp. 79–91.

Iwafuchi-Doi, M. and Zaret, K. S. (2014). "Pioneer Transcription Factors in Cell Reprogramming". *Genes & Development* 28.24, pp. 2679–2692.

Iyer, B. V., Kenward, M., and Arya, G. (2011). "Hierarchies in Eukaryotic Genome Organization: Insights from Polymer Theory and Simulations". *BMC Biophysics* 4.1, p. 8.

Jachowicz, J. W., Bing, X., Pontabry, J., (...), and Torres-Padilla, M.-E. (2017). "LINE-1 Activation after Fertilization Regulates Global Chromatin Accessibility in the Early Mouse Embryo". *Nature Genetics* 49.10, pp. 1502–1510.

Jindal, G. A. and Farley, E. K. (2021). "Enhancer Grammar in Development, Evolution, and Disease: Dependencies and Interplay". *Developmental Cell* 56.5, pp. 575–587.

Jolma, A., Yin, Y., Nitta, K. R., (...), and Taipale, J. (2015). "DNA-dependent Formation of Transcription Factor Pairs Alters Their Binding Specificity". *Nature* 527.7578, pp. 384–388.

Jopling, C., Boue, S., and Belmonte, J. C. I. (2011). "Dedifferentiation, Transdifferentiation and Reprogramming: Three Routes to Regeneration". *Nature Reviews Molecular Cell Biology* 12.2, pp. 79–89.

Jopling, C., Sleep, E., Raya, M., (...), and Belmonte, J. C. I. (2010). "Zebrafish Heart Regeneration Occurs by Cardiomyocyte Dedifferentiation and Proliferation". *Nature* 464.7288, pp. 606–609.

Kähärä, J. and Lähdesmäki, H. (2015). "BinDNase: A Discriminatory Approach for Transcription Factor Binding Prediction Using DNase I Hypersensitivity Data". *Bioinformatics* 31.17, pp. 2852–2859.

Kato, M., Hata, N., Banerjee, N., (...), and Zhang, M. Q. (2004). "Identifying Combinatorial Regulation of Transcription Factors and Binding Motifs". *Genome Biology* 5.8, R56.

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., (...), and Henikoff, S. (2019). "CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells". *Nature Communications* 10.1, p. 1930.

Kazemian, M., Pham, H., Wolfe, S. A., (...), and Sinha, S. (2013). "Widespread Evidence of Cooperative DNA Binding by Transcription Factors in Drosophila Development". *Nucleic Acids Research* 41.17, pp. 8237–8252.

Ke, Y., Xu, Y., Chen, X., (...), and Liu, J. (2017). "3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis". *Cell* 170.2, 367–381.e20.

Kern, F., Fehlmann, T., and Keller, A. (2020). "On the Lifetime of Bioinformatics Web Services". *Nucleic Acids Research* 48.22, pp. 12523–12533.

Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). "Chromatin Accessibility and the Regulatory Epigenome". *Nature Reviews Genetics* 20.4, pp. 207–220.

Koster, J. and Rahmann, S. (2012). "Snakemake–a Scalable Bioinformatics Workflow Engine". *Bioinformatics* 28.19, pp. 2520–2522.

Krivega, I. and Dean, A. (2017). "LDB1-mediated Enhancer Looping Can Be Established Independent of Mediator and Cohesin". *Nucleic Acids Research* 45.14, pp. 8255–8268.

Kubin, T., Pöling, J., Kostin, S., (...), and Braun, T. (2011). "Oncostatin M Is a Major Mediator of Cardiomyocyte Dedifferentiation and Remodeling". *Cell Stem Cell* 9.5, pp. 420–432.

Kulakovskiy, I. V., Vorontsov, I. E., Yevshin, I. S., (...), and Makeev, V. J. (2016). "HOCOMOCO: Expansion and Enhancement of the Collection of Transcription Factor Binding Sites Models". *Nucleic Acids Research* 44.D1, pp. D116–D125.

Lambert, S. A., Jolma, A., Campitelli, L. F., (...), and Weirauch, M. T. (2018). "The Human Transcription Factors". *Cell* 172.4, pp. 650–665.

Lee, M. T., Bonneau, A. R., Takacs, C. M., (...), and Giraldez, A. J. (2013). "Nanog, Pou5f1 and SoxB1 Activate Zygotic Gene Expression during the Maternal-to-Zygotic Transition". *Nature* 503.7476, pp. 360–364.

Leidenroth, A. and Hewitt, J. E. (2010). "A Family History of DUX4: Phylogenetic Analysis of DUXA, B, C and Duxbl Reveals the Ancestral DUX Gene". *BMC Evolutionary Biology* 10.1, p. 364.

Levitsky, V., Zemlyanskaya, E., Oshchepkov, D., (...), and Merkulova, T. (2019). "A Single ChIP-seq Dataset Is Sufficient for Comprehensive Analysis of Motifs Co-Occurrence with MCOT Package". *Nucleic Acids Research* 47.21, e139–e139.

Li, L., Baibakov, B., and Dean, J. (2008). "A Subcortical Maternal Complex Essential for Preimplantation Mouse Embryogenesis". *Developmental Cell* 15.3, pp. 416–425.

Li, N., Jin, K., Bai, Y., (...), and Liu, B. (2020). "Tn5 Transposase Applied in Genomics Research". *International Journal of Molecular Sciences* 21.21, p. 8329.

Li, Z., Schulz, M. H., Look, T., (...), and Costa, I. G. (2019). "Identification of Transcription Factor Binding Sites Using ATAC-seq". *Genome Biology* 20.1, p. 45.

Lu, F., Liu, Y., Inoue, A., (...), and Zhang, Y. (2016). "Establishing Chromatin Regulatory Landscape during Mouse Preimplantation Development". *Cell* 165.6, pp. 1375–1388.

Ma, S., Zhang, B., LaFave, L. M., (...), and Buenrostro, J. D. (2020). "Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin". *Cell* 183.4, 1103–1116.e20.

Macfarlan, T. S., Gifford, W. D., Agarwal, S., (...), and Pfaff, S. L. (2011). "Endogenous Retroviruses and Neighboring Genes Are Coordinately Repressed by LSD1/KDM1A." *Genes & development* 25.6, pp. 594–607.

Macfarlan, T. S., Gifford, W. D., Driscoll, S., (...), and Pfaff, S. L. (2012). "Embryonic Stem Cell Potency Fluctuates with Endogenous Retrovirus Activity". *Nature* 487.7405, pp. 57–63.

Madissoon, E., Jouhilahti, E.-M., Vesterlund, L., (...), and Kere, J. (2016). "Characterization and Target Genes of Nine Human PRD-like Homeobox Domain Genes Expressed Exclusively in Early Embryos". *Scientific Reports* 6.1, p. 28995.

Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A., and Landsman, D. (2005). "Histone Structure and Nucleosome Stability". *Expert Review of Proteomics* 2.5, pp. 719–729.

Martins, A. L., Walavalkar, N. M., Anderson, W. D., (...), and Guertin, M. J. (2018). "Universal Correction of Enzymatic Sequence Bias Reveals Molecular Signatures of Protein/DNA Interactions". *Nucleic Acids Research* 46.2, e9.

Mathelier, A., Xin, B., Chiu, T.-P., (...), and Wasserman, W. W. (2016). "DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo." *Cell systems* 3.3, 278–286.e4.

Maurano, M. T., Humbert, R., Rynes, E., (...), and Stamatoyannopoulos, J. A. (2012). "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA". *Science* 337.6099, pp. 1190–1195.

Mayran, A., Sochodolsky, K., Khetchoumian, K., (...), and Drouin, J. (2019). "Pioneer and Nonpioneer Factor Cooperation Drives Lineage Specific Chromatin Opening". *Nature Communications* 10.1, p. 3807.

McCulloch, K. J., Macias-Muñoz, A., Mortazavi, A., and Briscoe, A. D. (2022). "Multiple Mechanisms of Photoreceptor Spectral Tuning in *Heliconius* Butterflies". *Molecular Biology and Evolution* 39.4, msac067.

Melcer, S. and Meshorer, E. (2010). "Chromatin Plasticity in Pluripotent Cells." *Essays in biochemistry* 48.1, pp. 245–262.

Merika, M. and Thanos, D. (2001). "Enhanceosomes". *Current Opinion in Genetics & Development* 11.2, pp. 205–208.

Merrell, A. J. and Stanger, B. Z. (2016). "Adult Cell Plasticity in Vivo: De-Differentiation and Transdifferentiation Are Back in Style". *Nature Reviews Molecular Cell Biology* 17.7, pp. 413–425.

Moris, N., Pina, C., and Arias, A. M. (2016). "Transition States and Cell Fate Decisions in Epigenetic Landscapes". *Nature Reviews Genetics* 17.11, pp. 693–703.

Neph, S., Vierstra, J., Stergachis, A. B., (...), and Stamatoyannopoulos, J. A. (2012). "An Expansive Human Regulatory Lexicon Encoded in Transcription Factor Footprints". *Nature* 489.7414, pp. 83–90.

Nicetto, D. and Zaret, K. S. (2019). "Role of H3K9me3 Heterochromatin in Cell Identity Establishment and Maintenance". *Current Opinion in Genetics & Development* 55, pp. 1–10.

November, J. (2018). "More than Moore's Mores: Computers, Genomics, and the Embrace of Innovation". *Journal of the History of Biology* 51.4, pp. 807–840.

Ong, C.-T. and Corces, V. G. (2014). "CTCF: An Architectural Protein Bridging Genome Topology and Function". *Nature Reviews Genetics* 15.4, pp. 234–246.

Ouyang, N. and Boyle, A. P. (2020). "TRACE: Transcription Factor Footprinting Using Chromatin Accessibility Data and DNA Sequence". *Genome Research* 30.7, pp. 1040–1046.

Pajcini, K. V., Corbel, S. Y., Sage, J., (...), and Blau, H. M. (2010). "Transient Inactivation of Rb and ARF Yields Regenerative Cells from Postmitotic Mammalian Muscle". *Cell Stem Cell* 7.2, pp. 198–213.

Panne, D. (2008). "The Enhanceosome." *Current opinion in structural biology* 18.2, pp. 236–242.

Park, P. J. (2009). "ChIP–Seq: Advantages and Challenges of a Maturing Technology". *Nature Reviews Genetics* 10.10, pp. 669–680.

Payer, L. M. and Burns, K. H. (2019). "Transposable Elements in Human Genetic Disease". *Nature Reviews Genetics* 20.12, pp. 760–772.

Pearson, J. C., Lemons, D., and McGinnis, W. (2005). "Modulating Hox Gene Functions during Animal Body Patterning". *Nature Reviews Genetics* 6.12, pp. 893–904.

Peaston, A. E., Evsikov, A. V., Graber, J. H., (...), and Knowles, B. B. (2004). "Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos". *Developmental Cell* 7.4, pp. 597–606.

Pennacchio, L. A., Bickmore, W., Dean, A., (...), and Bejerano, G. (2013). "Enhancers: Five Essential Questions". *Nature Reviews Genetics* 14.4, pp. 288–295.

Perna, S., Pinoli, P., Ceri, S., and Wong, L. (2018). "TICA: Transcriptional Interaction and Coregulation Analyzer". *Genomics, Proteomics & Bioinformatics* 16.5, pp. 342–353.

Piper, J., Elze, M. C., Cauchy, P., (...), and Ott, S. (2013). "Wellington: A Novel Method for the Accurate Identification of Digital Genomic Footprints from DNase-seq Data". *Nucleic Acids Research* 41.21, e201–e201.

Platanitis, E., Demiroz, D., Schneller, A., (...), and Decker, T. (2019). "A Molecular Switch from STAT2-IRF9 to ISGF3 Underlies Interferon-Induced Gene Transcription". *Nature Communications* 10.1, p. 2921.

Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., (...), and Trapnell, C. (2018). "Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data". *Molecular Cell* 71.5, 858–871.e8.

Porrello, E. R., Mahmoud, A. I., Simpson, E., (...), and Sadek, H. A. (2011). "Transient Regenerative Potential of the Neonatal Mouse Heart". *Science* 331.6020, pp. 1078–1080.

Preussner, J., Zhong, J., Sreenivasan, K., (...), and Kim, J. (2018). "Oncogenic Amplification of Zygotic Dux Factors in Regenerating P53-Deficient Muscle Stem Cells Defines a Molecular Cancer Subtype". *Cell Stem Cell* 23.6, 794–805.e4.

Raj, A., Shim, H., Gilad, Y., (...), and Stephens, M. (2015). "msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding". *PLOS ONE* 10.9, e0138030.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., (...), and Kellis, M. (2015). "Integrative Analysis of 111 Reference Human Epigenomes". *Nature* 518.7539, pp. 317–330.

Rogerson, C., Ogden, S., Britton, E., (...), and Sharrocks, A. D. (2020). "Repurposing of KLF5 Activates a Cell Cycle Signature during the Progression from a Precursor State to Oesophageal Adenocarcinoma". *eLife* 9, e57189.

Rossi, M. J., Lai, W. K. M., and Pugh, B. F. (2018). "Simplified ChIP-exo Assays". *Nature Communications* 9.1, p. 2842.

Roy, A. L. and Conroy, R. S. (2018). "Toward Mapping the Human Body at a Cellular Resolution". *Molecular Biology of the Cell* 29.15, pp. 1779–1785.

Royall, A. H., Maeso, I., Dunwell, T. L., and Holland, P. W. H. (2018). "Mouse Obox and Crxos Modulate Preimplantation Transcriptional Profiles Revealing Similarity between Paralogous Mouse and Human Homeobox Genes". *EvoDevo* 9.1, p. 2.

Rubin, A. J., Parker, K. R., Satpathy, A. T., (...), and Khavari, P. A. (2019). "Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks". *Cell* 176.1-2, 361–376.e17.

Russell, P. H., Johnson, R. L., Ananthan, S., (...), and Carlson, N. E. (2018). "A Large-Scale Analysis of Bioinformatics Code on GitHub". *PLOS ONE* 13.10, e0205898.

Saksouk, N., Simboeck, E., and Déjardin, J. (2015). "Constitutive Heterochromatin Formation and Transcription in Mammals". *Epigenetics & Chromatin* 8.1, p. 3.

Sandelin, A. and Wasserman, W. W. (2004). "Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics". *Journal of Molecular Biology* 338.2, pp. 207–215.

Schneider, T. D. and Stephens, R. (1990). "Sequence Logos: A New Way to Display Consensus Sequences". *Nucleic Acids Research* 18.20, pp. 6097–6100.

Schulz, K. N. and Harrison, M. M. (2019). "Mechanisms Regulating Zygotic Genome Activation." *Nature Reviews Genetics* 20.4, pp. 221–234.

Sekiya, S. and Suzuki, A. (2012). "Intrahepatic Cholangiocarcinoma Can Arise from Notch-mediated Conversion of Hepatocytes". *Journal of Clinical Investigation* 122.11, pp. 3914–3918.

Sender, R., Fuchs, S., and Milo, R. (2016). "Revised Estimates for the Number of Human and Bacteria Cells in the Body". *PLOS Biology* 14.8, e1002533.

Sha, Q.-Q., Zhu, Y.-Z., Li, S., (...), and Fan, H.-Y. (2020). "Characterization of Zygotic Genome Activation-Dependent Maternal mRNA Clearance in Mouse". *Nucleic Acids Research* 48.2, pp. 879–894.

Siddharthan, R. (2010). "Dinucleotide Weight Matrices for Predicting Transcription Factor Binding Sites: Generalizing the Position Weight Matrix". *PLOS ONE* 5.3, e9722.

Skene, P. J. and Henikoff, S. (2017). "An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites". *eLife* 6, e21856.

Sönmezer, C., Kleinendorst, R., Imanci, D., (...), and Krebs, A. R. (2021). "Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo". *Molecular Cell* 81.2, 255–267.e6.

Soufi, A. (2014). "Mechanisms for Enhancing Cellular Reprogramming". *Current Opinion in Genetics & Development* 25, pp. 101–109.

Soufi, A., Donahue, G., and Zaret, K. S. (2012). "Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome". *Cell* 151.5, pp. 994–1004.

Soufi, A., Garcia, M. F., Jaroszewicz, A., (...), and Zaret, K. S. (2015). "Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming". *Cell* 161.3, pp. 555–568.

Spears, E., Serafimidis, I., Powers, A. C., and Gavalas, A. (2021). "Debates in Pancreatic Beta Cell Biology: Proliferation Versus Progenitor Differentiation and Transdifferentiation in Restoring β Cell Mass". *Frontiers in Endocrinology* 12, p. 722250.

Spitz, F. and Furlong, E. E. M. (2012). "Transcription Factors: From Enhancer Binding to Developmental Control". *Nature Reviews Genetics* 13.9, pp. 613–626.

Squair, J. W., Gautier, M., Kathe, C., (...), and Courtine, G. (2021). "Confronting False Discoveries in Single-Cell Differential Expression". *Nature Communications* 12.1, p. 5692.

Steiniger-White, M., Rayment, I., and Reznikoff, W. S. (2004). "Structure/Function Insights into Tn5 Transposition". *Current Opinion in Structural Biology* 14.1, pp. 50–57.

Sung, M.-H., Guertin, M. J., Baek, S., and Hager, G. L. (2014). "DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence". *Molecular Cell* 56.2, pp. 275–285.

Szabo, Q., Bantignies, F., and Cavalli, G. (2019). "Principles of Genome Folding into Topologically Associating Domains". *Science Advances* 5.4, eaaw1668.

Takahashi, K. and Yamanaka, S. (2006). "Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors". *Cell* 126.4, pp. 663–676.

Taubert, S., Gorrini, C., Frank, S. R., (...), and Amati, B. (2004). "E2F-Dependent Histone Acetylation and Recruitment of the Tip60 Acetyltransferase Complex to Chromatin in Late G1". *Molecular and Cellular Biology* 24.10, pp. 4546–4556.

The 1000 Genomes Project Consortium (2015). "A Global Reference for Human Genetic Variation". *Nature* 526.7571, pp. 68–74.

Thiel, G., Lietz, M., and Hohl, M. (2004). "How Mammalian Transcriptional Repressors Work." *European Journal of Biochemistry* 271.14, pp. 2855–2862.

Thomas, M. C. and Chiang, C.-M. (2006). "The General Transcription Machinery and General Cofactors". *Critical Reviews in Biochemistry and Molecular Biology* 41.3, pp. 105–178.

Thorel, F., Népote, V., Avril, I., (...), and Herrera, P. L. (2010). "Conversion of Adult Pancreatic α-Cells to β-Cells after Extreme β-Cell Loss". *Nature* 464.7292, pp. 1149–1154.

Töhönen, V., Katayama, S., Vesterlund, L., (...), and Kere, J. (2015). "Novel PRD-like Homeodomain Transcription Factors and Retrotransposon Elements in Early Human Development". *Nature Communications* 6.1, p. 8207.

Ulasov, A. V., Rosenkranz, A. A., and Sobolev, A. S. (2018). "Transcription Factors: Time to Deliver". *Journal of Controlled Release* 269, pp. 24–35.

van der Meulen, T., Mawla, A. M., DiGruccio, M. R., (...), and Huising, M. O. (2017). "Virgin Beta Cells Persist throughout Life at a Neogenic Niche within Pancreatic Islets". *Cell Metabolism* 25.4, 911–926.e6.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). "A Census of Human Transcription Factors: Function, Expression and Evolution". *Nature Reviews Genetics* 10.4, pp. 252–263.

Venter, J. C., Adams, M. D., Myers, E. W., (...), and Zhu, X. (2001). "The Sequence of the Human Genome". *Science* 291.5507, pp. 1304–1351.

Waddington, C. H. (1942). "The Epigenotype". *Endeavour*, pp. 18–20.

Waddington, C. (1957). *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology.* London: Allen & Unwin.

Wan, L.-B., Pan, H., Hannenhalli, S., (...), and Bartolomei, M. S. (2008). "Maternal Depletion of CTCF Reveals Multiple Functions during Oocyte and Preimplantation Embryo Development". *Development* 135.16, pp. 2729–2738.

Wasserman, W. W. and Sandelin, A. (2004). "Applied Bioinformatics for the Identification of Regulatory Elements". *Nature Reviews Genetics* 5.4, pp. 276–287.

Whitington, T., Frith, M. C., Johnson, J., and Bailey, T. L. (2011). "Inferring Transcription Factor Complexes from ChIP-seq Data". *Nucleic Acids Research* 39.15, e98–e98.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., (...), and Mons, B. (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship". *Scientific Data* 3.1, p. 160018.

Wu, J., Huang, B., Chen, H., (...), and Xie, W. (2016). "The Landscape of Accessible Chromatin in Mammalian Preimplantation Embryos". *Nature* 534.7609, pp. 652–657.

Wu, J., Xu, J., Liu, B., (...), and Sun, Y. (2018). "Chromatin Analysis in Human Early Development Reveals Epigenetic Transition during ZGA". *Nature* 557.7704, pp. 256–260.

Wu, W.-S. and Lai, F.-J. (2015). "Functional Redundancy of Transcription Factors Explains Why Most Binding Targets of a Transcription Factor Are Not Affected When the Transcription Factor Is Knocked Out". *BMC Systems Biology* 9.Suppl 6, S2.

Wunderlich, Z. and Mirny, L. A. (2009). "Different Gene Regulation Strategies Revealed by Analysis of Binding Motifs". *Trends in Genetics* 25.10, pp. 434–440.

Yamane, M., Ohtsuka, S., Matsuura, K., (...), and Niwa, H. (2018). "Overlapping Functions of Krüppel-like Factor Family Members: Targeting Multiple Transcription Factors to Maintain the Naïve Pluripotency of Mouse Embryonic Stem Cells". *Development* 145.10, dev162404.

Yang, T. and Henao, R. (2022). "TAMC: A Deep-Learning Approach to Predict Motif-Centric Transcriptional Factor Binding Activity Based on ATAC-seq Profile". *PLOS Computational Biology* 18.9, e1009921.

Yanger, K., Zong, Y., Maggs, L. R., (...), and Stanger, B. Z. (2013). "Robust Cellular Reprogramming Occurs Spontaneously during Liver Regeneration". *Genes & Development* 27.7, pp. 719–724.

Yao, Y. (2020). "Dedifferentiation: Inspiration for Devising Engineering Strategies for Regenerative Medicine". *npj Regenerative Medicine*, p. 11.

Ye, L., D'Agostino, G., Loo, S. J., (...), and Cook, S. A. (2018). "Early Regenerative Capacity in the Porcine Heart". *Circulation* 138.24, pp. 2798–2808.

Yin, Y., Morgunova, E., Jolma, A., (...), and Taipale, J. (2017). "Impact of Cytosine Methylation on DNA Binding Specificities of Human Transcription Factors". *Science* 356.6337, eaaj2239.

Yu, C., Ji, S.-Y., Dang, Y.-J., (...), and Fan, H.-Y. (2016). "Oocyte-Expressed Yes-Associated Protein Is a Key Activator of the Early Zygotic Genome in Mouse". *Cell Research* 26.3, pp. 275–287.

Zhang, W.-W., Li, L., Li, D., (...), and Lam, D. M.-K. (2018). "The First Approved Gene Therapy Product for Cancer Ad-*P53* (Gendicine): 12 Years in the Clinic". *Human Gene Therapy* 29.2, pp. 160–179.

Zhang, Y., Li, T.-S., Lee, S.-T., (...), and Marbán, E. (2010). "Dedifferentiation and Proliferation of Mammalian Cardiomyocytes". *PLOS ONE* 5.9, e12559.

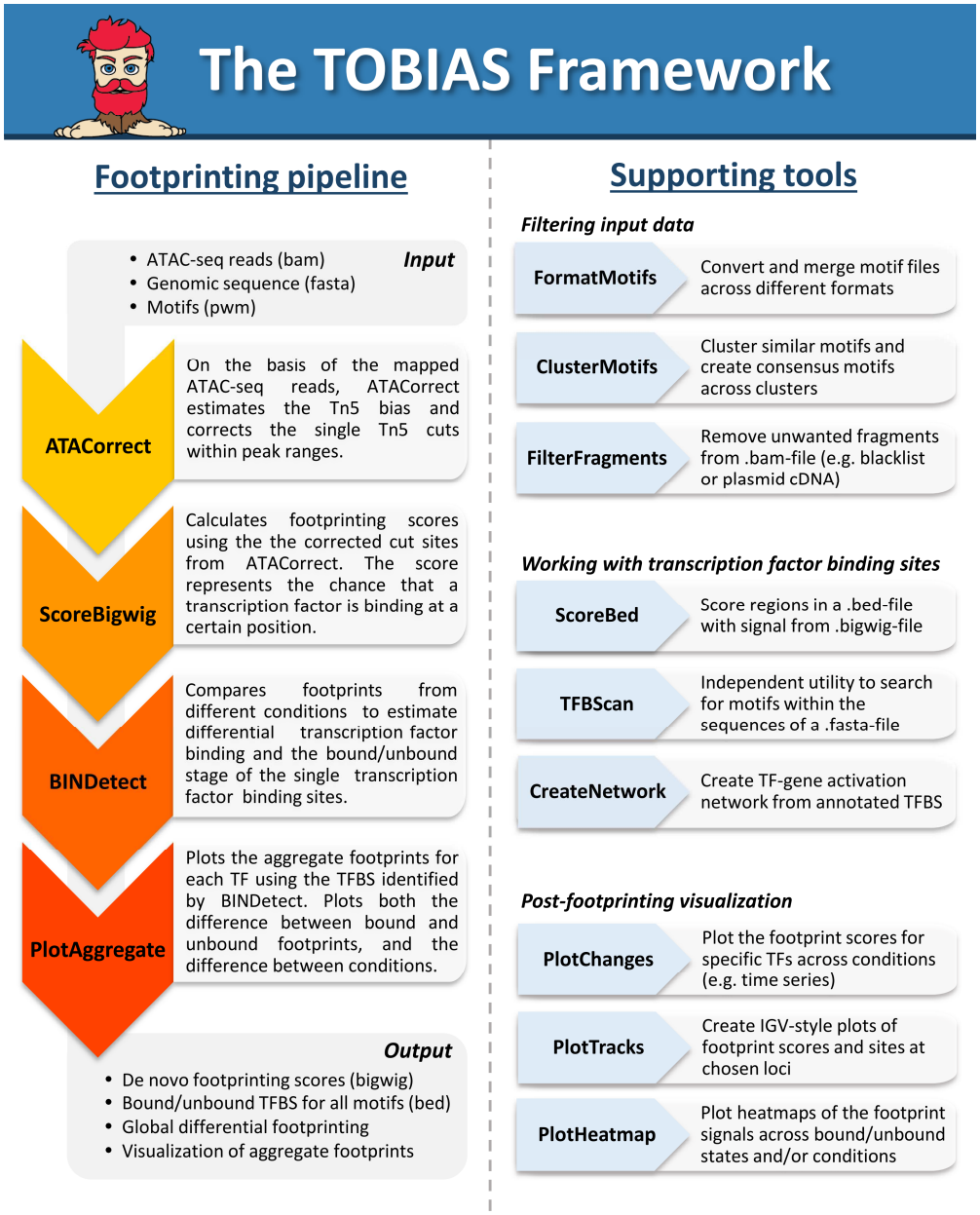
Zhang, Z., Chang, C. W., Goh, W. L., (...), and Cheung, E. (2011). "CENTDIST: Discovery of Co-Associated Factors by Motif Distribution". *Nucleic Acids Research* 39.Web Server issue, W391–W399.

Zhao, H., Chen, Z.-J., Qin, Y., (...), and Rajkovic, A. (2008). "Transcription Factor FIGLA Is Mutated in Patients with Premature Ovarian Failure". *The American Journal of Human Genetics* 82.6, pp. 1342–1348.

Zou, Z., Zhang, C., Wang, Q., (...), and Xie, W. (2022). "Translatome and Transcriptome Co-Profiling Reveals a Role of TPRXs in Human Zygotic Genome Activation". *Science* 378.6615, abo7923.

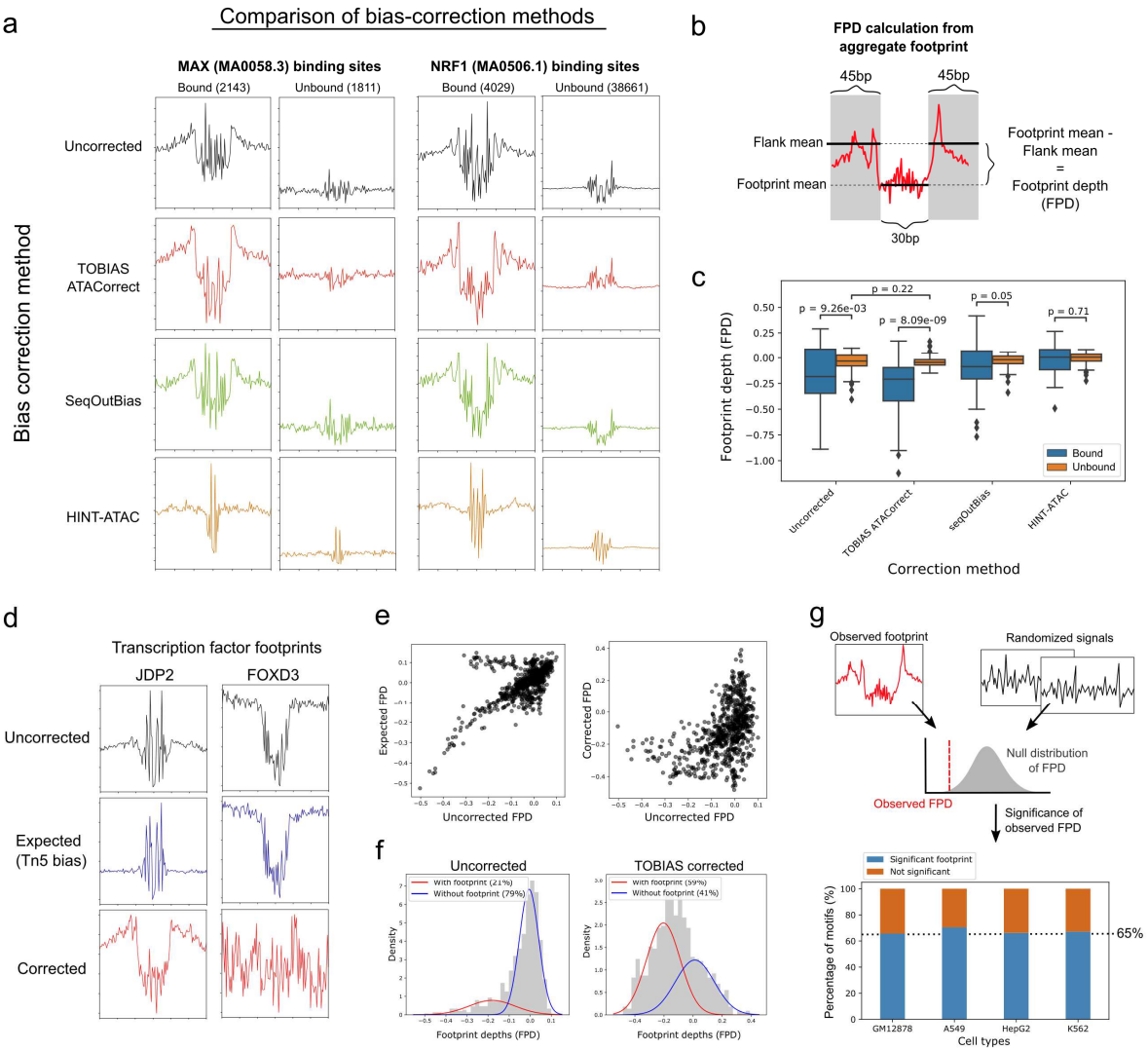
Appendices

A1 Publication 1: Supplementary figures



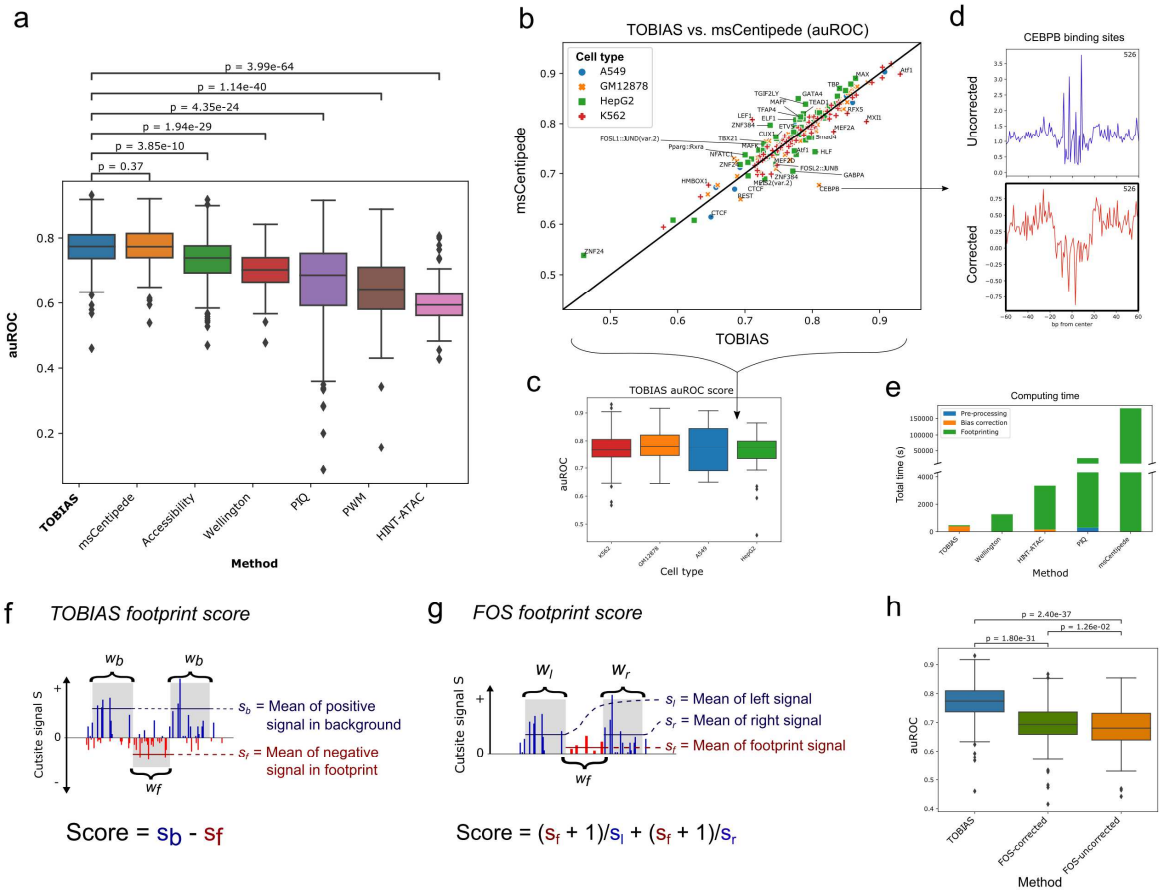
Supplementary Figure 1: Overview of the TOBIAS framework tools

The TOBIAS tools are intended for use in a standardized pipeline as shown on the left. ATACorrect and ScoreBigwig corrects Tn5 cuts and calculates footprint scores respectively. Next, BINDetect introduces information from different transcription factor binding motifs to predict binding sites both within and across conditions. PlotAggregate is used to visualize aggregate footprints. Furthermore, a large variety of supporting tools can be used at different stages of the pipeline (right), such as pre-filtering of .bam-files using FilterFragments or plotting of locus-specific footprints using PlotTracks.



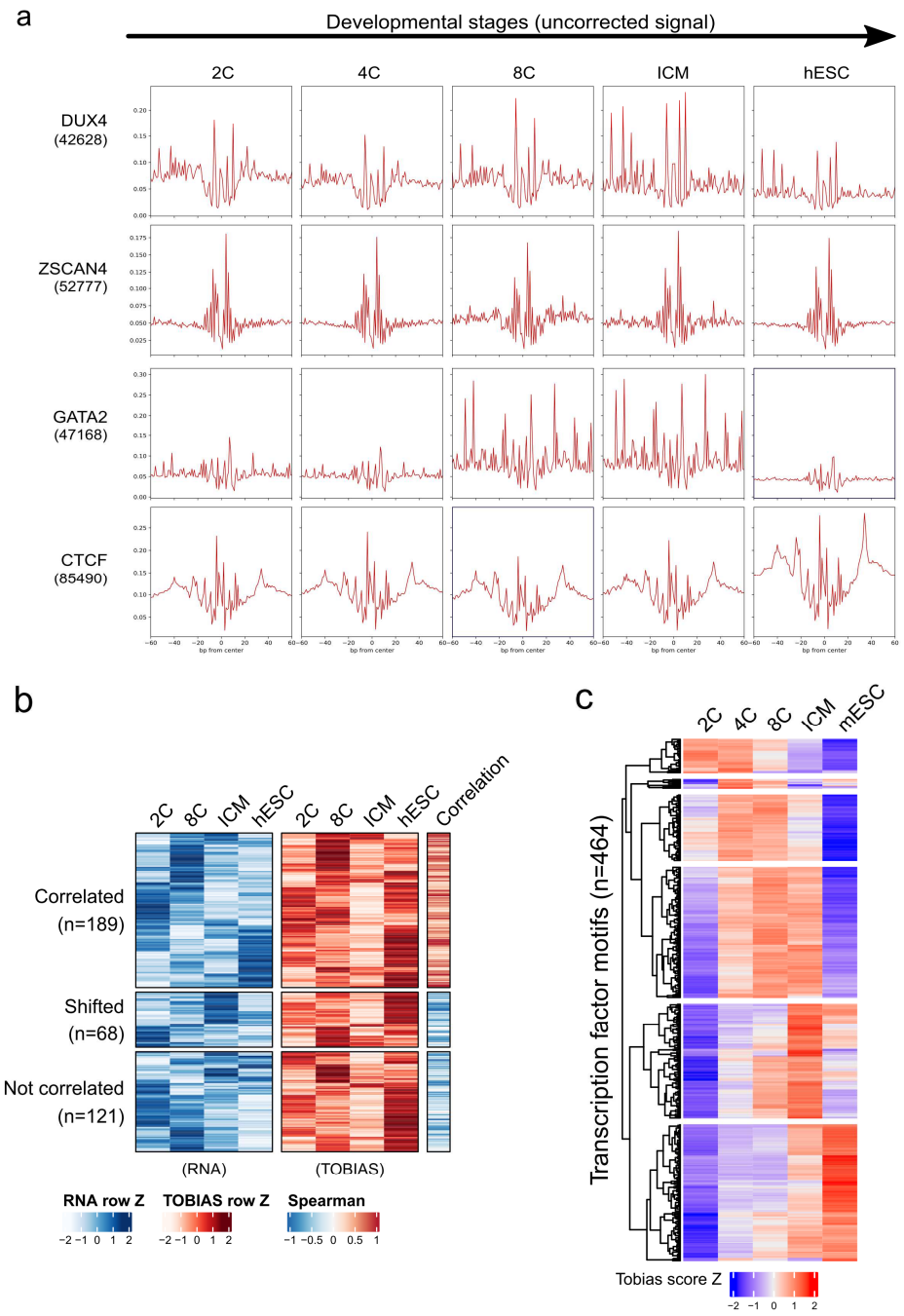
Supplementary Figure 2: Tn5-bias correction is crucial for visualization of footprints from ATAC-seq

(a) Comparison of aggregate footprints for transcription factors MAX and NRF1 across different bias-correction methods in cell type GM12878. Bound and unbound sites are defined by overlap of motif occurrences with ChIP-seq peaks. The aggregate signals are shown in a 120bp window using uncorrected signals (pileup of Tn5 insertions), TOBIAS ATACorrect, SeqOutBias and HINT-ATAC correction methods. Aggregate signals for all included TFs and cell types are found in the Supplementary Data 1. *(b)* The footprint depth (FPD) of an aggregate footprint is the difference between the mean of the flank signal and the mean of the footprint signal. Negative FPDs represent a stronger footprint. *(c)* Quantification of footprint depths between bound/unbound subsets (as explained in (a)) for different bias correction methods. N=54 transcription factors per boxplot where the bounds of the box are 25th and 75th percentiles (Q1 and Q3), the center is the median, and the whiskers are defined as Q1 - 1.5*(Q3-Q1) and Q3 + 1.5*(Q3-Q1). Significance is calculated by a two-sided Mann-Whitney U-test without adjustments. “Uncorrected” refers to the footprint depths estimated from uncorrected Tn5 signals. *(d)* The aggregate footprints for transcription factors JDP2 and FOXD3 across the uncorrected, expected and corrected Tn5 signals in GM12878. *(e)* Correlation between depth of footprints for uncorrected vs. expected footprints (left) and corrected vs. expected footprints (right). The plot consists of FPD values for N=746 motifs in cell type GM18278. *(f)* Mixture model of N=746 footprint depths in cell type GM12878. For uncorrected signals (left), the mixture model shows that 21% of all motifs generate a footprint. For corrected signals (right), 59% of all motifs to generate a footprint. *(g)* Significance of observed footprint depths. A null distribution of random aggregate footprints is created by shuffling the observed aggregate signal values. The observed footprint depth is compared to the null distribution for each TF, and the significance of the FPD being on the lower tail of the null distribution is calculated (Z-score; one-sided p-value < 0.01). Source data is provided in the Source Data file.



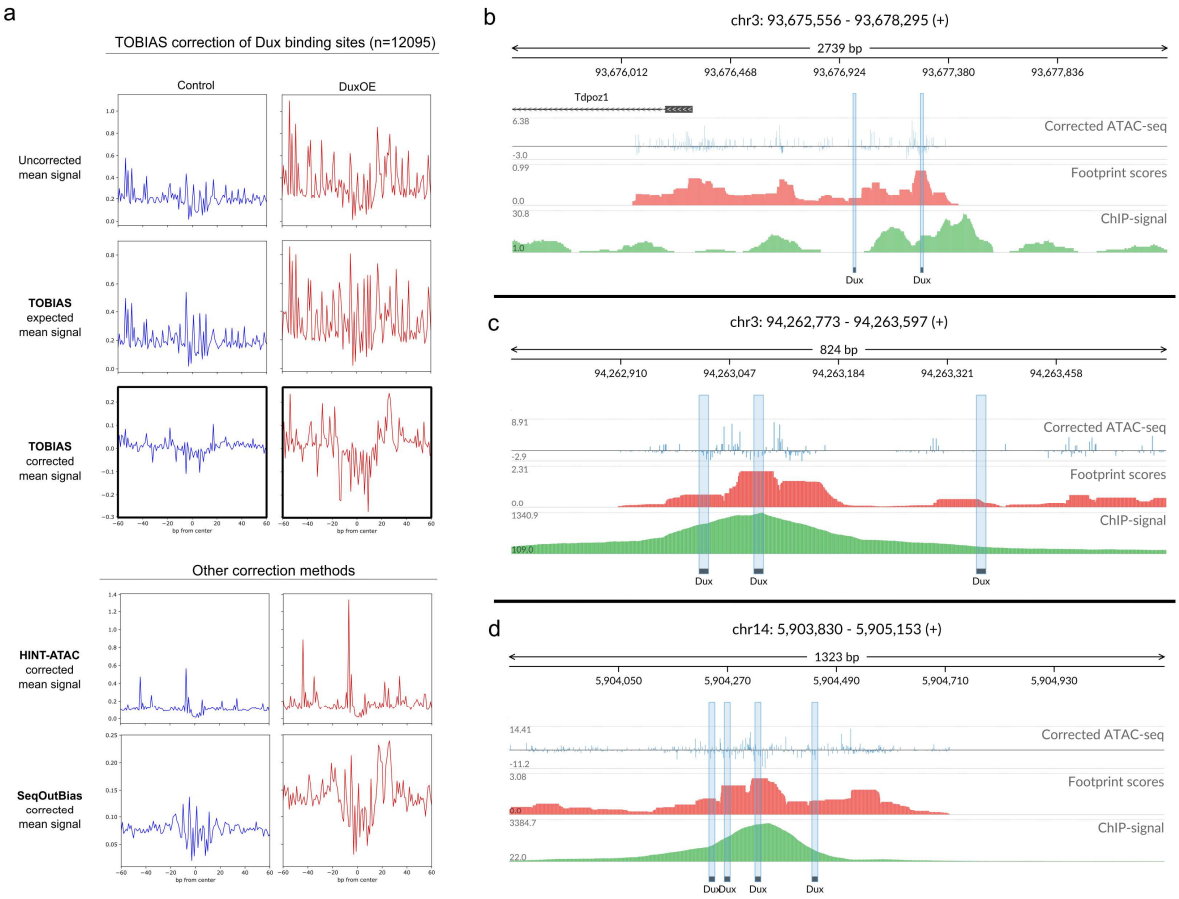
Supplementary Figure 3: Comparison of TOBIAS to existing footprinting methods

(a) Comparison of predictive power across different footprinting methods. The auROC (area under the Receiver Operating Curve) is used as a measure of predictive power for each method and is generated by overlap with ChIP-seq peaks (limited to chromosome 1). N=217 TF experiments across four cell types. Significance is defined by a two-sided Mann-Whitney U test without adjustment. *(b)* Scatterplot comparing the auROC of TOBIAS and msCentipede. Each point represents one TF based on N=217 paired ChIP-seq/ATAC-seq datasets, which are colored and marked dependent on the respective cell type the ChIP-seq was performed on. The diagonal line represents equal auROC between TOBIAS and msCentipede. *(c)* The auROC of TOBIAS predictions across cell types K562 (n=87), GM12878 (n=54), HepG2 (n=64) and A549 (n=12). *(d)* The aggregate footprints for true CEBPB binding sites (bound sites verified by ChIP-seq) in cell line GM12878. The number in the upper-right corner (526) represents the number of CEBPB binding sites included in the aggregate. *(e)* Comparison of computing times for footprinting tools. The CPU run time for each tool is measured across the three tasks of "pre-processing" (only for PIQ), "bias-correction" (only for TOBIAS and HINT-ATAC) and footprinting (all tools). *(f)* Calculation of the TOBIAS footprint score. The term w_b is the width of the flanking background left/right, and w_f is the width of the footprint. The score is calculated as the difference of the background mean (positive values) and the footprint mean (negative values). *(g)* Calculation of the FOS footprint score. The FOS (Footprint Occupancy Score) is calculated on the basis of the mean signals of the left flank (of width w_l), the right flank (of width w_r) and the footprint (of width w_f). *(h)* Comparison of score predictions (N=217 TFs) (as explained in (a)) for TOBIAS, FOS footprint score on corrected data, and FOS footprint score on uncorrected data. Significance is defined by a two-sided Mann-Whitney U test without adjustment. All boxplots are defined with the box being the 25th and 75th percentiles (Q1 and Q3), the center being the median, and the whiskers defined as Q1 - 1.5*(Q3-Q1) and Q3 + 1.5*(Q3-Q1).



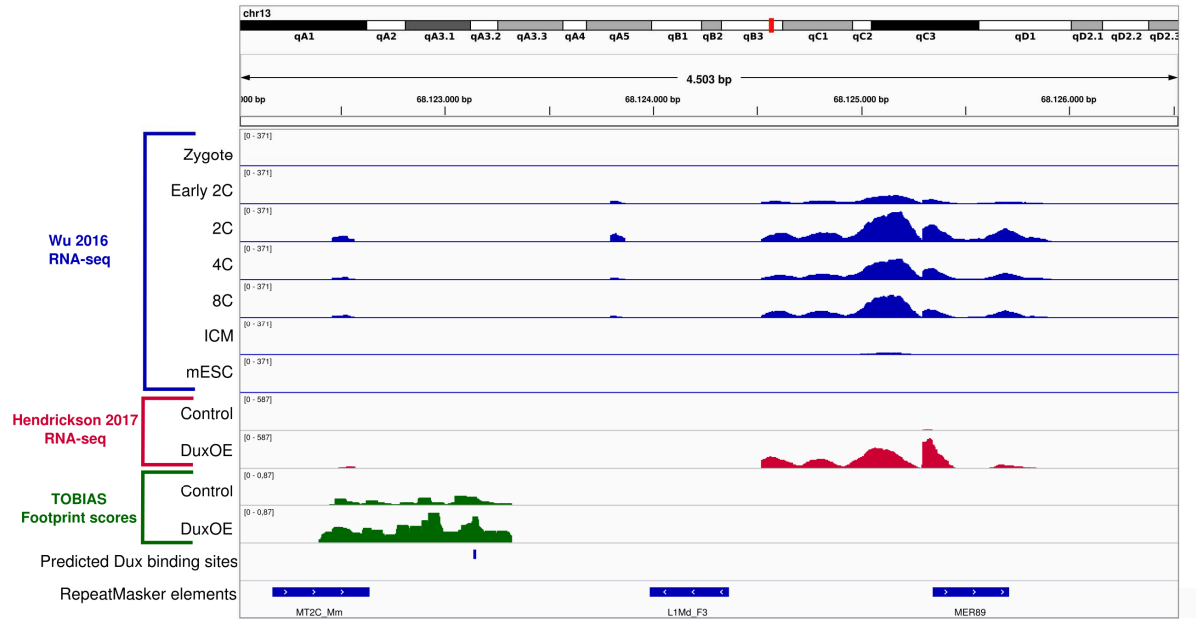
Supplementary Figure 4: Transcription factor activity and expression during mouse and human development

(a) Depiction of uncorrected footprint aggregates across time points for transcription factors DUX4, ZSCAN4, GATA2 and CTCF. Aggregated Tn5 signals are shown across time points 2C, 4C, 8C, ICM and hESC. The numbers below each TF name represent the number of binding sites included in the plots. *(b)* Correlation of footprints and RNA-seq. The left heatmap (blue) depicts expression of transcription factor clusters in the respective human developmental stages. The left heatmap (red) depicts the corresponding TOBIAS scores. Spearman column represents the spearman correlation between TOBIAS/RNA rows. The TF clusters are grouped into "Correlated" (Spearman≥0.2), "shifted" (RNA max value appears before TOBIAS max value) and "Not correlated" (Spearman<0.2 with no apparent shift in RNA). (c) Dynamic transcription factor binding during mouse embryonic development. The heatmap depicts the TOBIAS-predicted footprint scores for 464 motifs during the time points 2C, 4C, 8C, ICM and mESC. The rows are clustered into 6 clusters using hierarchical clustering. Source data for (b) and (c) is provided in the Source Data file.



Supplementary Figure 5: Dux binding is visible as footprints and correlate with ChIP-signal

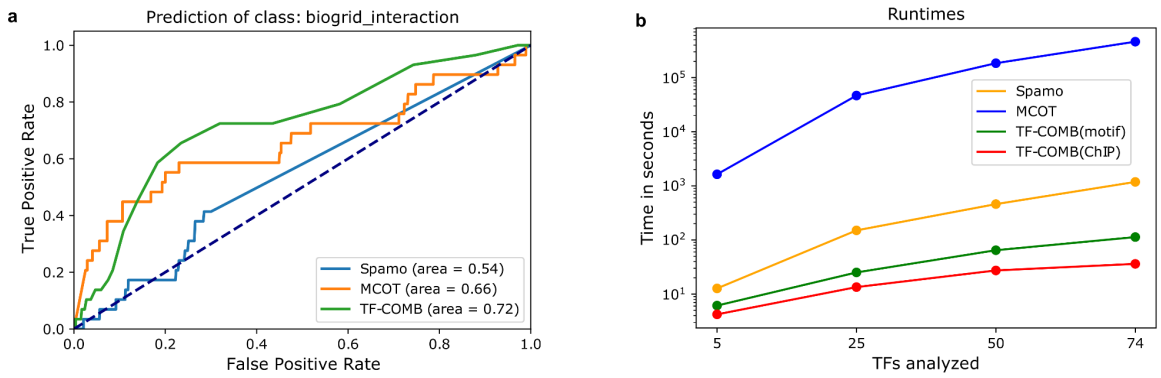
(a) Correction of the Dux footprint using different bias correction methods. The aggregate footprints for N=12095 Dux binding sites (within ATAC-seq peaks) are shown between Control and *Dux*OE conditions. The top three panels depict the uncorrected, expected and corrected signals as calculated by TOBIAS. The bottom panels depict the same sites corrected by either HINT-ATAC or SeqOutBias methods. *(b)* A view of the footprinting scores in the promoter of *Tdpoz1*. Genomic tracks show corrected ATAC-seq cutsites at 1bp resolution (blue), footprint scores as calculated by TOBIAS (red), and pileup of reads from Dux ChIP-seq from Hendrickson et al. 2017 (green). Potential Dux binding sites are highlighted in blue. *(c-d)* Footprinting correlates with ChIP-signal at multiple genomic loci. Genomic tracks are the same as described for (b).



Supplementary Figure 6: Predicted Dux binding site correlates with increase in expression

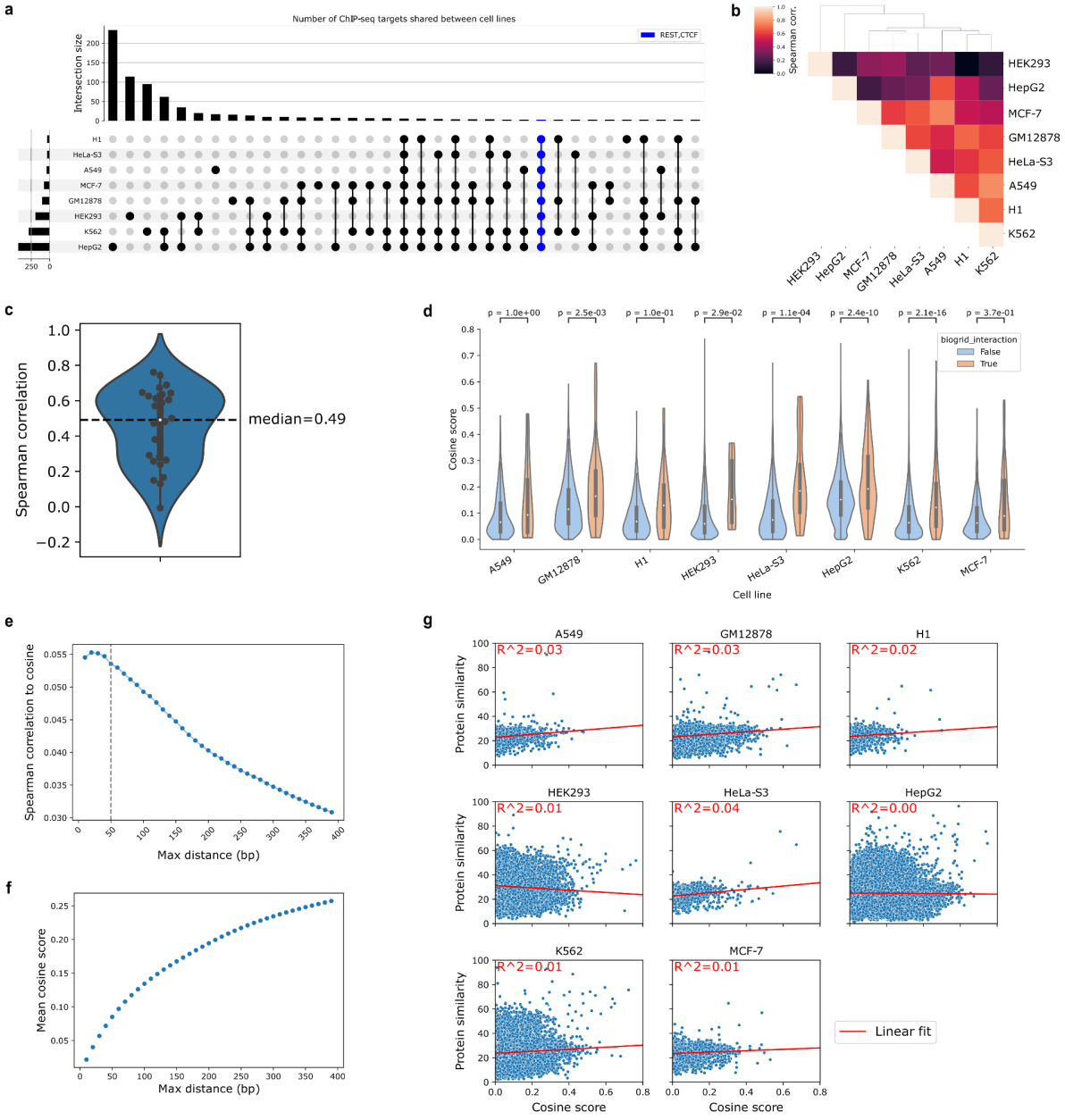
The figure shows genomic tracks of RNA-seq from embryonic cell stages (blue) and *Dux* overexpression (red), TOBIAS footprint scores predicted from *Dux* overexpression ATAC-seq (green), predicted Dux binding site as well as known repeats as annotated by RepeatMasker[1].

A2 Publication 2: Supplementary figures



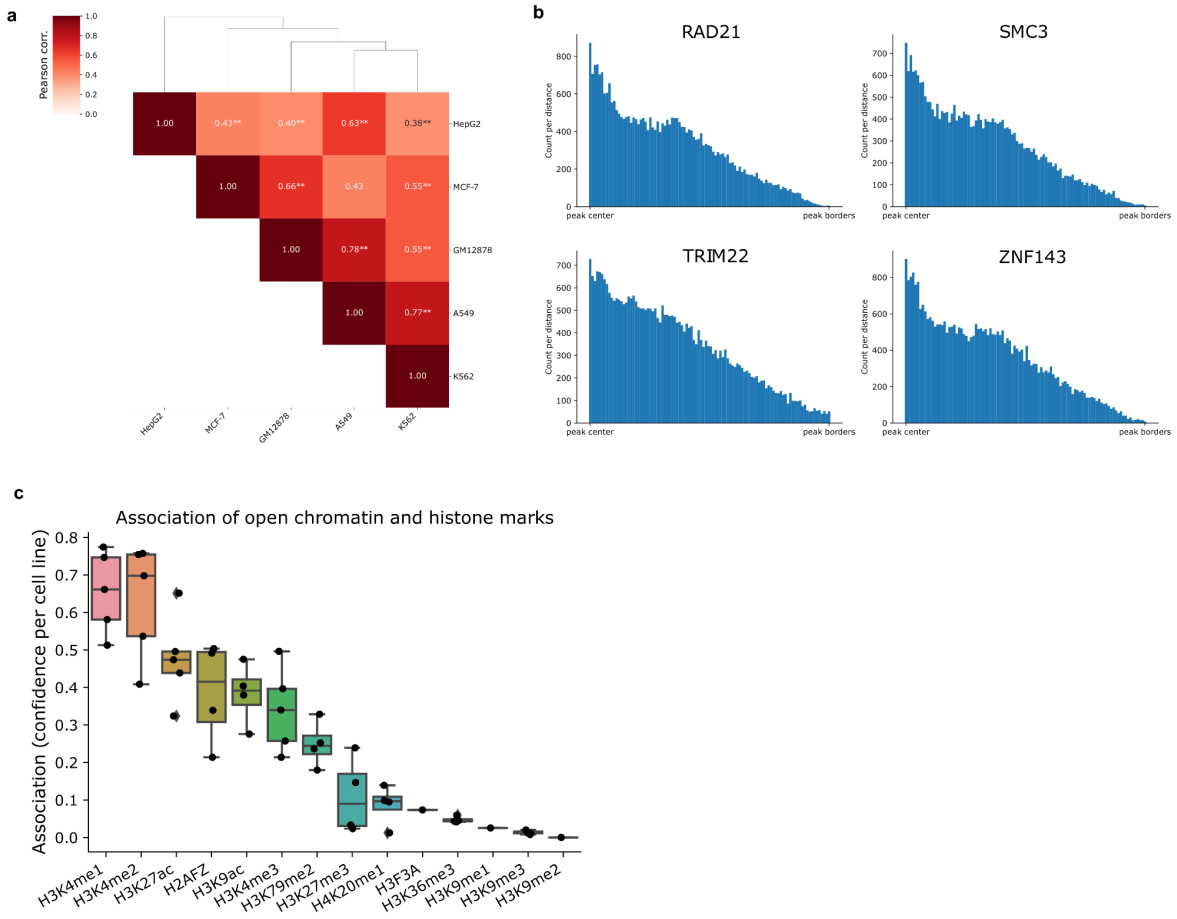
Supplementary Figure S1: Validation of TF-COMB and other tools

a) ROC curve of predictive ability of the assessed tools. Dashed line represents auROC=0.5.
b) Benchmark of runtime for assessed tools with increasing number of TFs analyzed.



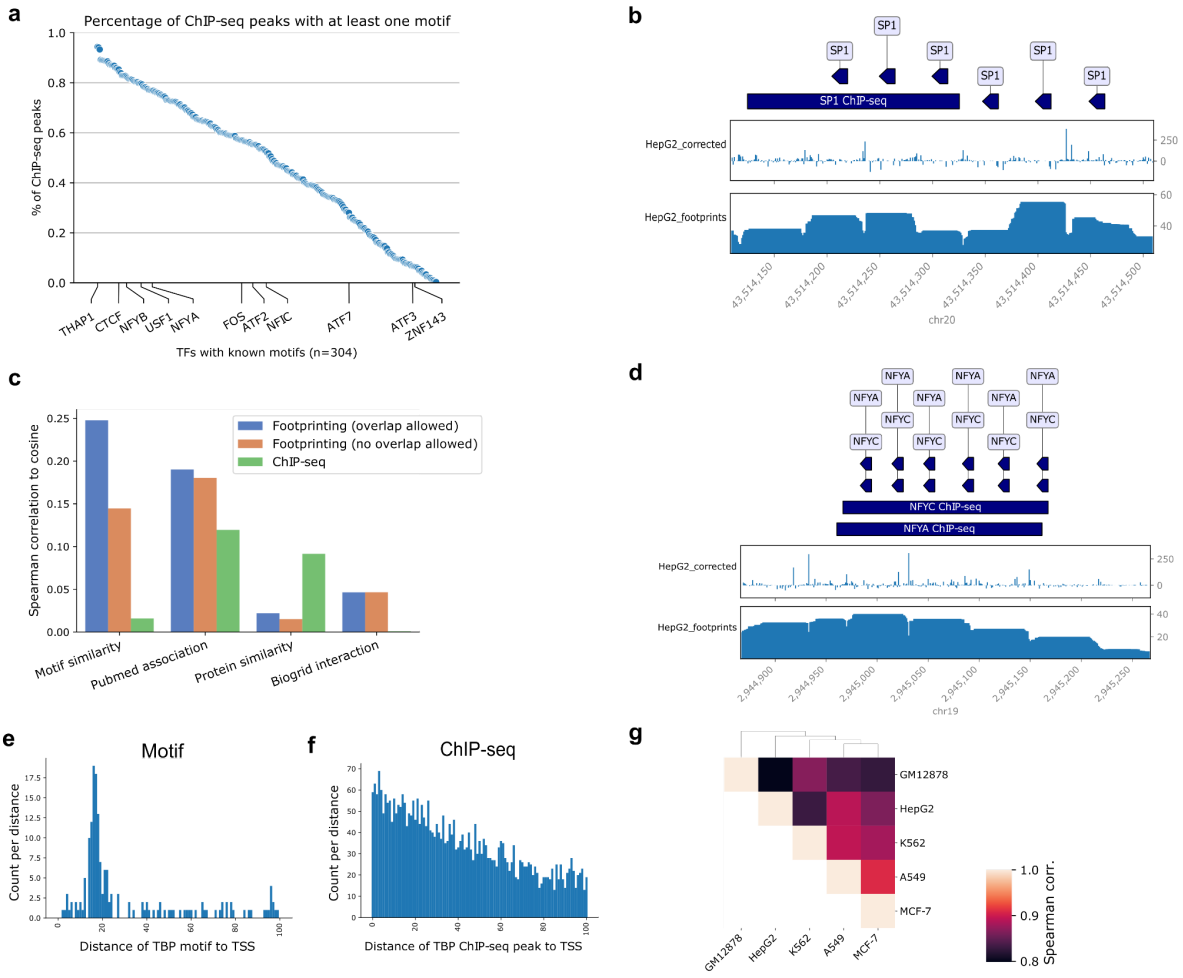
Supplementary Figure S2: Co-occurrence of ChIP-seq peaks across cell lines

a) Intersection count of available TFs ChIP-seq experiments for each cell line combination. Amount of TFs present in all cell lines highlighted in blue. Number of distinct TFs per cell line are on the left.
b) Correlation of TF-pair cosine values between cell lines.
c) Distribution of correlation values shown in b).
d) Distribution of cosine association scores for TF-pairs with and without PPIs across cell lines.
e) Correlation of cosine scores to PPIs with increasing allowed distance between TFs. Max distance of 50bp is marked with a dashed line.
f) Mean cosine association score with increasing allowed distance between TFs.
g) Cosine association compared to protein similarity of TF-pairs across cell lines. A linear fit is added in red and the R-squared is added at the left corner for each cell line.



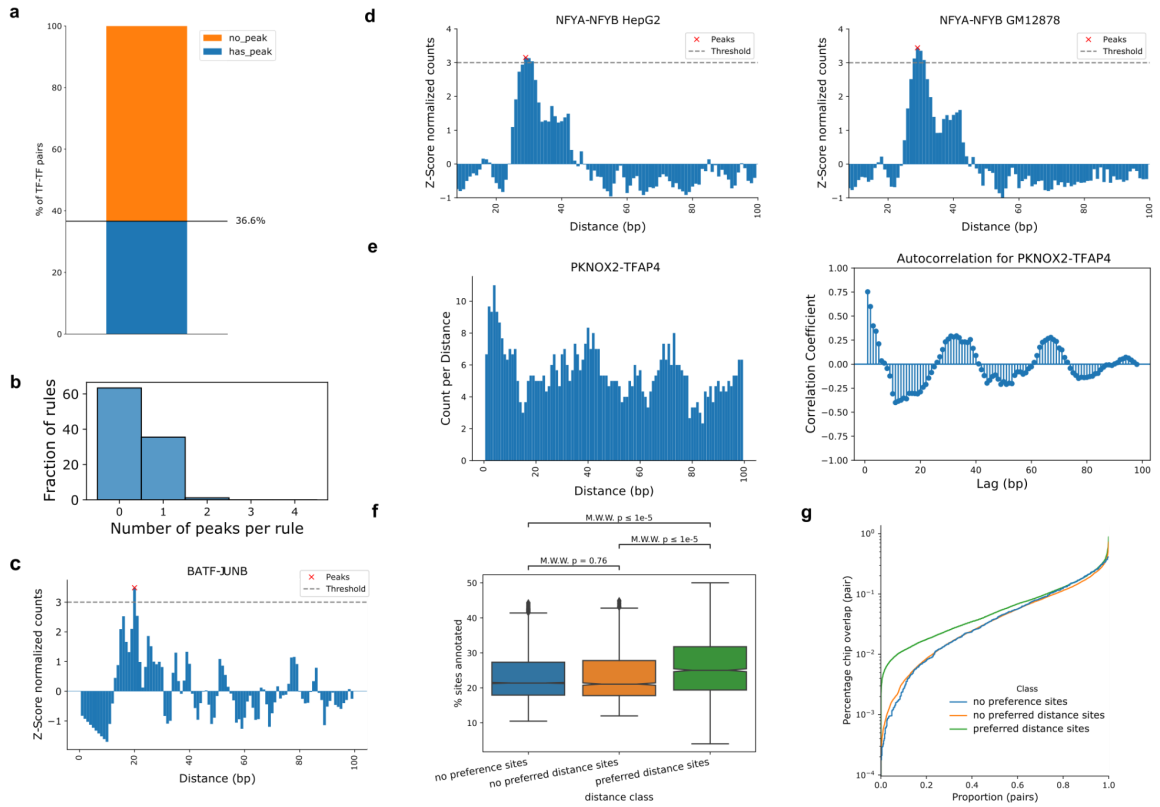
Supplementary Figure S3: Integration of epigenetic marks for co-occurrence

a) Correlation of TF binding distances(ChIP-seq) in relation to open chromatin (ATAC-seq) between cell lines. Significance of the correlation is marked with stars, where a p-value of less than 0.01 and 0.001 are marked with “*” and “**”, respectively.
b) Examples of relative TF binding locations within peaks. The x-axis represents the binding location from center (left) to border (right) of open chromatin peaks.
c) Association of open chromatin to locations of histone modifications and variants. Each box shows the association for all cell lines to a specific histone modification.



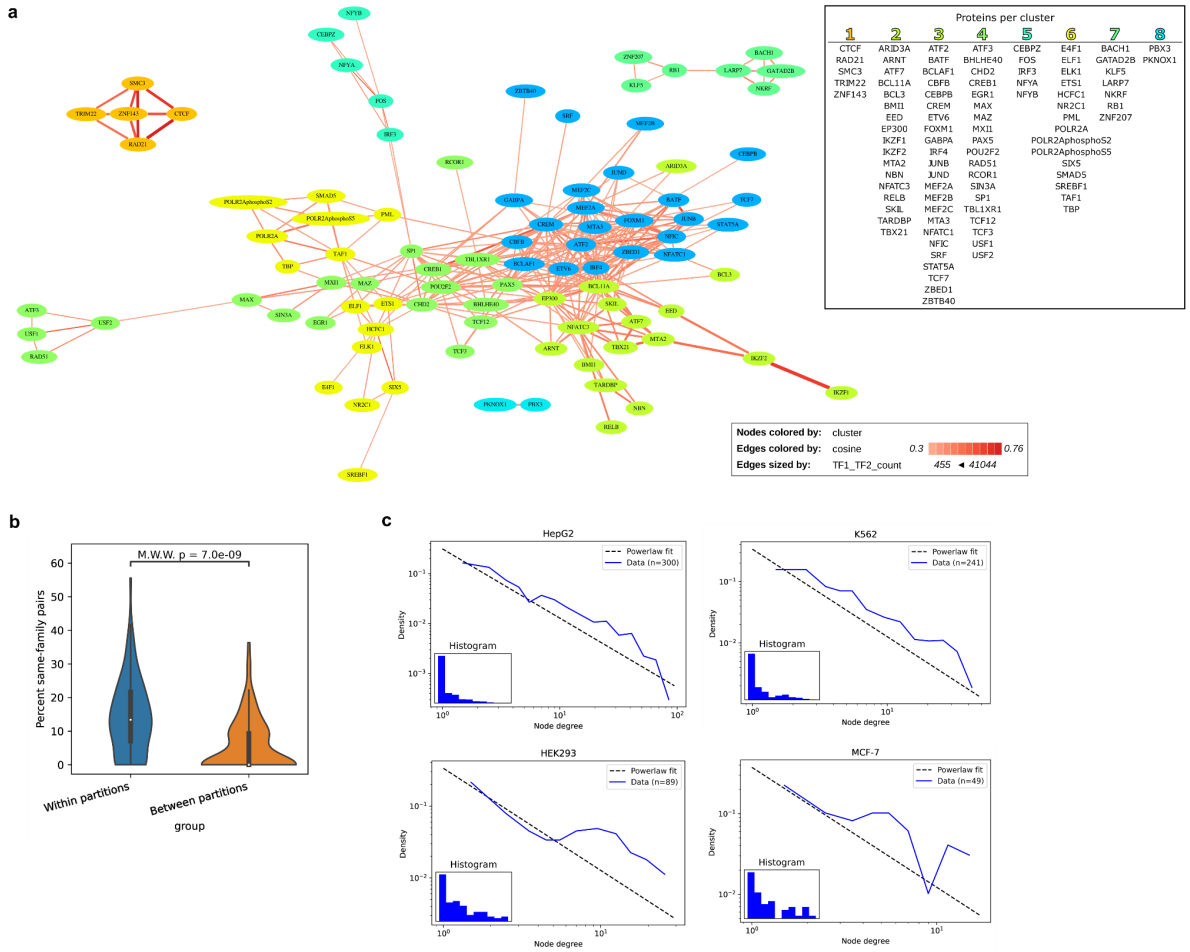
Supplementary Figure S4: Additional aspects of co-occurrence of footprinting

a) Fraction of ChIP-seq peaks containing sites of associated TF motifs.
b) Illustration of ChIP-seq peaks (upper blue boxes) in comparison to Tn5 signal (middle track) and TOBIAS derived footprinting score (lower track) for SP1. Individual TF motif locations are shown as triangles (upper track).
c) Correlation of cosine score with motif similarity, literature association score, protein similarity and protein interaction (BioGrid) score respectively per TF pair. Shown for ChIP-seq and footprinting with overlap allowed and excluded respectively.
d) Same plot as described in b) for NFYA/NFYC.
e-f) Distribution and gain of resolution shown for TBP binding site distances to TSS sites based on footprinting data analysis e) and ChIP-seq analysis f).
g) Correlation of TF pair cosine scores between cell lines from footprinting derived data.



Supplementary Figure S5: Additional aspects of binding grammar

a) Percent of TF-pairs that exhibit at least one preferred binding distance (peak) as displayed in c-d).
b) Distribution of TF-pairs (rules) on their number of predicted preferred binding distances.
c) Z-score normalized TF-pair binding counts sorted by distance. Peaks above threshold are considered preferred binding distance.
d) Difference in binding distance distribution for NFYA-NFYB in HepG2 (left) and GM12878 (right).
e) Periodic binding distance preference for PKNOX2-TFAP4. Left plot shows the distribution of binding distances for all co-occurring sites. Right plot shows the calculated autocorrelation, i.e. lag of binding distances, for the pair.
f) Percentage of sites annotated to genes (using UROPA) per distance class.
g) Proportion of ChIP-seq overlap per distance class.



Supplementary Figure S6: Network analysis of co-occurrence

a) The GM12878 co-occurrence network. Nodes are colored on the basis of Louvain community clustering. On the right a list of TFs per cluster is shown.
b) Distribution of same-family pairs randomly picked from within or between partitions (clusters of a)). Percent same-family is estimated per subset for 20 randomly selected pairs across n=1000 iterations.
c) Node degree (count) for all TFs of different cell line networks. Node degrees follow power-law distribution.