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

Authors

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

Affiliation

[1] Bioinformatics Core Unit (BCU), Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany

[2] Department of Cardiac Development and Remodeling, Max-Planck-Institute for Heart and Lung Research, Bad Nauheim, Germany

[3] German Centre for Cardiovascular Research (DZHK), Partner site Rhein-Main, Frankfurt am Main, 60596 Germany

Corresponding author email address

mario.looso@mpi-bn.mpg.de, @loosolab

Abstract

While footprinting analysis of ATAC-seq data can theoretically enable investigation of transcription factor (TF) binding, the lack of a computational method implementing both footprinting, visualization and downstream analysis has hindered the widespread application of this method. Here we present TOBIAS, a comprehensive footprinting framework enabling genome-wide investigation of TF binding dynamics for hundreds of TF simultaneously. As a proof-of-concept, we illustrate how TOBIAS can unveil complex TF dynamics during zygotic genome activation (ZGA) in both humans and mice, and explore how the TF Dux activates cascades of TF, binds to repeat elements and induces expression of novel genetic elements. TOBIAS is freely available at: https://github.com/loosolab/TOBIAS.

Keywords

Footprinting, ATAC-seq, epigenetics, transcription factors, ZGA, Dux

Background

Epigenetic mechanisms governing chromatin organization and transcription factor (TF) binding are critical components of transcriptional regulation and cellular transitions. In recent years, rapid improvements of pioneering sequencing methods such as ATAC-seq (Assay of Transposase Accessible Chromatin) [1], have allowed for systematic, global scale investigation of epigenetic mechanisms controlling gene expression. While ATAC-seq can uncover accessible regions where TFs might bind, true identification of specific TF binding sites (TFBS) still relies on chromatin immunoprecipitation methods such as ChIP-seq. However, ChIP-seq methods require high input cell numbers, are limited to one TF per assay, and are further restricted to TFs for which antibodies are readily available. Latest improvements of ChIP based methods [2] can circumvent some of these technical drawbacks, but the limitation of only being able to identify binding sites of one TF per assay persists. Therefore, it remains costly, or even impossible, to study the binding of multiple TFs in parallel.

The limitations of investigating TF binding become particularly apparent when investigating processes involving a very limited number of cells such as preimplantation development (PD) of early zygotes. PD encompasses the transformation of the fertilized egg that forms the zygote, which subsequently undergoes a series of cell divisions to finally constitute the blastocyst, a structure built by the inner cell mass (ICM) and trophectoderm (Figure 1a). Following fertilization, maternal and paternal mRNAs are degraded prior to zygotic genome activation (ZGA) (reviewed in [3]), which leads to the transcription of thousands of genes [4]. Integration of multiple omics-based profiling methods have revealed a set of key TFs that are expressed at the onset of and during ZGA including Dux [5, 6], Zscan4 [7], and other homeobox-containing TFs [8]. However, what genetic elements they directly bind to and/or regulate during PD remains poorly understood. Consequently, the global network of TF binding dynamics throughout PD remains almost entirely obscure.

A computational method known as *digital genomic footprinting* (DGF) [9] has emerged as an alternative means, which can overcome some limitations of investigating TF binding with ChIP-based methods. DGF is a computational analysis of chromatin accessibility assays such as ATAC-seq, which makes use of the intrinsic effect that DNA effector enzymes only cut accessible DNA regions. Similarly to nucleosomes, bound TFs hinder cleavage of DNA, resulting in defined regions of decreased signal strength within larger regions of high signal - known as *footprints* [10] (Figure 1b). This concept shows considerable potential as it theoretically allows to survey genome-wide binding of multiple TFs in parallel from a single experiment.

However, there are still a multitude of challenges to DGF methods [11, 12]. While ATAC-seq became very popular as it is simpler and require less starting material in comparison to DNAse-seq, only a few of the existing footprinting tools inherently support ATAC-seq analyses [13-16]. In this context the non-random behavior of cleavage enzymes that bind preferentially to certain sequence compositions (e.g. Tn5 bias for ATAC-seq) turned out to be a major challenge [17-20]. In addition, computational issues such as software availability, the use of non-standard file-formats, varying dependencies and lack of support for multiprocessing have made current footprinting tools hard to integrate into existing analysis pipelines. Aside from the identification of footprints, the challenge of integrating footprints, TF motifs and genomic location of genes to be able to fully investigate the epigenetic processes involving TF binding is not a trivial task.

While all of these factors significantly influence the outcome of footprinting analyses, previous investigations have been focused on improving individual computational steps such as estimating differential TF binding on a global scale [21-23], identifying footprints for specific TFs in a local genomic context [16, 24], or correcting the bias within the genomic signals [25, 26]. Few methods have included bias correction as an integrated part of footprint detection [16]. Essentially, a comprehensive framework that takes all of these parameters into account does not exist.

Here we describe and exploit application of TOBIAS (**T**ranscription factor **O**ccupancy prediction **B**y **I**nvestigation of **A**TAC-seq **S**ignal), a comprehensive computational framework that we created for footprinting analysis (Figure 1c). TOBIAS is a collection of computational tools utilizing a minimal input of ATAC-seq reads (.bam-format), TF motif information (in the form of PWMs) and genome information to enable Tn5 bias correction, footprinting, and comparison of TF binding even for complex experimental designs (e.g. time series). Furthermore, TOBIAS includes a variety of modules for downstream analysis such as TF network inference and visualization of footprints. In addition to the TOBIAS Python package, we provide scalable analysis workflows implemented in Snakemake [27] and NextFlow [28], including a cloud computing compatible version making use of the de.NBI cloud [29].

Results

Validation of TOBIAS

As a comprehensive framework for DGF analysis comparable to TOBIAS does not exist, we rated the individual TOBIAS modules in a well-studied system of paired ATAC-seq and ChIP-seq datasets (see Methods; Validation) against published methods where possible. In terms of Tn5 bias correction, we found that TOBIAS outperforms other tools in distinguishing between bound/unbound sites (Supp. Figure 1a, Supp. File 1). For detection of footprints, we also found that TOBIAS clearly outperforms other known methods capable of screening TFs in parallel (Supp. Figure 1b left and methods). By making use of another exemplary dataset of ATAC-seq data derived from hESC [30], we confirmed the obvious improvement of footprint detection after Tn5 bias correction (Supp. Figure 2a left). Importantly, we also identified a number of cases where the TF motif itself is a disfavored position for Tn5 integration, thereby creating a false-positive footprint if left uncorrected, which disappears after Tn5 bias correction (Supp. Figure 2a; right). Utilizing a footprint metric as described by [22] (Supp. Figure 2b) across different stages of Tn5 bias correction (uncorrected/expected/corrected signals), we found a high correlation between uncorrected and expected footprinting depths (Supp. Figure

2c). In contrast, this effect vanished after TOBIAS correction (Supp. Figure 2d), indicating the gain of a real footprint information superimposed by Tn5 bias. In a global perspective, taking 590 TFs into account, TOBIAS generated a measurable footprint for 64% of the TFs (Supp. Figure 2e). This is in contrast to previous reports wherein it has been suggested that only 20% of all TFs leave measurable footprints [22]. To summarize, we found that TOBIAS exceeded other software solutions in terms of correctly identifying bound TF binding sites.

Footprinting uncovers transcription factor binding dynamics in mammalian ZGA

To demonstrate the full potential of TOBIAS, in particular in the investigation of processes involving only few cells, we analyzed a series of ATAC-seq datasets derived from both human and murine preimplantation embryos at different developmental stages ranging from 2C, 4C, 8C to ICM in addition to embryonic stem cells of their respective species [30] [31]. Altogether, TOBIAS was used to calculate footprint scores for a list of 590 and 464 individual TFs across the entire process of PD of human and mouse embryos, respectively. After clustering TFs into co-active groups within one or multiple developmental timepoints (Human: Figure 2a and Supp. Table 1; Mouse: see next section), we first asked whether the predicted timing of TF activation reflects known processes in human PD. Intriguingly, we found 10 defined clusters of specific binding patterns, the majority of which peaked between 4C and 8C, fully concordant with the transcriptional burst and termination of ZGA (Figure 2a).

Two clusters of TFs (Cluster 1+2; n=83) displayed highest activity at the 2-4C stage and strongly decreased thereafter, suggesting that factors within these clusters are likely involved in ZGA initiation. We set out to classify these TFs, and observed a high overlap with known maternally transferred transcripts [32] (LHX8, BACH1, EBF1, LHX2, EMX1, MIXL1, HIC2, FIGLA, SALL4, ZNF449), explaining their activity before ZGA onset. Importantly, DUX4 and DUXA, which are amongst the earliest expressed genes during ZGA [5, 6], were also contained in these clusters. Additional TFs included HOXD1, which is known to be expressed in human unfertilized oocytes and preimplantation embryos [33] and ZBTB17, a TF mandatory to generate viable embryos [34]. Cluster 6 (n=67) displayed a particularly prominent 8C specific signature, that harbored well known TFs involved in lineage specification such as PITX1, PITX3, SOX8, MEF2A, MEF2D, OTX2, PAX5 and NKX3.2. Furthermore, overlapping TFs within Cluster 6 with RNA expression datasets ranging from the germinal vesicle to cleavage stage [5], 12 additional TFs (FOXJ3, HNF1A, ARID5A, RARB, HOXD8, TBP, ZFP28, ARID3B, ZNF136, IRF6, ARGFX, MYC, ZSCAN4) were confirmed to be exclusively expressed within this time frame. Taken together, these data show that TOBIAS reliably uncovers massively parallel TF binding dynamics at specific time points during early embryonic development.

Transcription factor scores correlate with footprints and gene expression

To confirm that TOBIAS-based footprinting scores are indeed associated with leaving *bona fide* footprints we utilized the ability to visualize aggregated footprint plots as implemented within the framework. Indeed, bias corrected footprint scores were highly congruent with explicitly defined footprints (Figure 2b) of prime ZGA regulators at developmental stages in which these have been shown to be active [7]. For example, footprints associated with DUX4, a master inducer of ZGA, were clearly visible from 2C-4C, decreased from 8C onwards and were completely lost in later stages, consistent with known expression levels [30] and ZGA onset in humans. Footprints for ZSCAN4, a primary DUX4 target [5], were exclusively visible at the 8C stage. Interestingly, GATA2 footprints were exhibited from 8C to ICM stages which is in line with its known function in regulating trophoblast differentiation [35]. As expected, CTCF creates footprints across all timepoints. Strikingly, we observed that these defined footprints were not detectable without TOBIAS mediated Tn5 bias correction (Supp. Figure 2f). These data show that footprint scores can be reliably confirmed by footprint visualizations, which further allow to infer TF binding dynamics.

To test if the global footprinting scores of individual TFs correlate with the incidence and level of their RNA expression, we matched them to RNA expression datasets derived from individual timepoints throughout zygotic development, taking TF motif similarity into account. Indeed, we found that TOBIAS scores for the majority of TFs either correlated well with the timing of their expression profiles or displayed a slightly delayed activity after expression peaked (Supp. Figure 3a). This is important because it shows that in conjunction with expression data, TOBIAS can unravel the kinetics between TF expression (mRNA) and the actual binding activity of their translated proteins. The value of this added information becomes particularly apparent when analyzing activities of TFs that did not correlate with the timing of their RNA expression (Supp. Figure 3a; not correlated).

For example, within the non-correlated cluster 13 TFs were identified which are of putative maternal origin [32] including SALL4. In mice, Sall4 protein is maternally contributed to the zygote, subsequently degraded at 2C and then reexpressed after zygotic transcription has initiated [36]. Consistent with this, SALL4 expression increased dramatically from 8C onwards (Supp. Table 2). Notably, TOBIAS predicted SALL4 to have the highest activity in 2C and second-highest activity in hESC (on-off-on-pattern). These data show that TOBIAS can predict true on-off-on-patterns, and can infer significant insight into TF activities, in particular for those where determining their expression patterns alone does not suffice to explain when they exert their biological function.

Differential footprint analysis reveals functional divergence between human and mouse ZGA

The timing of ZGA varies between mice (2C) and humans (4C to 8C) (reviewed in [37] [38]). By integrating the TOBIAS scores from human and mouse (Supp. Figure 3b and Supp. Table 3), and instrumentalizing the capability of TOBIAS to generate differential TF binding plots for all time points automatically, we investigated similarities and differences of PD between these species. Firstly, reflecting the shift of ZGA onset, we identified 30 TFs which appeared to be

ZGA specific in both human and mouse (Figure 2c) including several homeobox factors which already have described functions within ZGA [39] as well as ARID3A which has been shown to play a role in cell fate decisions in creating trophectoderm [40].

Next, we used the differential TF binding plots to display differences in ZGA at the transition between 2C and 4C in mouse (Figure 3a), and human 8C and ICM (Figure 3b) (Supp. File 2 + 3 for all pairwise comparisons). In mice, we observed a shift of Obox-factor activity in 2C to an activation of Tead (Tead1-4) and AP-2 (Tfap2a/c/e) motifs in 4C. Notably, AP2/Tfap2c is required for normal embryogenesis in mice [41] and was also recently shown to act as a chromatin modifier that opens enhancers proximal to pluripotency factors in human [42]. We observed a similar shift of TF activity for homeobox factors such as PITX1-3, RHOXF1, CRX and DMBX1 at the human 8C stage towards higher scores in ICM for known pluripotency factors such as POU5F1 (OCT4) and other POU-factors. Taken together, these results highlight the ability of TOBIAS to capture differentially bound TFs, not only across the whole timeline, but also between individual conditions and species.

Throughout the pairwise comparisons, we observed that TFs from the same families often display similar binding kinetics within species, which is not surprising since they often possess highly similar binding motifs (Figure 3a right). To characterize TF similarity, TOBIAS provides functionality to cluster TFs based on the overlap of TFBS within investigated samples (Figure 3c+3d). This enables quantification of the similarity and clustering of individual TFs that appear to be active at the same time. Thereby, we observed a group of homeobox motifs which cluster together with more than 50% overlap of their respective binding sites in mouse (Figure 3c). In contrast, other TFs such as Tead and AP-2 cluster separately, indicating that these factors utilize independent motifs (Supp. File 2+3). While this might appear trivial, this clustering of TFs in fact also highlights differences in motif usage between human and mouse. One prominent example is the RHOXF1 motif, which shows high binding-site overlap with Obox 1/3/5 and Otx2 binding sites in mouse (Figure 3c; ~60% overlap), but does not cluster with OTX2 in human (Figure 3d; ~35% overlap). This observation suggests important functional differences of RHOX/Rhox TFs between mice and humans. In support of this hypothesis RHOXF1, RHOXF2 and RHOXF2B genes are exclusively expressed at 8C and ICM in humans, whereas Rhox factors are not expressed in corresponding developmental stages of preimplantation in mouse (Supp. Table 4). Conceivably, this observation, together with the finding that murine Obox factors share the same motif as RHOX-factors in humans, suggests that Obox TFs might function similarly to RHOX-factors during ZGA. Altogether, the TOBIAS mediated TF clustering based on TFBS overlap allows for quantification of target-similarity and divergence of TF function between motif families.

Dux expression induces massive changes of chromatin accessibility, transcription and TF networks

We became particularly attracted to Dux/DUX4 which TOBIAS correctly predicted to be one of the earliest factors to be active in both human and mouse (Figure 2a and Supp. Figure 3b)[5-7, 43, 44]. Despite its prominent role in ZGA, there is however still a poor understanding of how Dux regulates its primary downstream targets, and consequently its secondary targets, during this process. We therefore applied TOBIAS to identify Dux binding sites utilizing an ATAC-seq dataset of Dux overexpression (DuxOE) in mESC [5].

Inspecting the differential TF activity predicted by TOBIAS, we observed an increase of activity of Dux, Obox and other homeobox-TFs as expected (Figure 4a, Supp. File 4). Interestingly, this went along with a massive loss of TF binding for pluripotency markers such as Nanog, Pou5f1 (OCT4) and Sox2 upon DuxOE, indicating that Dux renders previously accessible chromatin sites associated with pluripotency inaccessible.

Consistently, Dux footprints (Figure 4b; left) were clearly evident upon DuxOE. Importantly, TOBIAS discriminated ~30% of all potential binding sites within open chromatin regions to be bound in the DuxOE condition further demonstrating the specificity of this method (Figure 4b; right). To rank the biological relevance of the individually changed binding sites between control and DuxOE conditions, we linked all annotated gene loci to RNA expression. A striking correlation between the gain-of-footprint and gain-of-expression of corresponding loci was clearly observed and mirrored by the TOBIAS predicted bound/unbound state (Figure 4c). Amongst the genes within the list of bound Dux binding sites (Supp. Table 5 for full Dux target list) were well known Dux targets including *Zscan4c* and *Pramef25* [45], for which local footprints for Dux were clearly visible (Figure 4d). The high resolution of footprints is particularly pronounced for *Tdpoz1* which harbors two potential Dux binding sites of which one is clearly footprinted in the score track, while the other is predicted to be unoccupied (Figure 4d; bottom). In line with this, *Tdpoz1* expression is significantly upregulated upon DuxOE as revealed by RNA-seq (log2FC: 6,95). Consistently, *Tdpoz1* expression levels are highest at 2C in zygotes and decrease thereafter, strongly indicating that *Tdpoz1* is likely a direct target of Dux during PD both *in vitro* and *in vivo* [31, 46] (Supp. Table 5). Footprinting scores also directly correlated with ChIP-seq peaks for Dux in the Tdpoz1 promoter (Supp. Figure 4a), an observation which we also found at many other positions (Examples shown in Supp. Figure 4b+c).

Many of the TOBIAS-predicted Dux targets encode TFs themselves. Therefore, we applied the TOBIAS network module to subset and match all activated binding sites to TF target genes with the aim of inferring how these TF activities might connect. Thereby, we could model an intriguing pseudo timed TF activation network. This directed network uncovered a TF activation cascade initiated by Dux, resulting in the activation of 7 primary TFs which appear to subsequently activate 32 further TFs (first three layers depicted in Figure 4e). As Dux is a regulator of ZGA, we asked how the *in vitro* activated Dux network compared to gene expression throughout PD *in vivo*. Strikingly, the in vivo RNA-seq data of the resolved developmental dataset [31] confirmed an early 2C specific expression for Dux, followed by a slightly shifted activation pattern for all direct Dux targets except for *Rxrg* (Figure 4f). However, it is of note that *Rxrg* is significantly upregulated in the *in vitro* DuxOE from which the network is inferred (Supp. Table 5), pointing to both the similarities and differences between the *in vivo* 2C and *in vitro* 2C-like stages induced by Dux. In conclusion, these data show that beyond identifying specific target genes of individual TFs, TOBIAS can infer biological insight by predicting entire TF activation networks.

Notably, many of the predicted Dux binding sites (40%) are not annotated to genes (Figure 4g), raising the question what role these sites play in ZGA. Dux is known to induce expression of repeat regions such as LTRs [5] and consistently, we found that more than half of the DUX-bound sites are indeed located within known LTR sequences (Figure 4g) which were transcribed both *in vitro* and *in vivo* (Figure 4h). Interestingly, we found that 28% of all Dux binding sites overlap with genomic loci encoding LINE1 elements. Although LINE1 expression does not appear to be altered in mESC cells, there is a striking pattern of increasing LINE1 transcription from 4C-8C (Figure 4h) *in vivo*, pointing to a possible role of LINE1 regulation throughout PD. Finally, we found that 6% of the Dux binding sites do not overlap with any annotated gene nor with putative regulatory repeat sequences, even though transcription clearly occurs at these sites (Figure 4h bottom). One example is a predicted Dux binding site on chromosome 13, which coincides with a spliced region of increased expression between control mESC/DuxOE and comparable high expression in 2C, 4C and 8C (Supp. Figure 5). These data clearly indicate the existence of novel transcribed genetic elements, the function of which remains unknown, but which are likely controlled by Dux and could play a role during PD.

In conclusion, TOBIAS predicted the exact locations of Dux binding in promoters of target genes, and could unveil how Dux initiates TF-activation networks and induces expression of repeat regions. Importantly, these data further show that TOBIAS can identify any TFBS with increased binding, not only those limited to annotated genes, which aids in uncovering novel regulatory genetic elements.

Discussion

Footprint scores reveal true characteristics of protein binding

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

The power of this framework to handle time-series data becomes especially apparent when correlating the TOBIAS-based prediction of TF binding to RNA-seq data from the same time points. For instance, TOBIAS could infer when the maternally transferred TF SALL4 is truly active while its gene expression pattern alone does not allow to make such conclusions. Along this line, TOBIAS is also powerful in circumstances where gene expression of a particular TF appears to be anticorrelated with its binding activity. It is tempting to speculate that TFs for which footprinting scores are low, even though their RNA expression is high, might act as transcriptional repressors, because footprinting relies on the premise that TFs will increase chromatin accessibility around the binding site. In support of this hypothesis, recent investigations have suggested that repressors display a decreased footprinting effect in comparison to activators [23]. Therefore, the integration of ATAC-seq footprinting and RNA-seq is an important step in revealing additional information such as classification TFs into repressors and activators, as well as the kinetics between expression and binding.

Species-specific TFs use common ZGA motifs in mice and human

By integration of human and murine TF activities using both differential footprinting and species-specific TFBS overlaps, our analyses revealed that the majority of TF motifs are active at corresponding timepoints of human and mouse ZGA. This is not necessarily surprising since homologous TFs that exert the same functions usually use similar motifs (e.g Pou2f1/POU2F1, Otx1/OTX1 and/or Foxa3/FOXA3). Interestingly though, we found that this is not the case for all TF motifs. In this context, we found that the human RHOXF1 motif (Figure 2b) is likely not utilized by Rhox proteins in mice even though more than 30 Rhox genes exists. Evidently, throughout multiple duplications, Rhox genes seem to have obtained other functionalities in mouse [47] in comparison to the two human RHOX genes that are expressed in reproductive tissues [48]. Therefore, although we found the human RHOXF1 motif to be highly active in mice, this motif is most likely utilized by other proteins such as the mouse specific Obox proteins. In support of this conclusion, expression patterns of Obox proteins appear to be tightly regulated during PD [49] ([31]). High expression of Obox 1/2/5/7 is observed from the zygote to 4C stage, while Obox3/6/8 are expressed and peak at later stages (Supp. Table 4). Notably, there is a significant sequence similarity of the homeobox domains but not in the other parts of the RHOXF1 and Obox protein sequences, which supports the similarity in binding specificity. Although the potential functional overlap of RHOXF1 and Obox factors remains unresolved, our inter-species analysis suggests an unappreciated function of these factors and their targets during PD, clearly warranting an in depth investigation.

In the context of TF target prediction, the power of TOBIAS was particularly highlighted by the fact that the analysis could identify almost all known Dux targets. In addition to coding genes, our analysis disclosed novel Dux binding sites and significant footprint scores at LINE1 encoding genomic loci, which appear to be activated at the 4C/8C stage. This finding is especially interesting because a recent study has shown that LINE1 RNA can interact with Nucleolin and Kap1 to repress Dux expression [50]. Therefore, our findings give rise to a kinetics driven model in which Dux not only initiates ZGA but also regulates its own termination by a temporally delayed negative feedback loop. How this feedback loop is exactly controlled remains to be determined.

Limitations and outlook of footprinting analysis

Despite the striking capability of DGF analysis, some limitations and dependencies of this method still remain. Amongst these is the need of high-quality TF motifs for matching footprint scores to individual TFs with high confidence. In other words, while the binding of a TF might create an effect that can be interpreted as a footprint, without a known motif, this effect cannot be matched to the corresponding TF. This becomes evident in the context of DPPA2/4, a TF described by several groups to act in PD and even upstream of Dux [44]. DPPA2/4 targets GC rich sequences [44], but its canonical binding motif remains unknown. It also needs to be noted that footprinting analysis cannot take effects into account that arise from heterogeneous mixtures of cells wherein TFs are bound in some cells and in others not. Therefore, if not separated, the classification of differential binding will be an observation averaged across many cells, possibly masking subpopulation effects. Recent advances have enabled to perform ATAC-seq in single cells [51], but this generates sparse matrices, rendering footprinting approaches on single cells illusive. However, we speculate that by creating aggregated pseudo-bulk signals from large clustered SC ATAC datasets, DGF analysis might also become possible in single cells.

Conclusions

Here, we have illustrated the TOBIAS framework as a versatile tool for ATAC-seq analysis which helps to unravel transcription factor binding dynamics in complex experimental settings that are otherwise difficult to investigate. We showed that entire networks of TF binding, which have previously been explored using a combination of omics methods, can be recapitulated to a great extent by DGF analysis, which requires only ATAC-seq and TF motifs. From a global perspective, we provided new insights into PD by quantifying the stage-specific activity of specific TFs. Furthermore, we highlighted the usage of TOBIAS to study specific transcription factors as exemplified by our investigations on Dux. Finally, we used the specific TF target predictions to gain insights into the local binding dynamics of Dux in the context of TF-activation networks, repeat regions and novel genetic elements.

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The TOBIAS software is available on GitHub at: https://github.com/loosolab/TOBIAS. Excerpts of the data analyzed here are accessible for dynamic visualization at: http://loosolab.mpi-bn.mpg.de/tobias-meets-wilson. All raw data analyzed are available from GEO or ENCODE as described in Methods. The complete TOBIAS output for the analysis of the Dux overexpression dataset can be downloaded from: https://figshare.com/projects/Digital_Genomic_Footprinting_Analysis_of_ATAC-seq_dataset_from_preimplantation_timepoints_via_TOBIAS/69959.

Competing interests

None to declare

Funding

This work was funded by the Max Planck Society, the German Research Foundation (DFG), grant KFO309 (project number 284237345, epigenetics core unit) to ML, and by the Cardio-Pulmonary Institute (CPI), EXC 2026, Project ID: 390649896 to ML.

Authors' contributions

MB, CK, JK and ML wrote the manuscript. MB, PG, HS, AP, KK, AF and JP performed the bioinformatics analysis. JK, TB and ML directed, coordinated and supervised the work.

Acknowledgements

We would like to thank the IT-group at MPI-BN for continued support with IT-infrastructure. We would also like to thank Marius Dieckmann, the administrator of the Kubernetes cluster in Gießen, for his support and help in implementing the TOBIAS-Nextflow Cloud version.

Methods

Datasets

Organism	**Deposited data**	**Source**	**Identifier**
Mouse	ATAC-seq, RNA-seq and ChIP-seq from mESC control and Dux overexpression	[5]	GEO: GSE85632
Mouse	ATAC-seq and RNA-seq from various preimplantation stages	[31]	GEO: GSE66390

Human	ATAC-seq and RNA-seq from various preimplantation stages	[30]	GEO: GSE101571

For all public data sets used in this study (see table above), raw files were obtained from the European Nucleotide Archive [52] and processed as described in the methods section. See also methods section "Comparison of TOBIAS to existing methods" for links to the ENCODE data used for method validation.

Processing of ATAC-seq data

Raw sequencing fastq files were assessed for quality, adapter content and duplication rates with FastQC v0.11.7, trimmed using cutadapt [53] and aligned with STAR v2.6.0c [54] (parameters: "--alignEndsType EndToEnd --outFilterMismatchNoverLmax 0.1 --outFilterScoreMinOverLread 0.66 --outFilterMatchNminOverLread 0.66 --outFilterMatchNmin 20 --alignIntronMax 1 --alignSJDBoverhangMin 999 --alignEndsProtrude 10 ConcordantPair --alignMatesGapMax 2000 --outMultimapperOrder Random --outFilterMultimapNmax 999 --outSAMmultNmax 1") to either the mouse or human genome using Mus_musculus.GRCm38 or Homo_sapiens.GRCh38 versions from Ensembl [55]. Accessible regions were identified by peak calling for each sample separately using MACS2 (parameters: "--nomodel --shift -100 --extsize 200 --broad") [56]. Peaks from each sample were merged to a set of union peaks across all conditions using "bedtools merge". Each union peak was annotated to the transcriptional start site of genes (GENCODE [57]) in a distance of -10000/+1000 from the TSS using UROPA [58].

Processing of RNA-seq data

Raw reads were assessed for quality, adapter content and duplication rates with FastQC v0.11.7, trimmed using cutadapt [53] and aligned with STAR v2.6.0c [54] (parameters: "--outFilterMismatchNoverLmax 0.1 --outFilterScoreMinOverLread 0.9 --

outFilterMatchNminOverLread 0.9 --outFilterMatchNmin 20 --alignIntronMax 200000 --alignMatesGapMax 2000 --alignEndsProtrude 10 ConcordantPair --outMultimapperOrder Random --outFilterMultimapNmax 999") to either the mouse or human genome using Mus_musculus.GRCm38 or Homo_sapiens.GRCh38 versions from Ensembl [55]. Differentially expressed genes were identified using DESeq2 v1.22 [59]. Only genes with a minimum log2 fold change of ±1, a maximum Benjamini–Hochberg corrected P-value of 0.05 and a minimum combined mean of five reads were classified as significantly differentially expressed.

Processing of ChIP-seq data

Raw sequencing files in fastq format were quality assessed by Trimmomatic by trimming reads after a quality drop below a mean of Q15 in a window of 5 nucleotides [60]. All reads longer than 15 nucleotides were aligned versus the mouse genome version mm10, keeping just unique alignments (parameters: --outFilterMismatchNoverLmax 0.2 --outFilterScoreMinOverLread 0.66 --outFilterMatchNminOverLread 0.66 --outFilterMatchNmin 20 --alignIntronMax 1 --alignSJDBoverhangMin 999 --outFilterMultimapNmax 1 --alignEndsProtrude 10 ConcordantPair) by using the STAR mapper [54]. Read deduplication was done by Picard (http://broadinstitute.github.io/picard/).

Processing of transcription factor motifs

TF motifs were downloaded from JASPAR CORE 2018 [61], the JASPAR PBM HOMEO collection and Hocomoco V11 [62] databases. We further included the human ARGFX_3 motif from footprintDB [63] which originates from a HT-SELEX assay [64]. In annotation to the Dux/Dux4 motifs of JASPAR and Hocomoco, we also included two TF motifs for MDUX/DUX4 created using MEME-ChIP [65] with standard parameters on the ChIP-seq peaks of [45] (GSE87279).

JASPAR motifs were linked to Ensembl gene ids by mapping the provided "Uniprot id" to the "Ensembl gene id" through biomaRt [66]. Hocomoco motifs were likewise linked to genes through the provided HGNC/MGI annotation. Due to the redundancy of motifs between JASPAR and Hocomoco, we further filtered the TF motifs to one motif per gene, preferentially choosing motifs originating from mouse/human respectively. For each TOBIAS run, we created sets of expressed TFs as estimated from RNA-seq in the respective conditions. This amounted to 590 motifs for the dataset on human preimplantation stages, 464 motifs for the dataset on mouse preimplantation, and 459 for the DuxOE dataset.

Maternal genes

Maternal genes for human and mouse were downloaded from the REGULATOR database [32]. Entrez gene ids were converted to Ensembl gene ids using biomaRt [66] and subsequently matched to available TF motifs.

Overlap of Dux binding sites to repeat elements

Repeat elements for mm10 were downloaded from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/mm10/database/rmsk.gz). Overlap of Dux sites to individual repeat elements (as seen in figure 4G) was performed using "Bedtools intersect". The sum of overlaps were counted by repeat class (LINE1/LTR).

Visualization

All TF-score heatmaps were generated by R Version 3.5.3 and complex heatmap package version 3.6 [67]. Individual gene views were generated by loading TOBIAS output tracks into IGV version 2.6.2 [68] or using the svist4get visualization tool [69]. TF networks were drawn with Cytoscape version 3.7.1 [70]. Heatmaps of genomic signal density were generated using Deeptools version 3.3.0 [71]. All other figures, such as footprint plots, volcano plots and motif clustering dendrograms were generated by the TOBIAS visualization modules as described below.

The TOBIAS framework

In developing TOBIAS, we found that there were six main areas of DGF which had not been comprehensively addressed in the context of ATAC-seq footprinting analysis:

- All-in-one framework including bias correction, footprinting, quantification of protein binding and visualization
- Investigation of TF binding on a global level (which TFs are more bound globally) as well as the locus-specific level (which TF binds to which genomic locations including statistics on differential binding)
- Consideration of the redundancy and similarity of known TF binding motifs in the context of footprinting
- A scoring model for TF-DNA binding taking into account the potential lack of a canonical footprint effect
- Comparison and quantification of TF binding activity within complex experimental settings (multiple conditions or time series)
- All in one automated workflows for recurring analysis tasks

Modules enabling these individual analysis steps are included in the TOBIAS package, which is publicly available at Github (https://github.com/loosolab/TOBIAS) as well as on PyPI and Bioconda. Besides the examples given in the repository README, we also provide a Wiki (https://github.com/loosolab/TOBIAS/wiki) which introduces some of the individual software modules. We used the pre-defined workflows in Snakemake and NextFlow to run the full analysis. The single modules are explained in more detail below.

Bias correction (TOBIAS ATACorrect module)

Each Tn5-cut site is defined as the 5' end of the read shifted by +5 at the plus strand and -4 at the minus strand to center the transposase event. Using the mapped reads from closed chromatin, ATACorrect builds a dinucleotide weight matrix [72] representing the preference of Tn5 insertion. In contrast to the classical position weight matrix (PWM) the dinucleotide weight matrix (DWM) captures the inter-base relationships which arise due to the palindromic nature of the bias. A background model is similarly built by shifting all reads +100bp as described by [17].

Reads within open chromatin peaks are then corrected by estimating the expected number of cuts per base pair and subtracting this from the observed cut sites as follows (modified from [24]):

$$c_i = x_i - e_i$$

where

$$e_i = \hat{x}_i * \hat{b}_i \quad , \quad \hat{x}_i = \sum_{j=i-50}^{i+50} xj \quad , \quad \hat{b}_i = \frac{bj}{\sum_{j=i-50}^{i+50} bj}$$

where x_i is the observed cut sites, $e_i$is the expected cut sites, b_i is the calculated bias level, andc_i is the corrected cut sites at position i. To limit the influence of low-bias positions in the calculation of , a lower limit is set for b_i by calculating the fit of cutsites vs. bias to a rectified linear unit function (ReLu) in moving 100bp-windows and setting every b_i below the linear fit to 0. This calculation is performed for all base pairs within open chromatin, setting all other positions to 0. Lastly, each c_i is rescaled to fit the original sum of cuts $\hat{x}_i$ for each window.

Footprinting (TOBIAS ScoreBigwig module)

We estimate footprint scores across open chromatin regions by calculating:

$$FP = \bar{x}_{flank} - \bar{x}_{mid}$$

where

$$\bar{x}_{flank} = \frac{\sum_{i=j}^{j+wf} x_i + \sum_{i=j+wf+wm}^{j+2*wf+Wm} x_i}{2*wf} \; for \; x_i > 0 \quad ,$$

$$\bar{x}_{mid} = \frac{\sum_{i=j+wf}^{j+wf+wm} x_i}{wm} \; for \; x_i < 0$$

x_i is the number of cuts at position i, *wf* = width of flank in bp, *wm* = width of middle (footprint) in bp. The defaults used are: *wf* = [10;30], *wm* = [20;50].

The term $\overline{x}_{mid}$will be negative and will therefore raise the score if there is a high depletion of cuts in the footprint (middle). If there is no depletion, the score will simplify to the mean of cuts in the flanking regions, representing accessibility. It is therefore not necessary to see a canonical footprint shape for the footprint score to be high. The footprint score can be interpreted as higher scores being more evidence that a protein was bound at a given position. All calculations are done by the TOBIAS "ScoreBigwig" module.

Estimation of transcription factor states and pairwise comparison between conditions (TOBIAS BINDetect module)

To match the calculated footprint scores to potential binding sites, TOBIAS BINDetect integrates genomic sequence, footprint scores from several conditions and motifs to identify up- and down regulated TFs based on footprint scores.

In the first step of the algorithm, the MOODS library (https://github.com/jhkorhonen/MOODS [73]) is used to detect TF binding sites (within peaks) with a p-value threshold of 1e-4. Background base pair probabilities are estimated from the input peak set. Subsequently, each binding site is matched to footprint scores for each condition. Simultaneously, a background distribution of values is built by randomly subsetting peak regions at ~200bp intervals, and the scores from each condition are normalized to each other using quantile normalization. These values are used to calculate a distribution of background log2FCs for each pairwise comparison of conditions.

Overlaps between the TFBS identified in the first step are quantified by creating a distance matrix of TFs. The distance between a TF pair (TF1;TF2) is calculated as:

$$dist_{TF1;TF2} = 1 - max(\,overlap_{TF1;TF2} \,/\, total_{TF1}\,,\; overlap_{TF2;TF1} \,/\, total_{TF2})$$

where $total_{TF1}$and $total_{TF2}$ are the total basepairs of all *TF1* and TF2 sites respectively and $overlap_{TF1;TF2}$is the amount of base pairs of *TF1* which overlap with *TF2* sites. The max-statement ensures that the overlap is calculated with regards to the shortest TF motif.

In the second step of the algorithm, every TF binding site found (for each motif given as input) is split into bound and unbound sites based on a score threshold per condition. The threshold is set at the level of significance of a normal-distribution fit to the background distribution of scores (user-defined p-value). As well as the per-condition split, each site is assigned a log2FC (fold change) per comparison, which represents whether the binding site has larger/smaller footprint scores in comparison. The global distribution of log2FC's per TF is compared to the background distributions to calculate a *differential binding score*, which is calculated as:

$$\frac{(\overline{x}_o - \overline{x}_b)}{((std_o + std_b) \;/\; 2)}$$

where $\overline{x_o}, std_o$ and $\overline{x_b}, std_b$ are the means and standard deviations of the observed and background log2FC distributions respectively. A p-value is also calculated by subsampling 100 log2FCs from the background and calculating the significance of the observed change (Python's scipy.stats.ttest_1samp). By comparing the observed log2FC distribution to the background log2FC, the effects of any global differences due to sequencing depth, noise etc. are controlled.

The differential binding scores and p-values are visualized as a volcano plot per condition-comparison. All TFs with -log10(pvalue) above the 95% quantile or differential binding scores smaller/larger than the 5% and 95% quantiles (top 5% in each direction) are colored and shown with labels. Below the plot, hierarchical clustering of the TFBS-distance matrix is shown and all TFs with distances less than 0.5 (overlap of 50% of bp) are colored as separate clusters.

The result of BINDetect is a folder-structure containing an overview of all potential binding sites (as .bed as well as excel files), the predicted split into bound and unbound sites, and a global overview of differentially bound TFs per condition-comparison.

Visualizing aggregate plots and calculation of footprint depth (TOBIAS PlotAggregate module)

Footprints are visualized using the subtool “TOBIAS PlotAggregate”. Aggregate footprints are created by aligning genomic signals centered on all binding sites (taking into account strandedness), to create a matrix of (*n* sites) x (*n* bp). The aggregate signal is calculated as the mean of each column (each bp). The default of +/- 60bp from the motif center was used throughout this manuscript.

The aggregate footprinting depth (FPD), which is applied in Supp. Figure 2c-d, was calculated for each TF as:

$$FPD = \overline{signal_{flank}} - \overline{signal_{middle}}$$

where $\overline{signal_{middle}}$ is the mean of the signal centered on the TFBS (30bp) and $\overline{signal_{flank}}$ is the mean of the signal in the remaining flanks ([-60;-15] and [+15;+60] bp) (See Supp. Figure 2b).

Similarly to the investigations in previous literature [22], we applied a mixture model from the Mixtools R package [74] to estimate the fractions of TFs with/without measurable footprints (Supp. Figure 2e).

Transcription factor binding network (TOBIAS CreateNetwork module)

The TF-TF network for Dux was built by subsetting all binding sites on the following characteristics: Bound in the promoter of a target gene, labeled "Unbound" in Control, labeled "Bound" in DuxOE, and log2FC footprint score increasing for DuxOE vs. Control. All targets were further reduced to only include genes encoding TFs with available motifs. Motifs were matched to genes as explained in the methods section "Processing of transcription factor motifs". The network was then created using "TOBIAS CreateNetwork". The result is a network of source and target nodes with directed edges, which in words can be described as: (Source TF) binds in the promoter of (Target TF).

TOBIAS framework output structure

The output generated by the TOBIAS framework is organized in a hierarchical folder structure, which increases clarity of all steps of the analysis. The folder structure specifically organizes input data, pre-processing output like peak-calling and annotation, genomic tracks such as bias correction and footprints, as well as the local and global TF predictions. Particularly, the output for every individual TF investigated is arranged into separate folders containing TF specific plots, annotations and binding predictions. This structure makes it simple to use the output for further downstream analysis, as was showcased in this work. An exemplary output of the complete framework can be found at: https://figshare.com/projects/Digital_Genomic_Footprinting_Analysis_of_ATAC-seq_dataset_from_preimplantation_timepoints_via_TOBIAS/69959

Validation

Comparison of TOBIAS to existing methods

Although footprinting tools for DNase-seq exist [18, 75, 76] [24, 77-79] [80], we have focused our comparison on tools which are easily obtainable and installable, do not require ChIP-seq training-data, and are explicitly supporting ATAC-seq. We have additionally added two metrics for “peak strength” and “PWM score” to compare TOBIAS to other footprinting-free metrics. The validation datasets and usage of existing tools are described in the following sections.

Datasets

The TOBIAS framework was benchmarked using ATAC-seq data for the GM12878 cell line (GEO: GSE47753) and TF ChIP-seq data from ENCODE for the same cell line. ATAC-seq data was prepared as explained in the section “Processing of ATAC-seq data”. ChIP-seq peak peak regions were downloaded and associated to motifs from Jaspar CORE 2018 using “MEME Centrimo” [81]. Only ChIP-seq experiments with motif enrichment > 1.0e-100 (Centrimo E-value) were kept. The pairing of the remaining 36 motifs and ChIP-seq peaks is seen below:

ENCODE accession	TF name	JASPAR motif ID
ENCSR987MTA	BHLHE40	MA0464.2
ENCSR681NOM	CEBPB	MA0466.2
ENCSR839XZU	Crem	MA0609.1
ENCSR000DZN	CTCF	MA0139.1
ENCSR000DZQ	EBF1	MA0154.3
ENCSR841NDX	ELF1	MA0473.2
ENCSR000DZB	ELK1	MA0028.2
ENCSR000BKA	ETS1	MA0098.3
ENCSR626VUC	ETV6	MA0645.1
ENCSR331HPA	Gabpa	MA0062.2

ENCSR009MBP	HSF1	MA0486.2
ENCSR000DYS	JUND	MA0491.1
ENCSR000DYV	MAFK	MA0496.2
ENCSR000DZF	MAX	MA0058.3
ENCSR000BKB	MEF2A	MA0052.3
ENCSR000BNG	MEF2C	MA0497.1
ENCSR000DZI	MXI1	MA1108.1
ENCSR000DNM	NFYB	MA0502.1
ENCSR514VYD	NR2F1	MA0017.2
ENCSR000DZO	NRF1	MA0506.1
ENCSR000BHD	PAX5	MA0014.3
ENCSR000BGR	PBX3	MA1114.1
ENCSR711XNY	PKNOX1	MA0782.1
ENCSR000BGF	REST	MA0138.2
ENCSR000BRI	RUNX3	MA0684.1
ENCSR041XML	SRF	MA0083.3
ENCSR739IHN	TBX21	MA0690.1
ENCSR000BGZ	Tcf12	MA0521.1
ENCSR501DKS	Tcf7	MA0769.1
ENCSR000BGI	USF1	MA0093.2
ENCSR000DZU	USF2	MA0526.2
ENCSR000BNP	YY1	MA0095.2
ENCSR000BHC	ZBTB33	MA0527.1
ENCSR000DZL	ZNF143	MA0088.2
ENCSR072PWP	ZNF24	MA1124.1
ENCSR000DYP	ZNF384	MA1125.1

Bound binding sites per TF were defined as any TFBS within +/- 100bp from the paired ChIP-seq peak summit. In case of two or more binding sites per peak, the one closest to the summit was set to bound, and others were excluded. Unbound binding sites were defined as any TFBS not overlapping any ChIP-seq peak, as well as not overlapping bound sites from any other factors.

Bias correction approaches

TOBIAS was compared to the existing bias correction methods as follows:

- **seqOutBias** **([25])**
  The seqOutBias software was downloaded from GitHub (https://github.com/guertinlab/seqOutBias). Following the vignette for ATAC-seq, mappability files were created and ATAC-seq reads were corrected for plus/minus strand reads separately. After correction, we further shifted the positive and negative tracks +5 and -4bp respectively, as this was not performed by the tool itself.
- **HINT-ATAC** **([16])**
  The HINT software was downloaded from PyPI as part of the RGT software suite. Bias-correction was performed from the ATAC-seq reads using the command "rgt-hint tracks --bc --bigWig <bam>".

Aggregate footprints for each method across all (within peaks), bound and unbound binding sites (see explanation above) were visualized using "TOBIAS PlotAggregate".

Footprinting

For comparing TOBIAS to existing footprinting methods as follows:

- **msCentipede** **([14])**
  The msCentipede software was downloaded from GitHub (https://github.com/rajanil/msCentipede). For each TF, the binding model was built using the 5000 TFBS with highest PWM score genomewide. The resulting models were then used to infer the posterior binding-probability of TFBS in peaks.

- **Wellington ([76])**
  The pyDNase software was downloaded from PyPI. Footprints in ATAC-seq peaks were estimated using "wellington_footprints.py" with the "-A" option for ATAC-seq mode.

- **Peak strength**
  The "Peak strength" metric is defined as the mean number of Tn5 insertions in the ATAC-seq peak where the binding site is found. This score represents the accessibility of a certain region not taking into account local footprint information.
- **PWM score**
  The score of the motif-sequence match at the specific TFBS. As this is based on sequence alone, the PWM-score is independent of chromatin accessibility.

The area under the ROC curve (auROC) was used to evaluate the predictive power of each method.

Note on comparison: Overall, we find that TOBIAS performs at least equally well in comparison to msCentipede [14], a learning based approach, that demands high computational performance and individually trained models for every TF under investigation (Supp. Figure 1b). Of note, although this learning-based approach performs well overall, it exhibits a drastic loss of predictive power for some TFs, while the TOBIAS scoring model provides robust binding prediction scores even for those TFs that do not leave visible footprints at first glance (Supp. Figure 1b right).

Figures and figure legends

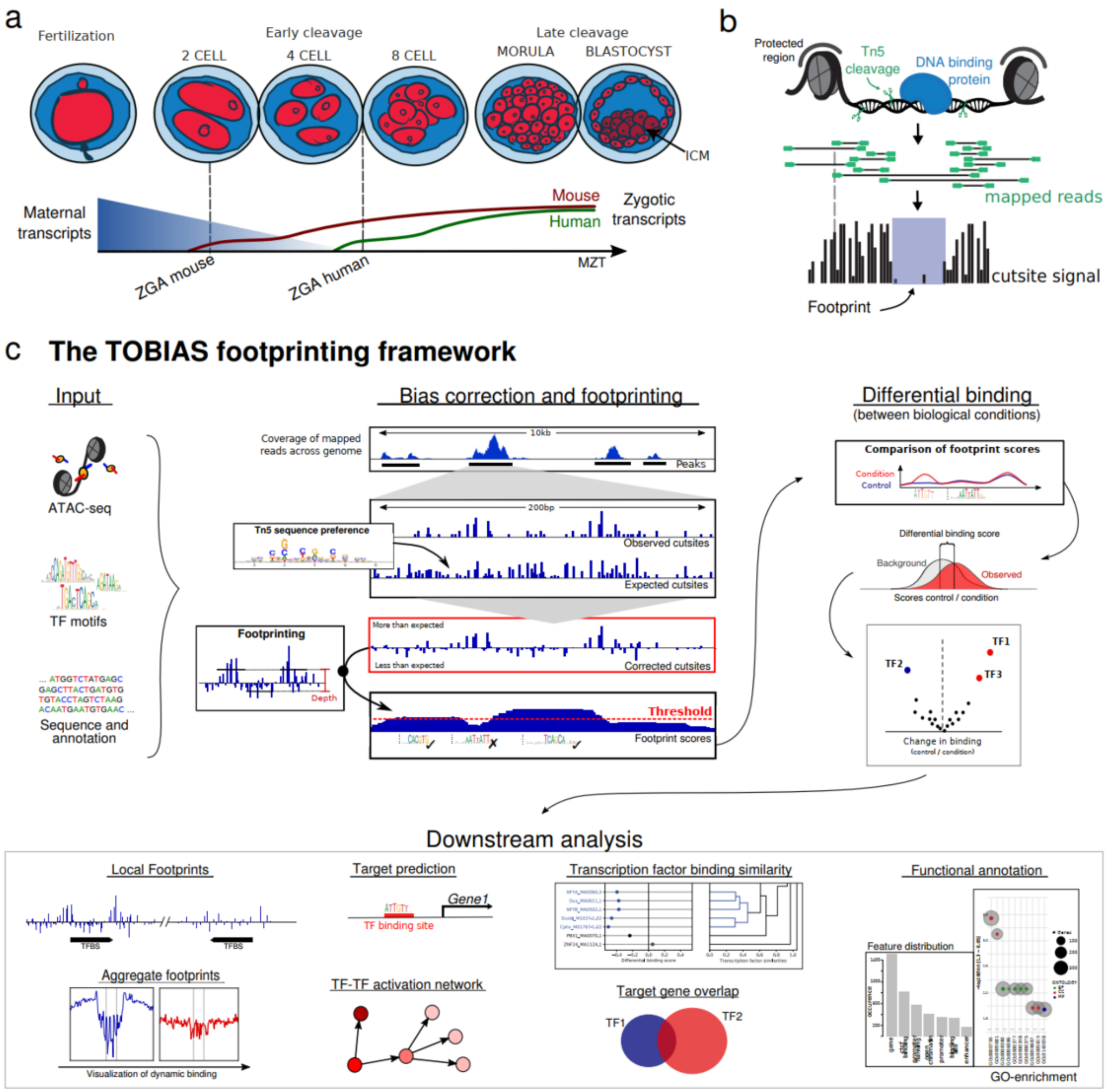


Figure 1: The use of chromatin accessibility assays to investigate early developmental processes

(a) Early embryonic development in human and mouse. *The fertilized egg undergoes a series of divisions ultimately creating the structure of the blastocyst. While maternal transcripts are depleted, the zygotic genome is activated in waves as indicated by the dark shading. ZGA initiates in mouse at 2-cell stage and in human at the 4-8-cell stage.*

(b) The concept of footprinting using ATAC-seq. *The Tn5 transposase cleaves and inserts sequencing adapters in open chromatin, but is unable to cut in chromatin occupied by e.g. nucleosomes* *or transcription factors. The mapped sequencing reads can be used to create a signal of single Tn5-events (cutsites), in which binding of transcription factors is visible as depletion of signal (the footprint).* ***(c) The TOBIAS digital genomic footprinting framework.*** *Using an input of sequencing reads from ATAC-seq, transcription factor motifs and sequence information, the TOBIAS footprinting framework detects local and global changes in transcription factor binding. Bias-correction of the Tn5 sequence preference enables detection of local chromatin footprints and matching to individual TFBS. Footprint scores can be compared between conditions and are used to define differential binding in pairwise comparisons. The global binding map allows for a variety of downstream analysis such as visualization of local and aggregated footprints across conditions, prediction of target genes for each TF as well as comparison of binding specificity between several transcription factors. Functional annotation such as GO enrichment can be used to infer biological meaning of target gene sets.*

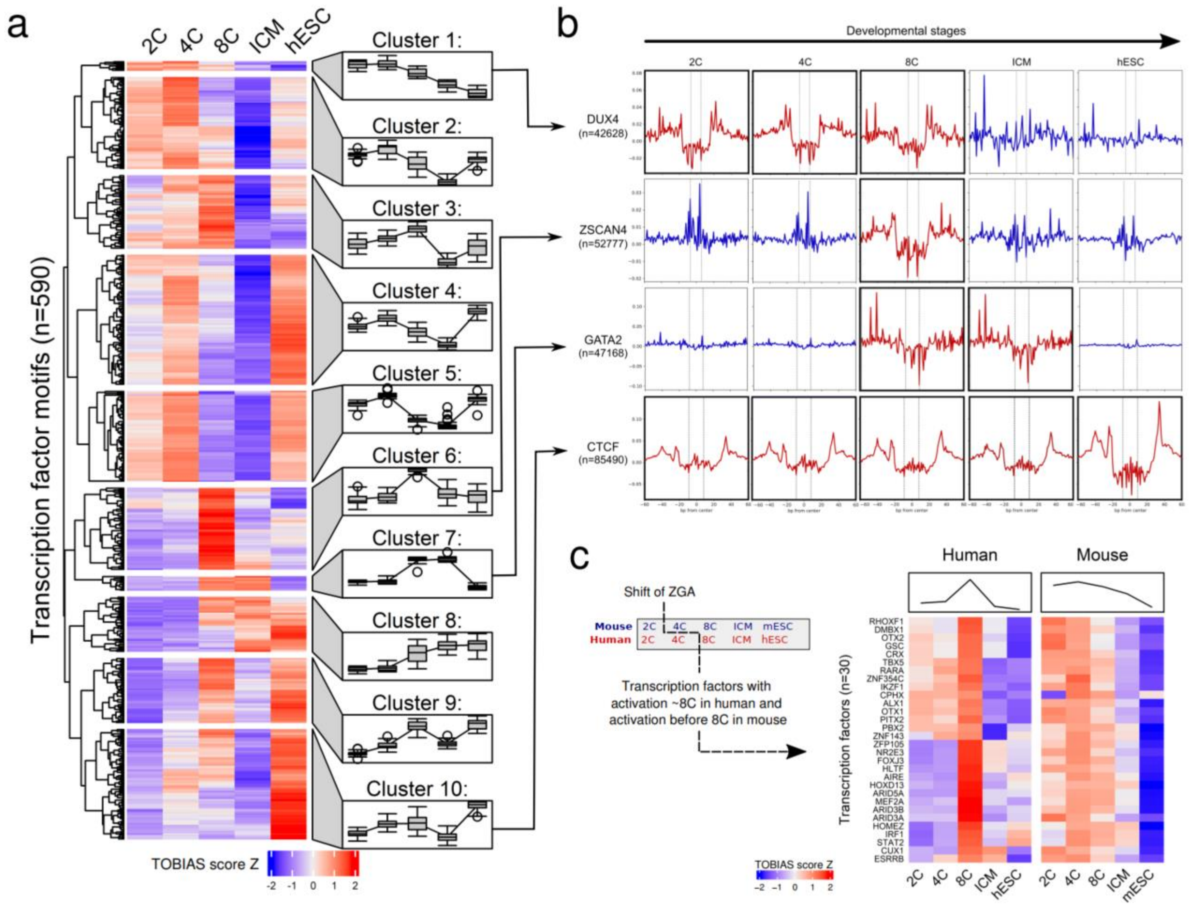


Figure 2: TOBIAS enables investigation of global changes in transcription factor binding

(a) Clustering of transcription factor activities throughout development*. Each row represents one TF, each column a developmental stage; blue color indicates low activity, red color indicates high activity. In order to visualize cluster trends, each cluster is associated with a trend line and time point specific boxplots.*

(b) Bias-corrected ATAC-seq footprints reveal dynamic TF binding. *Aggregated footprinting plot matrix for transcription factor binding sites. Plots are centered around binding motifs (n=* relates to the number of binding sites). Rows indicate TFs DUX4, ZSCAN4, GATA2, and CTCF; columns illustrate developmental stages from left to right. Active binding of the individual TFs at the respective timepoints is visible as a depletion in the signal around the binding site (highlighted in red). Upper three TFs are related to developmental stages, CTCF acts as a universal control, generating a footprint in all conditions. See Supplementary Figure 2A for uncorrected footprints.*

(c) TF activity is shifted by ZGA onset in human and mouse. *Heatmaps show activity of known ZGA-related TFs for human (left) and mouse (right) across matched timepoints 2C / 8C / ICM / hESC (mESC). Mean TF activity (top panel) peaks at 4-8C stage in human and is shifted to 2-4C stage in mouse by the earlier ZGA onset.*

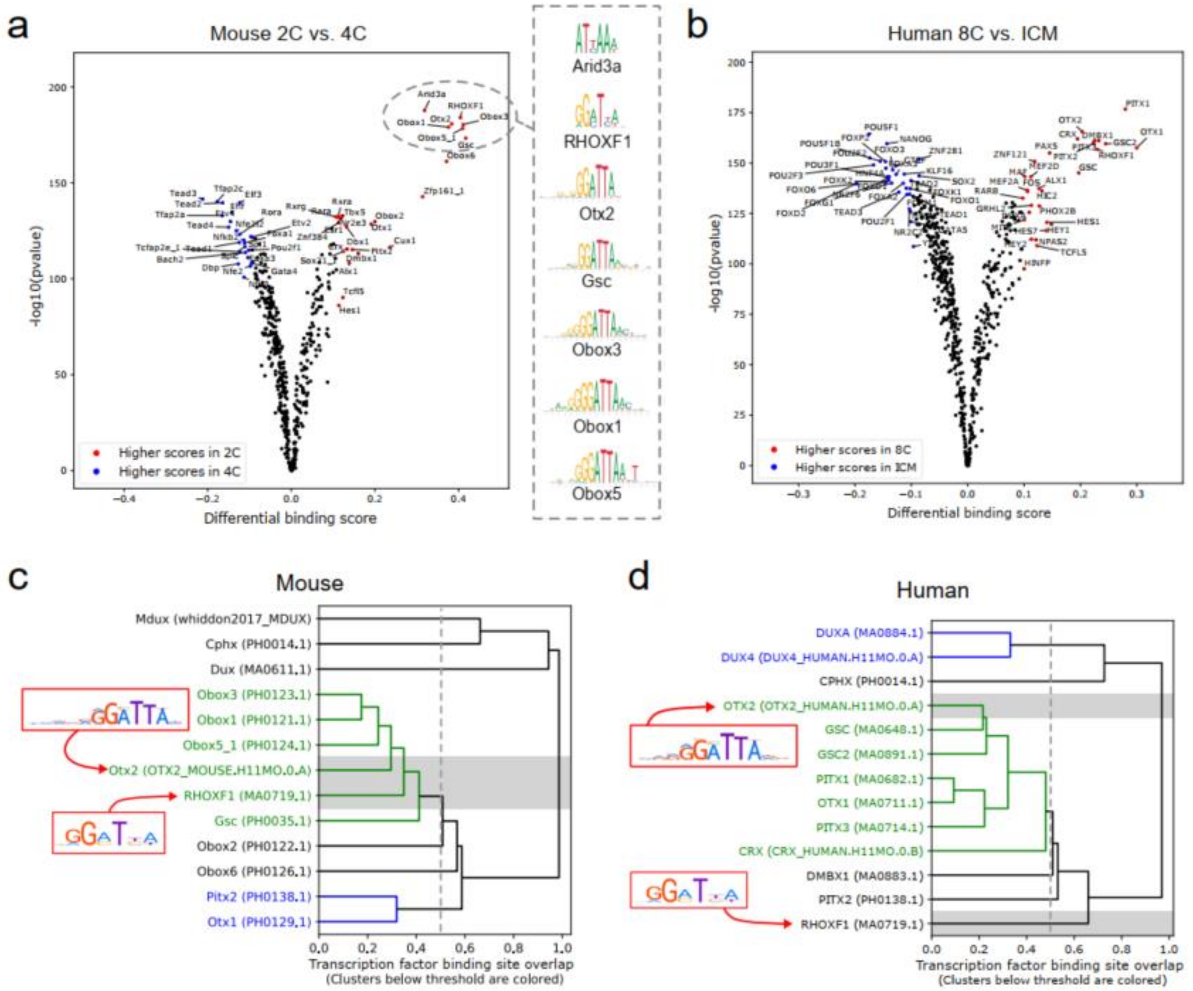


Figure 3: Comparison of binding site overlaps shows specification of ZGA functions between mouse and human

(a-b) Pairwise comparison of TF activity between developmental stages. *The volcano plots show the differential binding activity against the -log10(pvalue) (as provided by TOBIAS) of the investigated TF motifs; each dot represents one motif. For (A) 2C stage specific/significant TFs are labeled in red, 4C specific factors are given in blue. For (B) 8C stage specific/significant TFs are labeled in red, ICM specific factors are given in blue.*

(c-d) Clustering of TF motifs based on binding site overlap. *Excerpt of the global TF clustering based on TF binding location, illustrating individual TFs as rows. The trees indicate genomic positional overlap of individual TFBS with a tree-depth of 0.2 representing an overlap of 80% of motifs. Each TF is indicated by name and unique ID in brackets. Clusters of TFs with more than 50% overlap (below 0.5 tree distance) are colored. (C) shows overlap of motifs included in the mouse analysis, and (D) shows clustering of human motifs. Complete TF trees are provided in Supp. File 2 and Supp. File 3.*

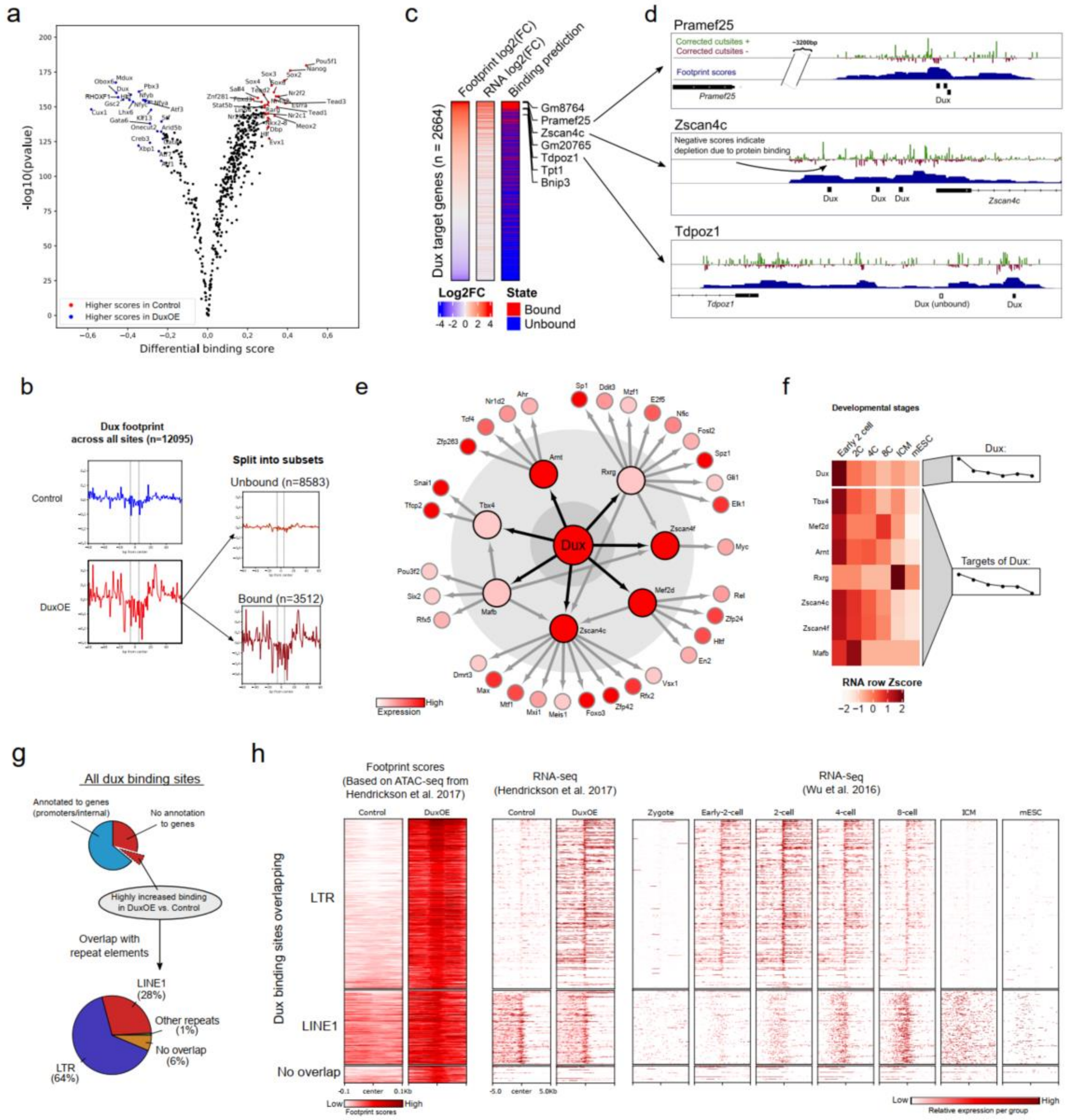


Figure 4: Dux binding induces transcription at gene promoters and LTR sequences in mouse

(a) Volcano plot comparing TF activities between mDux GFP- (Control) and mDux GFP+ (DuxOE). *Volcano plot showing the TOBIAS differential binding score on the x-axis and -log10 (p value) on the y-axis; each dot represents one TF. DuxOE specific/significant TFs are labeled in blue, Control specific/significant TFs are labeled in red.*

(b) Aggregated footprint plots for Dux. *The plots are centered on the predicted binding sites for Dux and shown for Control and DuxOE condition. The total possible binding sites for DuxOE (n=12095) are separated into bound and unbound sites (right). The dashed line represents the edges of the Dux motif.*

(c) Change in expression of genes near Dux binding sites. *The heatmap shows 2664 Dux binding sites found in gene promoters. Footprint log2(FC) and RNA log2(FC) represent the changes between Control and DuxOE for footprints and gene expression, respectively. Log2(FC) is calculated as log2(DuxOE/Control). The column "Binding prediction" depicts whether the binding site was predicted by TOBIAS to be bound/unbound in the DuxOE condition.*

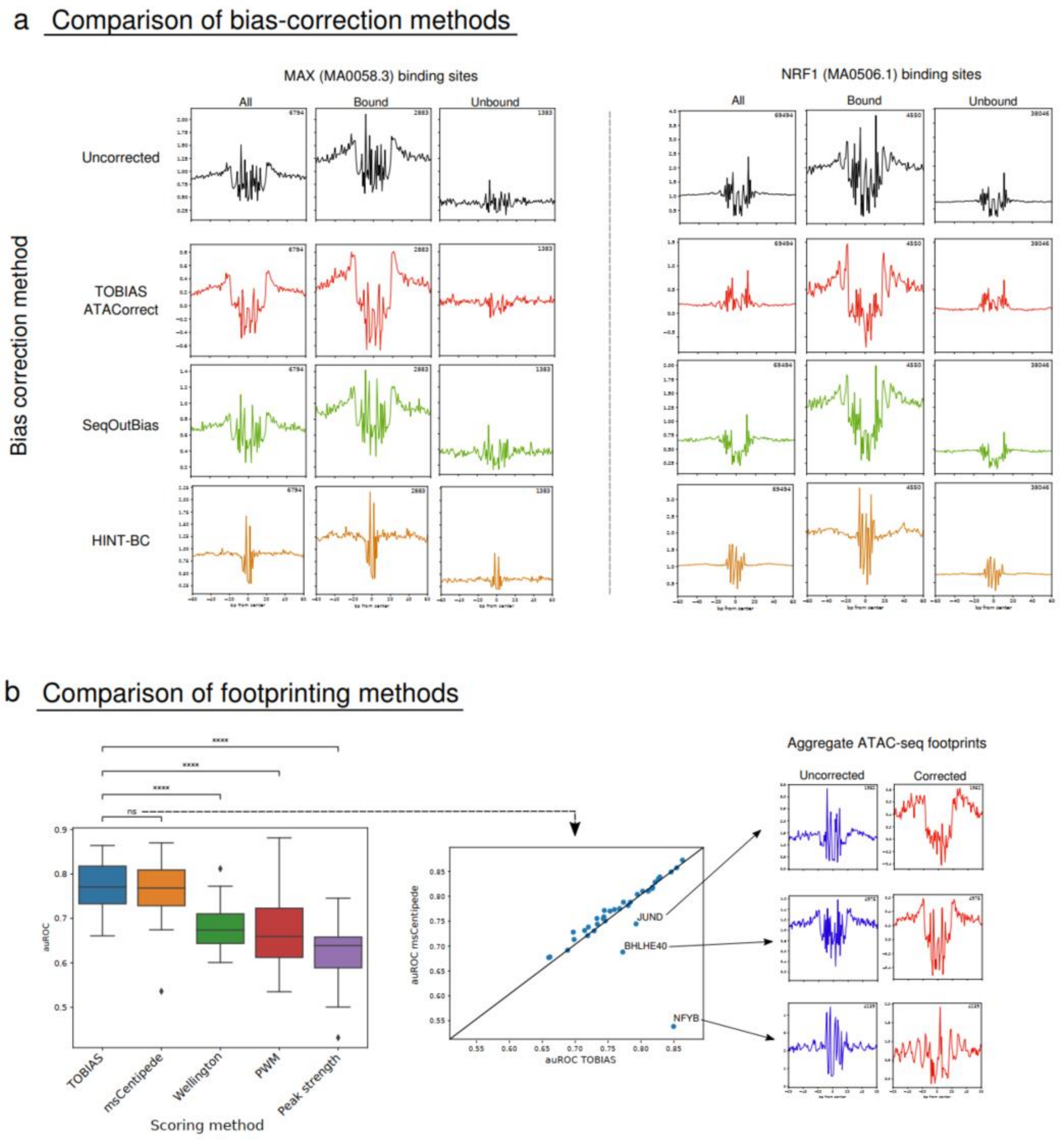
(d) Genomic tracks showing footprint scores of Dux-binding. *Genomic tracks indicating three DUX target gene promoters (one per row) and respective tracks for cut site signals (red/blue), TOBIAS footprints (blue), detected motifs (black boxes), and gene locations (solid black boxes with arrows indicating gene strand).*

(e) Dux transcription factor network. *The TF-TF network is built of all TFBS with binding in TF promoters with increasing strength in DuxOE (log2(FC)>0). Sizes of nodes represent the level of the network starting with Dux (Large: Dux, Medium: 1st level, Small: 2nd level). Nodes are colored based on RNA level in the OE condition* [5].

(f) Correlation of the Dux transcription factor network to expression during development. *The heatmap depicts the in vivo gene expression during developmental stages from [31]. The right-hand group annotation highlights the difference in mean expression for each timepoint. The heatmap is split into target genes of Dux, target genes of Arnt, Rxrg and Mef2d, as well as the pooled target genes from Tbx4, Mafb, Zscan4c and Zscan4f (Additional targets).*

(g) Dux binding sites overlap with repeat elements. *All potential Dux binding sites are split into sites either overlapping promoters/genes or without annotation to any known genes. The bottom pie chart shows a subset of the latter, additionally having highly increased binding (log2(FC)>1), and overlapping LTR/LINE1 elements.*

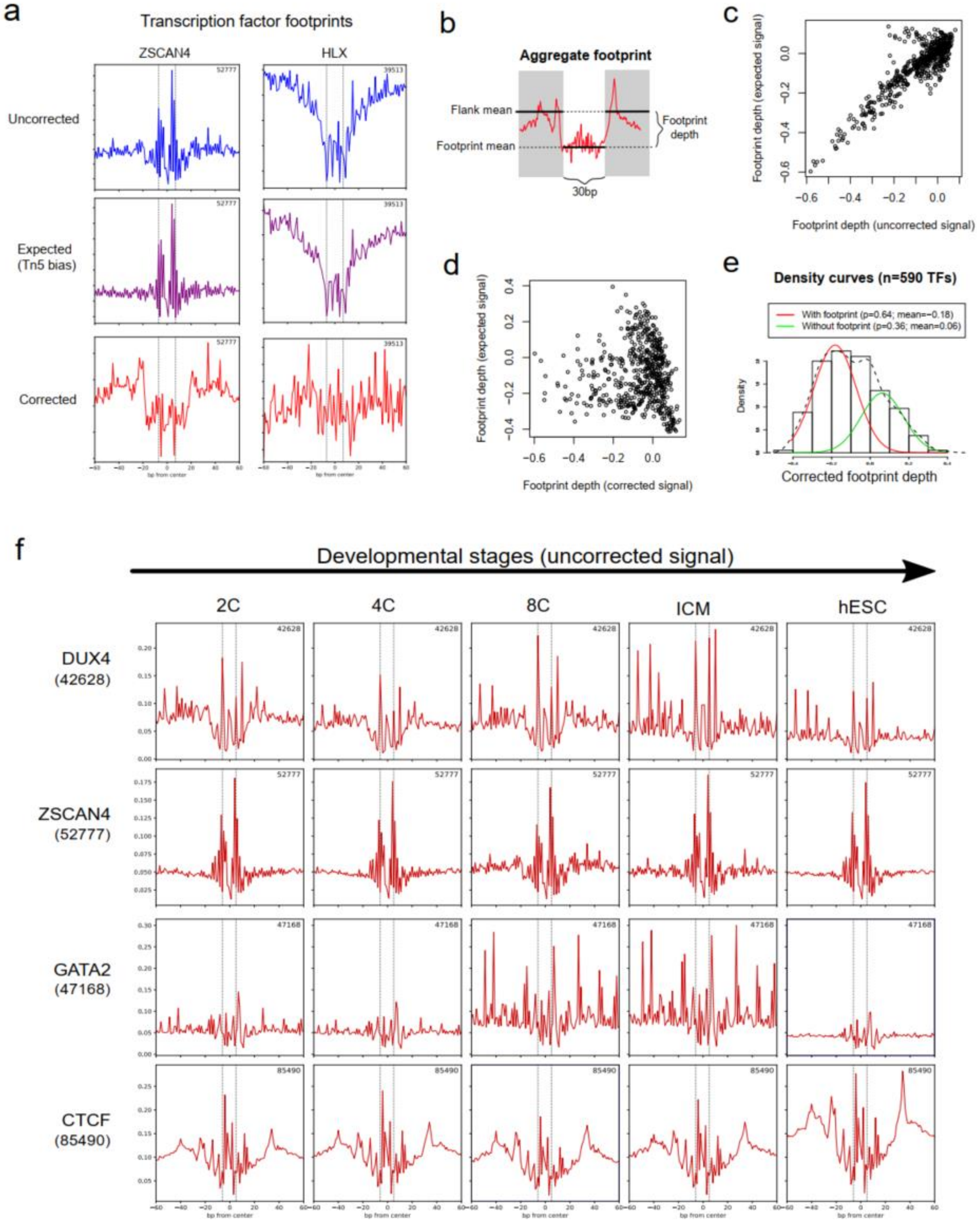
(h) Dux induces expression of transcripts specific for preimplantation. *Genomic signals for the Dux binding sites which are bound in DuxOE with log2(FC) footprint score >1 (i.e. upregulated in DuxOE) are split into overlapping either LTR, LINE1 or no known genetic elements (top to bottom). Footprint scores (+/- 100bp from Dux binding sites) indicate the differential Dux binding between control and DuxOE. RNA-seq shows the normalized read-counts from [5] and [31] within +/- 5kb of the respective Dux binding sites, while red color indicates high expression.*



Supplementary Figure 1: Comparison of existing bias-correction and footprinting methods

(a) Comparison of aggregate footprints for different bias-correction methods. *Bound and unbound transcription factor binding sites for MAX and NRF1 are shown across uncorrected signal (pileup of Tn5 insertions), TOBIAS ATACorrect, SeqOutBias and HINT-BC correction methods. An overview of all included TFs can be found in Supplementary File 1.*

(b) Comparison of predictive ability across different footprinting methods. *(left) auROC is calculated based on ENCODE ChIP-seq for 36 TFs and compared across methods. Boxes indicate quantiles, horizontal line indicates mean auROC of all TFs. Significance is indicated if applicable as asterisk. (center) TOBIAS and msCentipede are compared by pairwise dotplot, each dot represents one TF, TOBIAS has significant gains in auROC for JUND, BHLHE40 and NFYB, for which individual aggregated footprints are shown (left, uncorrected in blue, corrected in red)*



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

(a) Examples of Tn5-bias correction using "expected"-intermediates. *For ZSCAN4, the uncorrected signal is clearly influenced by the expected Tn5 bias, whereas the corrected signal* *uncovers the underlying effect of protein binding. Likewise, the uncorrected signal of HLX resembles a footprint which is mirrored by the expected signal. However, the corrected signal shows uniformity and uncovers that there is little effect of protein binding.*

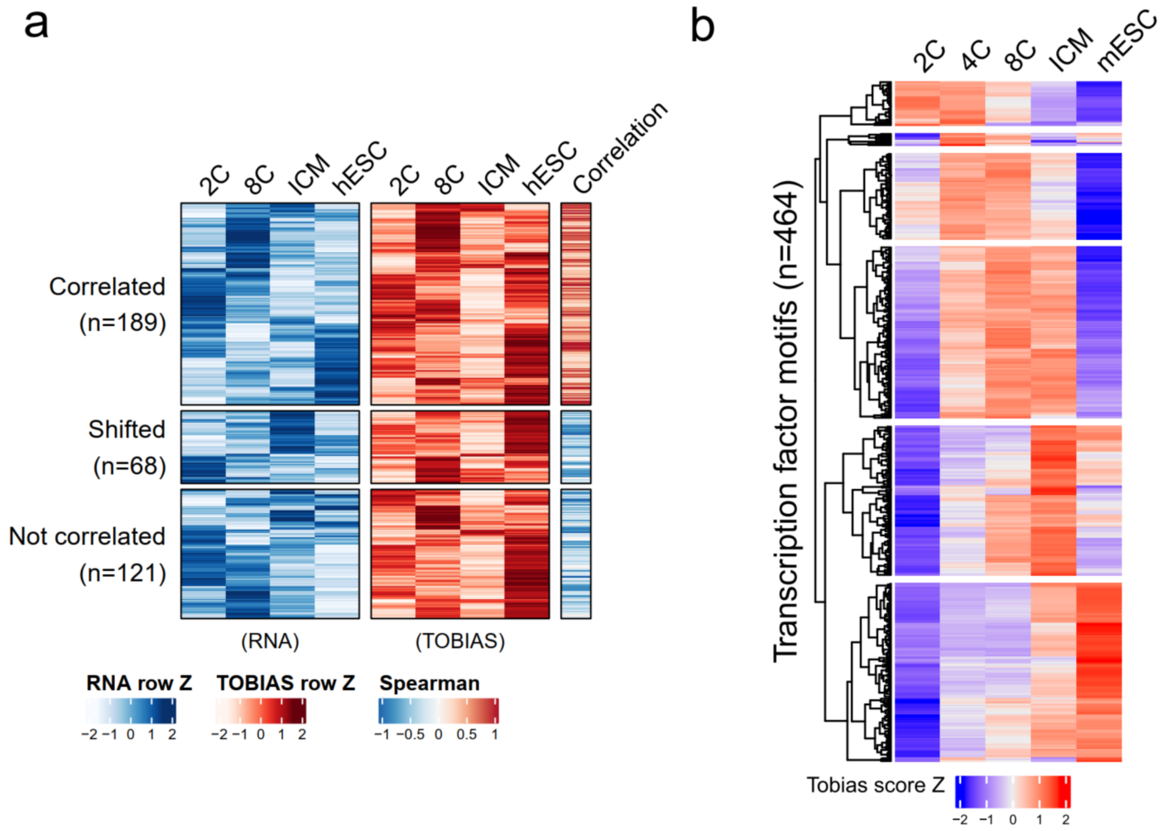
(b) Aggregate footprint depth model. *The footprint depth is calculated using a similar metric as described in [22].*

(c) Uncorrected Tn5-bias. *The scatter plot show the depth of footprints for uncorrected vs. expected footprints*

(d) Corrected Tn5-bias. The scatter plot *show the depth of corrected vs. expected footprints.*

(e) Mixture model *of all footprinting depths shows that 65% of motifs fall into the category of a measurable footprint in the aggregated profile. Data is based on 590 motifs in hESC.*

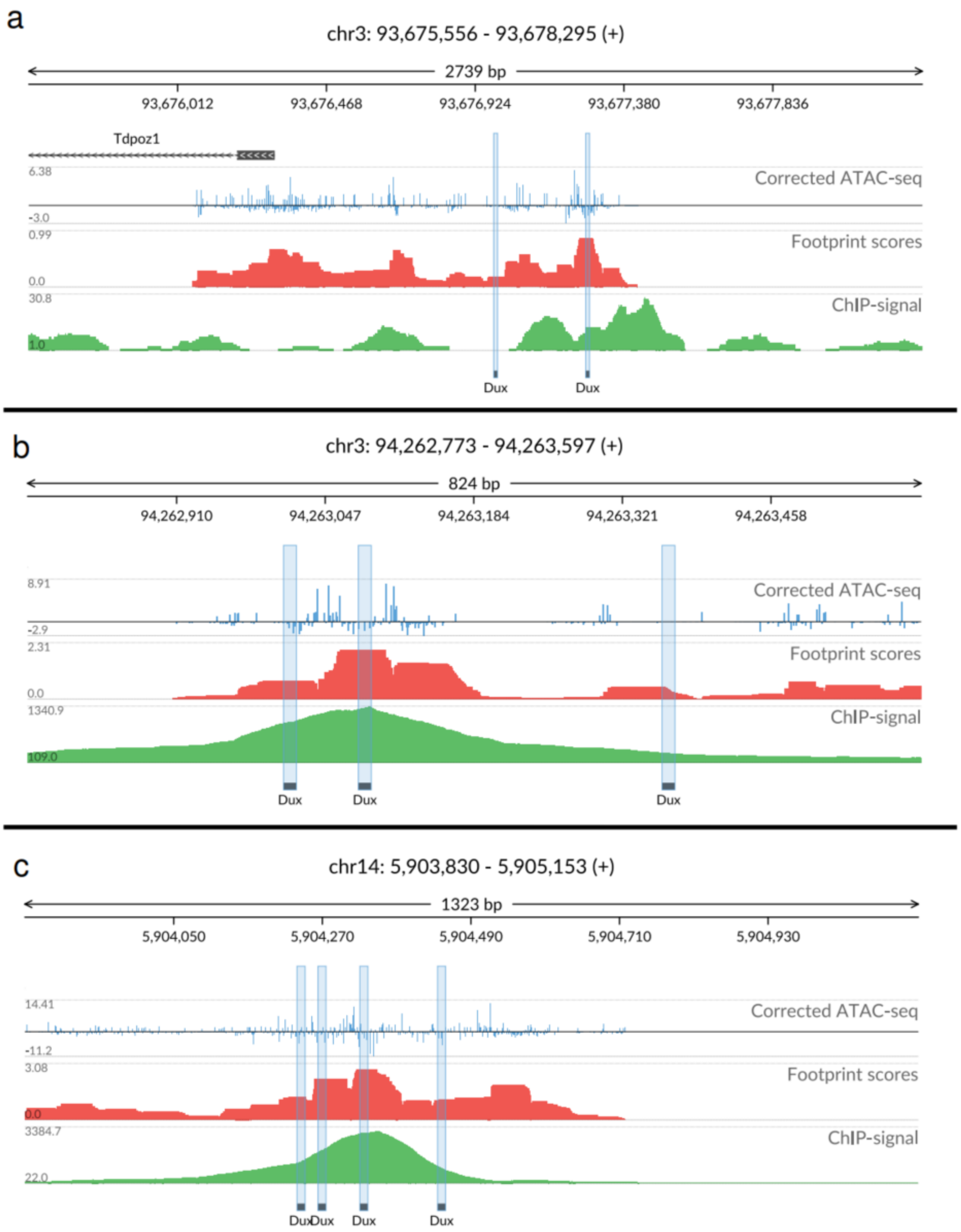
(f) A depiction of uncorrected footprint aggregates across timepoints for transcription factors DUX4, ZSCAN4, GATA2 and CTCF. *In contrast to the corresponding corrected signals seen in Figure 2A, the footprints are hardly visible in the uncorrected aggregates.*



Supplementary Figure 3: Transcription factor activity and expression during mouse and human development

(a) Correlation of footprints and RNA-seq. *The left heatmap (blue) depicts expression of transcription factor clusters in the respective developmental stages. The left heatmap (red) depicts the corresponding TOBIAS scores across human developmental stages. Spearman column represents the spearman correlation between TOBIAS/RNA. The TF clusters are grouped into "Correlated" (Spearman≥0.2), "shifted" (RNA max value appears before TOBIAS max value) and "Not correlated" (Spearman<0.2 with no apparent shift in RNA).*

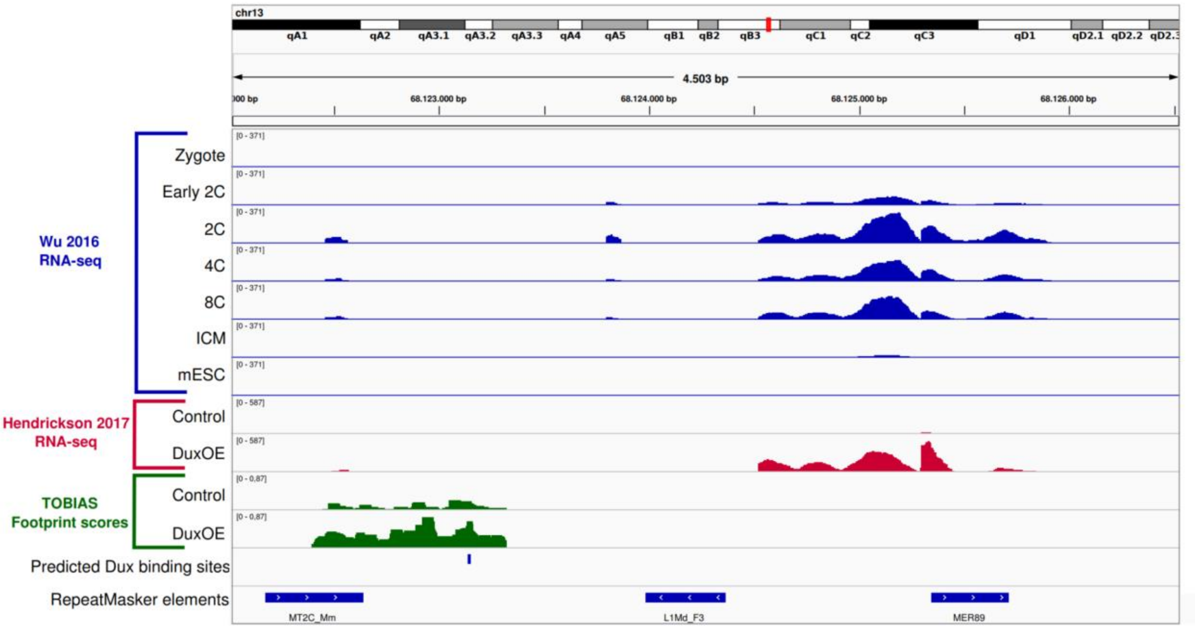
(b) Dynamic transcription factor binding during mouse embryonic development. *Similarly to figure 2A, the heatmap depicts the TOBIAS-predicted footprint scores for 464 motifs during the timepoints 2C, 4C, 8C, ICM and mESC. The rows are clustered into 6 clusters using hierarchical clustering. Individual cluster members are given in Supplementary Table 2.*



Supplementary Figure 4: Predicted footprinting scores correlate with ChIP-signal for Dux

(a) A view of the footprinting scores in the promoter of Tdpoz1. *Genomic tracks show corrected ATAC-seq cutsites at 1bp resolution (blue), footprint scores as calculated by TOBIAS (red), and pileup of reads from Dux ChIP-seq of [5] (green). Potential Dux binding sites are highlighted in blue.*

(b-c) Footprinting correlates with ChIP-signal at multiple genomic loci. *Genomic tracks are the same as described for (a).*



Supplementary Figure 5: Predicted Dux binding site correlates with increase in expression of closeby non-annotated regions

The figure shows genomic tracks of RNA-seq from [31] (blue) and [5] (red), TOBIAS footprint scores predicted from ATAC-seq (green) ([5]), predicted Dux binding site as well as known repeats as annotated by RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0).

Supplemental Information

List of Supplementary Files

Supplementary File 1: Visualization of different methods for Tn5 bias correction across all 36 TFs with matched ChIP-seq. Each page contains footprints for a specific TF across all binding sites (in peaks), bound sites (overlapping ChIP-seq) and unbound sites (not overlapping ChIP-seq) for uncorrected/expected/corrected signals from different bias correction methods.

Supplementary File 2: The direct output file of the "TOBIAS BINDetect"-module containing differential binding plots across all pairwise-comparisons of human developmental stages.

Supplementary File 3: The direct output file of the "TOBIAS BINDetect"-module containing differential binding plots across all pairwise-comparisons of mouse developmental stages.

Supplementary File 4: The direct output file of the "TOBIAS BINDetect"-module containing differential binding plots between control (mESC) and DuxOE samples.

List of Supplementary Tables

Supplementary Table 1: Prediction of transcription factor binding across human 2C/4C/8C/ICM/hESC clustered into co-active TFs. Each transcription factor is further linked to expression of the factor based on RNA-seq.

Supplementary Table 2: TOBIAS TF scores for human PD timepoints, correlated to corresponding RNA expression.

Supplementary Table 3: Prediction of transcription factor binding across mouse 2C/4C/8C/ICM/mESC clustered into co-active TFs. Each transcription factor is further linked to expression of the factor based on RNA-seq.

Supplementary Table 4: Human and Mouse RNA expression for Obox and RHOX/Rhox genes during preimplantation developmental stages.

Supplementary Table 5: Full list of the predicted Dux binding sites as well as their change between mESC and DuxOE as predicted by TOBIAS.

References

1. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ: **Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.** *Nat Methods* 2013, **10:**1213-1218.
2. Skene PJ, Henikoff S: **An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites.** *eLife* 2017, **6:**e21856.
3. Eckersley-Maslin MA, Alda-Catalinas C, Reik W: **Dynamics of the epigenetic landscape during the maternal-to-zygotic transition.** *Nat Rev Mol Cell Biol* 2018, **19:**436-450.
4. Jukam D, Shariati SAM, Skotheim JM: **Zygotic Genome Activation in Vertebrates.** *Dev Cell* 2017, **42:**316-332.
5. Hendrickson PG, Dorais JA, Grow EJ, Whiddon JL, Lim JW, Wike CL, Weaver BD, Pflueger C, Emery BR, Wilcox AL, et al: **Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons.** *Nat Genet* 2017, **49:**925-934.
6. De Iaco A, Planet E, Coluccio A, Verp S, Duc J, Trono D: **DUX-family transcription factors regulate zygotic genome activation in placental mammals.** *Nat Genet* 2017, **49:**941-945.
7. Eckersley-Maslin MA, Svensson V, Krueger C, Stubbs TM, Giehr P, Krueger F, Miragaia RJ, Kyriakopoulos C, Berrens RV, Milagre I, et al: **MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs.** *Cell Rep* 2016, **17:**179-192.
8. Madissoon E, Jouhilahti EM, Vesterlund L, Tohonen V, Krjutskov K, Petropoulos S, Einarsdottir E, Linnarsson S, Lanner F, Mansson R, et al: **Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos.** *Sci Rep* 2016, **6:**28995.
9. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al: **Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.** *Nat Methods* 2009, **6:**283-289.
10. Galas DJ, Schmitz A: **DNAse footprinting: a simple method for the detection of protein-DNA binding specificity.** *Nucleic Acids Res* 1978, **5:**3157-3170.
11. Sung MH, Baek S, Hager GL: **Genome-wide footprinting: ready for prime time?** *Nat Methods* 2016, **13:**222-228.
12. Vierstra J, Stamatoyannopoulos JA: **Genomic footprinting.** *Nat Methods* 2016, **13:**213-221.
13. Quach B, Furey TS: **DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter.** *Bioinformatics* 2017, **33:**956-963.
14. Raj A, Shim H, Gilad Y, Pritchard JK, Stephens M: **msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding.** *PLoS One* 2015, **10:**e0138030.
15. 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* 2017, **45:**4315-4329.
16. Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG: **Identification of transcription factor binding sites using ATAC-seq.** *Genome biology* 2019, **20:**45-45.
17. Koohy H, Down TA, Hubbard TJ: **Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme.** *PLoS One* 2013, **8:**e69853.
18. Sung MH, Guertin MJ, Baek S, Hager GL: **DNase footprint signatures are dictated by factor dynamics and DNA sequence.** *Mol Cell* 2014, **56:**275-285.

19. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, et al: **Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification.** *Nat Methods* 2014, **11:**73-78.
20. Madrigal P: **On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions.** *Front Bioeng Biotechnol* 2015, **3:**144.
21. Tripodi IJ, Allen MA, Dowell RD: **Detecting Differential Transcription Factor Activity from ATAC-Seq Data.** *Molecules* 2018, **23**.
22. Baek S, Goldstein I, Hager GL: **Bivariate Genomic Footprinting Detects Changes in Transcription Factor Activity.** *Cell Rep* 2017, **19:**1710-1722.
23. Berest I, Arnold C, Reyes-Palomares A, Palla G, Rasmussen KD, Helin K, Zaugg JB: **Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: &lt;em&gt;diffTF&lt;/em>.** *bioRxiv* 2018**:**368498.
24. Gusmao EG, Allhoff M, Zenke M, Costa IG: **Analysis of computational footprinting methods for DNase sequencing experiments.** *Nat Methods* 2016, **13:**303-309.
25. Martins AL, Walavalkar NM, Anderson WD, Zang C, Guertin MJ: **Universal correction of enzymatic sequence bias reveals molecular signatures of protein/DNA interactions.** *Nucleic Acids Res* 2018, **46:**e9.
26. Wang JR, Quach B, Furey TS: **Correcting nucleotide-specific biases in high-throughput sequencing data.** *BMC Bioinformatics* 2017, **18:**357.
27. Koster J, Rahmann S: **Snakemake-a scalable bioinformatics workflow engine.** *Bioinformatics* 2018, **34:**3600.
28. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol* 2017, **35:**316-319.
29. Belmann P, Fischer B, Krüger J, Procházka M, Rasche H, Prinz M, Hanussek M, Lang M, Bartusch F, Gläßle B, et al: **de.NBI Cloud federation through ELIXIR AAI [version 1; peer review: 2 approved, 1 not approved].** *F1000Research* 2019, **8**.
30. Wu J, Xu J, Liu B, Yao G, Wang P, Lin Z, Huang B, Wang X, Li T, Shi S, et al: **Chromatin analysis in human early development reveals epigenetic transition during ZGA.** *Nature* 2018, **557:**256-260.
31. Wu J, Huang B, Chen H, Yin Q, Liu Y, Xiang Y, Zhang B, Liu B, Wang Q, Xia W, et al: **The landscape of accessible chromatin in mammalian preimplantation embryos.** *Nature* 2016, **534:**652-657.
32. Wang K, Nishida H: **REGULATOR: a database of metazoan transcription factors and maternal factors for developmental studies.** *BMC Bioinformatics* 2015, **16:**114.
33. Adjaye J, Monk M: **Transcription of homeobox-containing genes detected in cDNA libraries derived from human unfertilized oocytes and preimplantation embryos.** *Mol Hum Reprod* 2000, **6:**707-711.
34. Adhikary S, Peukert K, Karsunky H, Beuger V, Lutz W, Elsasser HP, Moroy T, Eilers M: **Miz1 is required for early embryonic development during gastrulation.** *Mol Cell Biol* 2003, **23:**7648-7657.
35. Home P, Kumar RP, Ganguly A, Saha B, Milano-Foster J, Bhattacharya B, Ray S, Gunewardena S, Paul A, Camper SA, et al: **Genetic redundancy of GATA factors in the extraembryonic trophoblast lineage ensures the progression of preimplantation and postimplantation mammalian development.** *Development* 2017, **144:**876-888.
36. Xu K, Chen X, Yang H, Xu Y, He Y, Wang C, Huang H, Liu B, Liu W, Li J, et al: **Maternal Sall4 Is Indispensable for Epigenetic Maturation of Mouse Oocytes.** *Journal of Biological Chemistry* 2017, **292:**1798-1807.
37. Svoboda P: **Mammalian zygotic genome activation.** *Semin Cell Dev Biol* 2018, **84:**118-126.
38. Schulz KN, Harrison MM: **Mechanisms regulating zygotic genome activation.** *Nature Reviews Genetics* 2019, **20:**221-234.

39. Tohonen V, Katayama S, Vesterlund L, Jouhilahti EM, Sheikhi M, Madissoon E, Filippini-Cattaneo G, Jaconi M, Johnsson A, Burglin TR, et al: **Novel PRD-like homeodomain transcription factors and retrotransposon elements in early human development.** *Nat Commun* 2015, **6:**8207.

40. Rhee C, Edwards M, Dang C, Harris J, Brown M, Kim J, Tucker HO: **ARID3A is required for mammalian placenta development.** *Developmental Biology* 2017, **422:**83-91.

41. Winger Q, Huang J, Auman HJ, Lewandoski M, Williams T: **Analysis of Transcription Factor AP-2 Expression and Function During Mouse Preimplantation Development1.** *Biology of Reproduction* 2006, **75:**324-333.

42. Pastor WA, Liu W, Chen D, Ho J, Kim R, Hunt TJ, Lukianchikov A, Liu X, Polo JM, Jacobsen SE, Clark AT: **TFAP2C regulates transcription in human naive pluripotency by opening enhancers.** *Nat Cell Biol* 2018, **20:**553-564.

43. Eckersley-Maslin M, Alda-Catalinas C, Blotenburg M, Kreibich E, Krueger C, Reik W: **Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program.** *Genes Dev* 2019, **33:**194-208.

44. De Iaco A, Coudray A, Duc J, Trono D: **DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells.** *EMBO Rep* 2019, **20**.

45. Whiddon JL, Langford AT, Wong CJ, Zhong JW, Tapscott SJ: **Conservation and innovation in the DUX4-family gene network.** *Nat Genet* 2017, **49:**935-940.

46. Huang CJ, Chen CY, Chen HH, Tsai SF, Choo KB: **TDPOZ, a family of bipartite animal and plant proteins that contain the TRAF (TD) and POZ/BTB domains.** *Gene* 2004, **324:**117-127.

47. Lee S-E, Lee S-Y, Lee K-A: **Rhox in mammalian reproduction and development.** *Clin Exp Reprod Med* 2013, **40:**107-114.

48. Borgmann J, Tuttelmann F, Dworniczak B, Ropke A, Song HW, Kliesch S, Wilkinson MF, Laurentino S, Gromoll J: **The human RHOX gene cluster: target genes and functional analysis of gene variants in infertile men.** *Hum Mol Genet* 2016, **25:**4898-4910.

49. Royall AH, Maeso I, Dunwell TL, Holland PWH: **Mouse Obox and Crxos modulate preimplantation transcriptional profiles revealing similarity between paralogous mouse and human homeobox genes.** *Evodevo* 2018, **9:**2.

50. Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M: **A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity.** *Cell* 2018, **174:**391-405.e319.

51. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ: **Single-cell chromatin accessibility reveals principles of regulatory variation.** *Nature* 2015, **523:**486-490.

52. Harrison PW, Alako B, Amid C, Cerdeno-Tarraga A, Cleland I, Holt S, Hussein A, Jayathilaka S, Kay S, Keane T, et al: **The European Nucleotide Archive in 2018.** *Nucleic Acids Res* 2018.

53. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *2011* 2011, **17:**3.

54. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.

55. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al: **Ensembl 2018.** *Nucleic Acids Res* 2018, **46:**D754-D761.

56. Feng JX, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS.** *Nature Protocols* 2012, **7:**1728-1740.

57. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al: **GENCODE reference annotation for the human and mouse genomes.** *Nucleic Acids Research* 2018, **47:**D766-D773.

58. Kondili M, Fust A, Preussner J, Kuenne C, Braun T, Looso M: **UROPA: a tool for Universal RObust Peak Annotation.** *Sci Rep* 2017, **7:**2593.

59. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15:**550-550.

60. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30:**2114-2120.

61. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, Bessy A, Cheneby J, Kulkarni SR, Tan G, et al: **JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.** *Nucleic Acids Res* 2018, **46:**D1284.

62. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al: **HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis.** *Nucleic Acids Research* 2017, **46:**D252-D259.

63. Sebastian A, Contreras-Moreira B: **footprintDB: a database of transcription factors with annotated cis elements and binding interfaces.** *Bioinformatics* 2013, **30:**258-265.

64. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, et al: **Impact of cytosine methylation on DNA binding specificities of human transcription factors.** *Science* 2017, **356:**eaaj2239.

65. Machanick P, Bailey TL: **MEME-ChIP: motif analysis of large DNA datasets.** *Bioinformatics* 2011, **27:**1696-1697.

66. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nature Protocols* 2009, **4:**1184.

67. Gu L, Hitzel J, Moll F, Kruse C, Malik RA, Preussner J, Looso M, Leisegang MS, Steinhilber D, Brandes RP, Fork C: **The Histone Demethylase PHF8 Is Essential for Endothelial Cell Migration.** *PLoS One* 2016, **11:**e0146645.

68. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nature biotechnology* 2011, **29:**24-26.

69. Egorov AA, Sakharova EA, Anisimova AS, Dmitriev SE, Gladyshev VN, Kulakovskiy IV: **svist4get: a simple visualization tool for genomic tracks from sequencing experiments.** *BMC Bioinformatics* 2019, **20:**113.

70. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Research* 2003, **13:**2498-2504.

71. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis.** *Nucleic acids research* 2016, **44:**W160-W165.

72. Siddharthan R: **Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix.** *PLoS One* 2010, **5:**e9722.

73. Korhonen JH, Palin K, Taipale J, Ukkonen E: **Fast motif matching revisited: high-order PWMs, SNPs and indels.** *Bioinformatics* 2016, **33:**514-521.

74. Benaglia T, Chauveau D, Hunter DR, Young DS: **mixtools: An R Package for Analyzing Mixture Models.** *Journal of Statistical Software; Vol 1, Issue 6 (2010)* 2009.

75. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al: **An expansive human regulatory lexicon encoded in transcription factor footprints.** *Nature* 2012, **489:**83-90.

76. Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S: **Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data.** *Nucleic Acids Res* 2013, **41:**e201.

77. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS: **High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells.** *Genome Res* 2011, **21:**456-464.

78. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: **Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.** *Genome Res* 2011, **21:**447-455.

79. Luo K, Hartemink AJ: **Using DNase digestion data to accurately identify transcription factor binding sites.** *Pac Symp Biocomput* 2013**:**80-91.

80. Kahara J, Lahdesmaki H: **BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data.** *Bioinformatics* 2015, **31:**2852-2859.

81. Bailey TL, Machanick P: **Inferring direct DNA binding from ChIP-seq.** *Nucleic Acids Research* 2012, **40:**e128-e128.