HUMBOLDT-UNIVERSITÄT ZU BERLIN

Characterization of cis-regulatory elements via open chromatin profiling

Dissertation

zur Erlangung des akademischen Grades
Ph.D.
bzw. Doctor of Philosophy
im Fach Biologie
eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
von
MSc Aslihan Karabacak Calviello
Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst
Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen:

1 – Prof. Dr. Uwe Ohler

2 – Prof. Dr. Stein Aerts

3 – Dr. David Garfield

Tag der mündlichten Prüfung 24.01.2019

Contents

Table of Figures

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Berlin, den

Aslihan Karabacak Calviello

Abstract

Cis-regulatory elements such as promoters and enhancers, that govern transcriptional gene regulation, reside in regions of open chromatin. DNase-seq and ATAC-seq are broadly used methods to assay open chromatin regions genome-wide. The single nucleotide resolution of DNase-seq has been further exploited to infer transcription factor binding sites (TFBS) in regulatory regions through TF footprinting. However, recent studies have demonstrated the sequence bias of DNase I and its adverse effects on footprinting efficiency. Furthermore, footprinting and the impact of sequence bias have not been extensively studied for ATAC-seq.

In this thesis, I undertake a systematic comparison of the two methods and show that a modification to the ATAC-seq protocol increases its yield and its agreement with DNase-seq data from the same cell line. I demonstrate that the two methods have distinct sequence biases and correct for these protocol-specific biases when performing footprinting. The impact of bias correction on footprinting performance is greater for DNase-seq than for ATAC-seq, and footprinting with DNase-seq leads to better performance in our datasets. Despite these differences, I show that integrating replicate experiments allows the inference of high-quality footprints, with substantial agreement between the two techniques.

These techniques are further employed to characterize the cis-regulatory elements governing the embryogenesis of a complex organism, the fruit fly Drosophila melanogaster. Combining tight staging of embryos and tissue-specific nuclear sorting with open chromatin profiling, enables the definition of temporally and tissue-specifically resolved putative cis-regulatory elements. Large scale motif enrichment analyses of these elements confirm known associations between TFs and specific tissues or developmental time-points, as well as implicating novel links. Finally, DNase-seq signal and sequence features associated with putative TFBSs are combined in an integrative model that is highly predictive of tissue-specific TF binding.

Taken together, these analyses demonstrate the power of open chromatin profiling and computational analysis in elucidating the mechanisms of transcriptional gene regulation.

Zusammenfassung

Cis-regulatorische Elemente wie Promotoren und Enhancer, die die Regulation der Transkription von Genen steuern, befinden sich in Regionen des dekondensierten Chromatins. DNase-seq und ATAC-seq sind weit verbreitete Verfahren, um solche offenen Chromatinregionen genomweit zu untersuchen. Die einzel-Nukleotid-Auflösung von DNase-seq wurde des Weiteren genutzt, um Transkriptionsfaktor-Bindungsstellen (TFBS) in regulatorischen Regionen durch TF-Footprinting zu bestimmen. Kürzlich durchgeführte Studien haben jedoch gezeigt, dass DNase I einen Sequenzbias aufweist, welcher nachteilige Auswirkungen auf die Footprinting-Effizienz hat. Auch wurden das Footprinting und die Auswirkungen des Sequenzbias auf ATAC-seq noch nicht umfassend untersucht.

In dieser Arbeit nehme ich einen systematischen Vergleich der beiden Methoden vor und zeige, dass eine Modifikation des ATAC-seq-Protokolls die Ausbeute und die Übereinstimmung mit den DNase-seq-Daten derselben Zelllinie erhöht. Ferner zeige ich, dass die beiden Methoden unterschiedliche Sequenzbiases haben und korrigiere diese protokollspezifischen Biases beim Footprinting. Der Einfluss von Bias-Korrekturen der Footprinting Ergebnisse ist für DNase-seq größer als für ATAC-seq, und Footprinting mit DNase-seq führt zu besseren Ergebnissen in unserer Datensätze. Trotz dieser Unterschiede zeige ich, dass die Integration replizierter Experimente die Ableitung von qualitativ hochwertigen Footprints ermöglicht, wobei die beiden Techniken weitgehend übereinstimmen.

Diese Techniken werden ferner eingesetzt, um die cis-regulatorischen Elemente zu charakterisieren, die die Embryogenese der Fruchtfliege Drosophila melanogaster bestimmen. Durch die Verwendung von Embryonen die sich im richtigen Entwicklungsstadium befinden, sowie gewebespezifischer Kernsortierung mit offenem Chromatin-Profiling können zeitlich und gewebespezifisch aufgelöste vermeintliche cis-regulatorische Elemente definiert werden. Umfangreiche Motivanreicherungsanalysen dieser Elemente bestätigen bekannte Zusammenhänge zwischen TFs und spezifischen Geweben oder Entwicklungszeitpunkten und implizieren neue Verbindungen. Schließlich werden DNase-seq-Signal und Sequenzmerkmale, die mit mutmaßlichen TFBSs verbunden sind, in einem integrativen Modell kombiniert, das die gewebespezifische TF-Bindung in hohem Maße vorhersagt.

Zusammengenommen demonstrieren diese Analysen die Fähigkeit der offenen Chromatin-Profilierung und der Computeranalyse zur Aufklärung der Mechanismen der Genregulation.

1. Introduction

The orchestrated regulation of the activity of thousands of protein-coding genes is a fundamental feature of all living organisms, allowing for the maintenance of internal homeostasis and reaction to environmental stimuli. Coordinated gene expression across different cells further allows for organismal development, in which a zygote develops into an embryo containing dozens of different cell types with specialized functions.

While almost all cells of an organism share the same genomic DNA, not all elements in the genome are simultaneously used in all cells; temporal and spatial patterns of gene expression driven by temporally and spatially active cis-regulatory elements (CREs), shape organismal development and cell function. Such specificity is achieved by keeping different portions of the genome accessible in different conditions, facilitated by the organization of DNA around protein macromolecular complexes called nucleosomes, leading to chromatin formation. When accessible (or synonymously: residing in regions of open chromatin), CREs like promoters and enhancers can be bound by regulators such as transcription factors, which in turn are able to tune the transcription levels of target genes. Profiling tens of thousands of accessible regions in chromatin is thus crucial to understanding transcriptional regulation.

Two major experimental methods, DNase-seq and ATAC-seq, couple open chromatin profiling with next-generation sequencing to probe accessible regions genome-wide and have been instrumental in characterizing CREs in multiple organisms. In addition, these techniques harbor the potential to infer the binding of individual transcription factors within open chromatin regions, requiring tailored data analysis methods known as transcription factor footprinting approaches and a thorough understanding of experimental artifacts influencing the results.

Combining the power of open chromatin profiling techniques with established model organisms (such as *Drosophila melanogaster*) allows for detailed investigation of transcriptional regulatory networks, where both spatial and temporal patterns of promoter and enhancer activities are able shape the precise development of an entire organism.

1.1 Thesis outline

In Chapter 2, I present background regarding the molecular biology of transcription (Section 2.1), with a focus on chromatin structure, sequence elements and a short introduction to transcription factors. A brief description of *Drosophila* early embryo development is then presented in Section 2.2.

Section 2.3 details the experimental techniques used throughout this thesis, with an emphasis on ATAC-seq and DNase-seq. Computational analysis strategies are presented in Chapter 2.4, with a detailed discussion on methods used to infer transcription factor binding in open chromatin regions.

Chapter 3 contains a detailed description of the experimental and analytical methods used in this thesis, relating to experiments performed on human and *Drosophila* cells.

Results are presented in Chapter 4; in Section 4.1, I present results on comparing ATAC-seq and DNase-seq with respect to their ability in detecting open regions, using data from human cell lines. Inference of transcription factor binding using footprinting is presented, outlining the experimental biases affecting the two techniques. Different transcription factors show distinct footprint and background signal profiles in ATAC-seq and DNase-seq, where footprinting performance depends on the intrinsic experimental biases of the two techniques, in a protocol- and transcription factor-specific manner. Integrating information from biological replicates is presented as a strategy to infer high-confidence, reproducible transcription factor footprints in open chromatin.

Section 4.2 covers results associated with the application of open chromatin profiling in the early embryonic development of the fruit fly *Drosophila melanogaster*. Open chromatin is profiled for specific tissue subsets within precisely staged embryos (performed by the laboratories of Prof. Eileen Furlong at the EMBL and Dr. Robert Zinzen at the MDC), enabling the inference of both tissue- and time point-specific putative CREs. Different analysis strategies uncover the rich sequence content, including transcription factor binding motifs, underlying these regions. These analyses link transcription factors to putative CREs that are spatially and temporally active in neural and mesoderm development, finding known associations as well as implicating novel ones.

Section 5 includes a brief discussion, where I discuss the results presented in Chapter 4 and highlight the potential advantages and shortcomings of my work in the broader context of transcription factor binding and motif inference. Appendix and reference sections can be found at the end of the thesis.

2. Background

2.1. Transcriptional regulation of gene expression

2.1.1. Chromatin

In eukaryotic cells, nuclear DNA is assembled into a higher order structure, known as chromatin. Chromatin consists of a basic repeating unit named the nucleosome, which is an octamer made up of two copies each of histones H2A, H2B, H3 and H4 around which 147 base pairs (bp) of DNA is wrapped[1]. These histones are termed "core histones" and they consist of two domains: a "histone-fold" motif which is responsible for histone-histone and histone-DNA interactions, and an amino-terminal tail which can be subjected to posttranslational modifications (PTMs)[2]. The nucleosomes were found to be separated by 10-60 bp of linker DNA; the organization of this 10 nm chromatin fiber is termed the "nucleosomal array" or the primary structural unit[3]. Nucleosome-nucleosome interactions constitute the 30 nm fiber, also known as the secondary structural unit which is mediated by the linker histones such as H1 and H5 and further compaction of this 30 nm fibre forms the tertiary structural unit, resulting in the 240 nm metaphase chromosome seen in mitosis[3].

Two distinct chromatin structures were found via carmine acetic-acid staining of interphase nuclei[4]. The densely stained regions were found to maintain their condensed state throughout the cell cycle and these regions were collectively named "heterochromatin". Other regions (lightly stained in interphase) were observed to decondense as the cells progressed from metaphase to interphase; these were later found to exist in the form of the 30 nm fiber at this stage of the cell cycle[5] and these regions were termed "euchromatin".

Euchromatin is generally defined as early replicating in S phase, gene rich and transcriptionally permissive; whereas heterochromatin is late replicating in S-phase, gene-poor, transcriptionally inactive in general, low in recombination activity and associated with repeat sequences abundant in pericentric and telomeric regions[6]. Chromatin modifications were found to be associated with these different chromatin states, for example DNA methylation is abundant in heterochromatin whereas DNA in euchromatin is generally hypomethylated[7]. Many histone PTMs have also been characterised to date (acetylation, methylation, phosphorylation, ubiquitylation, sumoylation, ADP ribosylation, deimination and protein isomerisation) which are associated with heterochromatin or euchromatin, depending on the site of modification[2]. Regulation of chromatin structure, via the action of these modifications, or other chromatin-associated proteins, play crucial roles in the transcriptional regulation of gene expression, as explored in the next sections.

2.1.2. Cis-regulatory elements

Transcriptional regulation of gene expression is mediated by the collective action of cis-regulatory elements (CREs) such as promoters, enhancers, silencers and insulators (figure 1). At a basal level, gene transcription is achieved by the assembly of the transcription preinitiation complex (PIC), which is comprised of RNA polymerase II (Pol II, enzyme catalyzing transcription) and the basal transcription factors, at the core promoter regions[8]. The core promoter regions, which refer to the 50-100bp regions surrounding transcription start sites (TSSs), have two main classes: peaked (TSS spanning at most several nucleotides) and broad (several weak TSSs over an extended region)[9]. Core promoters are rich in sequence content and contain numerous motifs such as the TATA box, BRE, Inr, MTE, DPE, DCE, and XCPE1, some of which constitute binding sites for basal transcription factors, and the prevalence of which depends on the promoter class[9]. Enhancers, on the other hand, are TSS-distal elements that regulate expression by interacting specifically with promoters via looping[10]. These distal elements are bound by transcription factors (TFs, see next section), which in turn recruit coactivators/corepressors that regulate the target gene[10].

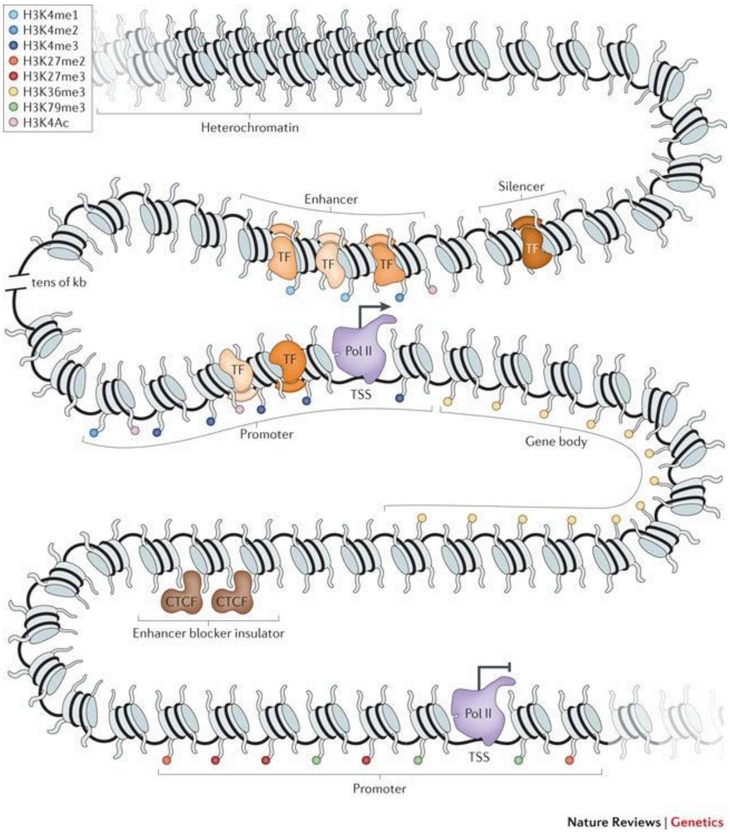


Figure 1: Transcriptional regulation of gene expression. Cis-regulatory elements (promoter, enhancer, silencer, insulator) are schematically shown around a gene locus. Factors and histone modifications associated with these elements are also shown. Adapted from reference[11].

Transcriptional gene regulation by these elements is closely linked to chromatin regulation, as active promoters and enhancers are found in open chromatin regions, where nucleosomes are locally displaced[10]. Therefore, locations of cis-regulatory elements can be assayed via open chromatin profiling techniques such as DNase-seq and ATAC-seq (see section 2.3.3). In accordance with this, active elements harbor active histone marks such as histone H3 lysine 4 trimethylation (H3K4me3) for promoters and H3K4me1 and H3K27 acetylation (H3K27ac) for enhancers[2]. Chromatin accessibility enables the binding of tissue-specific TFs to their cognate sequences (see next section).

2.1.3. Transcription factors

As mentioned above, TFs bound at short specific sequences called motifs (see section 2.4.3.1) within CREs are major regulators of gene expression[12]. TFs may be activators or repressors of gene expression and in most cases many factors act together to exert combinatorial control[12]. There are multiple mechanisms that affect factor binding, such as the affinity to target sequences and the number and arrangement of sites with respect to each other[13]. In addition, while pioneer transcription factors can bind nucleosome-associated inaccessible DNA and make the local chromatin structure more accessible by displacing nucleosomes, binding of other transcription factors require an already open/primed structure[14]. Many TFs bind enhancers in a tissue and time-point specific manner, and some act as master regulators that specify a given lineage[12].

TFs are classified according to their families, which in turn is based upon their specific DNA-binding domains (DBDs)[15]. There are around 100 characterized eukaryotic DBDs to date, examples of which include helix-turn-helix (HTH), homeodomain, zinc finger (ZF), leucine zipper (bZIP) and helix-loop-helix (bHLH)[15]. The motifs recognized by TFs are influenced by the specific interactions between the DBD and the underlying DNA sequence.

2.2. Drosophila melanogaster as a model organism of embryogenesis

Drosophila melanogaster is a widely studied model organism, and consequently the stages of embryonic development are well documented. At the onset of fertilization, a set of maternally deposited and localized factors, namely *bicoid*, *nanos* and *torso*, lead to patterning in the developing embryo, by forming gradients and activating zygotic gap genes, which in turn activate pair-rule genes[16]. This occurs within the first three hours of embryogenesis, where 13 rapid rounds of mitotic division takes place, without membrane formation, and the 6000 nuclei sharing the same cytoplasm[17]. This constitutes the blastoderm stage (stage 5), followed by gastrulation (stage 8), and time-points spanning 2-4hr post fertilization (pf) correspond to these stages. Upon gastrulation, the three germ layers (ectoderm, endoderm and mesoderm) are generated. As the datasets analyzed in this thesis are focused on the mesoderm and neurogenic ectoderm, the next stages will be summarized for these germ layers.

2.2.1. Specification of the mesoderm

The mesoderm gives rise to the somatic, visceral and heart muscle[18]. In the developing mesoderm, 2hr windows post fertilization correspond to the following stages: 4-6hr pf (stages 8-9) multipotent mesoderm, 6-8hr pf (stages 10-11) specification, 8-10hr pf (stages 12-13) diversification and 10-12hr pf (stages 13-15) terminal differentiation[19] (figure 2). The TF *twist* has a central role in mesoderm specification at multiple stages, starting as early as gastrulation[18]. It also regulates other central mesoderm-specific genes *dMef2*, which primes the differentiation into muscle[20], and *tinman*, which specifies the dorsal mesoderm[21]. *Tinman* further regulates *bagpipe*, which, in conjunction with *biniou*, specifies the visceral mesoderm[22] (figure 2). These TFs constitute the master regulators of mesoderm specification[19].

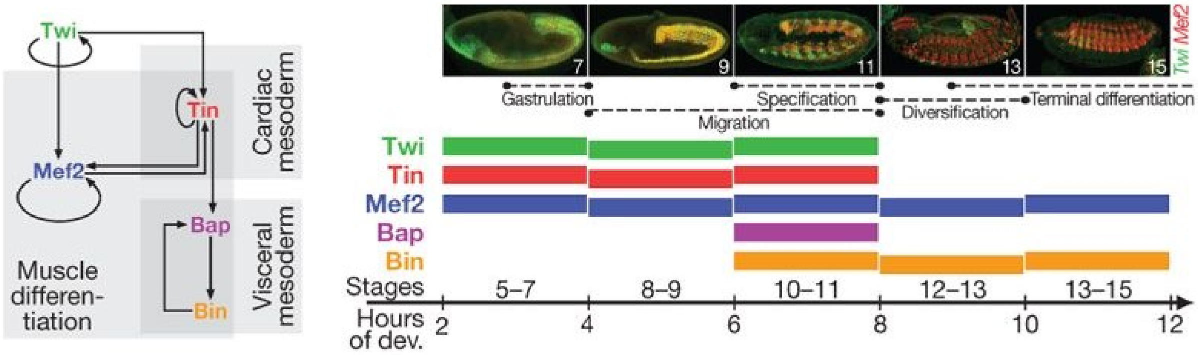


Figure 2: Expression of five main mesoderm TFs during early embryonic development. Regulatory relationships between the 5 TFs is shown on the left. Stage specific expression patterns of *twi* and *mef2* in the developing embryo is shown on the right panel (above), with the embryonic development stages (in 2hr windows) in which all 5 TFs are expressed, are stated below. Adapted from reference[19].

2.2.2. Specification of the neurogenic ectoderm

The neurogenic ectoderm gives rise to the central nervous system. In the developing neurogenic ectoderm, 2hr windows post fertilization correspond to the following stages: 4-6hr pf neuroblast formation, 6-8hr pf newborn neurons, 8-10hr pf neural patterning and 10-12hr pf terminal differentiation. The neurogenic ectoderm is made up of three columns along the dorsoventral (DV) axis: ventral, intermediate, and dorsal, where the cell fates for these columns are specified by the homeobox genes, *vnd*, *ind*, and *msh*, respectively[23]. Figure 3 shows the respective locations of the three columns in the developing embryo, marked by the expression of these genes. Proneural genes *achaete*, *scute*, *lethal of scute*, *wingless*, *hedgehog*, *gooseberry*, and *engrailed* also play important roles in neuroblast formation[23].

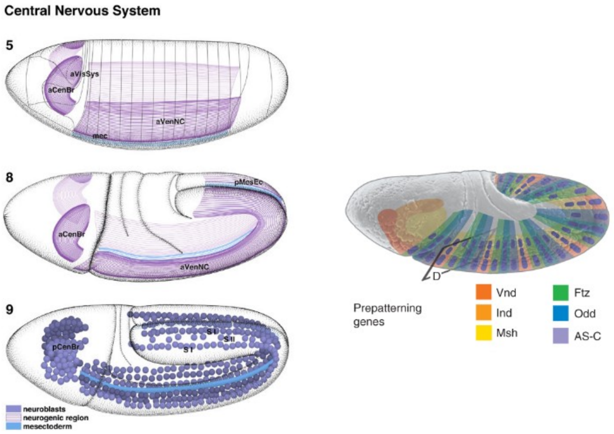


Figure 3: The developing CNS and the three columns of the neuroectoderm. The CNS development in embryos at stages 5, 8 and 9, is shown (the neurogenic region shown in light purple, left). The three columns of the neuroectoderm, as marked by vnd, ind and msh, are shown (together with segment polarity and pair rule genes that define segment boundaries, right). Adapted from http://www.sdbonline.org and reference[24].

2.3. Genome-wide sequencing techniques to identify and characterize cis-regulatory elements

2.3.1. Next-generation sequencing

Before the advent of next-generation sequencing technologies, the predominant methods for uncovering the sequence of DNA fragments have been Maxam-Gilbert[25] and Sanger sequencing[26]. Maxam-Gilbert method uses a set of chemical reactions that cleave a DNA template preferentially at one or two specific nucleotides (G+A, G, C+T, C). On the other hand, Sanger sequencing takes advantage of nucleotide analogues (dideoxynucleotides) that terminate the elongation of a template DNA molecule by DNA polymerase at a specific nucleotide (A, C, G or T). For both methods, conducting all four respective reactions on the same template and visualizing the resulting fragments side by side via gel electrophoresis, allows decoding the template sequence. Even though protocol modifications such as labeling the four reactions with different fluorophores to have direct fluorescent readout has increased automation in the case of Sanger sequencing[27,28], these methods have limited throughput. For instance, the human genome projects[29,30], where Sanger sequencing was employed, have been collective efforts of multiple groups, over the course of several years, with an estimated cost of 0.5-1 billion dollars[31]. Around the time the human genome projects were approaching completion, the National Human Genome Research Institute started an advanced DNA sequencing technology program, to fund the development of new technologies aimed at sequencing an individual human genome for 1000 dollars or less[32]. This set the stage for many next-generation sequencing technologies, including Illumina/Solexa that has been used in the generation of all datasets analyzed in this thesis.

The first step of the Illumina/Solexa sequencing workflow is library preparation[33]. In general, this involves DNA fragmentation, forward and reverse adapter ligation to the ends of the resulting fragments and amplification via PCR, although the experimental details depend on the protocol and starting material. For instance, RNA molecules need to be reverse transcribed into cDNA first. The following steps are cluster generation and sequencing, as illustrated in Figure 4. Cluster generation is achieved by a process called solid-phase amplification. This process starts with the adapter-ligated fragments annealing to complementary oligonucleotides covalently attached to the surface of a glass slide, known as the flow cell. The annealed oligonucleotides prime an extension reaction copying the original strand, creating flow cell-bound fragments. In turn, the free ends of the bound fragments anneal to other nearby complementary oligonucleotides, forming bridges. This initiates further rounds of extension and annealing (e.g. bridge amplification), generating many copies of the same fragment locally, called a cluster. All clusters on the flow cell are then sequenced via a reversible termination strategy. Sequencing primer anneals uniquely to the forward adapter and primes the extension reaction. Akin to Sanger sequencing, elongation-terminating nucleotide analogues are used. In contrast to the dideoxynucleotides used in Sanger sequencing, however, the fluorescently labeled 3´-O-azidomethyldNTPs used in this reaction enable the termination to be reversed[34]. In each round, the correct base is incorporated into the growing chain, visualized via the fluorescent signal and the termination reversed for the next base to be added. Cycles are repeated until the desired sequence length (e.g. read length) is achieved. In this way, sequences from one end of the fragments are uncovered, termed single-end sequencing. Both ends can be sequenced by repeating the whole process with a second sequencing primer complementary to the reverse adapter, termed paired-end sequencing.

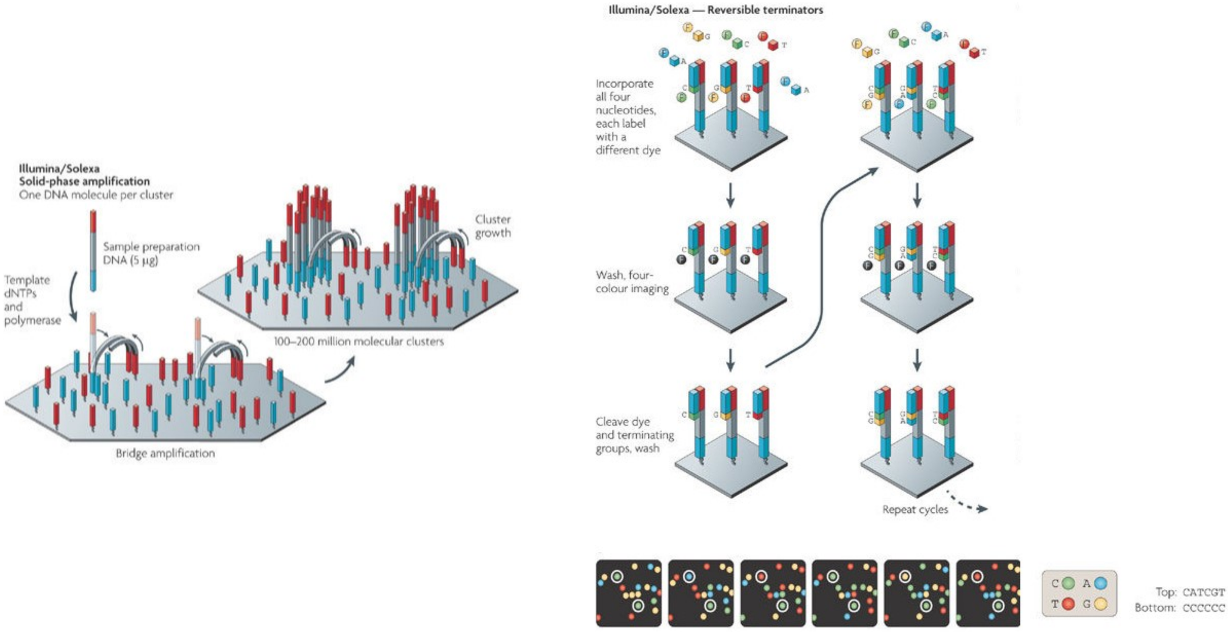

Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Bridge amplification
100–200 million molecular clusters
Cluster growth
Illumina/Solexa — Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
C A
T G
Top: CATCGT
Bottom: CCCCCC

Figure 4: Illumina/Solexa sequencing chemistry. Fragments are subjected to bridge amplification followed by cluster generation (left). Fragments are then sequenced via a cyclic reversible termination strategy (right). Adapted from reference[35].

Of the sequencers produced by Illumina, the HiSeq series and NextSeq 500 have been used in the production of the datasets analyzed in this thesis. To put the throughput into context, both HiSeq 2500 and NextSeq 500 machines can produce reads from a human genome at 30x coverage, in less than 30 hours[31]. This paves the way to probe the genome at unprecedented scale, using a multitude of genomics techniques.

2.3.2. Techniques to profile transcription factor binding sites and histone modifications

Transcription factors and histone modifications play crucial roles in the transcriptional regulation of gene expression. Therefore, techniques that enable them to be mapped genome-wide, have become instrumental in probing the complex transcriptional programs of cells.

2.3.2.1. ChIP-seq

Chromatin immunoprecipitation (ChIP) is perhaps the most widely used method to profile *in vivo* protein-DNA interactions[36,37]. In ChIP, proteins are covalently crosslinked to DNA using formaldehyde, to stabilize existing contacts[36,38]. The cells and nuclei are lysed, and the chromatin is then sheared by sonication to obtain shorter fragments, ideally in the 200-500 base pair range (Figure 5). Next, fragments bound to the protein of interest are enriched using an antibody specific to the protein (eg. the immunoprecipitation step). The DNA fragments are released from the protein-DNA complexes, by reversing the crosslinks. Finally, fragment sequences are determined to infer the binding locations of the protein of interest. In the early days of the ChIP assay, this was mostly achieved by designing probes specific to a locus of interest and assessing whether the probes hybridized to the obtained fragments[36,37]. Another method readily used is quantitative real-time polymerase chain reaction (qPCR) with primers against regions of interest[38]. Both approaches are low-throughput as they require pre-selection of candidate regions. Combining ChIP with next generation sequencing (ChIP-seq), overcomes this issue and profiles chromatin-associated proteins at a genome-wide scale.

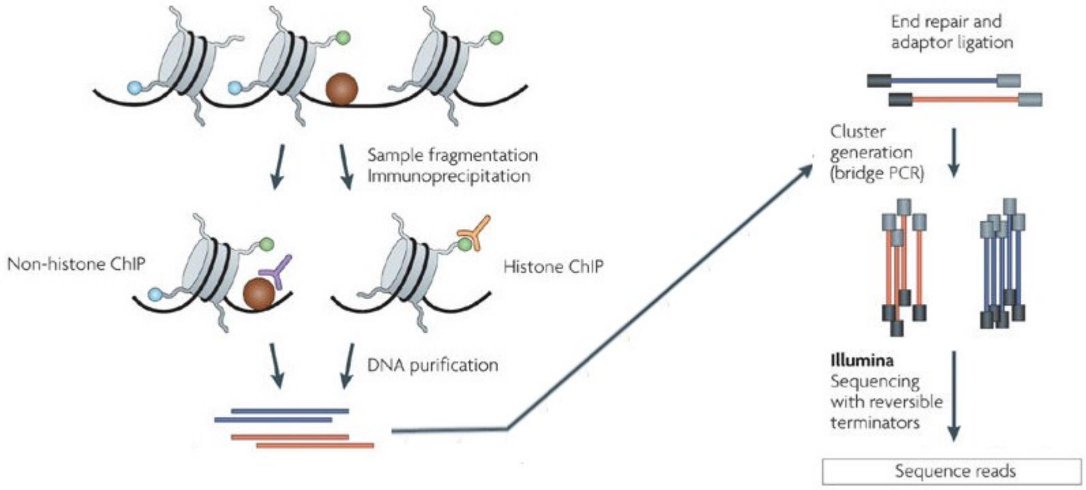


Figure 5: ChIP-seq. In ChIP chromatin is fixed, fragmented, and enriched for the bound protein or histone modification of interest by immunoprecipitation with a specific antibody (left). Adapters are ligated and libraries are sequenced (right). Adapted from reference[39].

In ChIP-seq, adapters are ligated to the ends of the fragments, followed by cluster generation and sequencing steps, as detailed in the previous section (Figure 5). In this way, millions of fragments can be sequenced in parallel. An important control in ChIP-seq is the input or total DNA, obtained by processing the samples in the same way, but omitting the immunoprecipitation step[38]. Input DNA is sequenced alongside the immunoprecipitated samples and gives insights into the fragmentation and sequencing biases. An additional control used in some cases is immunoprecipitation with a non-specific antibody (eg. IgG)[38].

Since the first ChIP-seq studies published in 2007[40–43], there have been many variations to the protocol[44]. For instance, nano-ChIP-seq[45] and linear DNA amplification (LinDA)[46] are two variations that require much less starting material than the original protocol: ~10000 cells as opposed to ~10 million. In another variation, ChIP-exo[47], the immunoprecipitated fragments are subjected to 5' to 3' exonuclease digestion, which brings the 5' ends of the fragments to the immediate vicinity of the bound protein, increasing the resolution from ~200 base pairs of the standard protocol to pinpointing the binding site. The next section will focus on batch isolation of tissue-specific chromatin for immunoprecipitation (BiTS-ChIP)[48], another technique that extends the standard ChIP-seq protocol and which has been partially used in the generation of some of the datasets analyzed in this thesis.

2.3.2.2. BiTS-ChIP

BiTS-ChIP[48,49] allows conducting ChIP-seq on a batch of tissue-specific nuclei isolated from a developing embryo (figure 6). The first step is to find a nuclear marker that is expressed exclusively in the tissue of interest and that can be used for fluorescent labeling of the nuclei. This can either be an endogenous protein or a transgene driven using a tissue-specific enhancer, coding for a tagged nuclear protein. Embryos expressing such a nuclear marker are collected, staged (eg. aged to the developmental stage of interest) and formaldehyde fixed to stabilize existing protein-DNA contacts. Individual nuclei are extracted by mechanical disruption of the fixed embryo. The nuclei are then fluorescently labeled by staining using a highly specific antibody against the chosen nuclear marker. Antibody staining is not needed if the nuclear marker is designed to have a fluorescent tag. Subsequently, a highly pure (typically >97%) population of fluorescently labeled, tissue-specific nuclei is obtained by fluorescence activated cell sorting (FACS). These nuclei are used in ChIP-seq experiments to infer tissue-specific patterns of histone modifications or binding of transcription factors and other chromatin-associated proteins.

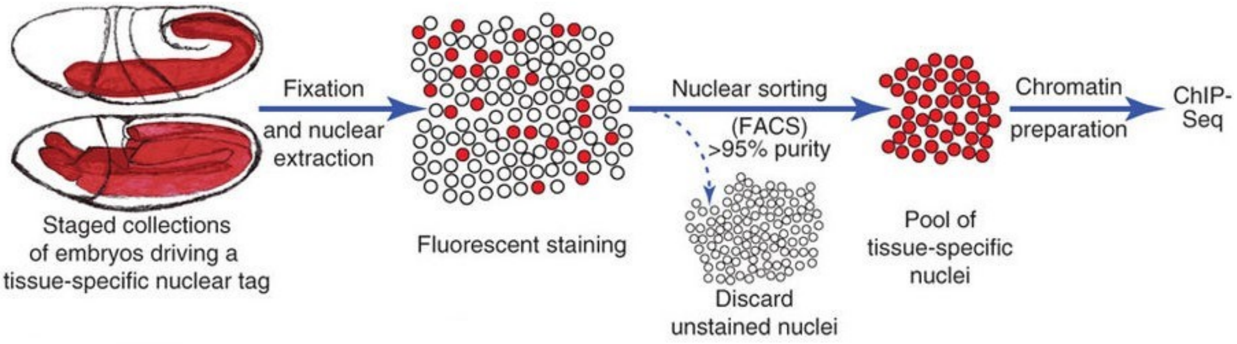


Figure 6: BiTS-ChIP protocol. Embryos expressing a tissue-specific nuclear marker are aged to the desired stage and formaldehyde fixed, followed by nuclear extraction, staining and FACS sorting. This results in a pool of tissue-specific nuclei that can then be assayed by ChIP-seq. Adapted from reference[49].

2.3.3. Techniques to profile open chromatin

As covered in section 2.1.2, tissue- and developmental stage-specific regulatory elements reside in nucleosome-free, accessible regions of the genome. These regions are hypersensitive to nuclease attack[50]. Digestion with the nuclease DNase I, coupled to high throughput sequencing (DNase-seq), is the first established genome-wide technique to probe such open chromatin regions[51,52], and is widely applied in research consortia such as ENCODE[53,54] or the Roadmap Epigenomics[55].

2.3.3.1. DNase-seq

Initially isolated from bovine pancreas[56], DNase I is a nuclease that can cleave DNA molecules by hydrolyzing the phosphodiester bonds of the sugar phosphate backbone[57]. In the mid-70s, studies have demonstrated that DNase I preferentially cleaves transcriptionally active chromatin[58,59]. Specifically, cells where selected gene loci (globin[58] and ovalbumin[59]) are transcriptionally active, are subjected to digestion by DNase I. The resulting DNA is observed to be depleted for fragments associated with the active genes, as shown by decreased annealing to gene-specific probes. This effect is not observed for cells where the genes are not transcribed, leading to the hypothesis that transcriptional activity is associated with an altered chromatin conformation, more susceptible to DNase I cleavage. These observations were extended some years later, in studies of Simian Virus 40[60] and *Drosophila* chromatin[61], which demonstrated that DNase I cleaves the underlying chromatin in a position-specific manner. These studies estimated two regions preferably digested by DNase I to be shorter than the length of DNA wrapped around 2 nucleosomes[60], and approximately 140 base pairs[61], respectively. Furthermore, both studies discussed that these regions could possibly be devoid of nucleosomes. They were defining for the first time, what we now know to be DNase hypersensitive sites (DHSs).

Historically, DHSs were mapped using a method called indirect end-labeling[62–64]. Briefly, the chromatin is first digested by DNase I, and then the isolated DNA cleaved further by a rare-cutting sequence-specific restriction endonuclease (RE). The resulting fragments are separated by gel electrophoresis, transferred to a membrane and hybridized to probes specific to the immediate flanks of the RE cut sites. The determined fragment lengths provide a direct measure of the distance between the RE and DNase I cleavage sites, from which the DHSs can be inferred. This method is low-throughput and can only be applied to a limited number of loci at a time, as it requires the loci of interest to be previously characterized (eg. knowledge of sequence and RE recognition sites). In contrast, DHSs are readily mapped today with the high-throughput, genome-wide DNase-seq method.

There are two variations of DNase-seq that are generally referred to as single-hit[51,65] and double-hit[52] protocols, as the resulting fragments represent a DNase I cut on one or both ends, respectively. In the single-hit protocol (figure 7, left), DNase I digestion is carried out and the first linker harboring an MmeI restriction enzyme recognition site is ligated to the digested ends. MmeI then cuts 20bp downstream from its recognition site, where the second linker is subsequently ligated. In the double-hit protocol (figure 7, right), the DNase I digested chromatin is subjected to size fractionation to get 100-500bp fragments. Illumina adapters are then ligated to the ends of the fragments. In both protocols, the linker-flanked fragments are PCR amplified, purified and sequenced according to the Illumina sequencing workflow. Computational analysis of DNase-seq datasets to infer DHSs and transcription factor footprints are discussed in sections 2.4.2 and 2.4.3.

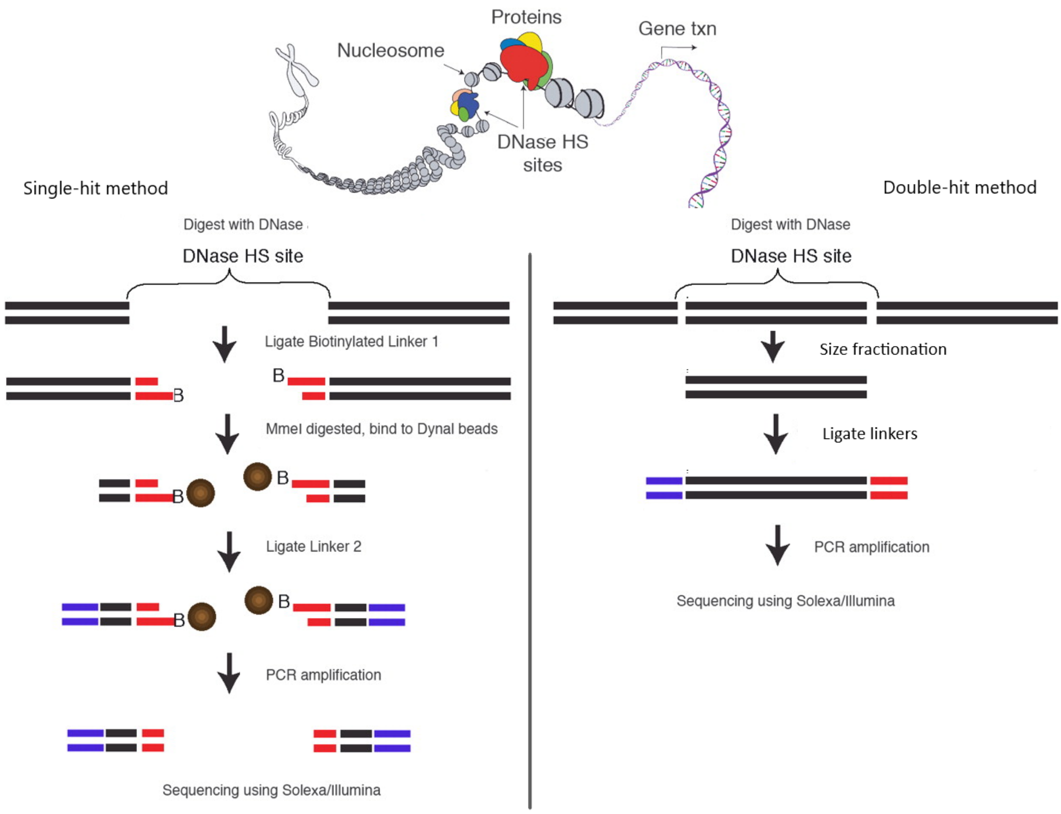


Figure 7: DNase-seq. In the single-hit DNase-seq protocol, a 20bp region representing one end of a DNase I-cut fragment is retrieved via an MmeI digestion step (left). In the double-hit protocol, the fragments are cleaved on both ends by DNase I (right). Adapted from reference[65].

2.3.3.2. ATAC-seq

A more recent technique to profile open chromatin regions is the assay for transposase-accessible chromatin using sequencing (ATAC-seq)[66]. Instead of a nuclease like DNase I, ATAC-seq employs Tn5 transposase enzymes. Transposases contribute to genomic rearrangements and consequently genome evolution, by mobilizing DNA elements called transposons[67]. Tn5 is a bacterial transposon normally functioning to confer antibiotic resistance to the host, through the three resistance genes it harbors (kanamycin, bleomycin and streptomycin)[68,69]. As with many others, mobilization of the Tn5 transposon is realized via a "cut-and-paste" mechanism where it is cleaved from its original location and inserted into the target DNA by the Tn5 transposase[69]. This depends on the specific interaction of the transposase with the 19bp sequences at the two ends of the Tn5 transposon. The Tn5 transposon-transposase complex then binds the target DNA and via a nucleophilic attack, the 3' ends of the transposon get covalently linked to the 5' ends of the cleaved target DNA[70]. During this process, the minus strand of the target DNA is cleaved at a position 9bps downstream of the plus strand, which leads to the duplication of these 9bps on either side of the inserted transposon. A better understanding of and modifications to the components of this system, led to its usage as an *in vitro* tool[70]. These include making the Tn5 transposase hyperactive through mutations[71], and using a modified version of the 19bp end sequence called the mosaic end (ME) with greater transposition efficiency[72]. Furthermore, it was found that pre-loading hyperactive Tn5 transposases *in vitro* with sequencing adapters harboring ME sequences, without the intervening transposon DNA, is sufficient for transposition[73,74]. Since there is no intervening DNA, this altered transposition reaction leads to the fragmentation of target DNA via transposase attack and cleavage while the resulting fragments are simultaneously tagged by adapter ligation to the 5' ends, a process known as "tagmentation"[74]. These developments and the reports of transposons preferentially integrating into nucleosome-free regions[75], set the stage for ATAC-seq as a Tn5 transposase-based method to profile open chromatin[66]. ATAC-seq has a fast and straightforward protocol that comprises of cell/nuclei isolation, lysis, tagmentation, PCR amplification and sequencing (figure 8). Much less starting material is required for ATAC-seq in comparison to DNase-seq, ~500-50,000[66] vs ~1-10 million[65] cells or nuclei, respectively, although recent protocol variations allow both techniques to be applied at the single cell level[76–78]. As for DNase-seq, computational analysis of ATAC-seq datasets are discussed in more detail in sections 2.4.2 and 2.4.3.

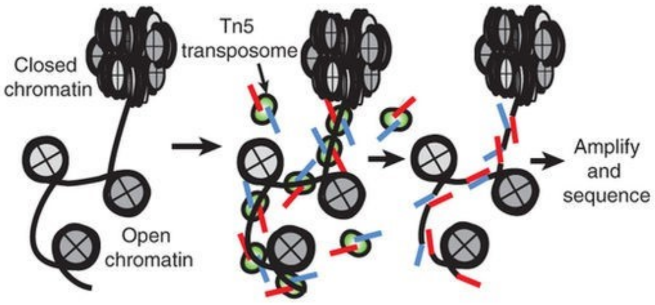


Figure 8: ATAC-seq. In ATAC-seq, Tn5 transposase dimers insert adapters into open chromatin regions, generating sequencing-ready fragments. Adapted from reference[66].

2.4. Computational analysis of genome-wide sequencing datasets

2.4.1. Read processing, alignment and filtering

As described in section 2.3.1, next generation sequencing technologies output sequences of defined length, also known as reads, belonging to fragments of interest from a given experimental setup. The output at this stage is generally in fastq format, which includes both the sequence information, as well as a per nucleotide quality metric for each read. A popular choice for the quality metric is called the phred score[79], which is essentially a measure of the probability of the base call during sequencing being correct for each nucleotide.

As adapters are ligated to the fragments for sequencing (see section 2.3.1), reads may contain adapter sequences which need to be trimmed, since otherwise they would interfere with proper alignment to the reference genome. Adapter sequences and where in the read they might be encountered are specific to the experimental setup and read length. For instance, an ATAC-seq library has a range of fragment lengths, starting from as short as 38bp[74]. The first nucleotide in the read corresponds to the 5’ end of the fragment. When the read is longer than the fragment, therefore, the 3’ end of the read will comprise of adapter sequences. Another example can be a single-hit DNase-seq library, where the fragment length is 20bp[51], and the reads would need to be trimmed down to this length. Tools for adapter trimming, can generally also be used to trim low quality bases from reads when needed.

The next step is to align the reads to the reference genome, in other words to determine which genomic region the read (and fragment) was originally derived from. This is essentially an approximate string matching problem[80]: the objective is to find where the read sequence matches the reference sequence, while allowing for mismatches and gaps. The main reasons for this are errors in sequencing and differences between the assayed sample and the reference genome due to sequence variation. Another consideration is the sheer amount of data. Aligning millions of reads to a reference genome millions of bases long, requires efficient algorithms. Two algorithmic ideas used by aligners are filtering and indexing[80]. Briefly, filtering eliminates regions of the reference genome where a match is not expected for a given read, by comparing the shorter subsequences within the read to the reference genome[81]. Indexing, on the other hand, refers to preprocessing the reference genome to allow matching sequences to be queried much faster. The main indexing approaches used are the enhanced suffix array[82] and the FM-index[83], which is based on the Burrows-Wheeler transform[84]. Aligners Bowtie2[85] and BWA[86], used in the analyses presented in this thesis, incorporate the FM-index. The output is in sam (short for sequence alignment/map) format which can be converted to the binary bam format.

Following the alignment, the .bam file can be further processed as needed. One of the first choices to be made is the handling of multimappers, i.e. reads that map to multiple locations in the genome. These may be filtered out to retain only reads that map uniquely to a single location, which increases certainty at the expense of coverage over repetitive regions. Another consideration is that, most library preparation methods include a PCR amplification step, which can lead to the same fragment to be sequenced multiple times. These PCR duplicates may be removed, with the choice depending on the experimental protocol: if independent fragments originating at the same location are expected, removing duplicates may discard some true fragments. Paired-end sequencing may aid in more confident removal of duplicates, as both 5' and 3' ends are considered. The final processed .bam file can then be used in downstream analyses.

2.4.2. Finding regions of enrichment via peak calling

Once discovering where the reads originated from in the reference genome via the alignment step, one common downstream analysis is to find the regions of enrichment, i.e. regions where the reads accumulate to generate enriched signal over background expectation. This is achieved by peak calling algorithms which have slight differences in application depending on the experimental protocol. For instance, in ChIP-seq (see section 2.3.2.1 and figure 5), the fragments of interest encompass a TF or histone modification. The sequenced reads belong to the 5' ends of these fragments, either on the plus or minus strand, shown as blue and red tags, respectively in figure 9, left. For the tag density to reflect the center of binding, many peak calling tools estimate fragment lengths and subsequently shift or extend the tags in the 3' direction (figure 9, right)[87]. The updated signal profile is then used to calculate regions of enrichment that reflect TF binding or locations of modified histones. On the other hand, in DNase-seq and ATAC-seq, the 5' ends of the reads represent the DNase I cut sites and the Tn5 transposition sites, respectively (see sections 2.3.3.1 and 2.3.3.2). Therefore, in this case peak calling algorithms are used to infer 5' read end pileups, without shifting. Open chromatin regions, and thus cis-regulatory elements, are inferred in this way.

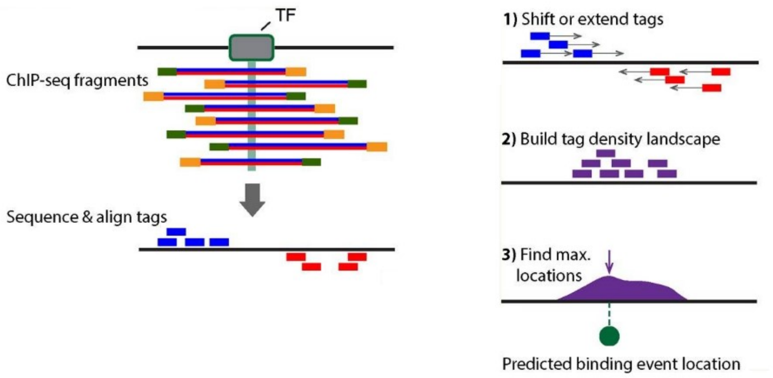


Figure 9: ChIP-seq peak calling. In ChIP-seq, the fragments span binding sites of the factor of interest (left). The 5' end of the fragments are shifted or extended to represent the center of binding (right). Adapted from reference[87].

The algorithmic approaches utilized by peak callers will be exemplified via two specific tools used in this thesis: MACS[88] and JAMM[89]. For ChIP-seq datasets, MACS estimates the fragment length, d, and shifts all reads by d/2 towards the 3' direction. To find peaks, MACS models read counts with a Poisson distribution, where mean and variance are expressed with a single parameter, lambda. As covered in section 2.3.2.1, input controls are essential for ChIP-seq experiments. By estimating lambda from the input control locally (i.e. from the same genomic regions as the peak candidates), MACS finds peaks that are significantly enriched over input. JAMM, also estimates the fragment length, d, however instead of shifting reads, it extends them towards the 3' direction, to match d. JAMM first determines broad windows of enrichment (over the control sample) throughout the genome, and then finds peaks within those windows by clustering the signal. Gaussian mixture models[90] with peak and noise components are used to this end. Both MACS and JAMM can be used with ChIP-seq data, as well as with DNase-seq and ATAC-seq data when the parameters are set accordingly.

A common method to find reliable peak sets is called the Irreproducible Discovery Rate (IDR)[91]. The IDR pipeline is comprised of calling peaks in replicates separately, comparing the results to assess both the extent of overlap among peak sets and the similarity of score ranks among overlapping peaks, and subsequently finding the number of reproducible peaks at a given IDR threshold.

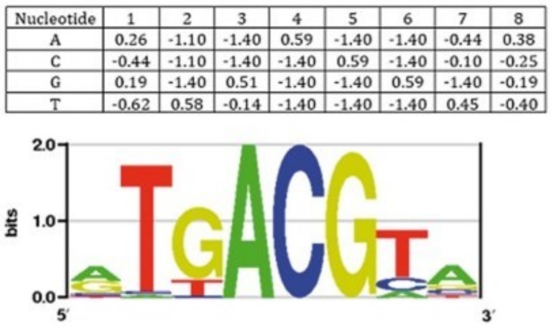
The identified peaks are subject to further downstream analyses depending on the type of experiment. Among these is differential signal analysis, which refers to methods such as EdgeR[92] and DESeq[93]. Initially designed to assay differential gene expression among two conditions, these methods can be employed in any case where regions of interest can be defined (e.g. peaks) and replicates are available for sets of conditions to be compared. In general, read counts within peaks are modeled using the negative binomial distribution, where the parameter for variance separates technical from biological variation, facilitating the identification of regions that show significantly higher signal in one condition compared to the other.

2.4.3. Finding transcription factor binding sites

As the previous section illustrated, peak calling on TF ChIP-seq datasets allows finding TF-bound regions genome-wide. However, this analysis has a relatively low resolution (i.e. the identified regions are much larger than the actual TF-DNA contact sites, 100-200bps vs 6-15bps, respectively) and would require a separate ChIP-seq experiment to be conducted for each TF of interest. Therefore, in the following sections, approaches to find putative TF binding sites at higher resolution is discussed, first focusing on sequence features defined by position weight matrices, followed by the data-driven TF footprinting.

2.4.3.1. Position weight matrices

TFs bind short (typically 6-15bps) and specific sequences throughout the genome. Individual binding sites of a given TF are not always identical; but display position-specific nucleotide preferences when aggregated, also known as binding motifs[94]. Motifs are most commonly represented via position weight matrices (PWMs)[95]. As shown in figure 10, each column in a PWM represents a base position, with the rows giving the weights of each of the four nucleotides at that position. The weights are usually log likelihoods (against a background model) of observing a given nucleotide at a given position, and an equivalent representation with plain probabilities or frequencies is called the position frequency matrix (PFM)[94]. PWMs (and PFMs) can also be represented visually, using motif logos (figure 10), where the nucleotides expected at a given position are drawn in proportion to their respective weights, with the total height representing the information content (IC)[96]. IC of a given position equates to the log likelihood (in log2 scale, against a background model) of each nucleotide multiplied by its frequency, summed over all four nucleotides. Thus, it ranges from 0, designating no specific nucleotide preference, to 2, where a single nucleotide is specifically preferred at a given position. One caveat of the PWM model is that it assumes all base positions to be independent of each other, which is not true for all TFs, where more complex approaches may be more appropriate[97]. Nevertheless, the simple PWM models are the most broadly used to date[94].



Nucleotide	1	2	3	4	5	6	7	8
A	0.26	-1.10	-1.40	0.59	-1.40	-1.40	-0.44	0.38
C	-0.44	-1.10	-1.40	-1.40	0.59	-1.40	-0.10	-0.25
G	0.19	-1.40	0.51	-1.40	-1.40	0.59	-1.40	-0.19
T	-0.62	0.58	-0.14	-1.40	-1.40	-1.40	0.45	-0.40

Figure 10: A position weight matrix and its motif logo visualization. In a PWM, each column represents the position, and each row represents the weights associated with each nucleotide (above). The PWM can be visualized via a motif logo, where the total height at each position corresponds to IC (below). Adapted from reference[94].

When the sequences of multiple binding sites for a given TF are known, and if the corresponding positions in the respective binding sites can be aligned to represent the binding motif, a PFM can be constructed simply by calculating nucleotide frequencies per position, which could then be converted to a PWM. As covered in the previous sections, ChIP-seq is an *in vivo* method that provides binding sites for a given TF genome-wide, which can then be used to look for the underlying PWM model. This constitutes the *de novo* motif discovery problem, where neither the precise locations of the binding sites within the ChIP-seq peaks, nor the expected motif parameters are known, and algorithms usually attempt to find the motifs that maximize IC[97]. There are also *in vitro* methods that assess TF-DNA binding, including protein-binding microarrays (PBMs), which provide robust binding scores for 8-mers[98] and high-throughput SELEX (HT-SELEX), in which 10-40bp long sequences are subjected to successive cycles of TF-binding, leading to increased specificity at each cycle[99]. PWMs are constructed from these *in vitro* approaches, using tailored computational methods[94]. Databases such as UniPROBE[100] and JASPAR[101] provide comprehensive PWM collections from such efforts.

If the binding model(s) of a TF is readily available, it becomes possible to scan a set of sequences or the whole genome, using the PWM model, to find motif matches that constitute putative TF binding sites (TFBSs). Many methods achieve this by sliding the PWM model one nucleotide at a time, and at each position, scoring the likelihood that the underlying sequence matches the PWM model, via summing the corresponding weights of the observed nucleotides[102]. The same likelihood calculation is carried out with a background model as well, and the log-likelihood ratio of the PWM model vs the background model, constitutes the match score. The background can be defined in different ways; one common way is to use a zero-order Markov model (i.e. frequencies of the four nucleotides) derived from the entire sequence set[103], whereas other methods choose more complex approaches such as local first-order Markov models, taking dinucleotide frequencies into account within a local window[104]. Putative TFBSs are usually analyzed further with data driven approaches to delineate true bound sites, such as TF footprinting covered in the next section.

2.4.3.2. Transcription factor footprinting

In the late 60s, it was observed that binding of RNA polymerase protects the underlying DNA from cleavage by DNase I[105], the first indication that protein bound DNA is less accessible to DNase I, compared to flanking regions. The emergence of sequencing methodologies a decade later[25,26], made it possible to infer the sequences of these protected stretches of nucleotides or shortly “footprints”[106]. Briefly, the first “DNase I footprinting” method[106] consisted of DNase I treatment of a DNA template (lac operator) in the presence of a specific binding protein (lac repressor), and electrophoresis of the resulting fragments on a nucleotide-resolution polyacrylamide gel. With this method, as the bound nucleotides cannot be cleaved by DNase I, the corresponding fragments cannot be obtained and the footprint appears as a gap on the gel. The products of standard Maxam-Gilbert sequencing[25] (see section 1.3.1) are run alongside the DNase I cleavage products on the same gel and the footprint sequence is inferred by comparison to the gap position. Variations of this *in vitro* method allow footprint inference *in vivo* as well[107,108], however these are low-throughput as they rely on probes specific to the region of interest.

The more recent high-throughput DNase-seq method, on the other hand, enables the inference of footprints genome-wide[109,110]. This is of special relevance to TFs, and a multitude of TF-footprinting methods have been developed to date[111], which can be grouped under three general categories: site-centric, segmentation based, and integrative site-centric methods. Site-centric methods model footprints specifically for candidate TFBSs, using the shape or magnitude of the DNase-seq signal around them[112–115]. Segmentation based methods, on the other hand, scan the DNase-seq signal for footprint-like signatures (eg. peak-trough-peak pattern) and subsequently match the identified footprints to putative TFs[116–120]. Integrative site-centric methods model bound sites using combinations of diverse features, such as motif match score, sequence conservation and variable length bins of DNase-seq signal around candidate TFBSs[121–126].

The efforts to assay bound sites genome-wide via TF-footprinting have come under scrutiny by studies demonstrating that DNase I cleaves the underlying DNA in a non-uniform manner, where sequence composition dictates the cleavage propensities (also known as sequence bias)[127,128]. This necessitates the discrimination of actual footprints from footprint-like signal profiles originating solely due to sequence bias[115]. To account for this, a number of TF-footprinting tools explicitly model and incorporate the bias background in their models or processing pipelines, by calculating the ratio of observed to expected DNase cuts for short sequences of fixed length[111,114,119]. 6-mers have been the primary choice, as they capture enough variation to represent the bias[115], in line with the finding that the main sequence information content around a DNase cut site is confined to the flanking 3 nucleotides on either side[127]. Open chromatin regions[111,115,119] or DNase-seq experiments conducted on deproteinized genomic DNA[111,114] have been used to infer these 6-mer cleavage propensities.

Recent efforts have explored the feasibility of TF-footprinting with ATAC-seq[123,124,129–131], however this is not yet studied as extensively as for DNase-seq. Furthermore, like DNase I, Tn5 transposase is reported to have specific target sequence preferences, which encompass the central 9bps that get duplicated during transposition, as well as ~5bp flanking regions on either side[69,74,132,133]. In line with this, a recent method aiming to correct sequence biases in high-throughput sequencing datasets, reported a 17bp long gapped k-mer with 8 meaningful positions as the optimal k-mer to correct ATAC-seq data[134], however, the signal was not smoothed completely in this setting. Taken together, the optimal way to correct for Tn5 sequence bias and its putative effects on TF-footprinting remain open questions in the field.

3. Materials and Methods

Each subsection within materials and methods is denoted part 1, part 2 or both, depending on which results section it refers to.

3.1 DNase-seq and ATAC-seq experimental procedures and data preprocessing

Part 1

DNase-seq and ATAC-seq assays were performed on human cell lines, K562 and HEK293 cells. K562 and HEK293 cells were cultured in Iscove's Modified Dulbecco's Medium (IMDM) and Dulbecco's Modified Eagle's Medium (DMEM), respectively, both complemented with 10% fetal bovine serum (FBS) and 1% Penicillin/Streptomycin.

Single-hit DNase-seq experiments were conducted on 50 million cells as previously described[65], with the minor modification of using 5' phosphorylated oligo 1b. Samples digested with 12U, 4U and 1.2U total DNase I were pooled. Libraries constructed from pooled digests were sequenced on the Illumina HiSeq2500 platform using the single-end sequencing mode with 50-bp reads. Analysis was conducted in line with the official ENCODE DNase-seq pipeline. Specifically, the reads were trimmed to the first 20 bases, as only this portion corresponded to the ends of DNase I-digested fragments, due to the MmeI cleavage step in the protocol. Trimmed reads were aligned to the hg19 build of the human genome, using the Burrows-Wheeler aligner (BWA)[86], tolerating up to two mismatches. Sequences aligning to more than four locations were discarded. Further processing was performed to filter out unwanted chromosomes and problematic regions such as alpha satellites. In order to remove PCR artifacts, reads that piled up (>=10 reads) at a single base were discarded, if they constituted at least 70 percent of all reads in the surrounding 30 base pair window.

ATAC-seq experiments were performed on 50000 cells for the K562 samples and 100000 cells for the HEK293 samples, following the published protocol[66] but increasing transposition time from 30 minutes to 1 hour for all samples. In addition, lysis conditions were varied in different experiments. For the K562 sample denoted "10 minute lysis", cell lysis was performed via a 10 minute centrifugation in lysis buffer, as described in the original protocol[66]. For the K562 sample denoted "5 minute lysis", a shorter lysis of 5 minutes was used. For the K562 sample denoted "no lysis buffer" and all HEK293 samples, the centrifugation in lysis buffer step was omitted altogether, and the cell pellets were taken directly to the transposition reaction. Libraries were sequenced on HiSeq2000 (Illumina), with 100-bp paired end reads. Since fragments as short as 38 base pairs were expected, adapter sequences were trimmed from the 3' end of the reads. Specifically, matches of any length to the reverse-complemented Nextera Transposase Adapters (CTGTCTCTTATACACATCTGACGCTGCCGACGA, CTGTCTCTTATACACATCTCCGAGCCCACGAGAC) were removed. Trimmed reads were aligned to the hg19 build of the human genome, using bowtie2[85] with parameter -X set to 1500, to allow correct alignment of paired-end fragments up to 1500 base pairs. Only reads that aligned uniquely to a single location were retained, by filtering out the multimappers marked with the XS:i flag in the sam file. PCR duplicates were removed using Picard (http://broadinstitute.github.io/picard/). Further processing was performed to filter out contigs as well as the Y and mitochondrial chromosomes, and retain only reads that aligned concordantly as a pair within the expected fragment length range (38-1500 bp).

Library complexity and saturation were calculated using the preseq program[135], using the c_curve and lc_extrap functionalities. Correlations of reads counts between libraries were calculated using the bamCorrelate bins command of the deepTools suite, with the parameters –corMethod pearson, -bs 100, --fragmentLength 1 and –doNotExtendPairedEnds.

Part 2

DNase-seq and ATAC-seq assays were performed on *Drosophila melanogaster* embryos. Transgenic *D. melanogaster* embryos carrying a mesodermal marker gene (histone H2B fused with streptavidin-binding protein) were described previously[48]. Staged embryos were collected and formaldehyde fixed as previously described[136]. In brief, embryos were collected on apple-agar plates in two-hour windows following three one-hour pre-collections for synchronization purposes. After ageing (at 25 °C) to the desired age, embryos were washed from the plates into a sieve using water, and dechorionated in 50% bleach (diluted from 6-14% sodium hypochlorite, Merck) for 2 min. Formaldehyde fixation was performed for 15 min with shaking (500 rpm) at room temperature in cross-linking solution (50 mM Hepes, 1 mM EDTA, 0.5 mM EGTA, 100 mM NaCl, pH 8, 1.8% formaldehyde v/v) with a heptane layer. Fixation was stopped by pelleting embryos by centrifugation at 500 g and exchanging the buffer for 125 mM glycine in PBS and shaking for a further 5 min. The embryos were washed in PBS, dried, snap frozen in liquid nitrogen, and stored at −80 °C in ~ 1 g aliquots.

Embryo dissociation and nuclear isolation were performed as described previously (steps 1–10)[48] using a dounce homogenizer and a 22G needle. The resulting nuclei were pelleted at 2,000g at 4 °C, resuspended in nuclear freezing buffer (50 mM Tris at pH 8.0, 25% glycerol, 5 mM Mg(OAc)2, 0.1 mM EDTA, 5 mM DTT, 1× protease inhibitor cocktail (Roche), 1:2,500 superasin (Ambion)) and flash frozen in liquid nitrogen.

Target populations of cell nuclei from staged fixed embryos were obtained by FACS as previously described[48] with the following modifications. Prior to incubation with primary antibodies, nuclei from 6–8-h embryos were incubated in PBS supplemented with 5% BSA, 0.1% TritonX-100 and 0.2% Igepal-630 on a rotator at 4 °C for 30 min. Primary antibody staining was performed overnight at 4 °C in 3 ml PBS supplemented with 5% BSA and 0.1% TritonX-100 per 1g frozen embryos. Primary antibodies used were monoclonal anti-Elav (Developmental Studies Hybridoma Bank 9F8A9 at 1:100 dilution) to mark postmitotic neurons and anti-Mef2 (produced and pre-cleared in the Furlong laboratory and used at 1:200 dilution) to mark myogenic mesoderm. Secondary antibody staining was performed for 1 h at 4 °C in the same buffer. Following each antibody staining, nuclei were washed twice by pelleting and resuspending in 10 ml PBS supplemented with 5% BSA. An aliquot of stained, unsorted nuclei was put aside to represent the whole embryo. For DNase digestion, nuclei were resuspended in R buffer (7.5mM Tris pH8, 45mM NaCl, 30mM KCl, 6mM MgCl2, 1mM CaCl2) and 10–20 million nuclei were digested using 5–20 U DNaseI at 37 °C for 3 min, and the reaction was stopped by adding 500 μ l stop buffer (50mM Tris pH8, 100 mM NaCl, 0.1% SDS, 100 mM EDTA pH8). A small control digest without DNaseI was performed to assess DNA integrity. Following addition of RNaseA, samples were incubated at 55 °C for 10 min, then 25 μ l proteinase K (25 mg/ml) was added and the samples were incubated overnight at 65 °C to reverse cross-links. A small aliquot was run on a 1% agarose gel to assess digestion levels, and optimal digests were size-fractionated using 10–40% sucrose gradients. DNA fragments ~ 100–500 bp in length were isolated from fractions using a Qiagen PCR clean up kit and checked for enrichment in known hypersensitive sites by qPCR. The digests with the highest qPCR enrichment were selected for library preparation using the NextFlex qRNA-seq Kit v.2 (Biooscientific #NOVA-5130-12). In brief, ~ 10–30 ng DNA consisting of ~ 100–500 bp fragments that result from DNase digestion was end-repaired and terminal adenosine residues were added. Adapters containing in-line molecular barcodes were ligated, after which the material was size selected using AMPure beads (negative selection with 0.6× beads, then positive selection with 0.98× beads). PCR amplification was performed using barcoded primers to introduce sample barcodes for 12–16 cycles, depending on input amount. The PCR-amplified library was purified using AMPure beads, quantified using a Qubit High-sensitivity DNA kit (Invitrogen), and sized on a Bioanalyzer High-Sensitivity DNA chip (Agilent). Libraries were pooled and sequenced in paired-end mode on a HiSeq2000 (Illumina). Reads were mapped to the dm3 reference genome using BWA aln, keeping only reads with a mapping quality score greater than 20. Duplicate reads originating from PCR were removed using the Je suite making use of the molecular indices.

S2 cells were either kept in their native (unfixed) chromatin state or fixed for 10 mins at room temperature with 1% formaldehyde. Cell lysis was done in 0.05% Igepal-640, incubating 5 mins on ice. DNase-seq was performed as explained above for the embryos.

For the embryo ATAC-seq datasets, embryos were staged, fixed, subjected to nuclear extraction and sorted as explained above. To sort intermediate column (IC) specific-nuclei, an appropriate primary antibody was used: (against a GFP specifically expressed in the IC) anti-GFP, rabbit mAb (#G10362 conc: 0.2 mg/ml), at 1:200 dilution for 1 hour, nutating at 4 °C. The secondary antibody used was goat anti-Rabbit IgG (H+L) Superclonal Secondary Antibody, Alexa Fluor 555 (Product # A27039) at 1:100 dilution. No staining was needed to sort ventral column (VC) specific-nuclei, since dsRED could be assayed directly in the VC-dsRED fly nuclei. ATAC-seq was performed on 200,000 sorted or unsorted nuclei, with the following protocol: nuclei were pelleted by centrifugation (3200g, 5min, 4 degrees), resuspended in 1ml lysis buffer (PBT + 0.2% NP40), lysed by rotating for 60min at 4 degrees, and then washed once with PBT (same centrifugation). Pellets were resuspended in tagmentation reaction (25ul TD buffer (Illumina), 5ul Tn5 enzyme (Illumina) and 20ul water) and incubated at 37 degrees for 1 hour. Reverse-crosslinking was carried out by adding 50ul STOP buffer (50mM Tris-HCl (pH 8.0), 100mM NaCl, 0.1% SDS, 100mM EDTA (pH 8.0), 1mM spermidine, 0.3mM spermine and 40ug/ml RNase A) to each reaction and incubating at 55 degrees for 10 min. 3ul Proteinase K (20 mg/ml) was then added and the samples incubated at 65 degrees overnight. Transposed DNA was purified using QIAquick PCR Purification Kit (Qiagen), in 10ul elution buffer. The following 50ul PCR reaction was prepared for each sample: 10ul water, 2.5ul forward primer, 2.5ul reverse primer (with barcodes), 25ul NEBNext

High-Fidelity 2X PCR Master Mix (NEB) and the 10ul transposed DNA. Then, a 15ul qPCR reaction was prepared with: 5ul from the prepared PCR reaction mix, 2.5ul water, 0.5ul forward primer, 0.5ul reverse primer, 1.5ul 10X Sybr Green, and 5ul NEBNext High-Fidelity 2X PCR Master Mix (NEB). We ran the qPCR with the following protocol: 72 degrees for 3 min, 98 degrees for 30 sec, followed by 25X (98 degrees for 10 sec, 63 degrees for 30 sec and 72 degrees for 2 min). For each sample we inferred the CT value and calculated CT+6 as the optimal number of PCR cycles. The main PCR reaction was then ran with the same protocol as for the qPCR, but using the optimal number of cycles. The PCR amplicons were ran on a 1.2% agarose gel, and gel extracted to exclude primer dimers, using QIAquick Gel Extraction Kit (Qiagen) and 20ul elution buffer, constituting the final ATAC-seq library. Libraries were sequenced on NextSeq (Illumina), with 75-bp paired end reads. Adapter trimming was done as explained for part 1. Trimmed reads were aligned to the dm6 reference genome, using Bowtie2, and aligned reads further processed as explained in part 1.

3.2 Peak calling

Part 1

In order to find open chromatin regions, peak calling was performed on the processed DNase-seq and ATAC-seq datasets using JAMM[89], with parameters -f 1 and -d y. Parameter -f 1 ensured taking only the 5' ends of the reads into account which corresponded to the actual cleavage/transposition sites. As duplicates were already removed prior to peak calling, parameter -d y was used to keep all processed reads.

Where replicates were available, peaks in agreement between the two replicates were found using the irreproducible discovery rate (IDR) pipeline[91]. Specifically, the "batch-consistency-analysis.r" script of the pipeline was executed using the "signal.value" parameter, ranking the peaks of the two replicates by signal intensity for comparison. The "half.width" and "overlap.ratio" parameters were set to -1 and 0, respectively, where true peak widths were used without alteration and two peaks were considered to be part of the same region if there was at least 1bp overlap between them. The number of peaks that were found to be concordant at the stringent 0.01 IDR threshold was noted. Then, JAMM was once again used, this time to call peaks on the two replicates together rather than individually, with the -f 1,1 parameter. In this way, peaks were called where both replicates consistently displayed signal enrichment. This peak set was further truncated using the number obtained from the IDR analysis, resulting in the final JAMM-IDR peaks.

For K562 ATAC-seq datasets, where replicates were not available, reads of the modified dataset with no lysis buffer was randomly subsetted to match the library depth of the original protocol with 10 minute lysis, and peaks were called using JAMM as described above, with the addition of the -e auto parameter for automatic estimation of a minimum fold enrichment. These K562 ATAC-seq peak sets were used to infer signal to noise ratios by calculating log2(average signal in the peaks/average signal in the 300bp upstream and downstream flanking regions).

Part 2

Peaks were called on *Drosophila* embryo DNase-seq datasets using MACS2[88] with the following parameters: -f BAMPE -g 1.2e8 --keep-dup all --call-summits. Parameter -f BAMPE ensured using real fragment sizes given by read pairs instead of the default fragment size modelling by MACS2. Parameter -g 1.2e8 specified the effective euchromatic genome size of *Drosophila melanogaster*. We kept all duplicates with --keep-dup all, as duplicates were removed prior to peak calling using molecular barcodes. Multiple summits of signal were found for each peak by --call-summits. Summits were extended 40bps upstream and downstream, and merged. The coverage of summits from all samples was calculated to identify the maximum coverage position of each cluster which was subsequently called the cluster summit. To call reproducible summits in each sample within these clusters, summits were slopped by around 20 bp and IDR was run, taking only those regions passing 10 percent IDR. (An alternative set of DHSs was also generated for these datasets using the JAMM-IDR approach outlined in part 1 and merging all peaks. In order to avoid confusion, these will be referred to as JAMM-IDR-DHSs whenever mentioned in this thesis.)

Peaks were called on *Drosophila* S2 cell DNase-seq datasets using JAMM[89] with the -f 1,1 and -d y parameters, to call peaks on replicates together, without removing duplicates. The procedure was followed separately for the native and crosslinked datasets. JAMM's filtered peak output was used as the final set of DHSs in both cases.

Peaks were called on *Drosophila* embryo ATAC-seq datasets using MACS2[88] with the following parameters: -f BED -g dm --nomodel --shift -50 --extsize 100 --keep-dup all -p 0.05.

Parameter -f BED specified the input format as bed and -g dm referred to the effective euchromatic genome size of *Drosophila melanogaster*, equivalent to 1.2e8 as above. The combination of parameters --nomodel --shift -50 --extsize 100 was used to inhibit the default fragment size modelling by MACS2 and instead shift and extend the reads by 50bp and 100bp, respectively. This corresponded to smoothing the ATAC-seq signal in a 100bp window, centered on the 5' end of reads (the transposition site). Duplicates were kept with --keep-dup and the p-value threshold for reporting peaks was set as 0.05 with -p 0.05. Peaks found in pairs of replicates were subjected to IDR analysis as explained in part 1 above, but using the "p.value" parameter to rank the peaks by p-value in this case. The number concordant peaks at the stringent 0.01 IDR threshold was noted. Next, reads in replicate datasets were pooled, and MACS2 was used to call peaks on the pooled datasets as explained above. This peak set was truncated using the number obtained from the IDR analysis, leading to the MACS2-IDR peaks. Finally, for each pooled dataset, peak calling was repeated with the addition of the --call-summits parameter, to call sub-peak resolution summits. Summits that overlapped the MACS2-IDR peaks were retained. As above, summits were extended 40bps upstream and downstream, merged across all datasets, resulting in the final ATAC-HSs. A weighted summit per ATAC-HS was defined as the average of original summit locations.

3.3 Sequence bias of DNase I and Tn5 transposase

Part 1

The sequence bias of the Tn5 transposase was calculated in the form of 6-mers, similar to the previous calculations of DNase bias[114]. To this end, libraries generated by Tn5 transposition on deproteinized genomic DNA (see supplementary table 2) were preprocessed in the same way as ATAC-seq datasets as detailed above. As the 5' ends of the reads corresponded to the transposition sites, the sequences of all 6-mers centered on these sites were retrieved (e.g. transposition between the third and fourth nucleotides). Occurrences of all these 6-mers in the data were counted and the relative frequencies were calculated for each. Similarly, background genomic frequencies were calculated by counting all 6-mers present in the mappable portion of the genome. The frequencies observed in the data were normalized to the background frequencies to obtain the final transposition propensities per 6-mer. Deviations from one indicated increased or decreased propensities, thus bias.

The average Tn5 transposition propensity in a candidate binding site of a given transcription factor was calculated by retrieving and counting all 6-mers associated with the site (without flanks). The counts were multiplied by the Tn5 transposition propensities of the associated 6-mers, summed and normalized by the total number of 6-mers in the site. The same calculation was applied for DNase, using the previously calculated DNase cleavage propensities per 6-mer[114].

Part 2

DNase I bias was calculated in the form of 6-mers, as described above; but using DHSs, instead of naked DNase-seq experiments. Specifically, reads within DHSs were taken into consideration when counting the 6-mer occurrences. In accordance with this, the background 6-mer frequencies were derived from the DHS sequences. For *Drosophila* embryo DNase-seq datasets, the JAMM-IDR-DHSs were used in conjunction with pooled reads from all datasets. This resulted in a single set of 6-mer bias values representing all embryo DNase-seq datasets. In the *Drosophila* S2 cell DNase-seq datasets, native DHSs were used to calculate bias for native datasets, and crosslinked DHSs for crosslinked datasets.

3.4 Scanning the genome for candidate binding sites

Part 1

The SpeakerScan Toolset[104] was used to scan the hg19 build of the human genome with position weight matrices (PWMs), to find candidate transcription factor binding sites (TFBS). PWMs contain expected frequencies for each nucleotide in a per-base fashion, modeling the binding sequence preferences of a given TF. A pseudocount of 0.0005 was added to each frequency in the PWMs, to ensure non-zero entries. At each PWM-sized window in the genome, a TFBS score was calculated, as the log-likelihood of the underlying sequence matching the PWM versus a background model. The background was modeled with a first order Markov chain in a 500 bp local window, centered on the considered position. The top scoring 50000 sites were taken along for transcription factor footprinting in this study. To validate the significance of motif matches for all PWMs, we simulated DNA sequences using the PWM model (positive set) and a background modeled with a first order Markov chain from hypersensitive sites (negative set) and sampled TFBS scores from these positive and negative sets. For a range of false positive rates (FPR, up to 1*10-6) we found the corresponding TFBS scores and reported the one closest to the lowest score in each motif set, and the associated FPR as an empirical p-value. This demonstrated that all our sets included significant motif matches, with the lowest empirical p-value being 5*10-5 (Appendix A: supplementary table 5).

Part 2

FIMO[103] from the MEME suite was used with default settings, to scan the dm3 build of the *Drosophila melanogaster* genome with a custom set of PWMs (see below), to find candidate transcription factor binding sites (TFBS). The background was modeled with a zero order Markov chain, using nucleotide frequencies derived from the total set of distal DHSs from the embryo datasets (A 0.279, C 0.221, G 0.221, T 0.279). All motif occurrences with a p-value less than 1e-4 (default FIMO output) were taken along for TF footprinting or integrative modelling of TF binding analyses.

The custom set of PWMs used in this thesis were previously published[137]. Briefly, *Drosophila* PWMs were collected from: the Furlong laboratory; the modENCODE consortium; Berkeley TF ChIP-Seq data; Berkeley Drosophila Transcription Network; Flyfactor/Flyreg database; and the Jaspar database. In addition, TF ChIP datasets from developmental stages of *Drosophila* embryogenesis were collected from: the Furlong laboratory; modENCODE ChIP-Seq and ChIP-chip and Berkeley ChIP-Seq and ChIP-chip. PWMs that were represented by the ChIP datasets were retained, and clustered (normalized Pearson correlation > 0.75) to get the final non-redundant set of 226 PWMs.

3.5 Identification of transcription factor footprints

Part 1

Transcription factor footprinting was performed with a site-centric method from our lab as previously described[114]. Specifically, candidate TFBSs were considered with 25bp flanks upstream and downstream (parameter PadLen=25). Parameter k=2 was used to model two components; one for the footprint and one for the background. Both components were modeled as multinomials along the considered window size (TFBS+50bps), where each value corresponded to the cleavage/transposition probabilities at a given nucleotide. For the footprint component, these probabilities were found by computing the aggregate DNase or ATAC-seq signal (from the 5' ends of the reads) around the TFBSs that overlap ChIP-seq peaks for that factor and re-estimating the signal via expectation maximization. For the background component, the probabilities were calculated as the signal that would be expected solely due to the protocol-specific bias values, given the sequences around the candidate TFBSs (parameter Background="Seq"). As we had previously not observed a distinct difference in performance, the background was kept fixed and not re-estimated (parameter Fixed=T). Once both components were learned, footprint scores were calculated for all candidate TFBSs, as the log-odds of footprint versus background (footprint log-likelihood ratio, FLR). To learn footprint models without bias correction, our method was applied as described above, but with a uniform, fixed background model that assumes equal cleavage probabilities at each nucleotide.

The IDR strategy was applied here as well where replicates were available, to find reproducible footprints. To this end, candidate TFBSs with positive FLRs in both replicates were chosen and ranked by FLR. IDR analysis was performed with the same parameters as explained for peak calling, where FLR values replaced signal intensities. Again, the number of sites that passed the stringent 0.01 IDR threshold was noted. Finally, TFBSs were ranked by the average FLR from the two replicates and truncated according to the IDR result. This led to the reproducible FLR-IDR footprints.

Footprint model AUCs (both area under the ROC and precision-recall curves) were calculated by 4-fold cross validation. Briefly, the data was split into 4 parts, and for TFBSs in each part, FLR was calculated using footprint and background models learned from the other 3 parts. TFBSs were ranked by FLR, and those intersecting ChIP-seq peaks were labeled as the true positives. The AUCs obtained from the four parts were averaged to obtain the final value.

Similarly, sensitivity and specificity measures were also obtained using models trained on 3/4 of the data and tested on the remaining 1/4.

Correction of Tn5 sequence bias in K562 ATAC-seq data with the seqOutBias software was carried out according to the guidelines provided in the vignette. Specifically, for plus strand reads, --kmer-mask NXNXXXCXXNNXNNNXXN and for minus strand reads --kmer-mask NXXNNNXNNXXCXXXNXN was used to correct the signal. The corrected data was then used to learn footprint models with our method, in conjunction with a uniform, fixed background model.

Part 2

For all *Drosophila* DNase-seq and ATAC-seq datasets (embryo and S2), footprinting was performed as explained in part 1 above; but with k=3 to model three components; two for the footprint and one for the background. In addition, for DNase-seq datasets, DHS-derived bias values were used (see section 3.3).

3.6 Differential signal analyses

Part 2

To conduct differential signal analysis in the *Drosophila* embryo DNase-seq datasets, we first counted reads in DHSs using the featureCounts function from the Bioconductor package Rsubread. DESeq2[93] was then used, with the default mean normalization, to identify statistically significant differences in read counts (signal) between different tissues or time-points. The Wald test was used to test significance, against the null hypothesis of fold change=1 (no difference in signal). To define our strict tissue-specific regions, we contrasted a given tissue, with the other tissues from the same time-point (e.g. meso-specific regions are defined by contrast to neuro and other datasets). DHSs were deemed tissue-specific if they were found only in the tissue of interest (and not the contrasted tissues) and if they had significantly higher signal here (fold change > 4, Benjamini-Hochberg adjusted p-value < 0.01). To define the time-point specific regions, we contrasted consecutive time-points of the same tissue as described above; but relaxing the fold change threshold to > 1.5. DHSs that were found in all relevant samples, with fold change values between -1.5 and 1.5, and an adjusted p-value > 0.05 were defined as the shared regions between those samples that show no differential signal.

To conduct differential signal analysis in the *Drosophila* embryo ATAC-seq datasets, we first counted reads in ATAC-HSs using the multiBamCov tool from the bedtools suite. DESeq2[93] was used, with the default mean normalization and Wald test, as described above. We defined tissue-specific ATAC-HSs by contrast to the unsorted datasets at the same time-point, requiring fold change > 2 and Benjamini-Hochberg adjusted p-value < 0.01.

3.7 Motif enrichment analyses

Part 2

Motif enrichment analyses were conducted on the TSS-distal tissue and time-point specific regions defined by differential signal analyses (see section 3.6). Each specific set was extended 100bp upstream and downstream from its summit and the resulting overlapping regions were merged. For each tissue specific set, a length, TSS-distance and GC content-matched background sets was selected from among all distal DHSs or ATAC-HSs (also +/-100bp extended and merged). For each time-point specific DHS, the background was selected according to the same criteria; but from among the set of shared regions (also +/-100bp extended and merged) between consecutive time-points. Matching was done with MatchIt, choosing the ratio parameter to have the same range of background regions across specific sets (6000-7000 for tissue-specific DHSs and ATAC-HSs, and 1000-2000 for time-point specific DHSs). Motifs used in enrichment analyses were obtained by combining all fly motif databases from the MEME suite (OnTheFly, Fly Factor Survey, dmmpmm, idmmpmm and flyreg) and the custom PWM list (see section 3.4) for a total of 1677 redundant PWMs. AME from the MEME suite was then used with default settings to assess the enrichment of these motifs in the specific sets vs the background sets. Specifically, Fisher's exact test was employed to test whether the number of motif matches in the specific set is significantly greater than the background set. Motif matches were defined using a p-value threshold of 0.0002 given by FIMO (see section 3.4). Motifs were reported as enriched, if they passed a Bonferroni-corrected p-value threshold of 0.05. All enriched PWMs per analysis were then combined (e.g. all tissue-specific set results and all time-point-specific set results were combined separately). Enriched PWMs were trimmed to remove uninformative edges using trimPWMedge function from the MotIV package with an information content threshold of 0.5. After trimming, the matrix-clustering tool from the RSAT suite was used to cluster the redundant PWMs with parameters

-w 4, -cor 0.7 and -Ncor 0.5, requiring at least 4 aligned bases, Pearson correlation of 0.7 and width-normalized correlation of 0.5 among PWMs, to cluster them.

Using the same specific and background sets as for enrichment, we trained models using GKM-SVM[138], to find sequence features that discriminate the specific sets from the background sets. GKM-SVM was used with default parameters where L=10 and K=6: feature weights were inferred for 10-mers, and the number of informative bases to estimate 10-mer counts was 6. AUROC values were derived from 5-fold cross validation. The top 300 10-mers (ranked by feature weights) associated with each specific set were then clustered to get PWM-like sequence features using RSAT as described above. These GKM-SVM derived PWMs were aligned to the enriched clusters, using TOMTOM from the MEME suite with default settings, and taking the best alignment per tested GKM-SVM-PWM.

3.8 Integrative model of TF binding

Part 2

Putative TFBSs obtained by FIMO, using the custom PWMs (see section 3.4), were filtered to retain only TSS-distal regions and eliminate regions with 0 or 1 count across all *Drosophila* embryo DNase-seq datasets. Furthermore, PWMs with <50 putative TBFSs overlapping ChIP-seq peaks were also eliminated. Three features were extracted for each remaining TFBS: 1) log2-transformed read counts in a +/-25bp region spanning the TFBS, 2) motif match score from FIMO and 3) sequence conservation as measured by the number of substitutions per base across the *Drosophila* phylogeny. All three were standardized to zero mean and unit variance and used as features to predict TF binding in an integrative model based on logistic regression (l1_logreg software: https://web.stanford.edu/~boyd/l1_logreg/). Models were trained in a supervised manner, where putative TFBSs overlapping ChIP-seq peaks constituted the positive sets. We subsetted equal numbers of positive and negative sites and calculated AUROCs with 4-fold cross validation. Models were trained for all time-points spanned by the ChIP-seq data (e.g. if a TF had ChIP-seq data for 0-12hr embryos, then we trained models for all embryo DNase-seq data spanning all tissues and time-points). Feature coefficients from best performing models per PWM were averaged to create the generic model.

4. Results

4.1. Part 1: Reproducible inference of regulatory regions and transcription factor footprints using open chromatin profiling data

Contribution Statement:

Aslihan Karabacak Calviello performed all the sequencing data analysis, supervised by Uwe Ohler. ATAC-seq and DNase-seq datasets were generated by Antje Hirsekorn. Material appearing in this Section has been copied or adapted from our preprint (doi: https://doi.org/10.1101/284364).

4.1.1. A modified ATAC-seq protocol decreases mtDNA contamination and improves agreement with DNase-seq

Early ATAC-seq libraries generated with the original protocol have large numbers of reads mapping to mitochondrial DNA (mtDNA) that need to be discarded, which severely impacts the final library depth[66]. For an ATAC-seq library where we followed this protocol, we made the same observation in K562 cells, with 75% of the reads mapping to mtDNA (figure 11A, supplementary table 1). To decrease the mtDNA contamination, we evaluated two different approaches: decreasing the time of cell lysis to 5 minutes in lysis buffer (from the original 10 minutes) and eliminating the lysis buffer step altogether by proceeding directly to the transposition reaction. Of these, particularly the approach where no lysis buffer was used, led to a substantial improvement, with only 18% percent of the reads mapping to mtDNA in this library (figure 11A, supplementary table 1). Avoiding the detergent lysis may help mitochondrial membranes to stay intact, with other forces such as osmotic pressure being adequate to permeabilize the nuclear membrane.

To adequately quantify the protocol-related differences of ATAC-seq vs. DNase-seq, we also generated a single-hit DNase-seq library in K562 cells, and compared this alongside three other publicly available single-hit DNase-seq datasets (supplementary table 2) with the ATAC-seq libraries. Avoiding the usage of lysis buffer also increased the read-level agreement between the two experimental approaches (figure 11B, Pearson correlations of read counts in 100 base pair bins; figure 12A). This effect was already partially visible in data from the short lysis protocol. To investigate whether this observation is also reflected at the region-level of open chromatin, we called peaks with JAMM[89] and identified the set of concordant peaks using the irreproducible discovery rate (IDR) pipeline for DNase-seq data where replicates were available (see methods)[91]. Using the peak signal values for ranking, at the stringent 0.01 IDR threshold, we found 80,300 JAMM-IDR peaks for DNase-seq. We also called peaks with JAMM in the ATAC-seq datasets; since replicates were not available for these libraries, the IDR procedure was not applied here. We found 134,761 and 90,973 peaks for the original protocol and the modified protocol with no lysis buffer usage, respectively. Compared to the original protocol, the open regions identified with the modified protocol are more TSS-proximal, with higher GC content, and, in line with previous reports[139], have a modestly reduced signal to noise ratio (figure 12B). The ATAC-seq peaks found with the original and modified protocols had 45,340 (figure 11C, left) and 37,934 (figure 11C, right) overlaps to DNase-seq peaks, respectively. Using an extended unfiltered set of open regions as background for Fisher's exact test, both overlaps were found to be highly significant (pval<2.2e-16), with a slightly higher odds ratio for the modified protocol (13.15 vs 10.75). This improved agreement at the open chromatin region-level, albeit moderate, provided further support that avoiding detergent lysis increases the concordance between ATAC-seq and DNase-seq.

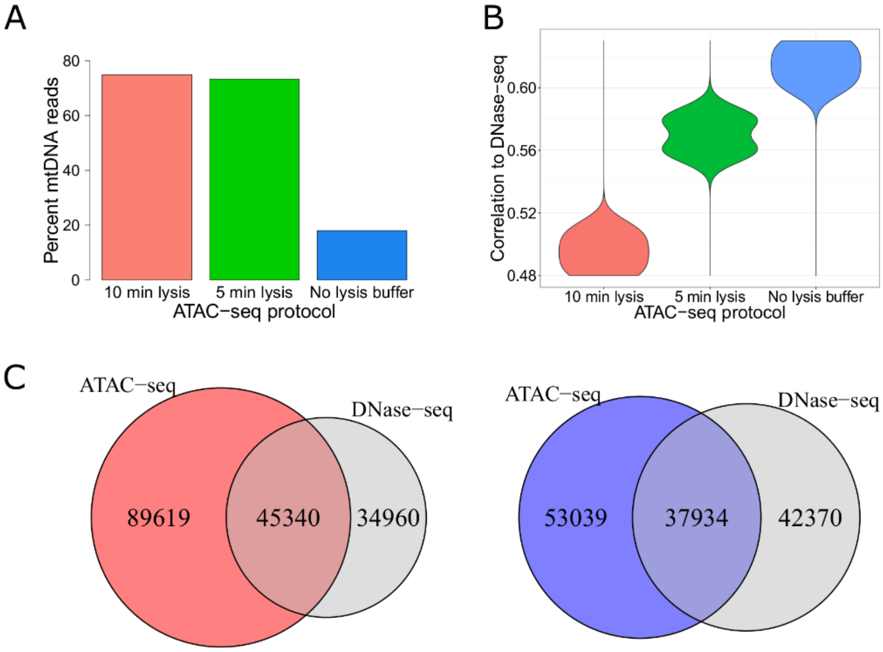


Figure 11: Generating ATAC-seq libraries without the usage of lysis buffer increases agreement with DNase-seq. (A) Percentage of all reads that align to the mitochondrial genome in K562 ATAC-seq libraries generated with the published protocol (10 min lysis), shorter lysis (5 min lysis) or without using lysis buffer (no lysis buffer). (B) Agreement of these libraries with all K562 DNase-seq libraries as measured by Pearson correlations of read counts in 100bp bins genome wide. (C) Overlap of peaks found in K562 DNase-seq data with peaks in ATAC-seq data generated using the published protocol (left) and peaks in ATAC-seq data generated without using lysis buffer (right).

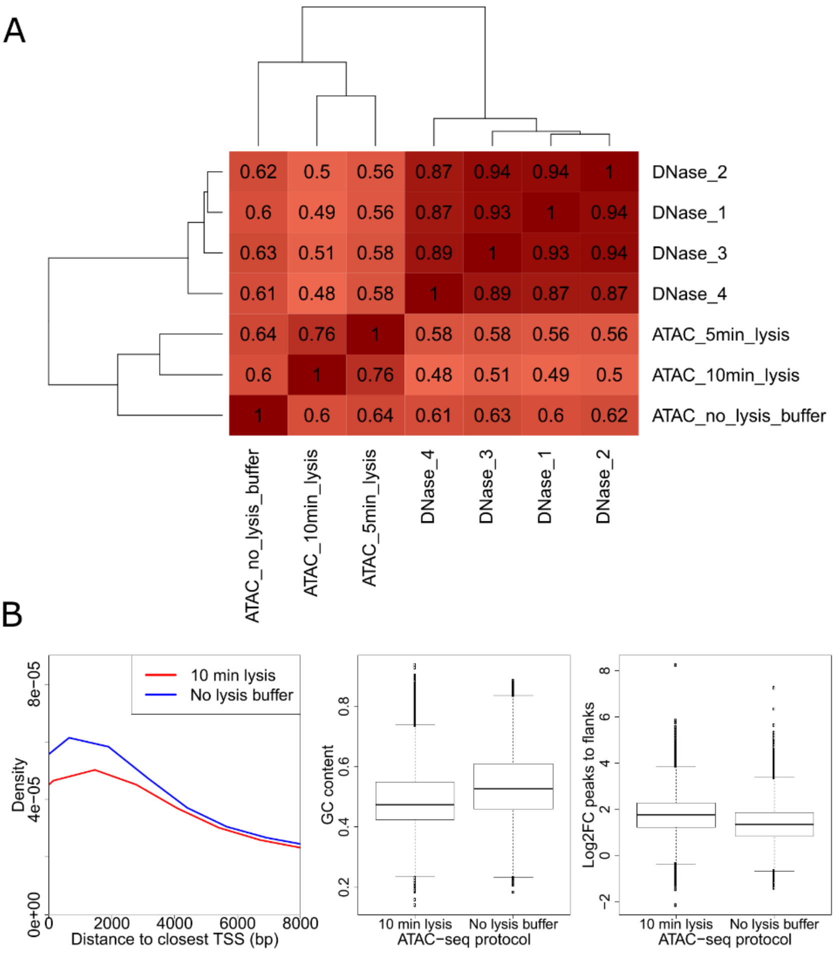


Figure 12: Characterization of ATAC-seq datasets generated with different protocols in K562 cells. (A) Pairwise Pearson correlations of read counts in 100bp bins genome-wide for all ATAC-seq and DNase-seq datasets in K562 cells. ATAC-seq datasets are labeled with the employed protocol: 10 min lysis (published protocol), 5 min lysis and no lysis buffer. DNase 1-3 are the replicates from the ENCODE project and 4 is the library newly generated for the study, all following the single-hit protocol. (B) Comparison of hypersensitive sites (HSs) found in K562 ATAC-seq datasets generated with the original (10 min lysis) and modified (no lysis buffer) protocols. HSs are compared with respect to distance to the nearest TSS (left), GC content (middle) and log2 fold change of read counts in HSs vs. flanking regions (right).

4.1.2. Open chromatin regions are found reliably at moderate library depths

The library depth of next-generation sequencing protocols that is required for a given downstream application is not always clear, especially when the regions of interest are not as clearly defined as e.g. protein-coding genes. To investigate the effect of library depth on uncovering open chromatin regions, we generated 11 ATAC-seq libraries with different depths in HEK293 cells using the protocol with no lysis buffer (four high, three medium and four low-depth libraries, figure 13A and supplementary table 1). The individual libraries were derived from two biological replicates. To obtain the highest possible depth representing these two samples (>300,000,000 read pairs each), all technical replicates were merged and denoted by "combined ATAC-seq replicates". Alongside the ATAC-seq experiments, we generated a single-hit DNase-seq library in HEK293 cells and additionally downloaded and processed two publicly available single-hit DNase-seq replicates (supplementary table 2). We observed strong positive correlations between all ATAC-seq and DNase-seq libraries at the level of genome-wide read counts (0.62-0.77 Pearson correlations of read counts in 100 base pair bins; figure 14), and JAMM-IDR peaks called for the combined ATAC-seq and DNase-seq replicates showed again a significant overlap (figure 15A).

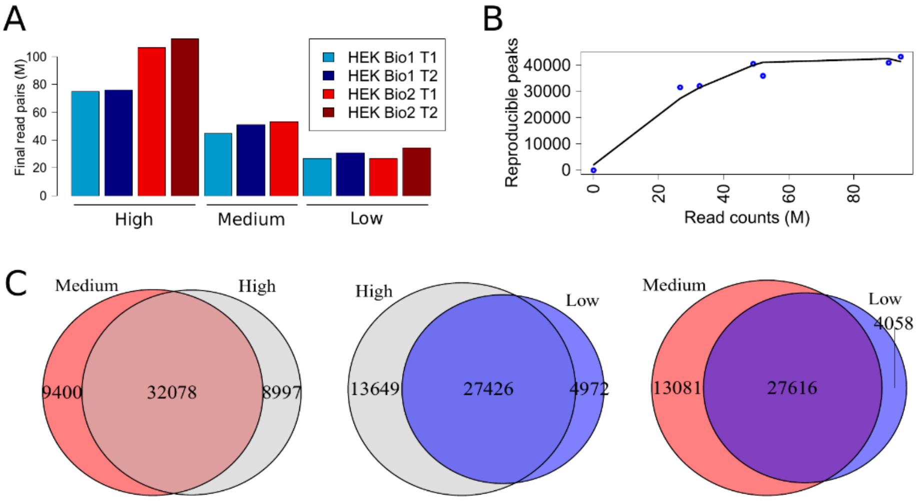
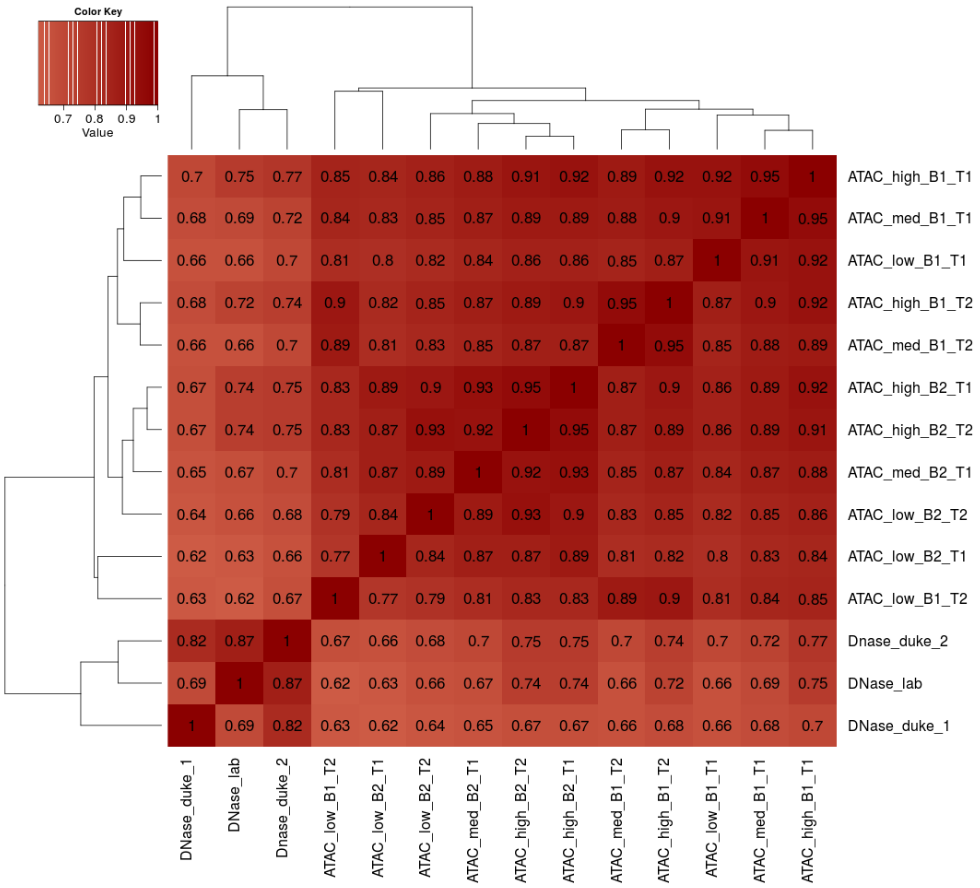


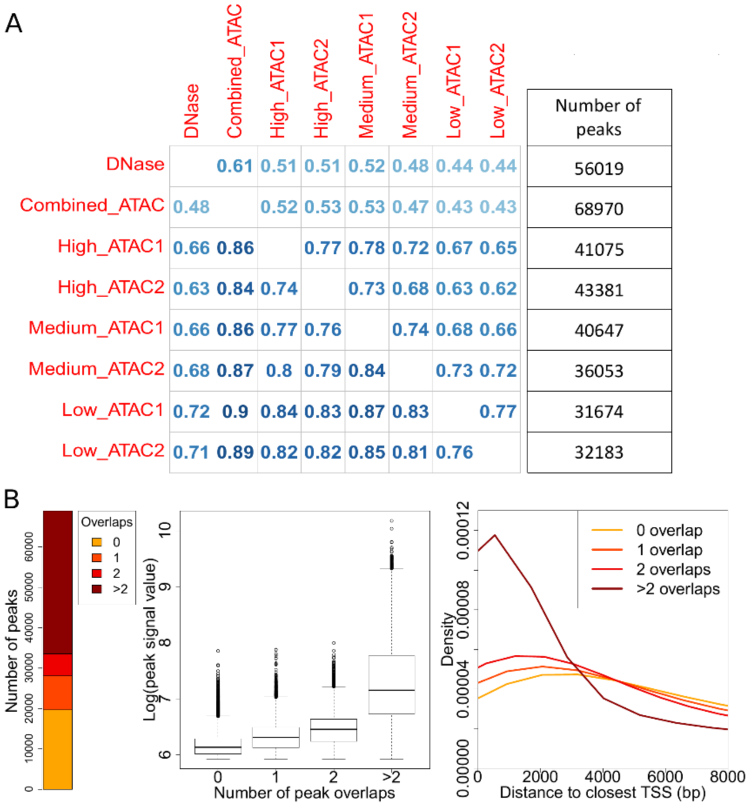
Figure 13: The task of finding open chromatin regions saturates at medium depth. (A) Number of reads after processing in the 11 HEK293 ATAC-seq libraries with different library depths. The two biological replicates are shown in blue and red, with the shades representing the technical replicates. (B) Numbers of reproducible peaks found with the JAMM-IDR strategy at different depths. (C) The overlaps between one set of peaks in (B) shown for high vs. medium (left), high vs. low (middle) and medium vs. low sets (right).



	DNase_duke_1	DNase_lab	Dnase_duke_2	ATAC_low_B1_T2	ATAC_low_B2_T1	ATAC_low_B2_T2	ATAC_med_B2_T1	ATAC_high_B2_T2	ATAC_high_B2_T1	ATAC_med_B1_T2	ATAC_high_B1_T2	ATAC_low_B1_T1	ATAC_med_B1_T1	ATAC_high_B1_T1
ATAC_high_B1_T1	0.7	0.75	0.77	0.85	0.84	0.86	0.88	0.91	0.92	0.89	0.92	0.92	0.95	1
ATAC_med_B1_T1	0.68	0.69	0.72	0.84	0.83	0.85	0.87	0.89	0.89	0.88	0.9	0.91	1	0.95
ATAC_low_B1_T1	0.66	0.66	0.7	0.81	0.8	0.82	0.84	0.86	0.86	0.85	0.87	1	0.91	0.92
ATAC_high_B1_T2	0.68	0.72	0.74	0.9	0.82	0.85	0.87	0.89	0.9	0.95	1	0.87	0.9	0.92
ATAC_med_B1_T2	0.66	0.66	0.7	0.89	0.81	0.83	0.85	0.87	0.87	1	0.95	0.85	0.88	0.89
ATAC_high_B2_T1	0.67	0.74	0.75	0.83	0.89	0.9	0.93	0.95	1	0.87	0.9	0.86	0.89	0.92
ATAC_high_B2_T2	0.67	0.74	0.75	0.83	0.87	0.93	0.92	1	0.95	0.87	0.89	0.86	0.89	0.91
ATAC_med_B2_T1	0.65	0.67	0.7	0.81	0.87	0.89	1	0.92	0.93	0.85	0.87	0.84	0.87	0.88
ATAC_low_B2_T2	0.64	0.66	0.68	0.79	0.84	1	0.89	0.93	0.9	0.83	0.85	0.82	0.85	0.86
ATAC_low_B2_T1	0.62	0.63	0.66	0.77	1	0.84	0.87	0.87	0.89	0.81	0.82	0.8	0.83	0.84
ATAC_low_B1_T2	0.63	0.62	0.67	1	0.77	0.79	0.81	0.83	0.83	0.89	0.9	0.81	0.84	0.85
Dnase_duke_2	0.82	0.87	1	0.67	0.66	0.68	0.7	0.75	0.75	0.7	0.74	0.7	0.72	0.77
DNase_lab	0.69	1	0.87	0.62	0.63	0.66	0.67	0.74	0.74	0.66	0.72	0.66	0.69	0.75
DNase_duke_1	1	0.69	0.82	0.63	0.62	0.64	0.65	0.67	0.67	0.66	0.68	0.66	0.68	0.7

Figure 14: Pairwise Pearson correlations of read counts in 100bp bins genome-wide for the ATAC-seq and DNase-seq datasets in HEK293 cells. All ATAC-seq datasets are generated with the protocol where no lysis buffer is used. The corresponding library depth (high, medium or low), biological (B1 or B2) and technical (T1 or T2) replicate status is indicated. DNase 1 and 2 are the replicates from the ENCODE project and lab refers to the library newly generated for the study, all following the single-hit protocol.

We then investigated to what extent the individual ATAC-seq libraries sequenced at different depths could capture the open chromatin regions uncovered by the combined replicates. To this end, libraries of similar depth from different biological replicates were matched in a pairwise manner to get JAMM-IDR peaks (supplementary table 3). This resulted in six total peak sets, corresponding to two of each of high, medium and low-depth library comparisons. Similar numbers of peaks were found at high and medium depth, with a slight decrease at low depth (figure 13B, figure 15A). Additionally, these peak sets displayed notable agreement among themselves and with the peaks of the combined ATAC-seq dataset (figure 13C, figure 15A). These observations suggested near-saturation for the task of defining open chromatin regions, even though none of the libraries were at saturation at these depths (figure 16). Moreover, these six IDR peak sets showed 63% to 72% overlap with the peaks of the DNase-seq data, which exceeded the 61% observed for the combined ATAC-seq data (figure 15A); even though a higher number of peaks was found in the combined dataset, IDR analysis of the individual datasets led to more reproducible subsets of the total pool. In support of this, the peaks found in the combined ATAC-seq dataset that did not overlap any of the peaks in the six individual sets, were predominantly low-signal, distal regions (figure 15B). Taken together, replicate libraries of low to medium depth of 25-50 million reads were sufficient for reliable identification of open chromatin regions in human cell lines.



	DNase	Combined_ATAC	High_ATAC1	High_ATAC2	Medium_ATAC1	Medium_ATAC2	Low_ATAC1	Low_ATAC2	Number of peaks
DNase		0.61	0.51	0.51	0.52	0.48	0.44	0.44	56019
Combined_ATAC	0.48		0.52	0.53	0.53	0.47	0.43	0.43	68970
High_ATAC1	0.66	0.86		0.77	0.78	0.72	0.67	0.65	41075
High_ATAC2	0.63	0.84	0.74		0.73	0.68	0.63	0.62	43381
Medium_ATAC1	0.66	0.86	0.77	0.76		0.74	0.68	0.66	40647
Medium_ATAC2	0.68	0.87	0.8	0.79	0.84		0.73	0.72	36053
Low_ATAC1	0.72	0.9	0.84	0.83	0.87	0.83		0.77	31674
Low_ATAC2	0.71	0.89	0.82	0.82	0.85	0.81	0.76		32183

Figure 15: Analysis of reproducible peaks in HEK293 cells. (A) Overlaps between all reproducible JAMM-IDR peaks found in HEK293 DNase-seq and ATAC-seq datasets. The number in each cell represents the ratio of the peaks in the row-dataset that overlap the peaks of the column-dataset. Total numbers of peaks are given on the right. (B) Number of JAMM-IDR peaks in the combined ATAC-seq replicates that overlap the union of peaks from the six individual datasets zero, one, two or more times (left). Peak signal values (middle) and distance to closest TSS (right) are shown for these four groups.

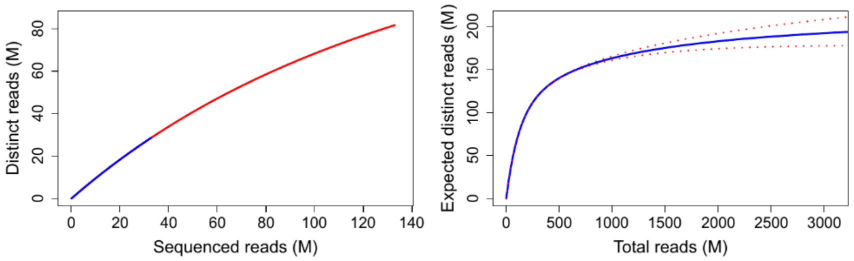


Figure 16: Library complexity and saturation plots for HEK293 ATAC-seq datasets. Complexity (left) and saturation plots (right) for biological replicate 1 technical replicate 1 (B1-T1). Library complexity is shown at high and low library depth levels, in red and blue, respectively. (B1-T2, B2-T1 and B2-T2 follow similar trends, not shown.)

4.1.3. Sequence bias of ATAC-seq deviates from that of DNase-seq

A multitude of studies have explored the efficacy of transcription factor footprinting with DNase-seq. These studies have demonstrated that the DNase I enzyme cleaves genomic DNA in a non-random fashion, where it has different cut propensities for different sequences, and this sequence bias has adverse effects on the quality of footprinting when left uncorrected[115]. Our lab has previously published a site-centric computational footprinting tool where 6-mer DNase bias has been incorporated into the model to estimate the bias background in a multinomial mixture framework[114]. In order to gain insights into the sequence bias of ATAC-seq data, we calculated the 6-mer cleavage propensities of the Tn5 transposase, using available data from libraries generated by Tn5 transposition on deproteinized genomic DNA[74] (supplementary table 2). Comparison of the cleavage propensities in libraries generated using human genomic DNA vs. *D.melanogaster* genomic DNA, revealed very similar results (figure 17A, Pearson correlation 0.94), indicating that the Tn5 transposase has specific sequence preferences which are consistent in data from the two species. The dynamic range of this bias is on the same order of magnitude as for DNase bias[114]. We next asked how the sequence preferences of the Tn5 transposase compare to those of DNase I. Using values inferred previously from a single-hit DNase-seq experiment of deproteinized K562 cells[114], we observed this correlation to be fairly low (figure 17B, Pearson correlation 0.30). This indicated that these enzymes have largely distinct sequence biases.

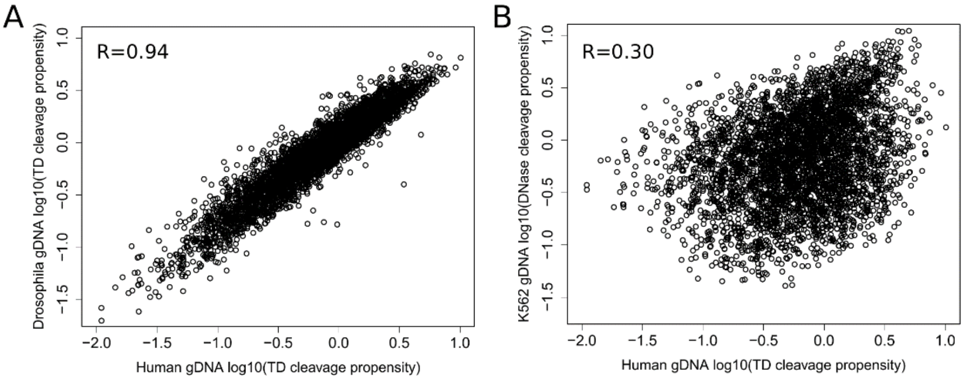


Figure 17: The sequence bias of the Tn5 transposase is distinct from that of DNase I. (A) Comparison of Tn5 transposition propensities of all 6-mers (log10 scale) in two libraries generated using deproteinized genomic DNA from human (YH1) and D.melanogaster. (B) 6-mer transposition propensities in the human library compared to cleavage propensities of DNase inferred previously from a single-hit DNase-seq experiment using deproteinized genomic DNA from K562 cells.

4.1.4. ATAC-seq and DNase-seq generate different footprint shapes for the same factor

In order to systematically examine how ATAC-seq compares to the more established DNase-seq method in transcription factor footprinting, we first focused on CCCTC binding factor (CTCF), a factor with a well-known, high information content binding site with substantial available ChIP-seq data including in HEK293 cells (supplementary table 4). We scanned the human genome for matches to the CTCF binding model obtained from the JASPAR database (supplementary table 5). As aggregate signal across all candidate CTCF motif matches is expected to be a mixture of footprint (bound sites) and background (unbound sites), our method[114] was applied to infer the bound subset by modeling the shapes of the CTCF footprints in the DNase-seq and combined ATAC-seq replicates. Shape of the aggregate signal at sites that overlap CTCF ChIP-seq peaks was used to initialize the footprint model. The background was modeled using protocol-specific bias values. The resulting footprint and background profiles revealed marked differences between ATAC-seq and DNase-seq (figure 18A, left and right, respectively). Most notable was a wider region of protection in the ATAC-seq data, in line with a previous study[74] which reported that the Tn5 transposase dimer needs circa 30 nucleotides to bind DNA and that cleavage occurs in the central 9 nucleotides. Another difference concerned the background profiles, attributable to the distinct sequence preferences of these two enzymes. In short, from the same set of CTCF motif matches, different footprint and background models were learned using ATAC-seq and DNase-seq datasets.

4.1.5. Footprinting using ATAC-seq and DNase-seq uncovers common bound sites

This observation led to the question whether the same sites would be identified as bound by a transcription factor when using ATAC-seq and DNase-seq in the same cell type. Using the protocol-specific footprint and background models learned for CTCF, we calculated the footprint scores for all considered motif matches, as the log-odds of footprint versus background per site (footprint log-likelihood ratio, FLR, see methods). The FLR is thus derived in a protocol-specific manner, solely from the single-nucleotide resolution signal around motif sites, without relying on additional features, and it accounts for sequence bias, making it an ideal metric to compare the footprints from the two protocols. As a positive FLR indicates a higher probability of being bound vs. unbound, we selected the motif matches that had a positive FLR in both replicates of the assayed method. We again used IDR to find the reproducible subset of CTCF footprints among these sites, ranked by FLR (FLR-IDR, see methods). Following this methodology for the combined ATAC-seq replicates, 12,651 motif sites had positive FLRs in both replicates, of which 8,298 were found to be reproducible by FLR-IDR (figure 19A). For the DNase-seq replicates, of the 13,592 sites with positive FLRs, 8,480 were reproducible. Nearly all of the reproducible footprints of ATAC-seq and DNase-seq overlapped CTCF ChIP-seq peaks (98% and 96% respectively, figure 19A). Furthermore, these reproducible footprints from the two experimental protocols were also concordant, with 6,170 sites (74%) overlapping (figure 18B). This analysis of ATAC-seq and DNase-seq data thus identified many common sites as bound, despite the difference in footprint shapes.

We next investigated the individual contributions of bias modeling and replicates to this increased concordance and accuracy. The contribution of the replicates comes from the application of IDR as mentioned above, which creates a systematic way to find relevant cutoffs for the footprint score. To elucidate the contribution of bias, we first trained CTCF footprint models in the combined ATAC-seq and DNase-seq replicates, as outlined above, but using a uniform background, which is equivalent to no bias correction (see methods). We then compared the sensitivity-specificity trade-off between the bias corrected and uncorrected models, for both DNase-seq and ATAC-seq (figure 19B; IDR thresholds agreed well with observed specificity). Bias correction increased the sensitivity of only DNase-seq, and the specificity was not affected for either method. Moreover, correcting bias in DNase-seq had a greater impact than correcting bias in ATAC-seq on the CTCF footprint score correlations between the two experimental methods (figure 19C). To investigate this further, we trained footprint models with and without bias correction for three additional transcription factors

(MAZ, REST and YY1) with available ChIP-seq data in HEK293 cells. We compared the model performances using area under the precision-recall curve for both ATAC-seq and DNase-seq (figure 19D). This again revealed a larger impact of bias correction on model performance for DNase-seq compared to ATAC-seq, including a rare case in which correction leads to decreased performance. This observation may result from the factor not leaving a footprint due to a short residence time on chromatin and thus true bound sites showing signals that resemble the bias background. In any case, DNase-seq bias correction had a more pronounced effect on TF footprinting than ATAC-seq bias correction.

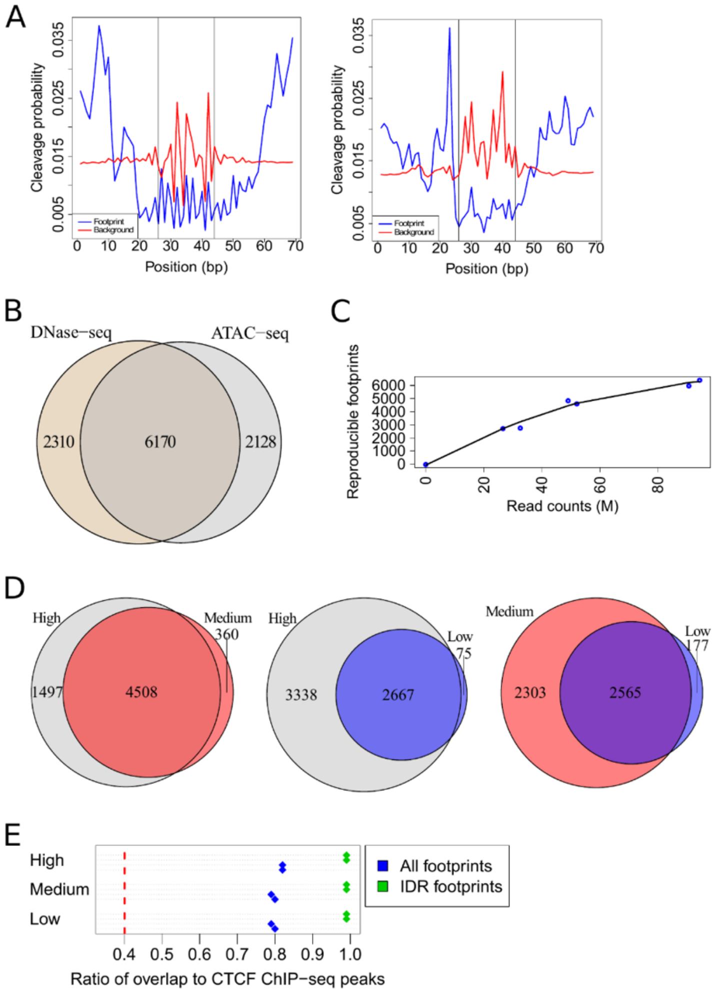
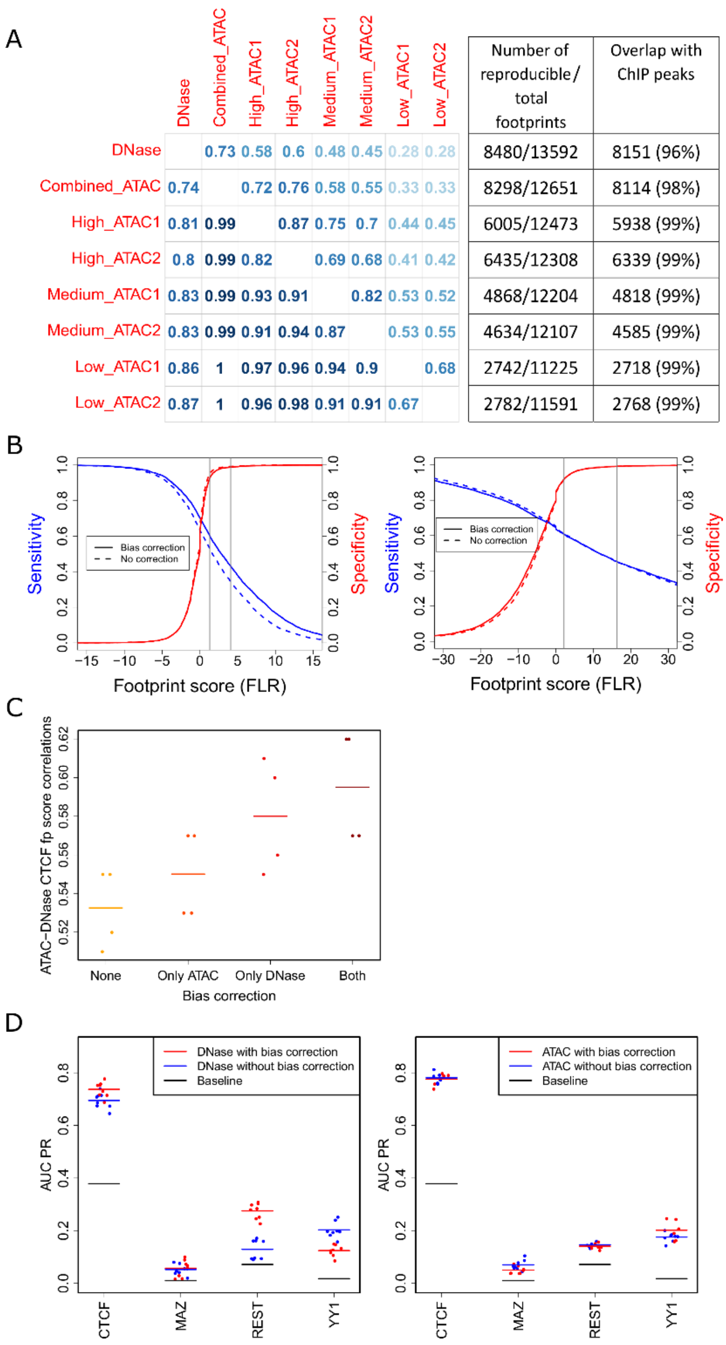


Figure 18: The number of reproducible footprints scales with library depth. (A) CTCF footprints inferred from HEK293 ATAC-seq data (left) and DNase-seq data (right). Vertical lines depict the edges of the motif match. (B) Overlap between reproducible CTCF footprints in the HEK293 DNase-seq and combined ATAC-seq replicates, found using the FLR-IDR strategy. (C) Numbers of reproducible CTCF footprints in HEK293 ATAC-seq datasets at different depths. (D) The overlaps between one set of footprints in (C) shown for high vs. medium (left), high vs. low (middle) and medium vs. low sets (right). (E) The ratio of reproducible CTCF footprints (IDR footprints) or all CTCF motif regions with positive footprint scores (all footprints) that overlap CTCF ChIP-seq peaks, in all six individual sets at different depths (supplementary table 3). Red dashed line indicates this ratio for all considered CTCF motif sites.



	DNase	Combined_ATAC	High_ATAC1	High_ATAC2	Medium_ATAC1	Medium_ATAC2	Low_ATAC1	Low_ATAC2	Number of reproducible/ total footprints	Overlap with ChIP peaks
DNase		0.73	0.58	0.6	0.48	0.45	0.28	0.28	8480/13592	8151 (96%)
Combined_ATAC	0.74		0.72	0.76	0.58	0.55	0.33	0.33	8298/12651	8114 (98%)
High_ATAC1	0.81	0.99		0.87	0.75	0.7	0.44	0.45	6005/12473	5938 (99%)
High_ATAC2	0.8	0.99	0.82		0.69	0.68	0.41	0.42	6435/12308	6339 (99%)
Medium_ATAC1	0.83	0.99	0.93	0.91		0.82	0.53	0.52	4868/12204	4818 (99%)
Medium_ATAC2	0.83	0.99	0.91	0.94	0.87		0.53	0.55	4634/12107	4585 (99%)
Low_ATAC1	0.86	1	0.97	0.96	0.94	0.9		0.68	2742/11225	2718 (99%)
Low_ATAC2	0.87	1	0.96	0.98	0.91	0.91	0.67		2782/11591	2768 (99%)

Figure 19: Performance of footprint models trained in HEK293 DNase-seq and ATAC-seq datasets. (A) Overlaps between all reproducible FLR-IDR CTCF footprints found in HEK293 DNase-seq and ATAC-seq datasets. The number in each cell represents the ratio of the footprints in the row-dataset that overlap the footprints of the column-dataset. Numbers of footprints and their overlaps with ChIP-seq peaks are given on the right. (B) The relationship between sensitivity and specificity measures of CTCF footprint models found in HEK293 DNase-seq (left) and ATAC-seq (right) datasets with and without bias correction. The vertical lines show the footprint scores that correspond to relaxed and stringent IDR thresholds, 0.1 and 0.01 respectively. (C) Correlations of CTCF footprint scores between HEK293 ATAC-seq and DNase-seq datasets with respect to their bias correction status. (D) Area under the precision-recall curve of footprint models learned for four factors (CTCF, MAZ, REST, YY1) in HEK293 ATAC-seq and DNase-seq datasets.

4.1.6. Number of reproducible footprints scales with library depth

Previous studies that inferred cell-type specific TF binding site annotations from DNase footprint data typically used very large datasets (with hundreds of millions of reads per cell type)[109,116]. To investigate the feasibility of footprinting at lower library depths, we next conducted the analysis on the 11 individual ATAC-seq libraries. We used the same setup for pairwise comparisons as for peak calling (supplementary table 3), this time to find reproducible CTCF footprints at different library depths. Even though the numbers of motif matches that had positive footprint scores were in the same range for all analyzed pairs, the numbers of reproducible footprints gradually declined with decreasing depth (figure 18C, figure 19A). This indicated that, unlike peak calling, footprinting efficiency did not saturate and rather followed the library complexities at these depths (figure 16). However, the footprints at distinct depths had substantial overlaps with each other and also constituted almost perfect subsets of the footprints found in the combined ATAC-seq data (figure 18D, figure 19A). Moreover, these reproducible footprint sets consistently showed 99% overlap with CTCF ChIP-seq peaks, compared to around 80% when considering all motif sites with positive FLRs (figure 18E). Taken together, even though deeper sequencing is beneficial to footprinting coverage, the assessment of reproducibility enables finding smaller but equally reliable sets of footprints at lower depths.

4.1.7. Properties of footprinting apply to larger sets of transcription factors

To elucidate whether the previous observations would also apply more generally beyond CTCF, we conducted the footprinting analysis on other factors. The limited availability of ChIP-seq data in HEK293 cells motivated an experimental setup to learn the footprint shapes in K562 cells, where ChIP-seq data is more abundant (supplementary table 4), and use these models to find footprints in HEK293 cells. To this end, all ATAC-seq data in K562 cells was merged to get adequate depth (supplementary table 1) and among the K562 DNase-seq datasets, the second ENCODE replicate was chosen (supplementary table 2). As proof of principle, we first confirmed that the CTCF footprint shapes were almost identical to those learned from HEK293 data (figure 20A). We then learned footprint models from K562 data for 19 additional transcription factors with available ChIP-seq data (supplementary tables 4 and 5). For a subset of these factors, namely NRF1, CREB1 and USF1, the footprint shapes reflected the expected protection pattern in both ATAC-seq and DNase-seq data; in line with the previous observations from CTCF motif regions, the ATAC-seq footprints displayed a wider region of protection compared to the DNase-seq footprints (shown for NRF1 in figure 21A). The footprint scores (FLR) for these three factors and CTCF were in close correspondence with the associated ChIP-seq signal values in K562 cells, conferring further confidence in these footprint models (figure 20B-E). Thus, we used these models to identify bound sites reproducibly with the FLR-IDR strategy in HEK293 cells. As for CTCF, reproducible footprints were found to be concordant between DNase-seq and combined ATAC-seq replicates; at the level of individual HEK293 ATAC-seq datasets, library depth and the numbers of reproducible footprints showed again a strong dependency (shown for NRF1 in figure 21B and C, respectively). As the observations could be replicated for multiple factors, these results likely provide insights into the general properties of the footprints.

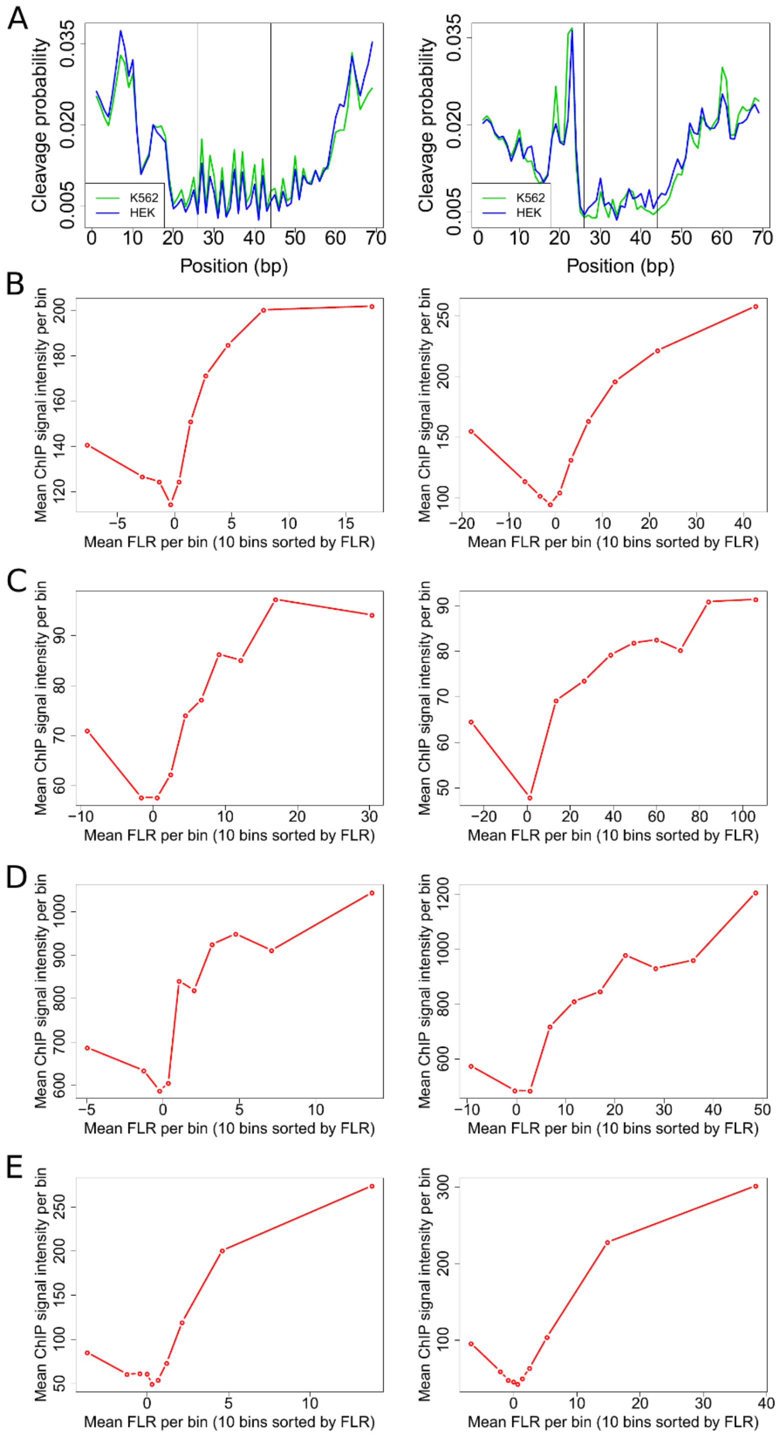


Figure 20: The relevance of the learned footprint models. (A) Identical CTCF footprint profiles in HEK293 and K562 ATAC-seq (left) and DNase-seq (right) datasets. (B-E) Concordance between ChIP-seq signal intensities and footprint scores (FLR) in K562 ATAC-seq (left) and DNase-seq (right) data for (B) CTCF, (C) NRF1, (D) CREB1 and (E) USF1. Motif sites that overlap ChIP-seq peaks are divided in ten bins according to FLR. The mean ChIP-seq signal intensity and FLR is plotted for each bin.

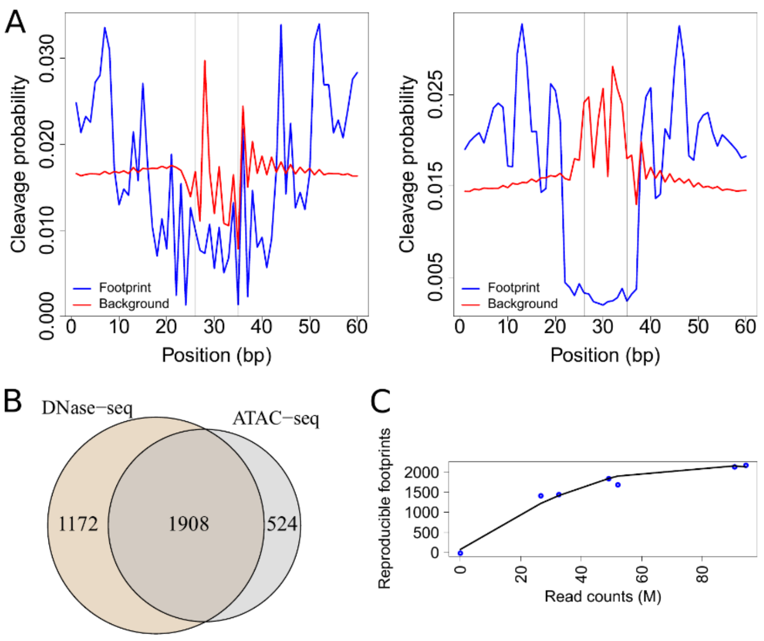


Figure 21: Analysis of NRF1 footprints. (A) NRF1 footprints inferred from K562 ATAC-seq data (left) and DNase-seq data (right). Vertical lines depict the edges of the motif match. (B) Overlap between reproducible NRF1 footprints in the HEK293 DNase-seq and combined ATAC-seq replicates, found using the footprint models learned from the K562 data. (C) Numbers of reproducible NRF1 footprints in HEK293 ATAC-seq datasets at different depths.

4.1.8. Protocol-specific sequence biases influence footprinting efficiency

Strong footprints that were concordant in both ATAC-seq and DNase-seq data were only found for four of the 20 assayed factors. For most factors, clear footprints were observed in one of the experimental methods, but not the other. Therefore, we asked whether the distinct sequence biases of the two methods play a role in the factor-dependent performance of footprinting. To get a continuous measure for performance (as opposed to the discrete visual assessment of footprint shapes), for all TFs in both experimental settings, we calculated the area under the receiver operating characteristic curve (AUC), ranking candidate sites by FLR and considering those that overlap ChIP-seq peaks to be true binding sites. In order to assess how performance is linked to the relationship between the footprint and background models, the Pearson correlations between these two models (eg. footprint-background model similarities) for each TF were calculated and compared to the AUCs. The AUCs negatively correlated with the footprint-background model similarities in both ATAC-seq and DNase-seq datasets (figure 22A and B, correlations of -0.36 and -0.6, respectively), indicating that when a footprint model is clearly distinguished from the background, it is more likely to explain transcription factor binding accurately. Moreover, the differences per TF between ATAC-seq and DNase-seq datasets for these two measures (AUCs and footprint-background model similarities), also had a negative correlation (-0.53, figure 22C), suggesting that the experimental protocol which achieves better separation between the footprint and background components, is also performing better for a given TF. Overall, DNase-seq footprinting had a clear advantage over ATAC-seq derived footprints (cf. figure 23A, which compares the area under the precision-recall curve values).

As the background component is derived directly from sequence bias and given our previous observation that DNase-seq bias correction shows a stronger positive effect compared to ATAC-seq bias correction, we once again explored the role of bias more explicitly. In particular, two of three factors for which ATAC-seq outperformed DNase-seq, MEF2A and STAT1, had the lowest DNase I cleavage propensities (eg. sequence bias) over their motif regions, among all assayed factors (figure 22D), whereas the Tn5 transposition propensities for these factors were average (figure 23B). Therefore, the background models learned from DNase bias for these factors had footprint-like shapes, impeding the clear separation between the two components, and thus explaining the poor performance of DNase-seq (shown for MEF2A in figure 23C). The equivalent scenario was not as clear to observe for ATAC-seq, possibly due to the difference in the efficiency of bias modeling, see discussion. In summary, due to the distinct sequence biases of ATAC-seq and DNase-seq, the sequence content of transcription factor binding sites can influence footprinting efficiency in a protocol-specific manner.

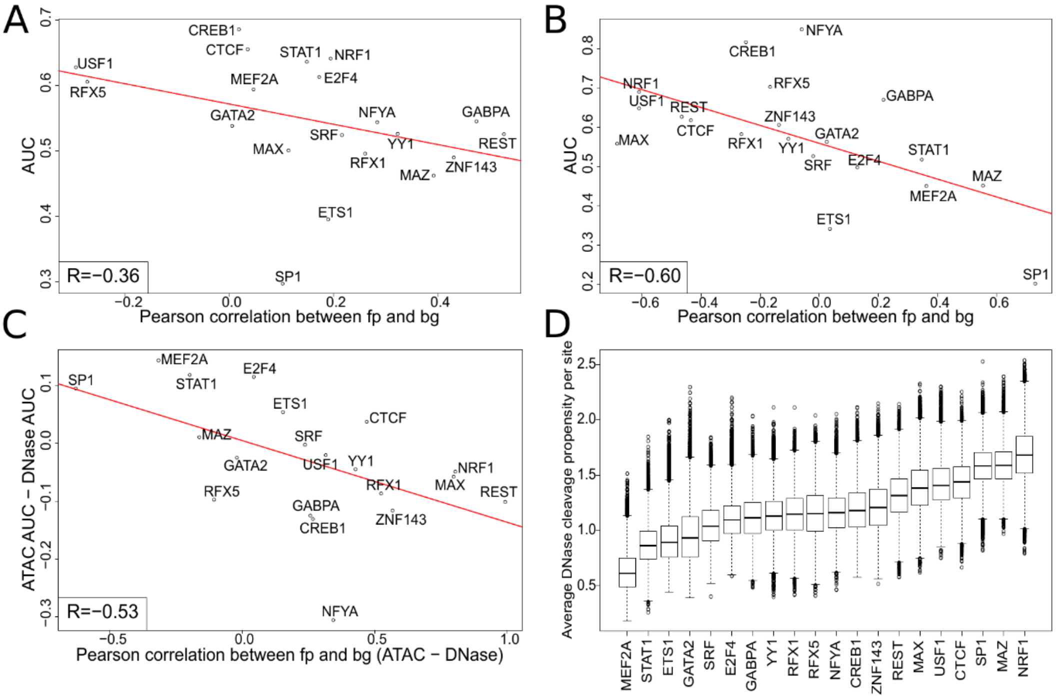


Figure 22: TF-footprinting accuracy is linked to clear discrimination of footprint from background. (A,B) AUCs vs footprint-background model similarities in (A) ATAC-seq data and (B) DNase-seq data. (C) Difference in AUCs (ATAC-DNase) vs difference in footprint-background model similarities (ATAC-DNase). (D) Average DNase I cleavage propensities over candidate TFBSs for all 20 assayed factors.

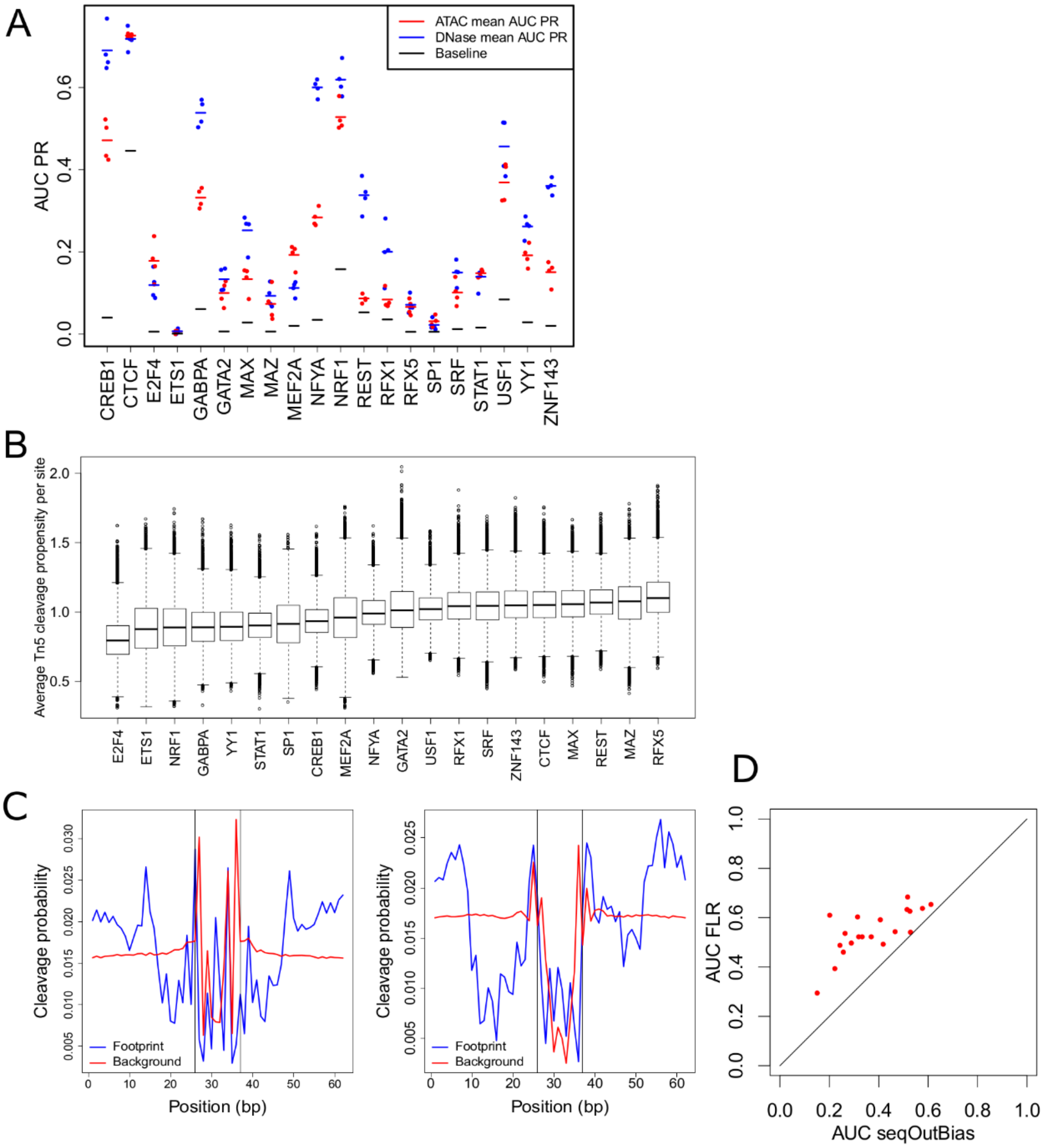


Figure 23: Method and TF-specific footprinting efficiency. (A) Area under the precision-recall curve of footprint models learned for all 20 assayed factors in K562 ATAC-seq and DNase-seq datasets. (B) Average Tn5 cleavage propensities over candidate TFBSs for all 20 assayed factors. (C) MEF2A footprints inferred from K562 ATAC-seq data (left) and DNase-seq data (right). Vertical lines depict the edges of the motif match. (D) Comparison of AUCs (area under the ROC curve) obtained with our method (FLR) vs the seqOutBias method.

4.2. Part 2: Characterization of tissue-specific cis-regulatory elements in Drosophila embryonic development

Contribution Statement:

Aslihan Karabacak Calviello performed the motif enrichment, GKM-SVM, TF footprinting, and integrative TF binding model analyses supervised by Uwe Ohler. Embryo ATAC-seq datasets were generated by Alexander Glahs and Aslihan Karabacak Calviello. Embryo DNase-seq datasets were generated and preprocessed (including peak calling and defining tissue and time-point specific DHSs) by Dr. James Reddington and Dr. David Garfield.

4.2.1. Nuclear sorting coupled to ATAC-seq or DNase-seq captures tissue-specific cis-regulatory elements

In this and the subsequent results sections, we utilize open chromatin profiling techniques to gain insights into the embryogenesis of the fruit fly *Drosophila melanogaster*. To this end, the BiTS-ChIP strategy (see section 2.3.2.2) is combined with either DNase-seq or ATAC-seq. Briefly, embryos in a chosen developmental stage are collected, formaldehyde fixed and subjected to nuclear extraction. Nuclear markers exclusively or predominantly expressed in the tissue of interest, are then used to fluorescently sort the nuclei to get tissue-specific subsets from the whole embryo (see methods for details). Therefore, the application of DNase-seq or ATAC-seq on these subsets enables assaying chromatin accessibility genome-wide in a tissue and time-point specific manner and elucidating the cis-regulatory networks in play during embryogenesis.

We profiled the open chromatin landscape of the developing embryo using this type of data, via two collaborative projects. The first one was in collaboration with Prof. Eileen Furlong's laboratory (specifically Dr. James Reddington and Dr. David Garfield) at EMBL. Here, embryos from five consecutive time-points (2-4hr, 4-6hr, 6-8hr, 8-10hr and 10-12hr) were collected. These span a wide spectrum of developmental stages, ranging from the blastoderm to terminally differentiated cells and collectively represent roughly half of the total embryogenesis time, which takes around 24hr. For all time-points, DNase-seq experiments were conducted on unsorted nuclei, representing the whole embryo at that stage. For the four time-points spanning 4-12hr, nuclei were sorted to get subsets specific to the mesodermal (shortly "meso") and neural (shortly "neuro") lineages. The myogenic TF, dMef2, was the mesodermal nuclear marker for all time-points, whereas TF Worniu and RNA-binding protein

Elav were the neural nuclear markers for stages 4-6hr, and from 6hr on, respectively. Nuclei that had neither the mesodermal nor the neural markers were also obtained from the sorts and are henceforth referred to as "other". Finally, only for the 6-8hr time-point, visceral and non-visceral mesoderm subsets were obtained, using the visceral mesoderm specific TF Biniou as the nuclear marker. These datasets are referred to as "binpos" and "binneg" in the rest of this thesis. Figure 24 shows normalized DNase-seq signal profiles of all datasets around an intronic region of the Mef2 gene locus, which is known to harbor enhancers with mesodermal activity. As expected, all mesodermal datasets (meso, binpos and binneg) have elevated signals in this region. Similar observations are made for other loci as well, e.g. neuro datasets show elevated signals at regions related to neural development (not shown). Taken together, this illustrates the power of combining the BiTS strategy with open chromatin profiling in characterizing tissue-specific regulatory elements that would otherwise not be detected when the whole embryo is profiled in bulk.

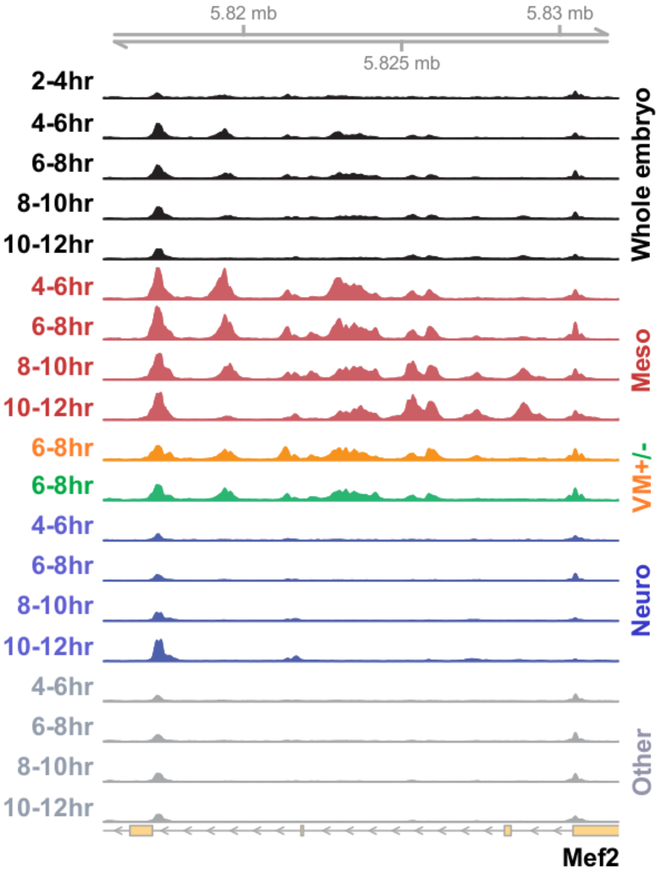


Figure 24: Tissue- and time-point-specific DNase-seq signal profiles around Mef2 gene locus. Depth normalized signal profiles from all embryo DNase-seq datasets are shown at the Mef2 gene locus. Significantly higher signals can be observed in the mesoderm-specific datasets (Meso and VM) (adapted from the figure kindly provided by J. Reddington).

The second project was in collaboration with Dr. Robert Zinzen's laboratory (specifically Alexander Glahs) at the MDC. In this case, we wanted to assay two specific subpopulations of the neural lineage: the intermediate and ventral columns of the developing neuroectoderm. Embryos from 4-6hr, 6-8hr and 8-10hr time-points were collected and for each time-point, TFs Ind and Vnd were used as nuclear markers for the intermediate and ventral column, respectively. Unsorted nuclei representing the whole embryo at these stages were also collected. In this project, we employed ATAC-seq to assay open chromatin regions. Figure 25 shows normalized ATAC-seq signal profiles of all datasets around the ind gene locus. As expected, ind-sorted datasets show elevated signals (especially at 4-6hr and 6-8hr) compared to vnd-sorted and unsorted datasets at this locus. This once again confirms the utility of our tissue-specific sorting strategy and demonstrates that either DNase-seq or ATAC-seq can be used in conjunction with it to assay tissue-specific cis-regulatory elements.

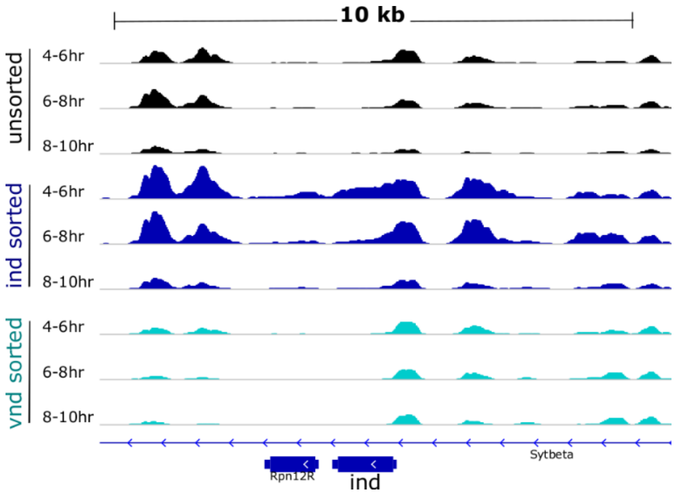


Figure 25: Tissue- and time-point-specific ATAC-seq signal profiles around ind gene locus. Depth normalized signal profiles from all embryo ATAC-seq datasets are shown at the ind gene locus. Significantly higher signals can be observed in the ind-sorted datasets.

4.2.2. Cis-regulatory elements open in the same tissue-specific context share common sequence signatures

Having established the potential of our strategy in finding tissue-specific cis-regulatory elements, we next set out to characterize the sequence features that define tissue-specificity in our DNase-seq datasets. We focused our attention to TSS-distal regions to assay known or putative enhancers. To this end, the total set of distal DNase hypersensitive sites (DHSs) were found via peak calling, and tissue-specific subsets were identified using differential signal analysis (see methods). Briefly, a DHS was deemed tissue-specific, if it was only found in a given tissue, and if it had significantly higher signal in that tissue compared to the other tissues at the same time-point. For example, 4-6hr meso specific DHSs were defined by contrast to 4-6hr neuro and 4-6hr other datasets. We first asked whether we could identify sets of TFs that were crucial for a given tissue. Therefore, we conducted motif enrichment analyses on the tissue-specific DHSs, using all fly-specific PWM databases (fly factor survey, dmmpmm, idmmpmm, OnTheFly and flyreg) from the MEME suite, combined with a custom list of fly PWMs from the Furlong laboratory, denoted "custom PWMs" henceforth (see methods), for a total of 1677 redundant PWMs. For each tissue-specific set, we defined a set of background regions as GC-content, length and TSS-distance matched subsets from all distal DHSs; and assessed the enrichment of all 1677 PWMs over background regions using AME from the MEME suite (see methods). Enriched PWMs were then clustered according to similarity to eliminate redundancy. Figure 26 (left) shows the resulting motif enrichment heatmap for each cluster and tissue-specific DHS set. For the annotations pertaining to each cluster (e.g. most enriched PWMs within the cluster and a summary motif representing the whole cluster), see appendix B: supplementary table 6. The heatmap clearly demonstrates that the motifs are enriched in a tissue-specific way, as the resulting clusters can generally be linked to a single tissue. Several TFs that are known to be important for the mesodermal lineage are enriched specifically in the meso datasets as expected: Mef2 (cluster 34), Tin (cluster 14), Bin (cluster 10) and Twi (cluster 3), indicating that the identified enrichments represent meaningful associations between TFs and tissues. In the neuro datasets, the motif for Ttk (clusters 12 and 13) is strongly enriched. Despite our clustering strategy to eliminate redundancy, some resulting clusters still represent similar motifs. For example, clusters 26, 38 and 45, which are all significantly enriched in the neuro datasets, might represent the same TF, likely belonging to the Sp1/Klf family. The neuro-specific enrichment of the G-rich clusters 6 and 27, likely represent true signal rather than an artifact, as our background sets were matched for GC-content. Finally, a specific E-box motif (cluster 1) and GATA (cluster 8) are enriched in the other datasets alongside an unknown cluster (cluster 64). In some of these neuro- and other-specific clusters, further efforts are needed to identify the true associated TFs.

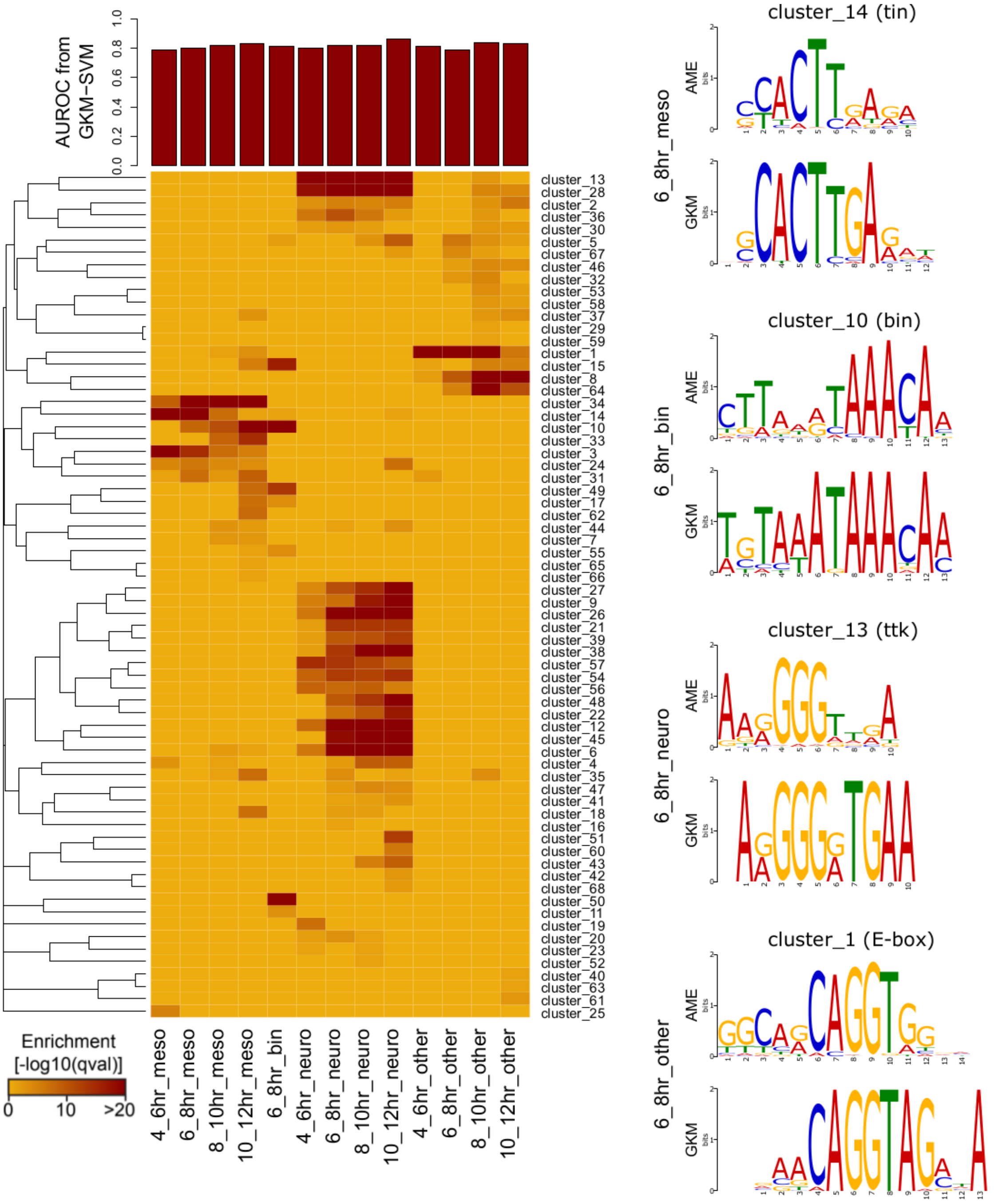


Figure 26: Motif enrichment results in tissue-specific DHSs. Heatmap shows the enriched motif clusters for the tissue-specific DHS sets, with the corresponding GKM-SVM results above (left). Alignments between enriched clusters and GKM-SVM derived PWMs are shown for clusters 14 (tin), 10 (bin), 13 (ttk) and 1 (E-box), respectively (right).

Having identified tissue-specific signatures using known motifs, we next asked whether sequence features that discriminate the specific sets from the background sets could be found without prior knowledge. We employed a computational method that combines gapped k-mer features with support vector machines (GKM-SVM, see methods), using 10-mers (short sequences of length 10) as features and the same set of background regions as for the motif enrichment analyses. This method successfully learns the 10-mer features that discriminate the specific sets from the background sets, as indicated by area under the receiver operating characteristic curve (AUROC) values around 0.8 for all datasets (figure 26, above heatmap). Therefore, we next clustered the top 300 10-mer features associated with each tissue-specific dataset, using the same clustering strategy as for the enriched motifs, in order to group similar 10-mers together to generate PWM-like profiles. We noticed that many of the GKM-derived PWMs were concordant with the AME clusters, as exemplified for clusters 14 (tin) for meso, 10 (bin) for binpos, 13 (ttk) for neuro and 1 (E-box) for other datasets, shown in figure 26, right. We then aligned the GKM-derived PWMs in each tissue-specific set to all 68 AME clusters, to identify the subset of AME clusters represented. These cluster subsets and the clusters known to be enriched for each tissue-specific set, showed significant overlaps for 10 of the 13 sets (figure 27), indicating that this method identifies meaningful sequence signatures, supported by known binding models for transcription factors. Finally, an interesting potential of this method is to learn novel sequence features, as it is not biased by known features such as PWMs. In line with this, some GKM-derived PWMs aligned to AME clusters that were not enriched in that tissue (figure 27), and others did not align to any of the AME clusters at all (not shown). A more detailed exploration of these sequence features might provide additional insights on achieving tissue-specificity, however we have not explored this further.

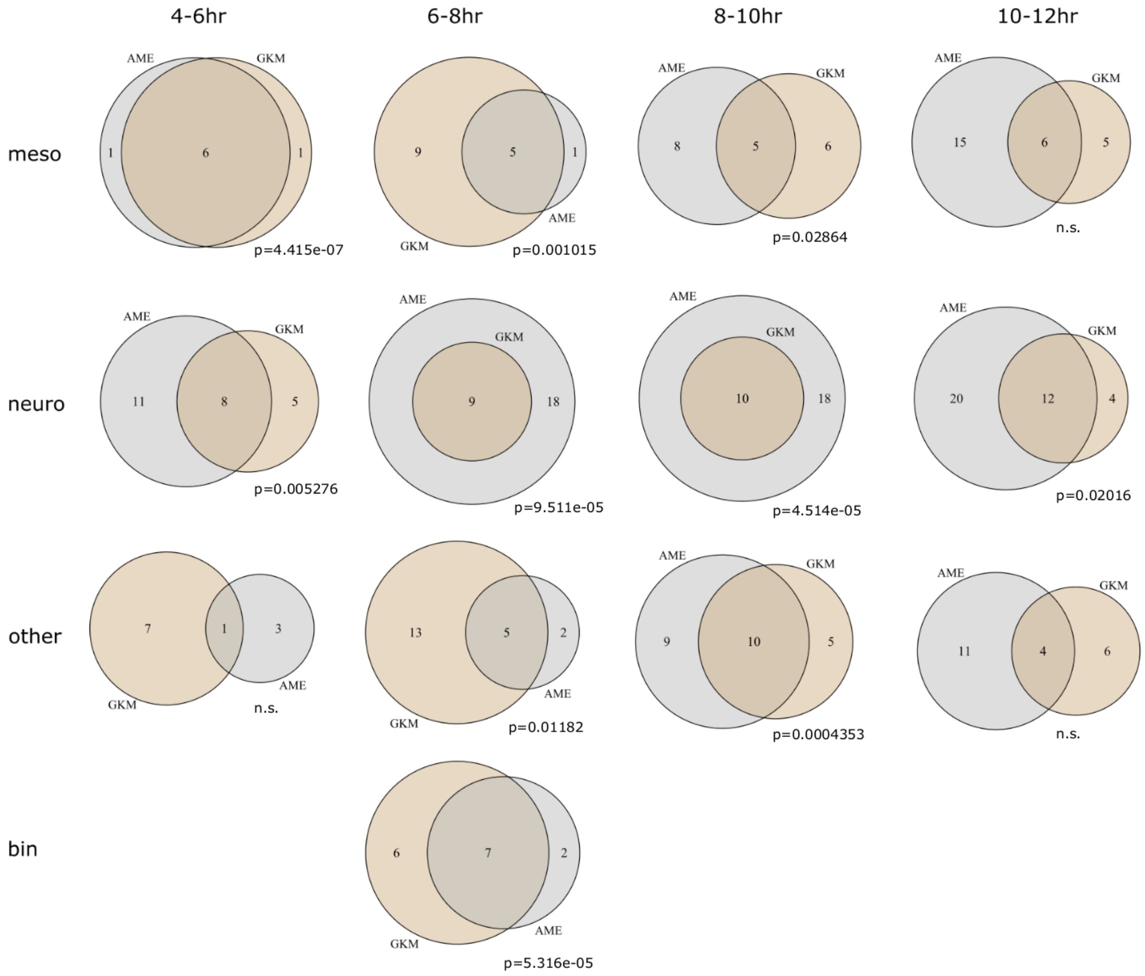


Figure 27: Overlaps between enriched motifs and GKM-SVM-derived PWMs. Venn Diagrams showing the agreement between enriched motif clusters and GKM-SVM derived PWMs for each tissue-specific DHS set. p-values are derived from Fisher's exact test (two-sided).

4.2.3. Contrasting regions open in the same tissue at consecutive time-points uncovers temporally resolved sequence features

In the previous section, we focused on sequence features defining tissue-specificity. As our DNase-seq dataset is also temporally resolved (i.e. belonging to specific time-points of embryonic development), we next asked whether sequence features that discriminate consecutive time-points of the same tissue could be identified. To this end, we defined time-point specific DHSs, by contrasting consecutive time-points, using a strategy similar to the definition of tissue-specific DHSs. A DHS is deemed time-point specific, if it is found at that time-point and not the other, and if it has significantly higher signal at that time-point. For example, the comparison 8-10hr vs 10-12hr meso datasets results in two sets of time-point specific DHSs: the "early" ones associated with the 8-10hr dataset (using 10-12hr as the contrast set), and the "late" ones associated with the 10-12hr dataset (with 8-10hr used as the contrast set this time). In this setting, we also defined DHSs that are shared between the two time-points as those that are found in both datasets and do not exhibit significantly higher or lower signal in either dataset. In order to calculate motif enrichments in the time-point specific DHSs, the background sets had to be defined. As consecutive time-point specific DHSs could have differences in sequence content (e.g. GC content), which in turn could influence the results, we did not directly contrast these sets. Instead, we selected GC-content, length and TSS-distance matched background sets from the shared regions associated with each tissue-specific set, as a proxy to directly comparing consecutive time-points. Following the 8-10hr vs 10-12hr meso example above, the background for both early and late sets are selected from the regions shared between these two datasets. Motif enrichment analyses were conducted using this background definition, with a similar approach to the analyses in the tissue-specific DHSs (i.e. with the same set of 1677 PWMs and same strategy to cluster the enriched motifs). The resulting heatmap is in figure 28A, where every corresponding early-late pair is shown as a group. Contrasting consecutive time-points of the same tissue is a more challenging task compared to defining tissue-specific DHSs, making the interpretation of the associated heatmap more difficult. One clear observation is the enrichment of motifs representing the Forkhead TF family (to which bin belongs), specifically in the 6-8hr binpos dataset, compared to the 4-6hr meso dataset (clusters 8, 39, 12 and 21). Twi (cluster 2) is enriched in 4-6hr meso datasets compared to 6-8hr meso, 6-8hr binneg and 6-8hr binpos datasets. Furthermore, cluster 19 is significantly enriched in 2-4hr whole embryo datasets, compared to both 4-6hr meso and 4-6hr neuro datasets, which is expected as this is an E-box motif cluster, including the TF Zld, essential for early stages of embryogenesis. Another interesting, but rather less interpretable cluster is 31 (Ttk and Ab), which shows enrichment in almost all early datasets, regardless of tissue. In a similar vein, clusters 11, 23 and 32 show a general late-specific enrichment. Whereas cluster 23 represents the Sp1/Klf TF family, interestingly, the other two clusters both implicate Lola, despite having very different motifs. In line with the motif enrichment results, that generally identified a given cluster as enriched in either the early or late time-point, GKM-SVM achieved AUROC values of 0.7-0.8 for all time-point specific sets (figure 28, above heatmap), also indicating the presence of sequence features that differentiate consecutive time-points.

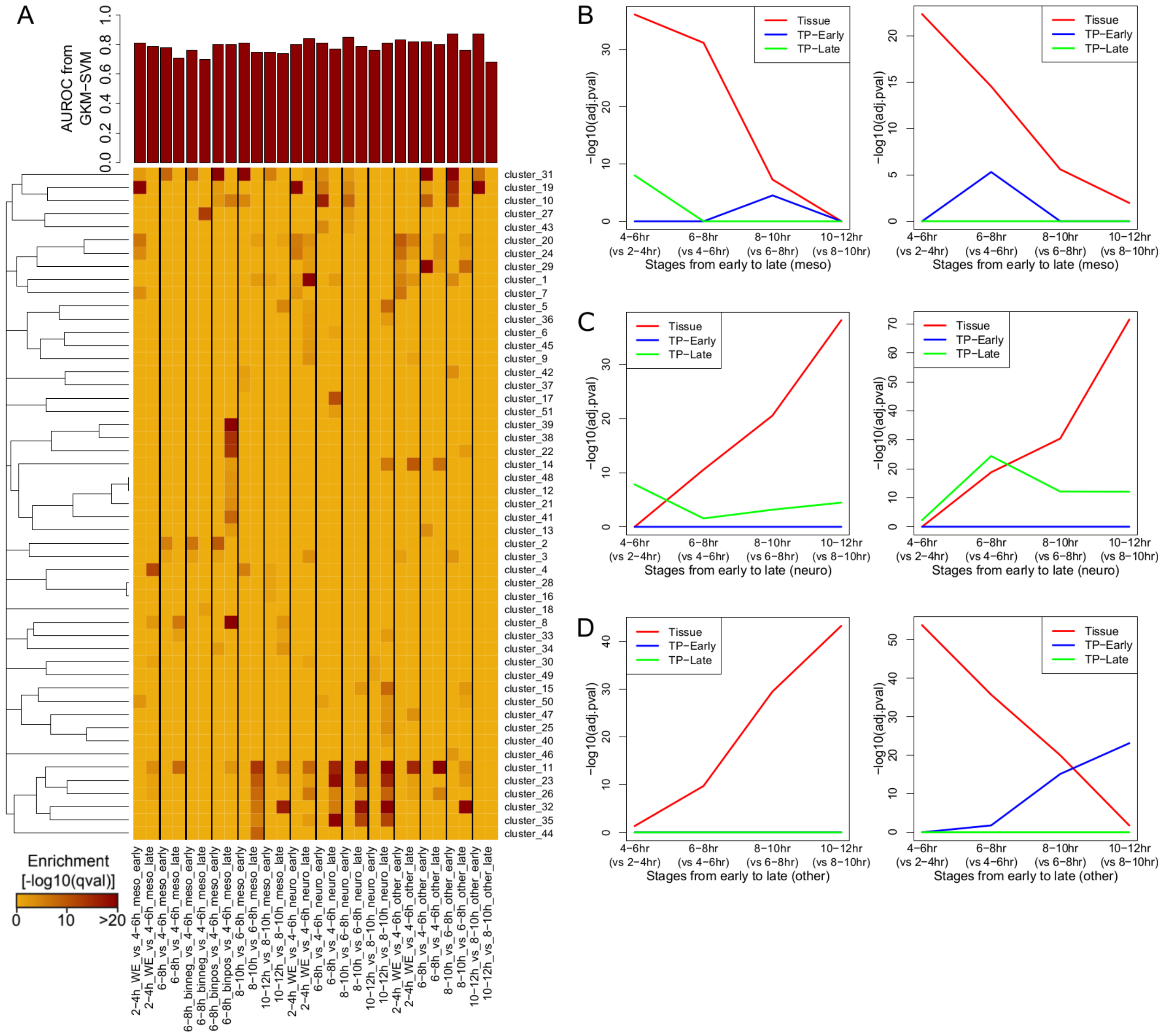


Figure 28: Motif enrichment results in time-point-specific DHSs. (A) Heatmap shows the enriched motif clusters for the time-point-specific DHS sets, with the corresponding GKM-SVM results above. (B-D) Patterns of tissue and time-point specific motif enrichment values over consecutive embryonic stages, shown for tin (FBgn0004110) and twi (berkeley_bdtnp_twi) in (B), kr (OTF0214.1) and klu (OTF0242.1) in (C) and GATA (FBgn0003507_2) and E-box (FBgn0259789) in (D), left and right, respectively.

Our motif enrichment analyses provide not only the presence, but also the significance of enrichment for a given PWM, via an adjusted p-value. Therefore, we combined the observations from the tissue and time-point specific analyses and assessed the change of enrichment significance over developmental time for multiple PWMs. Figure 28B-D shows six examples: tin and twi for meso (B), kr and klu for neuro (C) and GATA and E-box for other datasets (D, all left and right, respectively). Tin and twi both show a gradual decrease in meso-specific enrichment values over developmental time, in line with their crucial functions in specifying the mesodermal lineage early in embryogenesis. Time-point specific enrichments are generally in line with the tissue-specific enrichments; note the steep slope between 6-8hr and 8-10hr in the meso-specific enrichment values for tin, reflected in the 6-8hr vs 8-10hr early time-point enrichment. Kr and klu, both show a gradual increase of neuro-specific enrichment values over time. In the other-specific datasets, the GATA and E-box motifs show a gradual increase and decrease, respectively. Taken together, these analyses temporally resolve the motif enrichment results, with the potential to link TFs to certain stages of lineage specification.

4.2.4. Transcription factor footprinting poses challenges in this system

The utility of TF footprinting was demonstrated in the first part of the results presented in this thesis, successfully identifying bound sites for a set of TFs in human cell lines K562 and HEK293. Consequently, we next wanted to apply the same TF footprinting methods to our *Drosophila* DNase-seq data, to infer bound sites and gain further insights into the embryonic development of this complex organism. As the DNase-seq experiments were conducted on formaldehyde fixed embryos, we first wanted to explore the effects of the fixation on footprinting performance. To this end, we used further DNase-seq datasets generated in the Furlong laboratory, on fixed (crosslinked) and unfixed (native) *Drosophila* S2 cells. We focused on CTCF and Beaf32, due to the availability of ChIP-seq datasets in S2 cells for these factors (see supplementary table 4). We scanned the *Drosophila* genome using the PWMs obtained from the JASPAR database for these two factors (used PWMs are marked with an asterisk in Appendix B: supplementary table 9), getting motif matches genome-wide that represent putative transcription factor binding sites (TFBSs). We then examined the bp-resolution DNase I cut profiles around these putative TFBSs. Interestingly, the native and crosslinked S2 DNase-seq datasets displayed distinct cut profiles, shown for Beaf32 in figure 29A. This altered cut profile led to the hypothesis that formaldehyde crosslinking changed the properties of the DNase I-chromatin interaction, impacting the sequence preferences (bias) of DNase I. To test this hypothesis, we turned to a widely used alternative method to infer the bias: instead of using naked DNase-seq datasets as outlined in the previous sections of the thesis, we computed the 6-mer bias values from the DHSs of the datasets. The putative cut profiles inferred solely from these data-specific bias values, explained the observed profiles almost completely, for both native and crosslinked datasets (figure 29B and C, respectively). The cut profile inferred from naked DNase-seq derived bias showed a general resemblance to the observed profiles, albeit not explaining them completely (figure 29D). Therefore, we conducted footprinting analysis learning the background component from data-specific bias values, in order to more accurately identify footprints that differentiate from the background. This analysis performed quite differently for Beaf32 and CTCF. Whereas footprinting performance was rather high, with the native dataset outperforming the crosslinked dataset for Beaf32, performance was much lower for CTCF, with the crosslinked dataset outperforming the native dataset, as measured by area under the precision recall curve (AUC PR) values (Figure 29E).

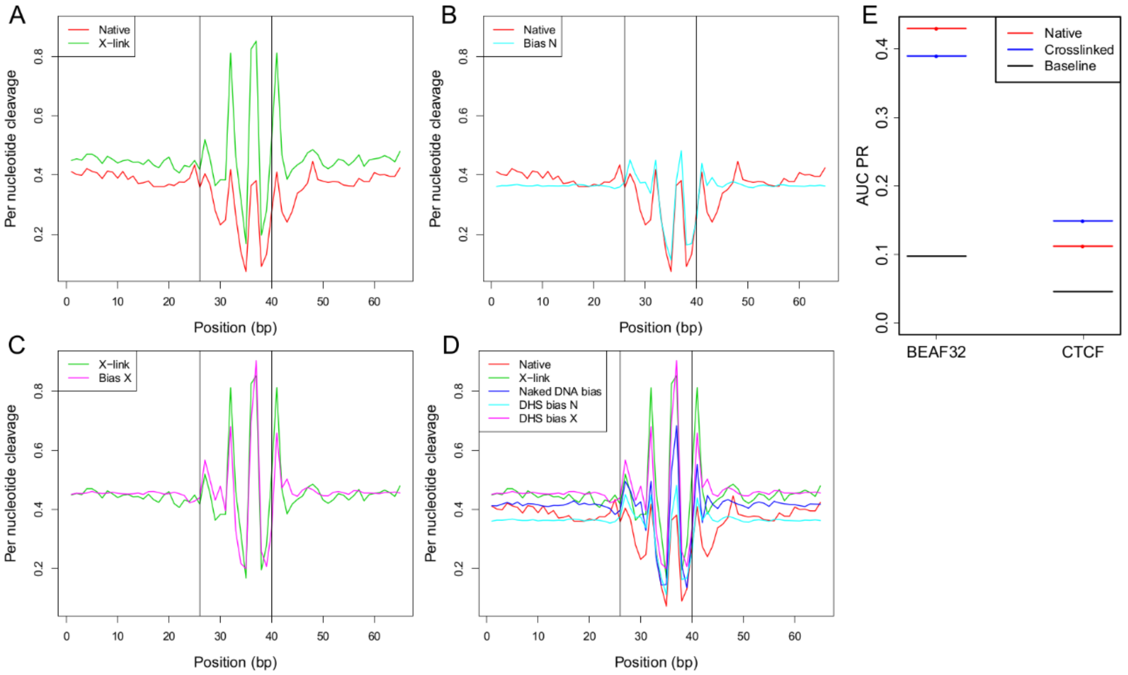


Figure 29: Differences in DNase I cut profiles and footprinting in native and crosslinked S2 DNase-seq data. (A) Aggregate DNase I cut profiles around Beaf32 motifs that overlap ChIP-seq peaks show differences among crosslinked and native datasets. (B) Aggregate cut profile in the native dataset matches the native DHS bias inferred profile. (C) Aggregate cut profile in the crosslinked dataset matches the crosslinked DHS bias inferred profile. (D) Naked bias inferred profile shows similarity to both DHS bias inferred profiles, not matching either one perfectly. (E) Footprinting performance for Beaf32 and CTCF, based on 4-fold cross validated AUC-PR (baseline denotes the fraction of positive examples in the training set).

We next returned to the embryo DNase-seq datasets to evaluate the performance of TF footprinting here. To this end, we selected 5 mesodermal TFs (bin, lmd, mef2, tin and twi) and scanned the *Drosophila* genome with their PWMs (the utilized PWMs are marked with an asterisk in Appendix B: supplementary table 9), to find genome-wide putative TFBSs. We then defined the 6-8hr meso dataset as the positive set, and the 6-8hr neuro dataset as the negative set. As explained for the S2 datasets above, data-specific bias values were also inferred in this setting, from the DHSs of embryo DNase-seq datasets (see methods). We first compared the library depth normalized DNase-seq cut profiles for the positive and negative datasets around the putative TFBSs that overlapped the ChIP-seq datasets for each TF (also see Appendix B: supplementary table 9). We also included the putative cut profiles that would result solely due to the embryo dataset-specific bias. Figure 30, panels A-E show these comparisons for bin, lmd, mef2, tin and twi, respectively. One of the main conclusions from these plots is that the shape of the DNase-seq signal around the TFBSs is very similar for the positive and negative datasets, with the bias-inferred profiles closely mirroring them, indicating that the shape of the signal might be not be very informative in identifying true bound sites (positive set). Another important conclusion is that the positive datasets show higher DNase-seq signals compared to the negative datasets for each TF, arguing that signal intensity might be more discriminatory than shape in this case. To test this further, we learned footprint models from the positive and negative datasets for each TF and compared the performance of these models to simply ranking the motif sites according to DNase-seq tag counts. Figure 31A, shows the performance (AUC PR) of the footprint models, which can discriminate the positive and negative datasets (i.e. achieve better performance in the positive datasets, as expected). However, the simple tag count ranking achieves better separation between the positive and negative datasets (Figure 31B), as well as significantly outperforming the footprint models (Figure 31C), confirming the utility of signal intensity in predicting bound sites.

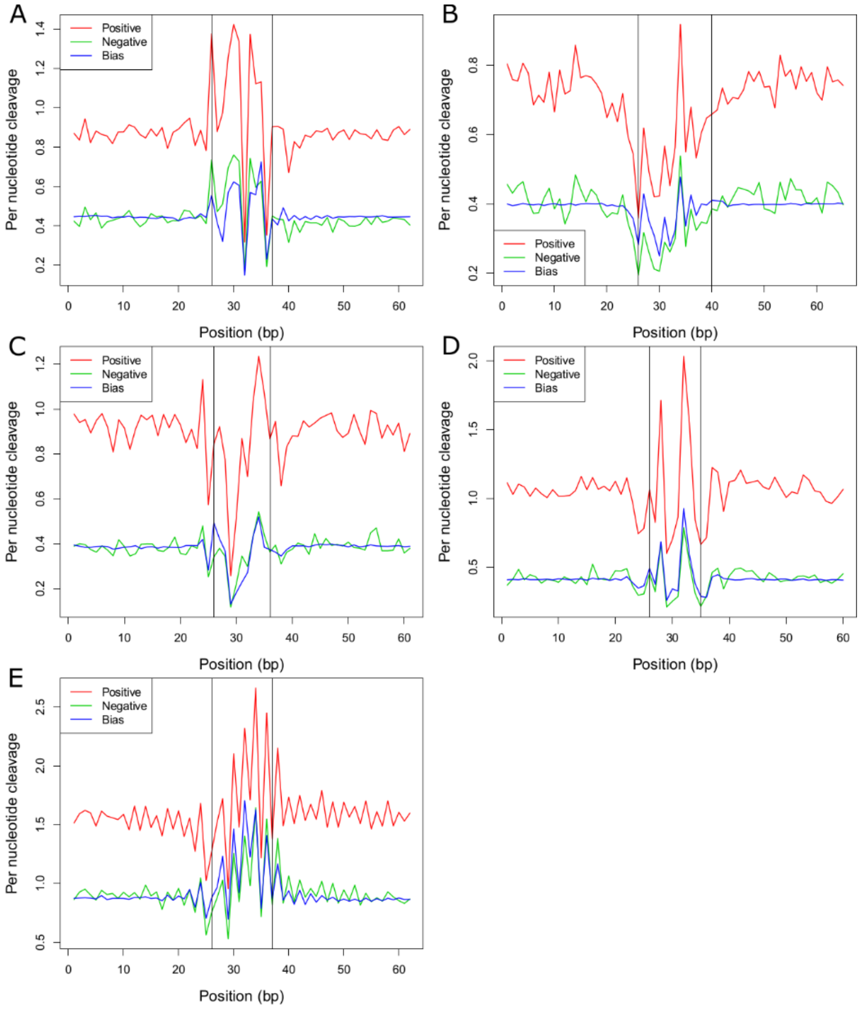


Figure 30: DNase I cut profiles around motif sites of 5 mesoderm specific TFs. Aggregate DNase I cut profiles around bin, lmd, mef2, tin and twi (A-E) motifs that overlap ChIP-seq peaks mirror each other in 6-8hr meso (positive) and 6-8hr neuro (negative) datasets, with these profiles matching the bias-inferred profiles in each case. The average signals in the positive datasets are always higher.

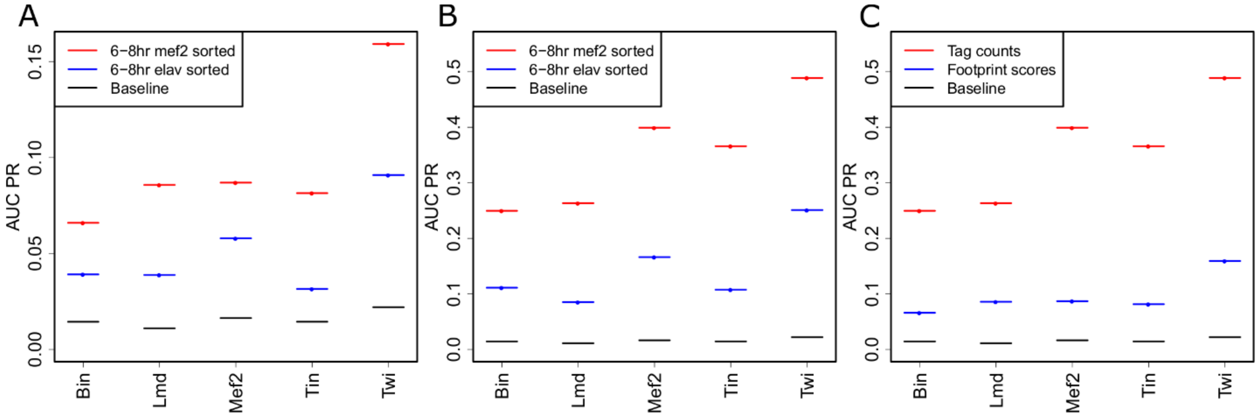


Figure 31: Performance of footprint models compared to ranking by tag counts. (A) Footprinting performance comparison for the 5 meso TFs using 6-8hr meso vs 6-8hr neuro datasets. (B) Tag count performance comparison for the 5 meso TFs using 6-8hr meso vs 6-8hr neuro datasets. (C) Direct comparison of footprinting and tag count results for the 6-8hr meso datasets (A-C, performance based on 4-fold cross validated AUC-PR, baseline denotes the fraction of positive examples in the training set).

In order to investigate whether these observations on the embryo DNase-seq datasets would also be replicated in the embryo ATAC-seq datasets, we applied a subset of the same analyses here. We defined TSS-distal tissue-specific ATAC-HSs for all ind and vnd-sorted datasets, as sites that exhibit significantly higher ATAC-seq signal in these sorted sets compared to the unsorted sets of the same time-points (see methods). For each tissue-specific set, we selected GC-content, length and TSS-distance matched background sets from all distal ATAC-HS sites. Motif enrichment analyses were conducted on the tissue specific sets, to find significant enrichment over the associated background, and the enriched motifs were subsequently clustered. The resulting motif enrichment heatmap in figure 32A, shows that the enriched clusters are generally ind or vnd-specific, with few exceptions, indicating that we identify motifs that potentially differentiate between these closely related tissues. This rich sequence content associated with the specific sets is also supported by the GKM-SVM results, which achieve AUROC values of around 0.8 in separating the tissue-specific sets from the background (figure 32A, above heatmap). As observed in the neuro-specific DNase-seq datasets previously, there are two highly enriched clusters implicating Ttk: clusters 7 and 16. Moreover, cluster 8 represents a G-rich motif, previously found to be enriched in the neuro-specific DNase-seq datasets as well, increasing the likelihood that a true signal is reflected rather than an artifact. Interestingly, this cluster is more specifically enriched in the vnd-sorted datasets, indicating the potential to link neuro-related sequence features to specific subsets of the developing neuroectoderm in this setting. Cluster 17 represents the TF Dichaete (D), highly enriched in the ind-sorted datasets in line with its functions in the developing neuroectoderm. Therefore, we next wanted to assess TF footprinting efficiency in our embryo ATAC-seq datasets, using this TF. We scanned the Drosophila genome with Dichaete PWM (marked with an asterisk in Appendix B: supplementary table 9), to find putative TFBSs. Akin to the analyses for the 5 meso TFs, we used the 6-8hr ind-sorted dataset as the positive set and contrasted it with the 6-8hr unsorted dataset. Figure 32B shows the library depth normalized ATAC-seq signals around the TFBSs, including the putative signal profile inferred solely from naked ATAC-seq 6-mer bias values. Once again, signal shape is similar among the positive and negative datasets, which is partially reflected in the bias profile (possibly due to inference of bias values from naked ATAC-seq rather than ATAC-HS sites). The positive dataset shows higher ATAC-seq signal, which leads footprint models being outperformed by the simple tag count ranking, shown in figure 32C, confirming the observations from our embryo DNase-seq datasets.

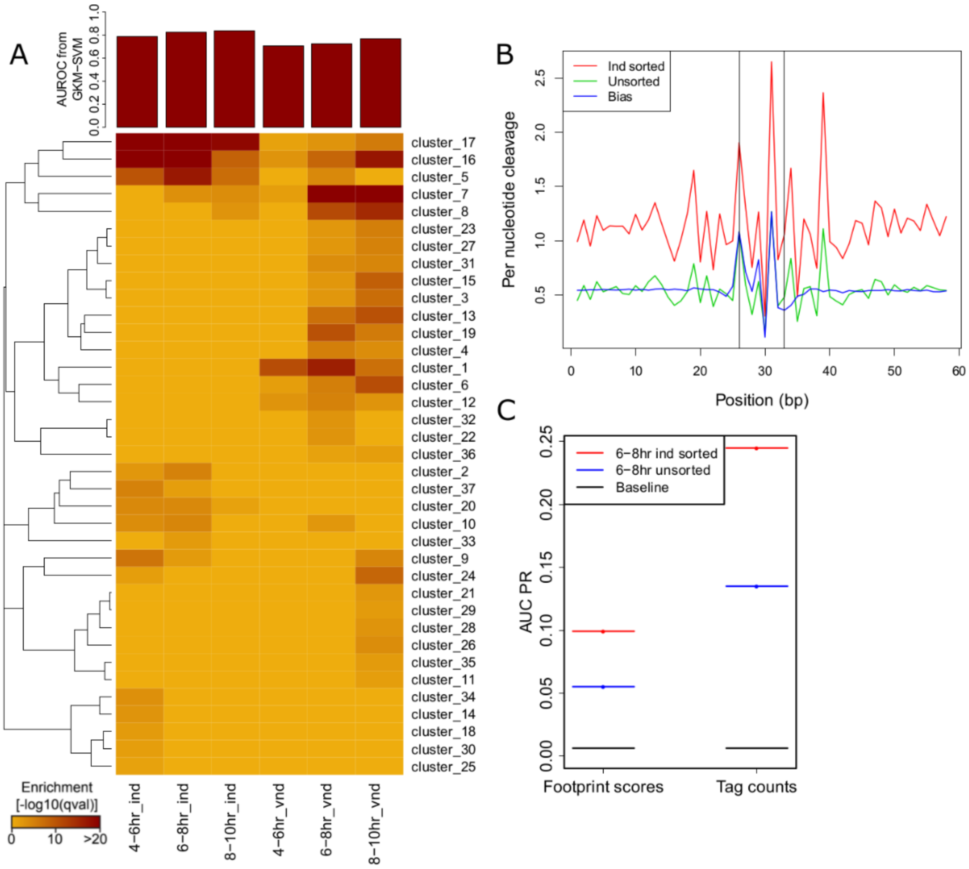


Figure 32: Motif enrichment and footprinting analyses in ATAC-seq data. (A) Heatmap shows the enriched motif clusters for the tissue-specific ATAC-HS sets, with the corresponding GKM-SVM results above. (B) Shape of the aggregate ATAC-seq signal around Dichaete motifs that overlap ChIP-seq peaks mirror each other in the ind-sorted and unsorted datasets, with the ind-sorted dataset showing higher signal. (C) Performance comparison of footprinting and tag count ranking in identifying Dichaete binding in ind-sorted vs unsorted datasets, based on 4-fold cross validated AUC-PR (baseline denotes the fraction of positive examples in the training set).

4.2.5. An integrative model allows finding putative binding sites for a multitude of factors

Upon observing that in both *Drosophila* open chromatin profiling projects, signal intensity was a better predictor of TF binding compared to signal shape (i.e. footprints), we wanted to predict bound sites for a multitude of TFs, using signal intensity as a feature. To this end, we selected a subset of the custom PWMs (see Appendix B: supplementary table 9, and methods for selection details), and scanned the *Drosophila* genome to find putative TFBSs for each. For each TFBS, we extracted the following three features: log-transformed DNase-seq tag counts at the TFBS with 25bp flanking regions upstream and downstream, motif match score and a score representing the sequence conservation (number of substitutions per base across the *Drosophila* phylogeny, SPH). We trained TF-specific logistic regression models using these three features, in a supervised manner, based on stage-matched ChIP-seq peaks for that TF (see methods). For example, for a TF with ChIP-seq data from 6-8hr embryos, we trained separate logistic regression models with tag counts from all 6-8hr DNase-seq datasets (meso, neuro, other and whole embryo) and calculated area under the receiver operating characteristic curve (AUC) values as a measure of performance. Figure 33 shows the best model AUCs per PWM, on the right panel (i.e. the TF-specific AUCs), and the corresponding model coefficients on the left panel. Tag counts are the most important features, with the highest model coefficients, with smaller contributions from motif scores and SPH to model performance. Appendix B: supplementary table 9 lists the best AUC per PWM, as well as the dataset in which it was achieved. In general, the best models are achieved in the relevant datasets, for example the 5 meso TFs bin, lmd, mef2, tin and twi achieve best AUCs in 6-8hr binpos, 6-8hr binpos, 8-10hr meso, 4-6hr meso and 6-8hr meso datasets, respectively, indicating that our integrative models correctly capture the relationship between tissues and TFs.

We next asked whether a common model could explain binding for most TFs. To this end, we created a generic model using the average coefficients for each feature (shown as the horizontal lines in figure 33, left panel). The generic model AUCs were calculated per TF, for the same datasets that achieved best TF-specific model AUCs. The generic model leads to only a mild decrease in overall model performance (Figure 33, right), indicating that it is applicable to a multitude of TFs. Thus, the generic model can be used to assay bound sites when open chromatin data is available, but ChIP-seq data is not.

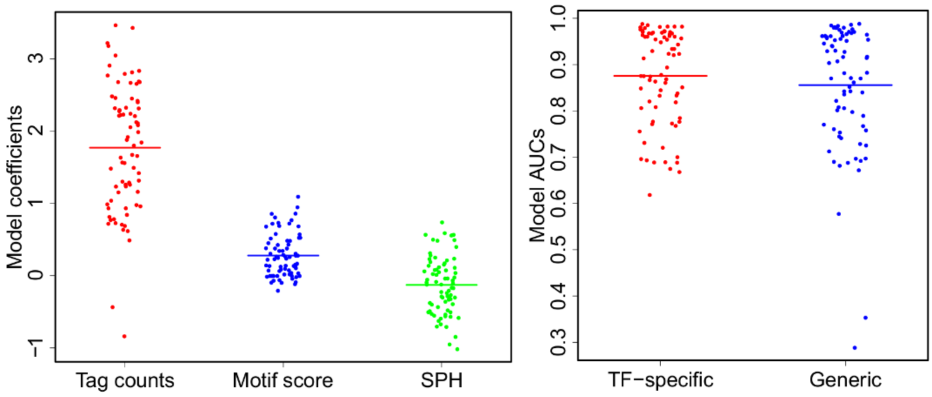


Figure 33: A logistic regression with three features is highly predictive of TF binding. Feature coefficients associated with tag counts, motif score and SPH, derived from the best-performing model per PWM (left). Best performing TF-specific model AUCs are compared to the AUCs derived from using the generic integrative model on the same datasets (right).

5. Discussion

Part 1

DNase-seq has been widely used to assay open chromatin regions and TF footprints. The emergence and increasing use of ATAC-seq, necessitates a systematic comparison of the two methods, especially for TF-footprinting. Here, in a comparative setting, we have shown that although the two methods have distinct sequence biases and generate different footprint shapes for the same TF, the sites they identify as bound are largely in agreement. However, the sequence content of TFBSs combined with protocol-specific sequence biases, impact footprinting efficiency for some TFs, leading to larger differences for these factors and making one method preferable to the other.

There are opposing views on the library depth required for TF-footprinting. Whereas some studies require at least 200 million reads[109], others demonstrate efficient TF-footprinting at moderate sequencing depths (50-60 million reads)[122,129], in agreement with our results. These moderate numbers were reported for both segmentation-based[129] and integrative site-centric[122] tools, challenging the view that these approaches have different depth requirements[109]. To get the highest possible depth, pooling all replicates has been a common practice in TF-footprinting studies. However, our results indicate that keeping the replicates separate to assess reproducibility may lead to more accurate footprint predictions. This is especially relevant for low-depth libraries, where this approach enables finding reliable subsets of the total footprint pool.

Although the sequence bias of DNase I is well characterized, there is still no consensus about the benefits of bias correction for TF-footprinting. Whereas some studies report increased accuracy upon bias correction[111], others do not make this observation[122]. One explanation for this might be the different approaches to DNase signal processing and TF-footprinting. Methods that extensively smooth the signal, or use features that diverge from single-nucleotide resolution (eg. binned signal) might be less affected by bias. Since our method has single-nucleotide resolution, we have used protocol-specific biases to model the background in our TF-footprinting approach, and we could demonstrate significant improvements on footprinting when using bias correction on DNase-seq data. While ATAC-seq footprinting showed also promising results on par with DNase-seq in HEK293 data (figure 19D), its performance in K562 data was significantly lower for almost all factors (figure 23A) where it outperformed DNase-seq for only three factors, two of which had low average DNase I cleavage propensities over their motif regions that resulted in a footprint-like background profile. The opposite was not as clear to observe, i.e. for factors where DNase-seq outperformed ATAC-seq, the average Tn5 cleavage propensities over the motif regions were not consistently at the lower end of the spectrum. Furthermore, the range of average cleavage propensities over all TFs was narrower for Tn5 (figure 22D vs figure 23B).

Recent studies have proposed several Tn5 bias correction methods and in order to rule out that this observation resulted from our 6-mer based approach, we used a different bias correction, in which a 17bp long gapped k-mer with 8 meaningful positions is used to correct ATAC-seq data[134]. This more sophisticated bias correction method did not improve the footprint model performance (figure 23D). Taken together, correcting for Tn5 sequence bias does either not have a strong impact on ATAC-seq footprinting, or neither of the approaches we used is comparable in its impact to DNase-seq bias correction.

Our comparative analysis clearly confirms previous reports that DNase cleavage bias might render footprints of some factors “invisible”, and that, in general, performance to identify footprints can vary significantly across assays and TFs. While an effective footprinting for all TFs, may in principle be achieved through a combination of assays with different sequence biases, our results do not suggest ATAC-seq for this purpose, due to its reduced performance; although it is possible to achieve better performance in deeper datasets as exemplified by our HEK293 data. Finally, in contrast to previous studies that reported no correlation between ChIP-seq signal values and footprint scores[119], we have previously observed and now observe again a strong link between these two measures, implying that the footprint score we have defined here is a quantitative measure of occupancy. In summary, we expect that the insights gained from this work will provide experimental design and computational analysis guidelines for future TF-footprinting studies.

Part 2

Drosophila melanogaster is a widely studied model organism, where many TFs that govern crucial stages of embryogenesis are known. Here, we combine open chromatin profiling with tissue and time-point specificity, to elucidate sequence features governing cell fate decisions during *Drosophila* embryogenesis. The pre-existing information regarding lineage-specifying master regulators confirms the relevance of our results.

The tissue-specific motif enrichment results presented here capture known TFs, and potentially implicate novel ones. One caveat of these results, however, is that it is not always a trivial task to link a single TF to an enriched cluster, due to multiple TFs having similar motifs. This can be exemplified via the neuro-specifically enriched Sp1/Klf family motif, where the exact associated TF is unclear. Furthermore, in some cases, where a given motif is clearly enriched, it can be difficult to associate the TF with the enrichment. For example, Ttk, the motif of which is significantly enriched in neural tissues in both DNase-seq and ATAC-seq datasets, is actually a repressor of neural fate[140,141]. In such cases, further elucidation of the link is necessary. In other cases, the relationship is much easier to appreciate; for example, the GATA motif enriched in other-specific datasets, clearly implicates Srp, which is known to be important for the specification of the endoderm[142]. As the sorted datasets are mesoderm and neuroectoderm specific, it is plausible to assume that the third layer, endoderm, should be enriched in the other-specific datasets. These results implicate the potential as well as the complications associated with tissue-specific motif analyses.

Comparing time-points, adds another layer to the motif enrichment analyses. Surprisingly, it is possible to find sequence features that discriminate these datasets, even though they correspond to consecutive time-points of the same tissues. One explanation for this might be the activity of slightly different sets of TFs per specific time-point. In line with this, we show that the motif enrichment values of TFs vary over our time-points. For example, both *twi* and *tin* show higher enrichments at earlier time-points, which is in line with their importance in specifying the early mesoderm specification. On the other hand, kr and klu, both of which are known to have functions in neural development[143,144], show an increase of motif enrichment in neuro datasets, as development progresses. The time-point analyses enable the observation of these trends. Another interesting observation from these analyses is the existence of motifs that show a "general-early" or "general late" enrichment regardless of the tissue. It is of interest to further investigate whether these correspond to biologically meaningful sequence signatures.

Finally, we show the caviats related to TF footprinting in this system and predict TF bound sites with an alternative integrative model. Furthermore, we propose a generic model that performs well for a majority of assayed TFs. We expect this model and our comprehensive motif enrichment analyses, to constitute a useful resource.

Appendix A: Supplementary tables (part 1)

Cell type	Sample description	Total mapped read pairs	Percent mtDNA	Percent uniquely aligned after removing mtDNA	Percent duplication after removing mtDNA	Final read pairs after processing
K562	10 minute lysis	98241437	74.9	60.86	36.1	11824634
K562	5 minute lysis	59725560	73.3	61.68	28.52	8293938
K562	No lysis buffer	64162804	18	76.09	28.83	26203527
HEK293	High depth, bio1-tech1	212332636	21.7	79.15	38.6	74957855
HEK293	High depth, bio1-tech2	215849442	16.7	79.41	42.41	75883012
HEK293	High depth, bio2-tech1	189055455	8.3	80.35	17.42	106390553
HEK293	High depth, bio2-tech2	212178995	3.4	80.84	25.81	112909794
HEK293	Medium depth, bio1-tech1	101177506	22	78.93	22.54	44903594
HEK293	Medium depth, bio1-tech2	115293922	16.9	79.12	27.43	50914321
HEK293	Medium depth, bio2-tech1	85731217	8.4	80.28	8.82	53211877
HEK293	Low depth, bio1-tech1	53199070	21.9	78.99	12.83	26607741
HEK293	Low depth, bio1-tech2	59968056	16.8	79.19	15.84	30798873
HEK293	Low depth, bio2-tech1	40964758	8.4	80.3	4.54	26613414
HEK293	Low depth, bio2-tech2	51835433	3.4	80.81	7.72	34364305

Supplementary table 1: General statistics of the ATAC-seq datasets generated in the study.

Cell type	Data type	Description	Accession code	Library depth after processing
K562	DNase-seq	Replicate 1 (ENCODE)	ENCFF000SWU	72166285
K562	DNase-seq	Replicate 2 (ENCODE)	ENCFF000SXA	138770111
K562	DNase-seq	Replicate 3 (ENCODE)	ENCFF000SWY	88033023
K562	DNase-seq	Replicate lab	Generated for the study	134851555
HEK293	DNase-seq	Replicate 1 (ENCODE)	ENCFF000SPK	68339552
HEK293	DNase-seq	Replicate 2 (ENCODE)	ENCFF000SQB	164469299
HEK293	DNase-seq	Replicate lab	Generated for the study	126253898
Human (YH1)	Tn5 transposition	Deproteinized genomic DNA	SRX030445	39753928
D. melanogaster	Tn5 transposition	Deproteinized genomic DNA	SRX030438	22705812

Supplementary table 2: Descriptions, accession codes and final read counts for the utilized DNase-seq datasets and libraries generated by Tn5 transposition of deproteinized genomic DNA.

Comparison name	Biological replicate 1	Biological replicate 2
High depth ATAC-seq 1	High depth, bio1-tech1	High depth, bio2-tech1
High depth ATAC-seq 2	High depth, bio1-tech2	High depth, bio2-tech2
Medium depth ATAC-seq 1	Medium depth, bio1-tech1	Medium depth, bio2-tech1
Medium depth ATAC-seq 2	Medium depth, bio1-tech2	Medium depth, bio2-tech1
Low depth ATAC-seq 1	Low depth, bio1-tech1	Low depth, bio2-tech1
Low depth ATAC-seq 1	Low depth, bio1-tech2	Low depth, bio2-tech2

Supplementary table 3: Scheme for ATAC-seq library comparisons for JAMM-IDR peak calls or FLR-IDR footprint calls.

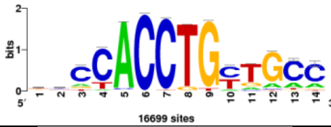
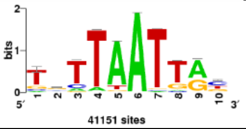
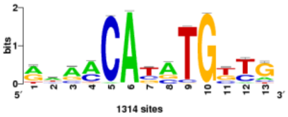
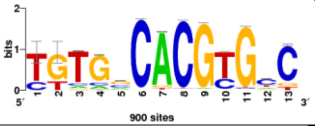
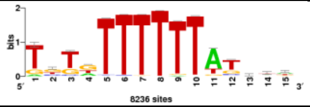
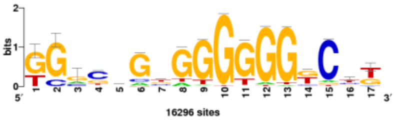
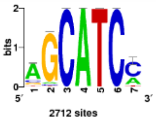
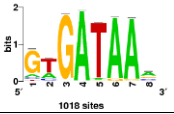
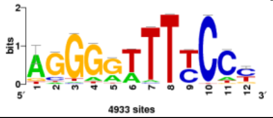
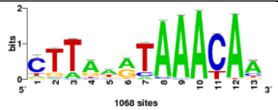
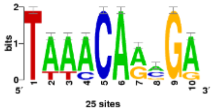
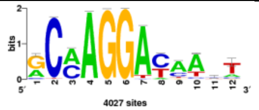
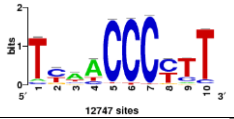
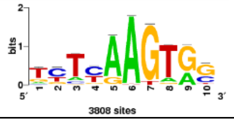
Cell line	Factor	Accession code
HEK293	CTCF	ENCFF002DCV
HEK293	MAZ	ENCFF834ZRT
HEK293	REST	ENCFF201ZGY
HEK293	YY1	ENCFF443TBN
K562	CREB1	ENCFF001UJI, ENCFF001UJJ
K562	CTCF	ENCFF002CEL, ENCFF002CLS, ENCFF002CWL, ENCFF002DBD, ENCFF002DDJ
K562	E2F4	ENCFF002CWM
K562	ETS1	ENCFF002CLX
K562	GABPA	ENCFF002CLZ
K562	GATA2	ENCFF002CMA, ENCFF002CWQ
K562	MAX	ENCFF002CXD
K562	MAZ	ENCFF002CXE
K562	MEF2A	ENCFF002CMD
K562	NFYA	ENCFF002CXI
K562	NRF1	ENCFF002CXK, ENCFF454OVP, ENCFF657YIC, ENCFF664FFU
K562	REST	ENCFF002CMF
K562	RFX1	ENCFF654RTP
K562	RFX5	ENCFF002CXV
K562	SP1	ENCFF002CMN, ENCFF191QSX
K562	SRF	ENCFF002CMP
K562	STAT1	ENCFF002CYB, ENCFF002CYC, ENCFF002CYD, ENCFF002CYE
K562	USF1	ENCFF002CMV
K562	YY1	ENCFF002CMW, ENCFF002CMX, ENCFF002CYQ
K562	ZNF143	ENCFF002CYR
S2	CTCF	GSM409078
S2	BEAF32	GSM1278639

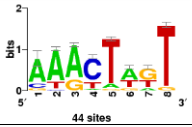
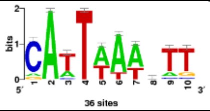
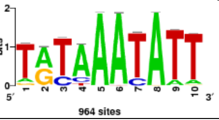
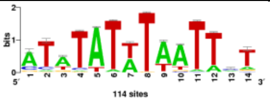
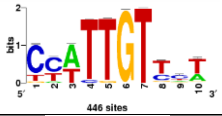
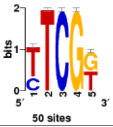
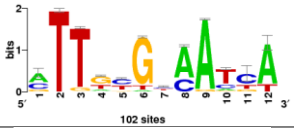
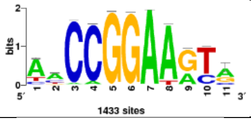
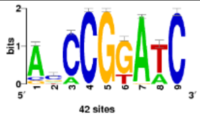
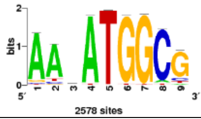
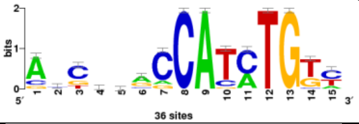
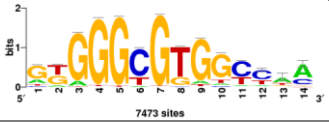
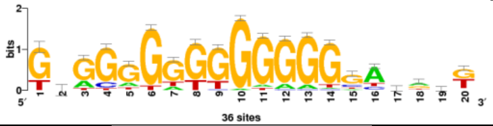
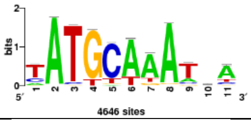
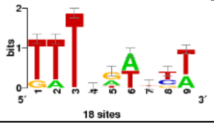
Supplementary table 4: ChIP-seq peaks used in the analysis.

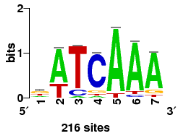
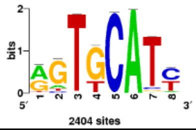
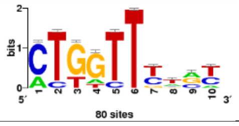
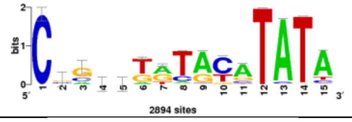
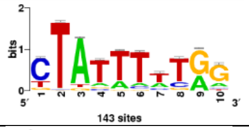
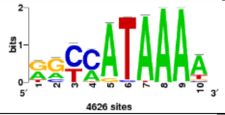
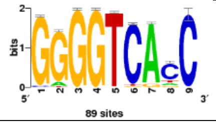
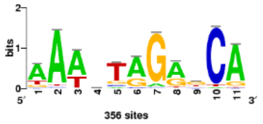
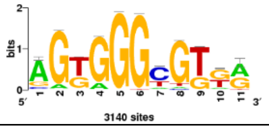
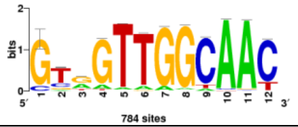
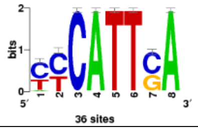
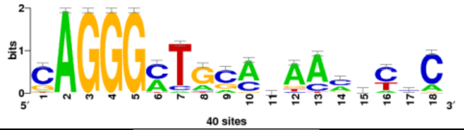
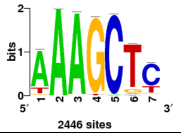
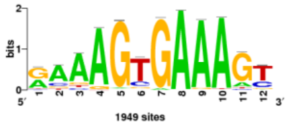
Factor name	PWM ID	Lowest PWM score in top 50K	Closest threshold PWM score	p-value associated with threshold
CREB1	MA0018.2	9.06	7.78555	1*10-6
CTCF	MA0139.1	8.09	7.89799	5*10-5
E2F4	M5180_1.01	1.71	1.78185	2*10-5
ETS1	MA0098.1	8.11	6.9036	1*10-6
GABPA	MA0062.2	8.42	8.43115	4*10-5
GATA2	MA0036.1	7.20	6.24233	1*10-6
MAX	M5613_1.02	5.45	3.68637	1*10-6
MAZ	M00649	9.32	8.22958	1*10-6
MEF2A	M5615_1.02	9.09	4.80812	1*10-6
NFYA	MA0060.1	8.83	8.46208	5*10-5
NRF1	M00652	3.90	1.39346	1*10-6
REST	MA0138.2	5.89	5.80754	3*10-5
RFX1	M00280	8.63	8.53351	5*10-5
RFX5	M5779_1.02	7.03	5.27345	1*10-6
SP1	MA0079.2	9.17	8.38317	1*10-6
SRF	MA0083.1	7.25	6.8771	1*10-6
STAT1	MA0137.2	9.39	9.0732	3*10-5
USF1	M5943_1.02	9.73	9.36591	1*10-6
YY1	M5954_1.02	8.56	7.33829	1*10-6
ZNF143	M5966_1.02	4.96	2.23431	1*10-6

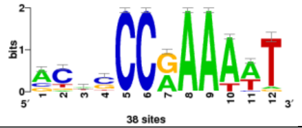
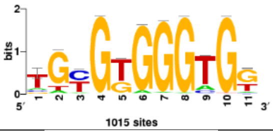
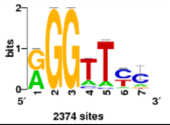
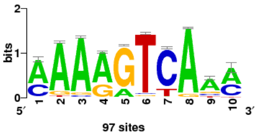
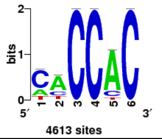
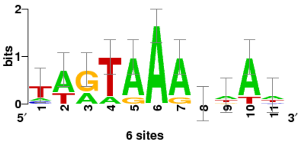
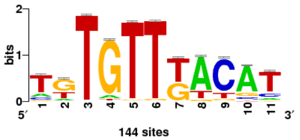
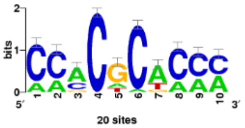
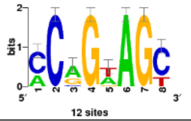
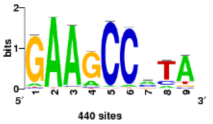
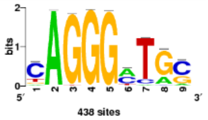
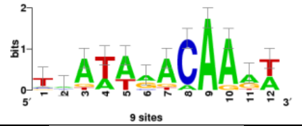
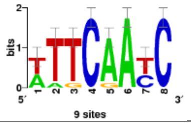
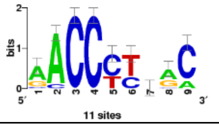
Supplementary table 5: PWM IDs used for genome-wide motif searches.

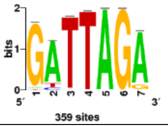
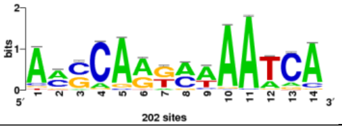
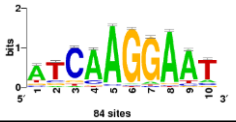
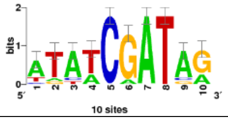
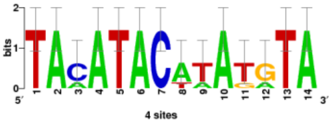
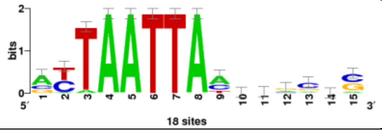
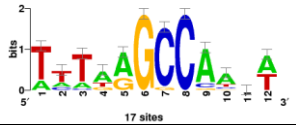
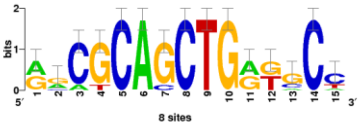
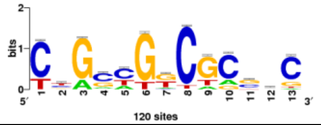
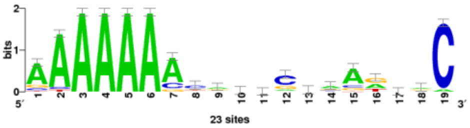
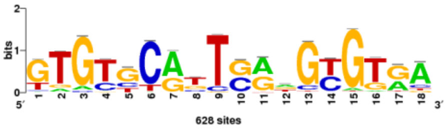
Appendix B: Supplementary tables (part 2)

Cluster #	TF Annotation	Cluster summary motif
cluster_1	E-box/bHLH (Da, Zld, Wor)	
cluster_2	Homeodomain (CG4328,Abd-A,H2.0,Dfd,Zen)	
cluster_3	E-box/bHLH (Twi)	
cluster_4	E-box/bHLH (Espl, Myc, Met)	
cluster_5	C2H2-ZF (Jim, Hb)	
cluster_6	C2H2-ZF (Lola, Crol, Sug)	
cluster_7	C2H2-ZF (Peb)	
cluster_8	GATA (So, Srp)	
cluster_9	NF-KB (Rel), C2H2-ZF (Lola)	
cluster_10	Forkhead (Bin, Croc)	
cluster_11	Forkhead (Bin)	
cluster_12	C2H2-ZF (Ttk)	
cluster_13	C2H2-ZF (Ttk, Kr)	
cluster_14	Homeodomain (Tin, Vnd, Bap)	

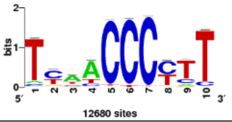
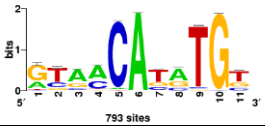
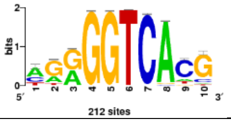
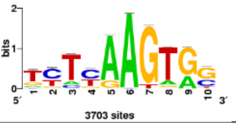
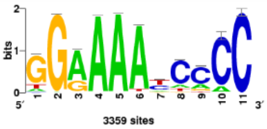
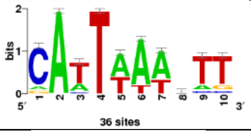
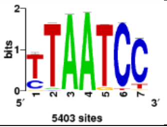
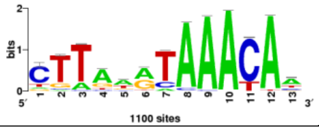
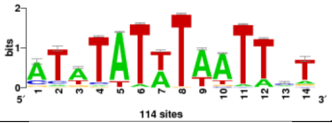
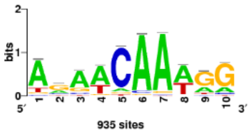
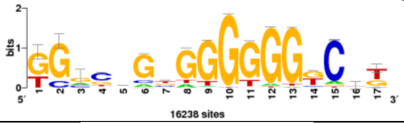
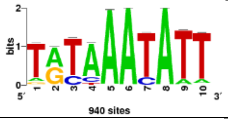
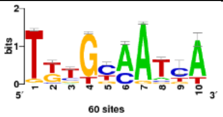
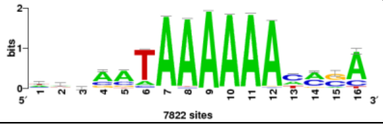
cluster_15	C2H2-ZF (Br)	
cluster_16	Homeodomain (Zen)	
cluster_17	Forkhead (Slp1)	
cluster_18	HTH (Bab1)	
cluster_19	HMG-box (D)	
cluster_20	Deaf1	
cluster_21	bZIP (Slbo)	
cluster_22	ETS-domain (Ets65A, Ets21C)	
cluster_23	Ets98B, Coop	
cluster_24	Da, Phol	
cluster_25	E-box/bHLH (Da)	
cluster_26	C2H2-ZF+Sp1/Klf (Luna)	
cluster_27	C2H2-ZF (L(3)neo38, CG7368)	
cluster_28	HTH (Rib), POU-homeodomain (Nub)	
cluster_29	Homeodomain (Exd)	

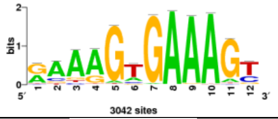
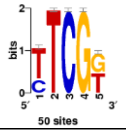
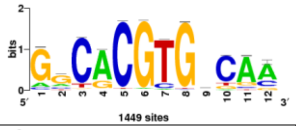
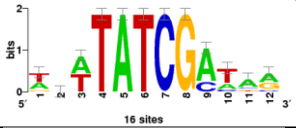
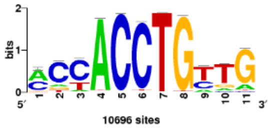
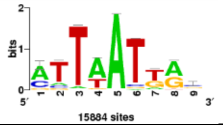
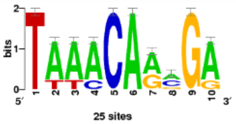
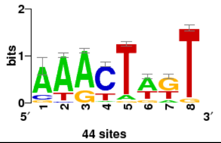
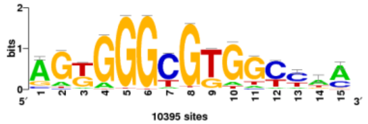
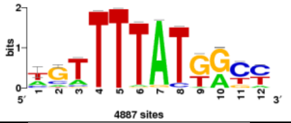
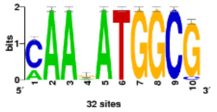
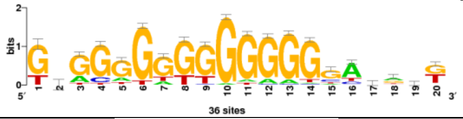
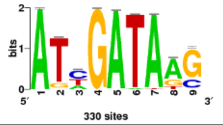
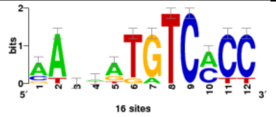
cluster_30	Retn, Pan	
cluster_31	C2H2-ZF (Chinmo, CG12236)	
cluster_32	Grh	
cluster_33	C2H2-ZF (Cf2)	
cluster_34	MADS-box (Mef2)	
cluster_35	Homeodomain (Abd-A, Cad)	
cluster_36	Usp, CG8319	
cluster_37	C4-ZF(Kni, Eip75B)	
cluster_38	C2H2-ZF+Sp1/Klf (Klf15, Btd)	
cluster_39	C2H2-ZF (CG4854)	
cluster_40	HMG-box (Cic)	
cluster_41	C2H2-ZF (ZIPIC)	
cluster_42	C2H2-ZF (Lola)	
cluster_43	C2H2-ZF (Blimp-1)	

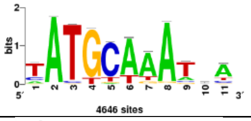
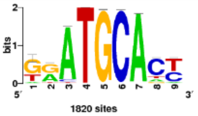
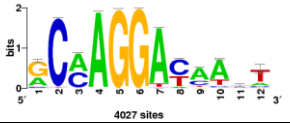
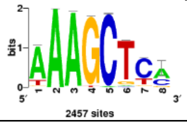
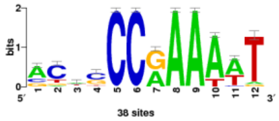
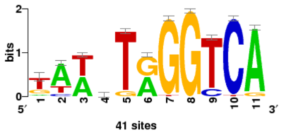
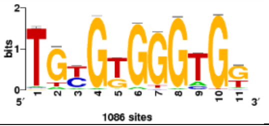
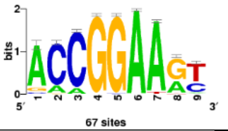
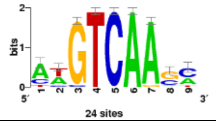
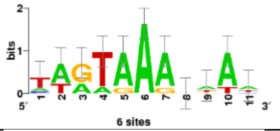
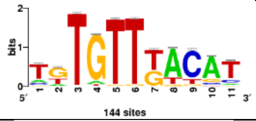
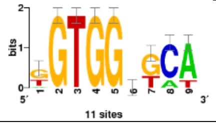
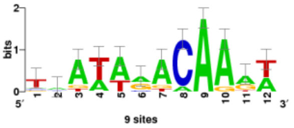
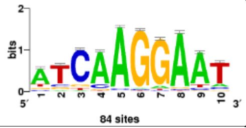
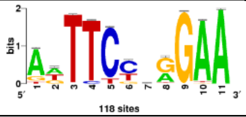
cluster_44	HTH (Eip93F)	
cluster_45	C2H2-ZF (Klu)	
cluster_46	C2H2-ZF (Lola)	
cluster_47	C4-ZF (Tll, ERR)	
cluster_48	CR43669/70/71	
cluster_49	C2H2-ZF (Br)	
cluster_50	Forkhead (Slp1)	
cluster_51	C2H2-ZF (Trl)	
cluster_52	C2H2-ZF (Bowl)	
cluster_53	C2H2-ZF (Gl)	
cluster_54	C2H2-ZF (ZIPIC)	
cluster_55	C2H2-ZF (Br)	
cluster_56	CG15601	
cluster_57	C2H2-ZF (Lola)	

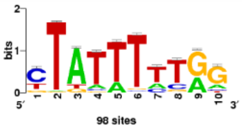
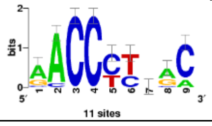
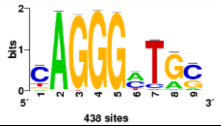
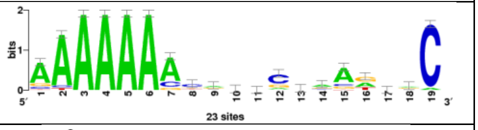
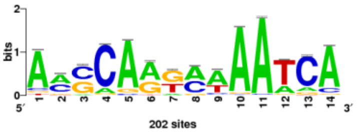
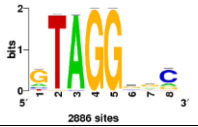
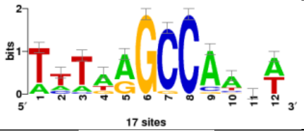
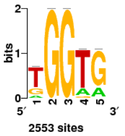
cluster_58	C2H2-ZF (Br)	
cluster_59	C2H2-ZF (Shn)	
cluster_60	Homeodomain (Ftz)	
cluster_61	Dref	
cluster_62	Top2	
cluster_63	CG7745	
cluster_64	Mes2	
cluster_65	E-box/bHLH (HLH4C)	
cluster_66	SMAD (Mad)	
cluster_67	C2H2-ZF (Rn)	
cluster_68	Homeodomain (Ey)	

Supplementary table 6: Motif clusters enriched in the tissue specific DHSs.

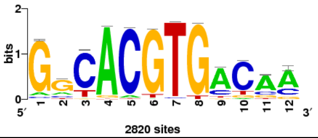
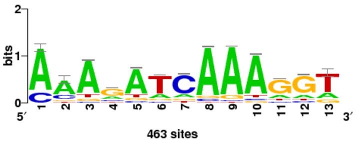
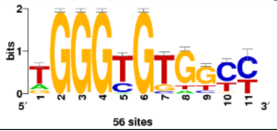
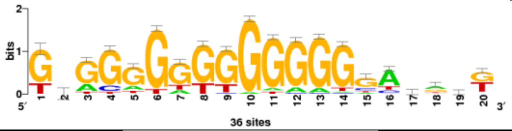
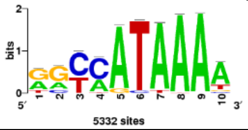
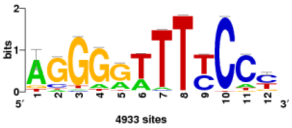
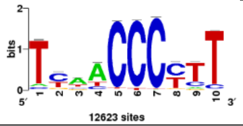
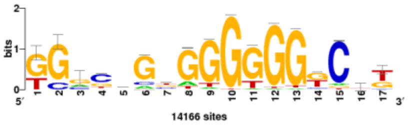
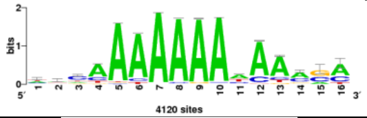
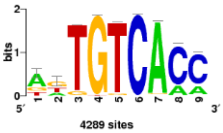
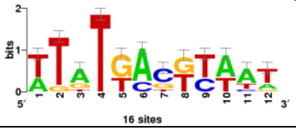
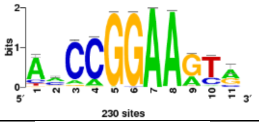
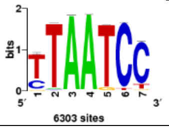
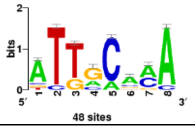
Cluster #	TF Annotation	Cluster summary motif
cluster_1	C2H2-ZF (Ttk, Kr)	
cluster_2	E-box/bHLH (Twi)	
cluster_3	C4-ZF (Usp, Hr4)	
cluster_4	Homeodomain (Vnd, Tin)	
cluster_5	C2H2-ZF (Lola), NF-KB (Rel)	
cluster_6	Homeodomain (Zen)	
cluster_7	Homeodomain (Bcd, Gsc)	
cluster_8	Forkhead (FoxP, Bin)	
cluster_9	HTH (Bab1)	
cluster_10	HMG-box (D, Sox15, Sox14)	
cluster_11	C2H2-ZF (Lola, Sug, CG7368)	
cluster_12	Forkhead (Slp1)	
cluster_13	bZIP (Slbo)	
cluster_14	C2H2-ZF (Jim, Hb, Rn)	

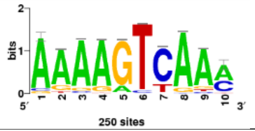
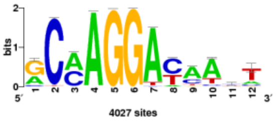
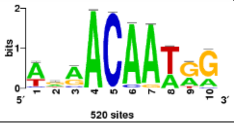
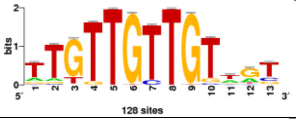
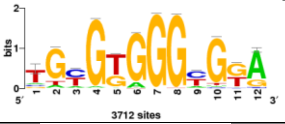
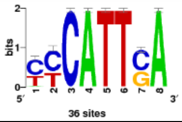
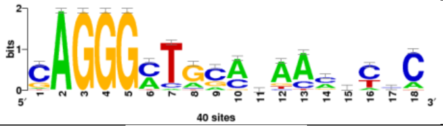
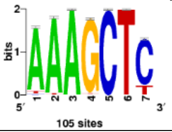
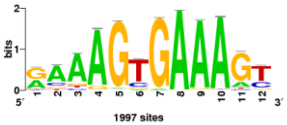
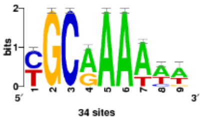
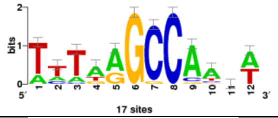
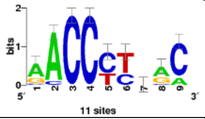
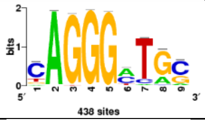
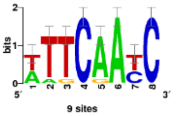
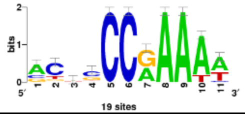
cluster_15	C2H2-ZF (Ken)	
cluster_16	Deaf1	
cluster_17	E-box/bHLH (E(spl)m8-HLH)	
cluster_18	Dref	
cluster_19	E-box/bHLH (Zld, Da)	
cluster_20	Homeodomain (Cad, Bsh, CG34031, Abd-A, Onecut)	
cluster_21	Forkhead (Bin)	
cluster_22	C2H2-ZF (Br)	
cluster_23	C2H2-ZF+Sp1/Klf (CG3065, Sr, Btd, Spps)	
cluster_24	Homeodomain (Cad), ARID-HTH (Retn)	
cluster_25	C2H2-ZF (Phol)	
cluster_26	C2H2-ZF (L(3)neo38, CG7368)	
cluster_27	GATA(Srp), Beaf32	
cluster_28	C2H2-ZF (Disco-r)	

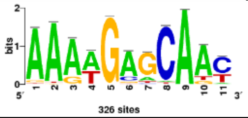
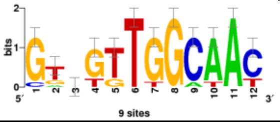
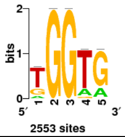
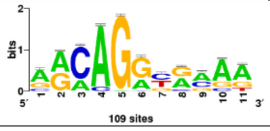
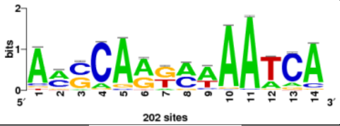
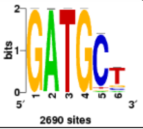
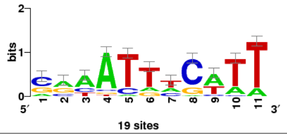
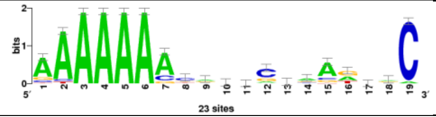
cluster_29	HTH (Rib), POU-homeodomain (Pdm2, Nub)	
cluster_30	C2H2-ZF (Peb, Chinmo)	
cluster_31	C2H2-ZF (Ab, Ttk)	
cluster_32	C2H2-ZF (Lola)	
cluster_33	HTH (Eip93F)	
cluster_34	ZF (Eip75B)	
cluster_35	C2H2-ZF (Klu)	
cluster_36	ETS-domain (Ets21C, Ets97D)	
cluster_37	E-box/bHLH (E(spl)mγ-HLH)	
cluster_38	C2H2-ZF (Br)	
cluster_39	Forkhead (Slp1)	
cluster_40	C2H2-ZF (Lola)	
cluster_41	C2H2-ZF (Br)	
cluster_42	Homeodomain (Ftz)	
cluster_43	STAT (Stat92E)	

cluster_44	MADS-box (Mef2)	
cluster_45	C2H2-ZF (Lola)	
cluster_46	C2H2-ZF (ZIPIC)	
cluster_47	C2H2-ZF (Rn)	
cluster_48	C2H2-ZF (Shn)	
cluster_49	C2H2-ZF (Lola)	
cluster_50	Mes2	
cluster_51	CR43669/70/71	

Supplementary table 7: Motif clusters enriched in the time-point specific DHSs.

Cluster #	TF Annotation	Cluster summary motif
cluster_1	E-box/bHLH (Espl), C2H2-ZF (CG7386)	
cluster_2	ZF(Hr51), HMG-box (Dtcf)	
cluster_3	C2H2-ZF+Sp1/Klf (Dar1, Luna)	
cluster_4	C2H2-ZF (CG7368)	
cluster_5	Homeodomain (CG4328, Cad)	
cluster_6	NF-KB (Dl, Rel)	
cluster_7	C2H2-ZF (Ttk, Pad)	
cluster_8	C2H2-ZF (Sug, Crol, Opa)	
cluster_9	C2H2-ZF (Jim, Dati)	
cluster_10	C2H2-ZF (Disco)	
cluster_11	bZIP (Gt)	
cluster_12	C2H2-ZF (Ab), ETS (Eip74EF)	
cluster_13	Homeodomain (Oc, Bcd)	
cluster_14	bZIP (Slbo)	

cluster_15	C4-ZF (Tll)	
cluster_16	C2H2-ZF (Ttk)	
cluster_17	HMG-box (D)	
cluster_18	C2H2-ZF (Aef1)	
cluster_19	C2H2-ZF (Klu), C2H2-ZF+Sp1/Klf (Klf15)	
cluster_20	HMG-box (Cic)	
cluster_21	C2H2-ZF (ZIPIC)	
cluster_22	C2H2-ZF (Lola)	
cluster_23	C2H2-ZF (Blimp-1)	
cluster_24	HTH (Rib)	
cluster_25	Mes2	
cluster_26	C2H2-ZF (Lola)	
cluster_27	C2H2-ZF (ZIPIC)	
cluster_28	CG15601	
cluster_29	HTH (Eip93F)	

cluster_30	C2H2-ZF (Erm)	
cluster_31	C2H2-ZF (CG4854)	
cluster_32	CR43669/70/71	
cluster_33	SMAD (Med)	
cluster_34	C2H2-ZF (Shn)	
cluster_35	C2H2-ZF (Peb)	
cluster_36	HTH+TEA (Sd)	
cluster_37	C2H2-ZF (Rn)	

Supplementary table 8: Motif clusters enriched in the tissue specific ATAC-HSs.

PWM	TF name	ChIP data	Dataset with best AUROC	AUROC	Generic AUROC
berkeley_BCD	bcd	2-3hr	2_4filt	0.8672	0.8621
berkeley_bdtnp_cad	cad	0-4hr	2_4WE	0.9211	0.9149
berkeley_bdtnp_dl	dl	2-3hr	2_4filt	0.6889	0.6926
berkeley_bdtnp_h	h	0-8hr	4_6WE	0.6893	0.6822
berkeley_bdtnp_run	run	2-3hr	2_4WE	0.9345	0.9176
berkeley_bdtnp_slp1	slp1	4-6hr, 6-8hr	6_8WE	0.9681	0.9626
berkeley_GT	gt	2-3hr	2_4WE	0.8192	0.7900
berkeley_HB	hb	2-3hr	2_4WE	0.7211	0.7259
dmmpmm2009_FBgn0025647_dmmpmm2009_0	Trl	8-16hr	8_10WE	0.8334	0.8223
flyfactor_bab1_sanger_5	bab1	0-12hr	4_6meso	0.9237	0.9307
flyfactor_bap_optimized_furlong	bap	6-8hr	6_8binpos	0.9698	0.9713
flyfactor_bin_optimized_furlong*	bin	6-8hr, 8-10hr, 10-12hr	6_8binpos	0.9811	0.9803
flyfactor_chinmo_solexa	chinmo	0-12hr	2_4WE	0.6754	0.3540
flyfactor_disco_solexa_5	disco	0-8hr, 8-16hr	4_6DN	0.8690	0.8707
flyfactor_dl_flyreg	dl	2-3hr	2_4WE	0.6939	0.6907
flyfactor_doc2_sanger_5	Doc2	4-6hr, 6-8hr	4_6DN	0.9734	0.9815
flyfactor_dtcf_furlong	dtcf/pan	4-6hr, 6-8hr	6_8WE	0.9611	0.9620
flyfactor_eve_solexa	eve	8-16hr	8_10WE	0.8761	0.8715
flyfactor_lmd_solexa_5*	lmd	6-8hr	6_8binpos	0.9678	0.9603
flyfactor_lola-pd_solexa	lola	0-12hr	2_4WE	0.7734	0.2890
flyfactor_mef2_optimized_furlong*	Mef2	2-4hr, 4-6hr, 6-8hr, 8-10hr, 10-12hr	8_10meso	0.9827	0.9803
flyfactor_pan_flyreg	dtcf/pan	4-6hr, 6-8hr	4_6DN	0.9551	0.9578
flyfactor_phol_sanger_5	phol	4-12hr	8_10meso	0.9709	0.9544
flyfactor_pmad_furlong	Mad	4-6hr, 6-8hr	4_6WE	0.9820	0.9769
flyfactor_pnr_furlong	pnr	4-6hr, 6-8hr	4_6DN	0.9610	0.9533
flyfactor_pnr_sanger_5	pnr	4-6hr, 6-8hr	4_6WE	0.9722	0.9665
flyfactor_sens_sanger_10	sens	4-8hr	4_6DN	0.7771	0.7678
flyfactor_slp1_nar	slp1	4-6hr, 6-8hr	6_8WE	0.9676	0.9711
fly_factor_survey_FBgn0002521_fly_factor_survey_1	pho	0-16hr, 4-12hr, 6-12hr	6_8WE	0.8943	0.8830
fly_factor_survey_FBgn0003300_fly_factor_survey_1	run	2-3hr	2_4filt	0.9345	0.9296
fly_factor_survey_FBgn0011655_fly_factor_survey_0	Med	2-3hr	2_4WE	0.8385	0.8430
fly_factor_survey_FBgn0267821_fly_factor_survey_0	da	2-3hr	2_4WE	0.6898	0.6885
flyfactor_tin_optimized_furlong*	tin	2-4hr, 4-6hr, 6-8hr	4_6meso	0.9795	0.9816
flyfactor_ttk-pa_sanger_5	ttk	6-8hr	6_8DN	0.9712	0.9652
flyfactor_twi_optimized_furlong*	twi	2-4hr, 4-6hr, 6-8hr	6_8meso	0.9827	0.9851
flyfactor_vfl_sanger_5	vfl/zld	1hr, 2hr, 3hr	2_4filt	0.7316	0.7131
flyfactor_vfl_solexa_5	vfl/zld	1hr, 2hr, 3hr	2_4filt	0.7720	0.7292
flyreg_FBgn0024766_flyreg_0	eve	8-16hr	8_10WE	0.8758	0.8427
idmmpmm2009_FBgn0000411_idmmpmm2009_0*	D	0-8hr	4_6DN	0.9883	0.9886

idmmpmm2009_FBgn0001150_idmmpmm2009_0	gt	2-3hr	2_4WE	0.8339	0.8085
idmmpmm2009_FBgn0003870_idmmpmm2009_0	ttk	6-8hr	6_8WE	0.9695	0.9796
idmmpmm2009_FBgn0260632_idmmpmm2009_0	dl	2-3hr	2_4WE	0.7007	0.6978
jaspar_bap	bap	6-8hr	6_8meso	0.9299	0.9040
jaspar_BEAF-32*	BEAF-32	0-12hr	6_8DN	0.7774	0.7418
jaspar_CAD	cad	0-4hr	2_4WE	0.9239	0.9171
JASPAR_CORE_2014_insects_FBgn0004862_JASPAR_CORE_2014_insects_0	bap	6-8hr	6_8meso	0.9439	0.9281
jaspar_CTCF*	CTCF	0-12hr	8_10neuro	0.8613	0.7982
jaspar_eve	eve	8-16hr	10_12DN	0.8783	0.8822
jaspar_exd	exd	0-8hr	4_6DN	0.6190	0.5780
jaspar_gt	gt	2-3hr	2_4WE	0.8066	0.7453
jaspar_h	h	0-8hr	4_6WE	0.6686	0.6724
jaspar_Mad	Mad	4-6hr, 6-8hr	4_6WE	0.9730	0.9756
jaspar_pnr	pnr	4-6hr, 6-8hr	4_6DN	0.9624	0.9622
jaspar_Trl	Trl	8-16hr	8_10WE	0.8514	0.8362
OnTheFly_FBgn0000411_OnTheFly_0	D	0-8hr	4_6DN	0.9857	0.9842
OnTheFly_FBgn0003448_OnTheFly_1	sna	2-4hr	2_4WE	0.8492	0.8523
OnTheFly_FBgn0004870_OnTheFly_0	bab1	0-12hr	4_6meso	0.9142	0.9077
OnTheFly_FBgn0015602_OnTheFly_0	BEAF-32	0-12hr	8_10neuro	0.7850	0.7585
OnTheFly_FBgn0039039_OnTheFly_0	lmd	6-8hr	6_8binpos	0.9661	0.9697
OnTheFly_FBgn0045759_OnTheFly_0	bin	6-8hr, 8-10hr, 10-12hr	6_8binpos	0.9826	0.9867
OnTheFly_FBgn0267821_OnTheFly_0	da	2-3hr	2_4WE	0.6966	0.6974
pouya_Dfd_disc_1	Dfd	0-8hr	6_8binneg	0.9546	0.9402
repeatMasked_beaf32_Modencode_Negre_et_al.2.repeatMasked	BEAF-32	0-12hr	4_6DN	0.7683	0.7709
repeatMasked_beaf32_Modencode_Negre_et_al.6.repeatMasked	BEAF-32	0-12hr	2_4WE	0.7563	0.7540
repeatMasked_cp190_Modencode_Negre_et_al.0.repeatMasked	Cp190	0-12hr	10_12DN	0.8452	0.8064
repeatMasked_cp190_Modencode_Negre_et_al.3.repeatMasked	Cp190	0-12hr	4_6DN	0.8088	0.8019
repeatMasked_ctcf_Modencode_Negre_et_al.1.repeatMasked	CTCF	0-12hr	10_12neuro	0.8210	0.7612
repeatMasked_dfd_Modencode_Boyle_et_al.1.repeatMasked	Dfd	0-8hr	4_6DN	0.9643	0.9519

repeatMasked_disco_Modencode_Negre_et_al.3.repeatMasked	disco	0-8hr, 8-16hr	4_6DN	0.8747	0.8708
repeatMasked_eve_Modencode_Boyle_et_al.0.repeatMasked	eve	1-6hr, 8-16hr	4_6WE	0.9489	0.9462
repeatMasked_hr78_Modencode_Negre_et_al.4.repeatMasked	Hr78	8-16hr	8_10meso	0.9764	0.9688
repeatMasked_kni_Modencode_Negre_et_al.3.repeatMasked	kni	8-16hr	10_12WE	0.9608	0.9412
repeatMasked_pnr_FurlongLab.3.repeatMasked	pnr	4-6hr, 6-8hr	4_6WE	0.9595	0.9625
repeatMasked_trl_Modencode_Boyle_et_al.3.repeatMasked	Trl	8-16hr	8_10WE	0.8642	0.8406
repeatMasked_usp_Modencode_Boyle_et_al.0.repeatMasked	usp	0-12hr	8_10WE	0.9572	0.9561

Supplementary table 9: Performance of PWM-specific vs generic logistic regression models in predicting TF binding.

References

1. Kornberg, R. D. Chromatin structure: A repeating unit of histones and DNA. *Science (80-.).* **184,** 868–871 (1974).

2. Zhang, T., Cooper, S. & Brockdorff, N. The interplay of histone modifications - writers that read. *EMBO Rep.* (2015). doi:10.15252/embr.201540945

3. Horn, P. J. & Peterson, C. L. Chromatin higher order folding: Wrapping up transcription. *Science (80-.).* **297,** 1824–1827 (2002).

4. Heitz, E. Das Heterochromatin der Moose. *Jahrbücher für wissenschaftliche Bot.* **69,** 762–818 (1928).

5. Pederson, D. S., Thoma, F. & Simpson, R. T. Core Particle, Fiber, and Transcriptionally Active Chromatin Structure. *Annu. Rev. Cell Biol.* **2,** 117–147 (1986).

6. Dillon, N. Heterochromatin structure and function. *Biol. Cell* **96,** 631–637 (2004).

7. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).

8. Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41,** 105–178 (2006).

9. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* **339,** 225–229 (2010).

10. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–286 (2014).

11. Romero, I. G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* **13,** 505–516 (2012).

12. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13,** 613–626 (2012).

13. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* **15,** 453–468 (2014).

14. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: Establishing competence for gene expression. *Genes Dev.* **25,** 2227–2241 (2011).

15. Pabo, C. O. & Sauer, R. T. Transcription Factors: Structural Families and Principles of DNA Recognition. *Annu. Rev. Biochem.* **61,** 1053–1095 (1992).

16. Rivera-Pomar, R. & Jäckle, H. From gradients to stripes in Drosophila embryogenesis: Filling in the gaps. *Trends Genet.* **12,** 478–483 (1996).

17. Gregor, T., Garcia, H. G. & Little, S. C. The embryo as a laboratory: Quantifying transcription in Drosophila. *Trends Genet.* **30,** 364–375 (2014).

18. Furlong, E. E. M., Andersen, E. C., Null, B., White, K. P. & Scott, M. P. Patterns of gene expression during Drosophila mesoderm development. *Science (80-.).* **293,**

1629–1633 (2001).

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

20. Bour, B. A. *et al.* Drosophila MEF2, a transcription factor that is essential for myogenesis. *Genes Dev.* **9,** 730–741 (1995).

21. Azpiazu, N. & Frasch, M. Tinman and bagpipe: Two homeo box genes that determine cell fates in the dorsal mesoderm of Drosophila. *Genes Dev.* **7,** 1325–1340 (1993).

22. Zaffran, S., Küchler, A., Lee, H. H. & Frasch, M. biniou (FoxF), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in Drosophila. *Genes Dev.* **15,** 2900–2915 (2001).

23. McDonald, J. A. *et al.* Dorsoventral patterning in the Drosophila central nervous system: The vnd homeobox gene specifies ventral column identity. *Genes Dev.* **12,** 3603–3612 (1998).

24. Hartenstein, V. & Wodarz, A. Initial neurogenesis in Drosophila. *Wiley Interdiscip. Rev. Dev. Biol.* **2,** (2013).

25. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–4 (1977).

26. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74,** 5463–5467 (1977).

27. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).

28. Prober, J. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (80-.).* **238,** 336–341 (1987).

29. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

30. Venter, J. C. *et al.* The sequence of the human genome. *Science (80-.).* **291,** 1304–1351 (2001).

31. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Molecular Cell* **58,** 586–597 (2015).

32. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nature Biotechnology* **26,** 1113–1115 (2008).

33. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).

34. Guo, J. *et al.* Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Natl. Acad. Sci.* **105,** 9145–9150 (2008).

35. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

36. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53,** 937–947 (1988).

37. Gilmour, D. S. & Lis, J. T. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 4275–9 (1984).

38. O'Geen, H., Echipare, L. & Farnham, P. J. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods Mol. Biol.* **791,** 265–286 (2011).

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

40. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-.).* **316,** 1497–1502 (2007).

41. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).

42. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4,** 651–657 (2007).

43. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).

44. Furey, T. S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* **13,** 840–852 (2012).

45. Adli, M. & Bernstein, B. E. Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat. Protoc.* **6,** 1656–1668 (2011).

46. Shankaranarayanan, P. *et al.* Single-tube linear DNA amplification (LinDA) for robust ChIP-seq. *Nat. Methods* **8,** 565–567 (2011).

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

48. Bonn, S. *et al.* Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. *Nat. Protoc.* **7,** 978–994 (2012).

49. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44,** 148–156 (2012).

50. Gross, D. S. & Garrard, W. T. Nuclease Hypersensitive Sites in Chromatin. *Annu. Rev. Biochem.* **57,** 159–197 (1988).

51. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132,** 311–322 (2008).

52. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital

genomic footprinting. *Nat. Methods* **6,** 283–289 (2009).

53. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44,** D726–D732 (2016).

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

55. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

56. Kunitz, M. CRYSTALLINE DESOXYRIBONUCLEASE. *J. Gen. Physiol.* **33,** 363–377 (1950).

57. Suck, D. & Oefner, C. Structure of DNase I at 2.0 å resolution suggests a mechanism for binding to and cutting DNA. *Nature* **321,** 620–625 (1986).

58. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science (80-.).* **193,** 848–856 (1976).

59. Garel, A. & Axel, R. Selective digestion of transcriptionally active ovalbumin genes from oviduct nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **73,** 3966–3970 (1976).

60. Scott, W. A. & Wigmore, D. J. Sites in simian virus 40 chromatin which are preferentially cleaved by endonucleases. *Cell* **15,** 1511–1518 (1978).

61. Wu, C., M. Bingham, P., Livak, K. J., Holmgren, R. & Elgin, S. C. R. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. *Cell* **16,** 797–806 (1979).

62. Wu, C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286,** 854–860 (1980).

63. Keene, M. A., Corces, V., Lowenhaupt, K. & Elgin, S. C. DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc. Natl. Acad. Sci.* **78,** 143–146 (1981).

64. McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D. & Felsenfeld, G. A 200 base pair region at the 5′ end of the chicken adult β-globin gene is accessible to nuclease digestion. *Cell* **27,** 45–55 (1981).

65. Song, L. & Crawford, G. E. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **5,** (2010).

66. 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).

67. Munoz-Lopez, M. & Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genomics* **11,** 115–128 (2010).

68. Berg, D. E., Davies, J., Allet, B. & Rochaix, J.-D. Transposition of R factor genes to

bacteriophage lambda. *Proc. Natl. Acad. Sci. U. S. A.* **72,** 3628–3632 (1975).

69. Reznikoff, W. S. Transposon Tn5. *Annu. Rev. Genet.* **42,** 269–286 (2008).

70. Goryshin, I. Y. & Reznikoff, W. S. Tn5 in vitro transposition. *J. Biol. Chem.* **273,** 7367–7374 (1998).

71. Wiegand, T. W. & Reznikoff, W. S. Characterization of two hypertransposing Tn5 mutants. *J. Bacteriol.* **174,** 1229–1239 (1992).

72. Zhou, M., Bhasin, A. & Reznikoff, W. S. Molecular genetic analysis of transposase-end DNA sequence recognition: Cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn5 transposase. *J. Mol. Biol.* **276,** 913–925 (1998).

73. Syed, F., Grunenwald, H. & Caruccio, N. Next-generation sequencing library preparation: Simultaneous fragmentation and tagging using in vitro transposition. *Nat. Methods* **6,** i–ii (2009).

74. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11,** (2010).

75. Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S. J. & Craig, N. L. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21966–72 (2010).

76. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).

77. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-.).* **348,** 910–914 (2015).

78. Jin, W. *et al.* Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528,** 142–146 (2015).

79. Ewing, B., Hillier, L. D., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).

80. Reinert, K., Langmead, B., Weese, D. & Evers, D. J. Alignment of Next-Generation Sequencing Reads. *Annu. Rev. Genomics Hum. Genet.* **16,** 133–151 (2015).

81. Burkhardt, S. *et al.* q-gram based database searching using a suffix array (QUASAR). in *Proceedings of the Third Annual International Conference on Computational Molecular Biology - RECOMB '99* 77–83 (1999). doi:10.1145/299432.299460

82. Abouelhoda, M. I., Kurtz, S. & Ohlebusch, E. Replacing suffix trees with enhanced suffix arrays. *J. Discret. Algorithms* **2,** 53–86 (2004).

83. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Proceedings 41st Annual Symposium on Foundations of Computer Science* 390–398 (2000). doi:10.1109/SFCS.2000.892127

84. Burrows, M. & J. Wheeler, D. A Block-Sorting Lossless Data Compression Algorithm. *Digit. Syst. Res. Cent. Res. Reports* (1994).

85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

87. Mahony, S. & Pugh, B. F. Protein–DNA binding in high-resolution. *Crit. Rev. Biochem. Mol. Biol.* **50,** 269–283 (2015).

88. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

89. Ibrahim, M. M., Lacadie, S. A. & Ohler, U. JAMM: A peak finder for joint analysis of NGS replicates. *Bioinformatics* **31,** 48–55 (2015).

90. Banfield, J. D. & Raftery, A. E. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics* **49,** 803–821 (1993).

91. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).

92. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

93. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

94. Orenstein, Y. & Shamir, R. Modeling protein-DNA binding via high-throughput in vitro technologies. *Brief. Funct. Genomics* **16,** 171–180 (2016).

95. Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the 'perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.* **10,** 2997–3011 (1982).

96. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188,** 415–431 (1986).

97. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quant. Biol.* **1,** 115–130 (2013).

98. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24,** 1429–1435 (2006).

99. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20,** 861–873 (2010).

100. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **43,** 117–122 (2015).

101. Portales-Casamar, E. *et al.* JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **38,** 105–110 (2010).

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

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

104. Megraw, M., Pereira, F., Jensen, S. T., Ohler, U. & Hatzigeorgiou, A. G. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res.* **19,** 644–656 (2009).

105. Novak, R. L. Deoxyribonuclease resistance of DNA-RNA polymerase complexes. *BBA Sect. Nucleic Acids Protein Synth.* **149,** 593–595 (1967).

106. 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).

107. Zinn, K. & Maniatis, T. Detection of factors that interact with the human β-interferon regulatory region in vivo by DNAase I footprinting. *Cell* **45,** 611–618 (1986).

108. Jackson, P. D. & Felsenfeld, G. A method for mapping intranuclear protein-DNA interactions and its application to a nuclease hypersensitive site. *Proc. Natl. Acad. Sci. U. S. A.* **82,** 2296–2300 (1985).

109. Vierstra, J. & Stamatoyannopoulos, J. A. Genomic footprinting. *Nat. Methods* **13,** 213–221 (2016).

110. Sung, M.-H., Baek, S. & Hager, G. L. Genome-wide footprinting: ready for prime time? *Nat. Methods* **13,** 222–228 (2016).

111. Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods* **13,** 303–309 (2016).

112. Cuellar-Partida, G. *et al.* Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28,** 56–62 (2012).

113. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32,** 171–178 (2014).

114. Yardimci, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* **42,** 11865–11878 (2014).

115. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11,** 73–78 (2013).

116. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489,** 83–90 (2012).

117. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21,** 456–464 (2011).

118. Piper, J. *et al.* Wellington: A novel method for the accurate identification of digital

genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** (2013).

119. Sung, M. H., Guertin, M. J., Baek, S. & Hager, G. L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol. Cell* **56,** 275–285 (2014).

120. Gusmao, E. G., Dieterich, C., Zenke, M. & Costa, I. G. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* **30,** 3143–3151 (2014).

121. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21,** 447–455 (2011).

122. Kähärä, J. & Lähdesmäki, H. BinDNase: A discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics* **31,** 2852–2859 (2015).

123. Quach, B. & Furey, T. S. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics* **33,** 956–963 (2017).

124. Chen, X., Yu, B., Carriero, N., Silva, C. & Bonneau, R. Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res.* **45,** 4315–4329 (2017).

125. Raj, A. *et al.* msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding. *PLoS One* **10,** (2015).

126. Luo, K. & Hartemink, A. J. Using DNase digestion data to accurately identify transcription factor binding sites. *Pac. Symp. Biocomput.* 80–91 (2013).

127. Lazarovici, A. *et al.* Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci.* **110,** 6376–6381 (2013).

128. Koohy, H., Down, T. A. & Hubbard, T. J. Chromatin Accessibility Data Sets Show Bias Due to Sequence Specificity of the DNase I Enzyme. *PLoS One* **8,** (2013).

129. Lu, Z., Hofmeister, B. T., Vollmers, C., DuBois, R. M. & Schmitz, R. J. Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. *Nucleic Acids Res.* **45,** (2017).

130. Baek, S., Goldstein, I. & Hager, G. L. Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity. *Cell Rep.* **19,** 1710–1722 (2017).

131. Pranzatelli, T. J. F., Michael, D. G. & Chiorini, J. A. ATAC2GRN: Optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. *BMC Genomics* **19,** (2018).

132. Goryshin, I. Y., Miller, J. A., Kil, Y. V., Lanzov, V. A. & Reznikoff, W. S. Tn5/IS50 target recognition. *Proc. Natl. Acad. Sci.* **95,** 10716–10721 (1998).

133. Madrigal, P. On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. *Front. Bioeng. Biotechnol.* **3,** 144 (2015).

134. Martins, A. L., Walavalkar, N. M., Anderson, W. D., Zang, C. & Guertin, M. J. Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions. *Nucleic Acids Res.* **46,** (2018).

135. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10,** 325–327 (2013).

136. Sandmann, T., Jakobsen, J. S. & Furlong, E. E. M. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. *Nat. Protoc.* **1,** 2839–2855 (2006).

137. Cannavò, E. *et al.* Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541,** 402–406 (2017).

138. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLoS Comput. Biol.* **10,** (2014).

139. Montefiori, L. *et al.* Reducing mitochondrial reads in ATAC-seq using CRISPR/Cas9. *Sci. Rep.* **7,** (2017).

140. Badenhorst, P., Finch, J. T. & Travers, A. A. Tramtrack co-operates to prevent inappropriate neural development in Drosophila. *Mech. Dev.* **117,** 87–101 (2002).

141. Badenhorst, P. Tramtrack controls glial number and identity in the Drosophila embryonic CNS. *Development* **128,** 4093–4101 (2001).

142. Reuter, R. The gene serpent has homeotic properties and specifies endoderm versus ectoderm within the Drosophila gut. *Development* **120,** 1123–1135 (1994).

143. Romani, S. *et al.* Kruppel, a Drosophila segmentation gene, participates in the specification of neurons and glial cells. *Mech. Dev.* **60,** 95–107 (1996).

144. Yang, X., Bahri, S., Klein, T. & Chia, W. Klumpfuss, a putative Drosophila zinc finger transcription factor, acts to differentiate between the identities of two secondary precursor cells within one neuroblast lineage. *Genes Dev.* **11,** 1396–1408 (1997).

Acknowledgements

First of all, I'd like to thank my supervisor, Prof. Uwe Ohler, for giving me the chance to work in his laboratory and for his support throughout this PhD. Thanks also go to the members of the Ohler laboratory, for creating a nice atmosphere to work in.

I'd like to thank Dr. James Reddington, Dr. David Garfield, Alexander Glahs and Dr. Sabrina Krueger, for being great collaborators, as well as amazing people. Thanks also go of course, to Dr. Robert Zinzen and Prof. Eileen Furlong, doing joint work with their laboratories has been a great experience.

This thesis has been written in three different countries, amongst changes and constantly being on the move, and well, suffice it to say that it has been challenging at times. Here's to all the people that have been my support circle:

The gang of building 89! I'm afraid to say all your names in case I forget anyone. But a special thanks go to Doro for reading and translating my abstract. I'd also like to thank Rebecca for helping me with the manuscript (and of course her friendship, and always being kind and helpful).

A very big special thanks go to my dear friend Basak, I am so grateful to have known you! Your presence in my life is much appreciated (and also the fact that you're about to print and submit this thesis on my behalf! I don't even know how to begin to thank you…). I'm also thankful to my friend Yesim for kindly offering help with publishing this thesis.

Very close friends that made life in Berlin beautiful: especially Ricardo, Jackie, Nese, Can!

My friends since decades (sisters at this point): Leyla, Zeynep, Ezgi and Duygu!

None of this would have been possible without the support of my amazing family: mom, dad and brother. Lots of love to you all, Karabacak family (or should I have said Famiglia della Gambanera)!

And last but not least: I was fortunate enough to find my partner in life, love, crime, humor and ridiculousness. Like you said in your acknowledgements a year ago: let's crack this code of life together, Lore. I love you!