Single-cell multi-omic analysis reveals principles of transcription-chromatin interaction during embryogenesis

Vivek Bhardwaj*# [1,2,3], Alberto Griffa* [1,2], Helena Viñas Gaza [1,2], Peter Zeller [1,2], Alexander van Oudenaarden# [1,2]

[1] Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Oncode Institute, Utrecht, The Netherlands.

[2] University Medical Center Utrecht, Utrecht, The Netherlands.

[3](present address) Institute of Biodynamics and Biocomplexity, Utrecht University, the Netherlands

*equal contribution

#corresponding authors

E-mail: a.vanoudenaarden@hubrecht.eu; v.bhardwaj@uu.nl

Abstract

Establishing a cell-type-specific chromatin landscape is critical for the maintenance of cell identity during embryonic development. However, our knowledge of how this landscape is set during vertebrate embryogenesis has been limited, due to the lack of methods to jointly detect chromatin modifications and gene expression in the same cell. Here we present a multimodal measurement of full-length transcriptome and chromatin modifications in individual cells during early embryonic development in zebrafish. We show that before the formation of germ layers, the chromatin and transcription states of cells are uncoupled, and become progressively connected during gastrulation and somitogenesis. Silencing of key developmental genes is achieved by local spreading of repressive chromatin as development proceeds. We use a joint analysis of transcription factor (TF) expression and chromatin states to predict lineage-specific activators and repressors and identify a subset of TFs that are themselves epigenetically regulated. Altogether, our data resolves the dynamic relationship between chromatin and transcription during early embryonic development and clarifies how these two layers interact to establish cell identity.

Introduction

Early embryonic development in animals is characterized by the controlled movement and positioning of cells, establishment of a body plan, and specification of tissue-specific cell states. While the spatial gradients of morphogens dominate the former two events [1], the formation of stable cell states is mostly believed to be regulated by epigenetic mechanisms [2]. Similar to the morphogens that regulate the patterning of an embryo, the epigenetic state can also be transgenerationally inherited [3]. This might play an important role in predefining the spatiotemporal expression of genes during early development and regulate cell fates. The inheritance of epigenetic signals, as well as tissue-specific signatures in mature cells, have been predominantly studied using genomic methods applied to cultured cell populations, whole tissues, or enriched cell populations sorted using cell surface or transgenic markers [4,5]. Recently, chromatin and DNA methylome mapping techniques have been developed to resolve cellular heterogeneity of epigenetic states at the level of individual cells. Major progress has been made using DNA methylome profiling of single cells, which have mostly been applied to study adult tissues [6–8]. More recently, we and others have applied single-cell methods to study chromatin states in adult tissues [9–12]. However, there is still a gap in our understanding of how these adult chromatin states arise, i.e. the process of establishment and propagation of cell-type-specific chromatin states during early embryonic development. Genome-wide studies mapping temporal chromatin changes of single cells during early embryogenesis are rare [13,14].

In this study, we asked how two specific (active and inactive) chromatin states of cells are shaped during early vertebrate embryogenesis, using zebrafish as a model system. As the chromatin and transcription state of cells in a developing embryo are highly dynamic and widely vary between cell types, bulk assays would average out these biologically important differences. Similarly, a single-cell assay profiling either chromatin or transcriptome cannot measure how these properties influence each other during development in a single cell. We, therefore, applied a novel single-cell multi-omic assay termed “T-ChIC” [15] to jointly profile the genome-wide active and inactive chromatin state along with the full-length transcriptome from the same single cells during zebrafish development spanning gastrulation and somitogenesis (4-24 hpf). Using these data we infer continuous developmental trajectories and ask how the chromatin state of genomic regions correlates with the expression of transcription factors and other developmentally important genes during cell fate commitment.

Results

Paired profiling of histone modifications and transcriptome of single cells during zebrafish embryogenesis

We developed a single-cell multi-omics method, termed T-ChIC (transcriptome and chromatin immuno-cleavage), which extends the previously described sortChIC [9] and VASA-seq [16] protocols, by integrating them in a single workflow (**Fig. 1a, Methods**) [15]. This allows us to quantify the pattern of histone modifications at kilobase resolution, while simultaneously providing full-length transcriptome coverage in single cells. We first applied T-ChIC to quantify the polycomb complex-mediated histone mark, H3K27me3, in zebrafish embryos collected at six selected time points post-fertilization, obtaining a total of 18,432 cells. This dataset provides us with complete coverage of gastrulation (4, 6, 8, 10 hpf), along with the beginning and the end of somitogenesis (12 and 24 hpf, respectively, **Fig. 1b**). A subset of these data contained cells without a functional antibody (labeled “T-noChIC”), which show a similar number of detected transcripts, and co-clustered with T-ChIC cells, confirming that the transcriptome quality of T-ChIC is independent of the antibody used (**Fig. S1a, b**). After removing cells with low MNase cuts, as well as potentially over-fragmented cells, we observed a strong enrichment of chromatin signal over specific genomic regions (**Fig. S1c, d**).

Early zebrafish embryos contain a high load of maternal transcripts required for early embryonic development, which are temporally replaced with newly transcribed, zygotic RNA [17]. Consistent with this transition, we observed a substantial decrease in unique fragment counts from spliced reads compared to unspliced reads with developmental time (**Fig. S1e**). Despite this decrease in detected transcripts, our overall number of detected genes with both spliced and unspliced counts is higher than previously reported in scRNA-seq studies [18,19], due to the increased sensitivity and full-length RNA recovery (**Supplementary Table 1**). At the chromatin level, we observed increased MNase cuts (representing H3K27me3 signal) per cell as development progresses (**Fig. S1f**). This corroborates previous observations of increasing H3K27me3 abundance during development based on bulk chromatin assays [20,21]. Overall, we retained 9275 cells with both transcriptome and H3K27me3 signal for further analysis.

We divided our data into early (4-6 hpf), intermediate (8, 10, 12 hpf), and late (24 hpf) time points, and compared our H3K27me3 signal with publicly available datasets at the corresponding time points (**Fig. 1c, methods**). While our pseudo-bulk H3K27me3 profiles showed a high genome-wide correlation with publicly available bulk ChIP data from matched time points (**Fig. S2a**), the analysis of genomic bins ranked by H3K27me3 signal shows improved signal enrichment of our data relative to the publicly available bulk sequencing data at a comparable sequencing depth (**Fig. S2b**). Moreover, the H3K27me3 levels show a clear relationship with the silencing of associated genes in single cells at all time points (**Fig. 1d, S2c**). We observed high H3K27me3 levels associated with the silencing of gene expression as early as 4 hpf for *hoxc3a*, involved in anterior-posterior patterning, as well as silencing of genes such as *pcdh1b* late during development (**Fig. 1d, Fig. S2c**). Furthermore, we also observed a transient association of K27me3 on genes. For example, *rfx4*, expressed in the central nervous system and neural rod, was silenced in non-neural ectoderm cells by H3K27me3 during gastrulation (**Fig. S2c,d**). These results suggest that our data allows us to gain quantitative insight into the relationship between H3K27me3 and gene expression during development.

Spatiotemporal spreading of H3K27me3 correlates with the silencing of gene expression during development

To annotate cell types in our data, we performed Leiden clustering of cells using their gene expression signal, followed by canonical correlation analysis of gene expression with that of a previously published time-course scRNA-seq data set [19] (**Fig. S3a**, **Methods**). Virtually all our cells matched one of the annotated cells from Wagner et al. with high confidence, allowing successful label transfer into our data (**Fig. S3b-d**). We further refined these labels based on cell ontologies from the Zebrafish Information Network [22], to categorize our cells into 34 cell types (**Fig. S3e**). To get a fine-grained view of cellular heterogeneity while reducing signal dropouts, we aggregated cells that are transcriptionally similar to each other, into “metacells” [23] (**Fig. S3d-e, methods**). Interestingly, most of the cell types annotated based on their gene expression profiles also show a clear separation based on their H3K27me3 enrichment as early as 8-12 hpf, suggesting that distinct, cell-state-specific H3K27me3 patterns already start to appear during gastrulation (**Fig. 2a**).

Next, we asked how the global H3K27me3 landscape is established in the cells during lineage commitment. We observed that with time, the number of genomic bins with H3K27me3 signal increased. In contrast, the average signal in detected bins plateaued at 24 hpf, indicating new regions acquiring the H3K27me3 signal instead of enrichment of signal on pre-marked regions (**Fig. S4a**). Therefore, we asked whether this increase in signal comes from a *de novo* gain in H3K27me3 or as a result of spill-over (indicating “cis-spreading”) of increasing H3K27me3 density from previously enriched regions. At least a subset of enriched regions in pluripotent cells display a cis-spreading of signal with differentiation, covering developmentally important genes such as the *zic* locus (**Fig. 2b**). To quantify cis-spreading genome-wide, we first subsetted the genomic bins, which had signal in at least 5% of filtered *Pluripotent* cells (at 4 hpf). Apart from tightly repressed genes such as *gata3, nr2f1a, six3b, pax9a, foxc1b, zic1/4, hox* clusters, and the *pcdh1/2* cluster with a broadly distributed signal, all other H3K27me3 signal was localized within 5 kb bins and the majority of these bins (65%) overlapped with a promoter region. We then calculated the signal on these bins compared to the background signal (averaged over 100 kb region) surrounding these bins in single cells. These two signals correlated positively for about 30% of the 5 kb bins suggesting a spill-over from the main signal peak to the surrounding background. In contrast, the remaining 70% displays a low correlation with background indicating a localized enrichment without spill-over to the surrounding background (**Fig. 2c, Fig. S4b**).

We confirmed this enrichment with an alternative approach based on domain calling on the pooled, pseudo-bulk dataset (**Methods**). While wider H3K27me3 domains detected on the pooled data correspond to the signal at 24 hpf, the sharper subpeaks within those domains were observed at early (4-6 hpf) time points (**Fig. S4d**). Relatively mature cell types, particularly from the neural ectoderm (such as differentiating neurons) show a higher correlation of subpeaks to the background, suggesting a wider spread in signal (**Fig. S4e**). We further stratified this signal in search of bins with a significant difference between lineages (**see Methods**). We only detected a handful of bins with statistically significant differences between lineages, and the mean H3K27me3 signal indicated that these results are not robust (**Fig. S4c**). Therefore the spread of H3K27me3 signal does not appear to be lineage-specific.

To understand how this spread of H3K27me3 relates to gene expression in time, we plotted the expression of the “host” gene (genes with promoter enrichment of H3K27me3 in pluripotent cells), and the “nearby” genes (with promoter within 100 kb region) over single cells arranged in pseudotime (**Fig. 2d**). Interestingly, the “host” genes displayed an increased expression before the spread of H3K27me3 signal (**Fig. 2d**). In contrast, the “nearby” genes displayed relatively smaller changes in transcription during early differentiation but were also downregulated at a later stage (**Fig. S4f**).

Finally, we identified genes where a functional silencing is achieved after spreading of H3K27me3, by applying linear regression to predict gene expression as a function of H3K27me3 density (defined as the number of reads per kb) on their nearest, or overlapping domains in metacells (see **Methods**). Silenced genes showed a strong negative correlation of H3K27me3 density with their expression, with the strongest targets being *hox* and *pcdh1* gene clusters (**Fig. S4g**). Genes with a cell-type specific expression such as *gata2a* and *dlx3b* (non-neural ectoderm) and *shha* (endoderm) were stably repressed in alternative lineages during gastrulation, while the DNA methyltransferase *tet2a* showed a transient silencing outside of ectodermal lineage during gastrulation (**Fig. S4h**). Overall our analysis suggests that for a specific set of genes, gene silencing is achieved once a threshold of H3K27me3 coverage is reached via cis-spreading, a process seemingly uncoupled with the transcription of these genes.

Chromatin state of cells is decoupled from transcription during early development

Considering the heterogeneity in the repressive chromatin landscape of cells observed as early as gastrulation, we asked how the interplay between active and inactive chromatin is established at this stage. To map the active chromatin, we focussed on H3K4me1; a histone modification associated with active and poised enhancers and promoters, which, unlike H3K27me3, has been observed before zygotic genome activation (ZGA) in zebrafish [24]. We generated T-ChIC data for H3K4me1 at 4, 6, 8, 10, and 12 hpf. To mitigate the interference of maternally contributed RNA, we modified our cell preparation protocol with a harsher detergent (see **Methods**), which leads to the expulsion of cytoplasmic RNA from the cells (hereafter referred to as “nuclei" batch). As expected, our nuclei dataset shows a 4-fold higher ratio of unspliced RNA compared to spliced RNA, and an overall lower number of detected genes compared to the whole-cell data, in line with the expected lack of spliced maternal RNA in the nuclei (**Fig. S5a, b**, **Supplementary Table 1**). The chromatin quality was unaffected, exemplified by the similar number and pattern of H3K27me3 MNase cuts with time from the "nuclei" and "whole cell" batch (**Fig. S5c**). Finally, we integrated our nuclei dataset with the 4-12 hpf subset of the whole-cell H3K27me3 T-ChIC dataset, creating a high-quality multi-omic dataset of 15961 cells (H3K27me3: 9197, H3K4me1: 6764) covering zebrafish gastrulation (**Fig. 3a, methods**).

With our integrated dataset, we first asked how the global chromatin environment of the cells changes with time. Comparing total MNase cuts for H3K4me1 and H3K27me3 with time, we observed that while the H3K27me3 signal globally increases in cells with time, the H3K4me1 signal decreases (**Fig. S5c**). To understand whether this global change stems from a change in the activity of cis-regulatory regions (CREs), we separated the data into H3K4me1 enriched regions (representing active or poised promoters and enhancers), H3K27me3-enriched regions (mostly observed near genes/promoters in earlier analysis), and other (mostly intergenic) regions. While the majority (84%) of H3K27me3-enriched regions were found to overlap with an H3K4me1 domain and show increasing H3K27me3 signal with time, this increase is not accompanied by a decrease in H3K4me1 on these regions (**Fig. 3b, Fig. S6a**). Instead, the decrease in signal was observed in a minor fraction of H3K27me3-unique sites, and random genomic regions away from enriched sites (**Fig. S6b**), suggesting that this global change in signal does not represent a change in CRE activity. Further, the ratio of H3K4me1 to H3K27me3 suggests that most promoters remain in a "bivalent" chromatin state during 4-12 hpf in all lineages, with a small fraction showing increased H3K4me1 activity in any specific lineage (**Fig. S6c**).

To obtain a more fine-grained view of cellular differentiation time and lineages on our integrated data, we took advantage of the high unspliced counts from our protocol. We applied the RNA velocity model [25], which uses the ratio between spliced and unspliced reads of genes to obtain the cell's differentiation path, and assigns a "latent time" to the cell, indicating their differentiation stage (**Fig. S7a-e**). We then asked how the change in the cell's chromatin state relates to the transcription of genes during their differentiation. For this, we aggregated transcriptionally similar cells into metacells, and correlated the H3K4me1 and H3K27me3 signals with unspliced (i.e. newly transcribed) RNA signals for all genes in each metacell. Interestingly, we observed that the correlation between the gene-body H3K4me1 and transcription increases with the average latent time of a metacell (**Fig. 3c-d**). Promoter regions, however, did not show this trend (**Fig. S6d**). For H3K27me3, the global signal was mostly uncorrelated with transcription on both promoters and gene bodies (**Fig. 3c, S6d)**. This suggests that despite increasing heterogeneity of chromatin signal, the overall chromatin state of a cell is decoupled from its transcriptional state during early development, and this coupling increases as the cells mature.

The chromatin state of transcription factors predicts their function during gastrulation

We next asked whether our integrated dataset could inform us about the regulation of transcription factor (TF) networks and their role in lineage specification. While many lineage-defining TFs are biochemically predicted to have both activation and silencing functions, we argued that the global level of H3K4me1 on transcription factor binding sites (TFBS) might indicate which function the TFs play in a cell. For example, if a TF expression is correlated to a gain in H3K4me1 on its binding sites, it could indicate its role as a transcriptional activator, while a loss in H3K4me1 on TFBS might indicate a silencing function in that cell. Further, if a TF function is epigenetically regulated, then the chromatin state of the TF itself would also be predictive of its function. Based on this notion, we built a prediction model that combines TF chromatin state and TF expression activity of TFBS within cells (**Methods**). With a combined model, we aim to classify TFs based on both their own chromatin regulation (regulated/independent), as well as their function based on the regulation of their targets (activator/repressor) (**Fig. 4a**).

Our model predicted the H3K4me1 activity at TFBS with high accuracy ($R^2 > 0.6$) for 45 TFs. Our classification captured the well-established developmental function of TFs, such as the activating function of *tbx16* in regulating paraxial mesoderm formation [26], and that of *tfap2a,* a transcriptional activator shown to be important for neural crest induction [27] (**Fig. 4b**). Further, it helped resolve the cell-type-specific functions of TFs predicted to be activators or repressors based on their protein domains (**Fig. S8a, Supplementary Table 2**). For example, *zbtb16a*, predicted to have a DNA-binding transcriptional repressor activity, and *zeb1a*, speculated as a context-dependent activator/repressor [28], were both revealed as a repressor during neural ectoderm (hindbrain) specification. Next, we asked whether the gain/loss of H3K4me1 activity is reflected in the gene transcription, measured as a change in nascent (unspliced) transcripts on the genes nearest to the TFBS. For our top predicted activators and repressors, we observed the expected up and downregulation of average nascent (unspliced) RNA signal of the target genes, corresponding to the change in TF expression and activity (**Fig. S8c**). This indicated that our model can capture new cell type-specific activation/repression functions of TFs during gastrulation. Additionally, our model also detects TFs that are epigenetically regulated (**Fig. S8b, Supplementary Table 2**). For TFs such as *sox13, tbx16, lhx1a*, *tfap2a*, a gain of H3K4me1, or a loss of H3K27me3, or a combination of both was associated with their respective TFBS activity. This indicates that while the majority of TFs expressed early in development appear to be regulated by alternative mechanisms, the chromatin state could play an important role in establishing the transcriptional memory for a subset of developmentally important TFs.

Discussion

In this study, we applied our novel single-cell multi-omic method named T-ChIC [15], to study the dynamics of active (H3K4me1) and repressive (H3K27me3) chromatin states during early vertebrate embryogenesis. We observe a dynamic spatiotemporal localization of these histone modifications, previously unresolved by bulk chromatin profiling assays. This data allows for a direct comparison of the chromatin state and the expression of genes in individual cells and helps to understand the role of this interaction in regulating cell fates during early development.

We observe that H3K27me3 shows a promoter-anchored spread during development. At the start of differentiation, selected genomic loci with multiple TSS are pre-marked with a broadly distributed H3K27me3 (such as the *hox* and *pcdh1* loci), while loci with single TSS (such as *pax7b*) appear as focussed H3K27me3 domains. A recent study has shown that such pre-marking is established by a non-canonical interaction between the two polycomb (PcG) complexes, PRC1 and PRC2 [29]. Here, we find that a selected set of these loci shows the spreading of H3K27me3 with differentiation, which eventually confers the silencing of host genes. While this spread appears to be mostly not lineage-specific, we do observe lineage-specific demethylation of many genes which are activated later in development, suggesting that the global landscape of H3K27me3 could be established through a lineage-agnostic spread, followed by lineage-specific demethylation. A recent study using mouse embryonic stem cells proposes nucleation and spreading as a way to maintain PcG silencing [30]. Based on our observations, we propose that this mechanism could also help to propagate the spread of silencing during development. Finally, we see that in the absence of H3K27 de-methylation, important developmental genes such as *hox, pax* and *shh* genes are silenced in a spatiotemporal manner. A mis-localized expression of these known PcG targets has been observed after a deletion of the core PRC2 enzyme, *ezh2* [31,32]. Apart from confirming these known targets, we additionally identify developmental genes such as *rfx4,* important for neural tube formation [33], *dlx3b,* important for placode development [34], among others, as novel PcG targets.

Although a rather large number of gene promoters are marked by H3K27me3 in pluripotent cells, only a minority of these show a genomic spread and silencing of the genes. By comparing the H3K4me3 to H3K27me3 signal on promoters, we find that most of these promoters are co-marked at 4-12 hpf, suggesting that they might serve as a “placeholder” for activation or silencing later in development. This provides an explanation to why the cis-spreading, and not the promoter enrichment of H3K27me3 is linked to gene silencing during development.

In line with previous studies [24,35], we find that the H3K4me1 is widespread in the genome of pluripotent cells, marking a large number of TF motifs and other genomic regions. While this chromatin mark systematically disappears in regions outside of cis-regulatory elements (CREs) during development, its activity on the CREs does not show a monotonous change with time. In fact, a systematic increase in H3K27me3 without a loss of H3K4me1 leads to a bivalent chromatin state on most CREs, together with a lineage-specific gain or loss of H3K4me1 on selected CREs. We show that these lineage-specific changes in H3K4me1 can be leveraged to predict the lineage-specific activator or repressor functions of TFs, by correlating this activity with the TF's own expression and chromatin states. Using this approach, we find novel functions of TFs in lineage specification, such as the role of *zbtb16a/b, zeb1a/b* as negative regulators during ectoderm specification, and the *tfap2a/b* as a positive driver of non-neural ectoderm during gastrulation. We also find selected lineage-specifying TFs such as *zfhx3, foxc1a,* and *irx3a* whose activity seem to be regulated by their own chromatin state during gastrulation. These results might point to a new pathway through which the chromatin states of the cells play a role in specifying cell fates, i.e. by establishing a transcriptional memory on key lineage regulators.

Overall, comparing the active and inactive chromatin states of cells, we observe that the active state is a better predictor of a cell's functional (transcriptomic) state in early development. A caveat is that we have not mapped other important silencing chromatin states, such as H3K9me3 or DNA methylation, which may turn out to be highly dynamic in early development.

We also see that both active and inactive states are rather uncoupled from nascent transcription in pluripotent cells, and get correlated as the cells mature in development. Note that this maturation time is not necessarily the same as the developmental time (hpf) of the embryo, as the transcriptionally mature cells collected from early time points also show a high correlation of active chromatin state and transcription. Therefore, we propose that a correlation of chromatin and transcriptional state of cells could be a hallmark of cell identity formation during development. Future studies to systematically map the overall chromatin state of single cells and gene expression would further explain how cell fates are established during embryogenesis.

Methods

Embryo collection and single-cell dissociation

Wild-type TL embryos were collected 20 minutes after fertilization in a petri dish with E3 medium (13.3 mg NaCl, 0.63 mg KCl, 2.4 mg $CaCl_2$, 4.0 mg $MgSO_4$, diluted in a final volume of 1 L of H_2O) and kept at 28.5°C in an incubator. During the first hour, the unfertilized embryos were discarded. At the desired stage (checked visually), embryos were transferred to a glass beaker with a small amount of E3 medium (avoiding contact with air). They were dechorionated by incubation in 5mL of 1 mg/mL of pronase (Pronase from Streptomyces griseus, Sigma Aldrich, cat. number: 10165921001) solution at 28.5°C for 3 to 5 minutes. When the chorions started to blister, part of the pronase solution was decanted and the embryos were washed several times with the E3 medium to stop the reaction. Dechorionated embryos were transferred with a glass pipette to 1.5-mL protein low-binding Eppendorf tubes, such that each tube only contained between 30 and 50 embryos. 200 µL of Ca^{2+}-free Ringer's solution (116 mM NaCl, 2.9 mM KCl, 5.0 mM pH 7.2 HEPES) was added, and embryos were left to deyolk for 5 minutes at room temperature (RT) after resuspending up and down gently a few times. Deyolked embryos were pelleted at 400g for 3 minutes using a swinging-bucket centrifuge (Eppendorf 5430R with swinging buckets). Supernatant was discarded by aspiration and the pellet was washed and resuspended with 500 µL of PBS + 10% FBS before being spun down again. For early time points (4, 6, and 8 hpf), cells were dissociated with the addition of 200 µL of pre-warmed FACSmax cell dissociation solution (Genlantis T200100) for 5 minutes resuspending gently up and down at RT. For later time points (10, 12 and 24 hpf) cells were dissociated with the addition of 200 µL of pre-warmed Protease solution (100 µL Trypsin-EDTA (0.5%), no phenol red, Thermofisher 15400054, 20 µL 10X PBS0, 80 µL H_2O) for 6 minutes on a shaker (Eppendorf Thermomixer Comfort) at 28°C and 400 rpm, resuspending vigorously every 2 minutes. Cells were checked under the microscope to confirm proper single-cell dissociation. After dissociation, cells were filtered with a 35 µL sieve (Corning, 352235), and washed with 500 µL of PBS + 10% FBS. Finally, cells were counted and resuspended in Wash Buffer 1 (WB1) (47.5 mL RNAse-free H_2O, 1 mL 1 M HEPES pH 7.5 (Invitrogen), 1.5 mL 5M NaCl, 3.6 mL pure spermidine solution (Sigma Aldrich, S2626-1G), 0.05% Tween20 (Sigma Aldrich, P1379-25ML), 1 tablet of cOmplete™, EDTA-free Protease Inhibitor Cocktail, (Merck, 5056489001), 0.2 U/µL RNasin® Plus Ribonuclease Inhibitor (Promega, N2615) (1:200), 4 µL/mL of 0.5 M EDTA (Thermofisher, AM9261) and kept at +4°C before starting the CellTracer staining.

CellTracer staining for timepoints

Cells were washed once in 200 µL of Wash Buffer (WB) (47.5 mL RNAse-free H_2O, 1 mL 1 M HEPES pH 7.5 (Invitrogen), 1.5 mL 5 M NaCl, 3.6 mL pure spermidine solution (Sigma Aldrich, S2626-1G), 0.05% Tween20 (Sigma Aldrich, P1379-25ML), 1 tablet of cOmplete™, EDTA-free Protease Inhibitor Cocktail, (Merck, 5056489001), 0.2 U/µL RNasin® Plus Ribonuclease Inhibitor (Promega, N2615) (1:200)) without spermidine solution. After centrifugation (400 g for 3 min) they were resuspended in 1 mL of WB1 without spermidine but with RNasin® (1:40) and 1 µL of dye (either CellTrace™ CFSE Cell Proliferation Kit, for flow cytometry, Thermofisher C34570, CellTrace™ Yellow Cell Proliferation Kit, for flow cytometry, Thermofisher C34573, CellTrace™ Far Red Cell Proliferation Kit, for flow cytometry, Thermofisher C3457, or combinations of two of these). Cells were vortexed very well and kept in the dark at +4°C for 20 minutes to stain. The staining was stopped with the addition of 70 µL of rat serum (Sigma Aldrich, R9759-5ML) and a 5 min incubation at RT. Lastly, cells were washed and resuspended in WB1 with spermidine solution (0,072 µL/mL) and 4 µL/mL 0,5 M EDTA. Once all time points had been stained with their appropriate dye/dyes combinations they were pooled in a 0.5-mL protein-low binding tube with approximately 1 million cells in total.

Nuclei isolation

Nuclei can be isolated following the same procedure as cells but using 10% Saponin (Sigma, 47036-250G-F) instead of 0.05% Tween20 in the WB and WB1.

T-ChIC library preparation and sequencing

The detailed, step-by-step scT-ChIC protocol applied to Zebrafish embryos is available at: https://www.protocols.io/private/714FFF81809D11EE958A0A58A9FEAC02. The Pa-MN fusion protein was produced as described earlier [36]. Final DNA libraries are sequenced paired-end 100 bp, on either a NovaSeq or NextSeq2000, at a sequencing depth between 15 and 25 million reads per sample (384-well plate).

Processing and quality control of T-ChIC data

The first-in-pair reads from the T-ChIC protocol contain an RNA or ChIC barcode in the following format "*RNA: 6N7X, ChIC: 3N8X*"; where N=UMI nucleotide and X=Cell barcode nucleotide. We used a custom Python script to split the raw *.fastq* files into the ChIC and RNA fractions based on which one of the 2 barcode patterns is observed at the start. The two fractions are then independently mapped to the GRCz11/danRer11 genome. A complete processing workflow with all parameters is available at: https://github.com/bhardwaj-lab/scChICflow (v 0.4) and is briefly described below.

The **RNA fraction** was trimmed using cutadapt (v2.1) [37] with parameters `*-e 0.1 -q 16 -O 3 --trim-n --minimum-length 10 --nextseq-trim=16 -A W{'10'}*`, along with Illumina truseq barcodes provided as `-a and -b ` options. The trimmed reads are mapped to the genome using STAR (v 2.7.11)[38], using the "*StarSolo*" mode, with these important parameters `--sjdbGTFfile <dr11_ens104.gtf> --outFilterIntronMotifs RemoveNoncanonical --soloCBmatchWLtype Exact --soloType CB_UMI_Simple`, where "*dr11_ens104.gtf*" refers to the ENSEMBL annotation version 104 (GRCz11) [39]. Secondary and supplementary alignments and low-quality mappings (<MAPQ 255) were removed using samtools (v1.21) [40] and reads were de-duplicated with UMI-tools (v.1.0.0) [41] using cell barcode and UMI position, along with options `*--method unique --spliced-is-unique*`. Coverage files were created using deepTools *bamCoverage* with CPM normalization [42].

For the **ChIC fraction**, barcodes were moved into the read header using UMI-tools *extract* (v.1.0.0). Reads were trimmed using cutadapt (v2.1) with parameters `*-e 0.1 -O 5 -u 1 -u -2 -U -2 -a W{10} -A W{10} -q 30 --trim-n --minimum-length 20 --nextseq-trim=30*`, along with illumina truseq barcodes provided as `-a and -b ` options. The trimmed reads are mapped to the genome using hisat2 (v2.2.1) [43], with parameters `--sensitive --no-spliced-alignment --no-mixed --no-discordant --no-softclip -X 1000`. Reads were de-duplicated with UMI-tools (v.1.0.0) using cell barcode and UMI position, along with options `*--method unique --spliced-is-unique --soft-clip-threshold 2*`. Quality control was performed using deepTools. Reads were counted on 50-kb windows in the genome using sincei [44].

Comparison of ChIC signal with publicly available data

We downloaded the raw *.fastq* files for 6hpf bulk CUTnRUN data of H3K27me3 from Ozdilek et. al. (GSE178343) [45], and raw *.fastq* files of 12hpf and 24hpf ChIP-seq data from the danio-code portal (accessions - 12hpf H3K27me3: DCD003854SQ, 12hpf H3K4me1: DCD003854SQ, 24hpf H3K27me3: DCD003200SQ). All *.fastq* files were mapped to the GRCz11 genome using snakePipes' DNA-mapping workflow, with parameters `--trim --fastqc --mapq 5 --dedup --bwBinSize 1000` [46]. The de-duplicated BAM files were subsampled to match the sequencing depth of the corresponding pooled time points (early vs 6hpf, middle vs 12hpf, late vs 24hpf), and the read coverage was compared using deepTools [42] multiBAMSummary (with parameter `*-bs 50000*`) and plotFingerPrint (with parameters `*--skipZeros -bs 10000 -n 50000*`).

Cell clustering and annotation using RNA signal

For clustering of single cells based on RNA signal, we used the "spliced" count matrices for "whole-cell" T-ChIC data, and "total" (spliced+unspliced+ambiguous) counts for "nuclei" T-ChIC data. Filtering and clustering of cells were performed in scanpy [47]. We removed cells with `*total_counts < 1000, or n_genes_by_counts > 10000, or pct_counts_in_top_100_genes > 0.6*`). We also removed cells with < 70% of counts on protein-coding genes. We selected genes present in at least 1% of cells (or at least 50 cells, whichever is smaller), and selected the top 4000 variable genes based on their analytical Pearson residuals [48]. We used the Pearson residuals to calculate principal components (PCs) and built a neighbor graph using 50 PCs and 30 neighbors (20 for nuclei data). We then used it to build a paga graph [49] based on Leiden clusters (*paga threshold = 0.1, leiden resolution = 1.5*). For 2D representation, UMAPs were initiated with the paga graph, along with additional parameters `*min_dist=1, spread=5*` (*spread = 1* for nuclei).

For the annotation of cell types and all other analyses, we calculated the normalized ChIC and RNA signal using the “shifted log transform” method (`*1/sqrt(alpha) log(4 * alpha * x + 1)*`) [50], with a fixed overdispersion (alpha) of 0.05 and total counts (“normed_sum”) as library size factors. For annotation of cells, we obtained the raw count matrices from Wagner et. al [19] and subsetted for the 4, 6, 8, 10, 14, and 24 hpf timepoints (for the “nuclei" batch, we also excluded 24 hpf). We normalized the counts in the same manner as our counts and selected the top 4000 variable genes (using `*FindVariableFeatures(selection.method = “vst”)*` in seurat). We then performed CCA-MNN analysis in seurat using `*FindTransferAnchors(method=”cca”)*` and used the transfer score to predict labels for single cells (**Fig S3A**). For top predicted labels for each cluster, we then manually confirmed the marker gene expression from ZFIN in the respective cluster, followed by renaming the cluster to suitable ZFIN cell ontology [22].

Integrated analysis of nuclei and whole cell data

To integrate the “nuclei" and “whole-cell” subsets of data, we merged the cells from the “nuclei" batch with that of 4-12 hpf subset of “whole-cell” batch, resulting in 15961 cells. We then removed genes with total spliced or unspliced counts < 100, or genes detected in < 100 cells, from the merged data, and calculated the top 4000 variable genes (HVGs) based on their analytical Pearson residuals [48] and used the intersection of the HVGs from the 2 batches (3091 genes) to perform PCA based on the pearson residuals of the “unspliced” counts from the 2 batches. Top 50 PCs were then used to align the 2 batches using harmony [51]. The harmony-corrected PCs were used for further analysis of the joint dataset (clustering, annotation, metacells and RNA velocity).

For calculation of latent time and lineages of cells on the integrated 4-12 hpf data, we combined the RNA velocity and diffusion pseudotime approach, using cellrank [52]. We calculated RNA-velocity using the “dynamical” model as described in the scVelo package [53] . The cell-specific moments *Ms* and *Mu* were calculated using the HVGs and PCs from the above analysis, and top 1000 genes were used for the dynamical model to calculate the gene-shared latent time and cell-specific velocities (*scv.tl.recover_dynamics* with parameters: fit_connected_states = False, max_iter = 50, t_max = 12, fit_basal_transcription=True, *scv.tl.velocity* with parameters: *min_r2=0.2, groups_for_fit=<8-12 hpf>*). The latent time was combined with connectivities using cellrank (*cr.tl.transition_matrix* parameter: *weight_connectivities=0.2*), and 1 initial and 6 terminal states were calculated using cellrank's GPCCA estimator.

Metacell analysis

To obtain a detailed view of cellular heterogeneity while reducing dropouts, we aggregated transcriptionally similar cells into the so-called "metacells", based on archetype analysis [23] . We used the SEACells python package with parameters `*n_SEACells=nc, n_waypoint_eigs=15, convergence_epsilon = 1e-5*`, where nc = 180 whole-cell data, and 160 for the integrated 4-12 hpf data. Each metacell was then annotated with the mean latent time or pseudotime of underlying single-cells, and the max number of cells belonging to an annotated cell type or collection time (hpf). For analysis involving a comparison of H3K4me1 and H3K27me3, the underlying number of single-cells were downsampled for each metacell, such that equal number of cells from both the histone modifications are assigned to each metacell, and only metacells with a minimum of 20 cells from both histone modifications were kept, to assure robust results.

Cell clustering using the ChIC signal

For the clustering of single cells based on ChIC signal, we performed latent semantic analysis (LSA) using the gensim package in python [54]. The Cells*Regions sparse matrix was treated as a vector of documents (Cells), where region counts represent word frequency. The documents were then transformed with log term-frequency (TF), inverse document frequency (IDF) as follows:

$$tf_{td} = 1 + \log_2 f_{i_k}$$

$$idf_{(}t, D) = \log_2 \left(\frac{N}{n_k} \right)$$

$$TF - IDF_{(}t, d, D) = tf_t, d * idf_t, D$$

Where $f_i k$ refers to the count frequency of a (50 kb) genomic bin T_k in a cell D_i, and nk refers to the number of cells containing non-zero counts for the bin T_k . The output is subjected to a pivoted unique normalization [55] to take into account the difference in total number of detected regions per cell.

$$pivotednorm = (1.0 - slope) * pivot + slope * TF - IDF_(t, d, D)$$

In our case, we calculated pivot as the average number of non-zero bins across all cells, and fixed the slope to *0.25* (recommended by [55]). The resulting matrix is subjected to a truncated SVD (singular value decomposition) [56], yielding Cell*Topic and Region*Topic matrices. Similar to scRNA-seq, we calculated 30 nearest neighbors using the 50 topics from the LSA output (dropping Topic-1, which strongly correlates with read depth), and used it to build the paga graph (with *threshold* = *0.1*). UMAPs were initialized using the PAGA graph, with parameters `*min_dist=0.1, spread=5*`. Leiden clusters were calculated on the neighborhood graph with `*resolution=1.5*`.

Peak calling and annotation

For the detection of regions with both narrow and broad enrichment in the genome, we used a two-step peak calling approach. We first pooled all our filtered cells from 4 to 24hr time points into a BAM file and used histoneHMM (v1.7) function `*call_regions*` with parameters `*-bs 750 -P 0.1*` [57], and further removed the detected regions with average posterior probability < 0.4, and referred to them as “domains”. Next, we performed peak-calling using MACS2 (v2.2.4) [58] on the same file, with parameters `--mfold 0 50 --extsize 200 --broad --keep-dup all`, and overlapped these peaks with the domains detected from histoneHMM. Peaks overlapping with the histoneHMM domains, and having a local enrichment score >=50, were referred to as “subpeaks”. For further analysis, we replaced the domains containing subpeaks with their respective subpeaks, resulting in 11221 enriched domains for H3K27me3, and 74004 domains for H3K4me1.

For peak annotation, we used the *genomicRanges* R package [59] to classify these domains into “promoter” (within +300 or -200 bases of a transcription start-site), genic (overlapping a gene body, but not promoter), and “intergenic” (outside promoters or gene body). The “genic” domains were re-classified into “gene covering”, if they covered >= 80% of a gene. All domains were annotated with the gene(s) which overlapped with, or (in the case of intergenic domains) were nearest to them. To detect which cis-regulatory elements are present inside these domains, we overlapped them with the location of “consensus PADREs” annotated by the danio-code project [60]. We used the gimmemotifs [61] (v0.18.0), with parameters `scan -*N 30*` to annotate these peaks for the associated transcription factor binding sites (TFBS), using the `vertebrate.v5.0` motif database. The detected motifs were then filtered for zebrafish TF motifs that are also present in the Swiss regulon (dr11) database [62], resulting in 590 motifs belonging to 912 TFs.

Linear regression analysis for H3K27me3

To detect the regions in the genome showing H3K27me3 cis-spreading, we used the table of 5-kb bin counts in single cells. We defined the “center” bin as the bins showing non-zero counts in >= 5% of “pluripotent” cells (784 bins), and “neighbor” bins as the 10 up and downstream bins to the center bins. We then applied linear regression to predict the counts in “neighbor” bins ($\hat{Y}$) as a function of the sum of counts in the “center” bins ($\hat{X}$) across metacells.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i$$

To test if the spread is germ layer-specific, we compared this model to a second model including germ layer covariate ($\hat{X}_2$) via a likelihood ratio test.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_2 + \hat{\epsilon}_i$$

Similarly, we used a linear regression model formulation above for the prediction of H3K27me3 silenced genes, where

$$\hat{Y} = log2((sum(counts_{un}) * sum(introns_{un})/10000))$$

and

$$\hat{X} = log2(counts_{k27} * length_{k27}/10000)$$

$counts_{un}$ represent unspliced counts of genes overlapping a H3K27me3 peak, and $introns_{un}$ is the length of their introns. $counts_{k27}$ are ChIC counts on the H3K27me3 domain and $length_{k27}$ is the length of the domain.

TF activity prediction and classification

To calculate the TF activity using H3K4me1 signal, we filtered our annotated H3K4me1 peaks (with assigned cPADREs) for peaks uniquely enriched for H3K4me1. Zebrafish TF motifs from the Swiss regulon (dr11) database [62], (590 motifs belonging to 912 TFs) were assigned to the peaks, based on motif match score inside the cPADREs within those peaks. Next, we obtained the H3K4me1 counts per peak per cell and assigned these counts to each TF motifs annotated with these peaks. Finally, we converted these raw counts into bias-corrected TF motif deviance scores per cell using chromVar [63]. We also calculated deviance scores for metacells, using the aggregated counts per metacell, instead of single cells.

For TF activity prediction, we calculated the normalized H3K4me1 and H3K27me3 and (spliced) RNA counts per metacell, along with the metacell annotations (germ layer, latent time) and used them to predict the TF activity using lasso-penalized regression [64]. We first divided the data into a 70-30 (training/test) set and used the training set to tune the penalty parameter (λ) using grid search on 10-fold resamples. The top model was then run on each TF separately, and evaluated on the test set based on R-squared estimates. For classification of TFs, we extracted the weights for the final model for each TF at the highest λ, and interpreted them based on previous knowledge about these histone modifications. For example, since H3K4me1 activity represents active or poised enhancer, a positive correlation (weight > 0) between a TF expression and H3K4me1 activity suggests its action as an activator, while a negative correlation (weight < 0) suggests a repressor. Similarly, a non-zero weight for a TFs H3K4me1 or H3K27me3 level would indicate that a TF activity is regulated by either, or both of the marks.

Data availability

Raw sequencing data, along with annotated .loom formatted files from this study are available at GEO (GSE265874).

Code availability

The code for processing of tChIC data from raw (fastq) files up to count tables is available at https://github.com/bhardwaj-lab/scChICflow. Scripts for downstream analysis are available at github.com/vivekbhr/fishchic_2024.

Acknowledgements

We acknowledge the Utrecht Sequencing Facility (USEQ) for providing sequencing service and data. USEQ is subsidized by the UMC Utrecht and The Netherlands X-omics Initiative (NWO project 184.034.019). We thank Reinier van der Linden (Hubrecht FACS facility) for performing single-cell sorting. This work was supported by European Research Council Advanced under grant ERC-AdG 101053581--scTranslatomics, and the NWO consortium grant OCENW.GROOT.2019.017 to A.v.O. The SNF (P2BSP3-174991), HFSP (LT000209/2018-L) and Marie Skłodowska-Curie Actions (798573) supported P.Z. EMBO LTF (ALTF 1197–2019) supported V.B.

Author information

Contributions

V.B and P.Z designed the experiments with input from A.v.O. P.Z contributed the T-ChIC protocol. A.G and H.V.G performed the experiments with inputs from P.Z and V.B. V.B performed the analysis and interpreted results with input from A.v.O, P.Z and A.G. V.B and A.v.O wrote the manuscript with input from other authors.

References

1. Xu, P.-F., Houssin, N., Ferri-Lagneau, K. F., Thisse, B. & Thisse, C. Construction of a vertebrate embryo from two opposing morphogen gradients. *Science* **344**, 87–89 (2014).
2. Bogdanović, O., van Heeringen, S. J. & Veenstra, G. J. C. The epigenome in early vertebrate development. *Genesis* **50**, 192–206 (2012).
3. Fitz-James, M. H. & Cavalli, G. Molecular mechanisms of transgenerational epigenetic inheritance. *Nat. Rev. Genet.* **23**, 325–341 (2022).
4. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
5. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

6. Liu, H. *et al.* DNA methylation atlas of the mouse brain at single-cell resolution. *Nature* **598**, 120–128 (2021).

7. Nichols, R. V. *et al.* High-throughput robust single-cell DNA methylation profiling with sciMETv2. *Nat. Commun.* **13**, 7627 (2022).

8. Bai, D. *et al.* Simultaneous single-cell analysis of 5mC and 5hmC with SIMPLE-seq. *Nat. Biotechnol.* (2024) doi:10.1038/s41587-024-02148-9.

9. Zeller, P. *et al.* Single-cell sortChIC identifies hierarchical chromatin dynamics during hematopoiesis. *Nat. Genet.* (2022) doi:10.1038/s41588-022-01260-3.

10. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).

11. Wu, S. J. *et al.* Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.* **39**, 819–824 (2021).

12. Cheung, P. *et al.* Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *Cell* **173**, 1385–1397.e14 (2018).

13. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

14. Guo, F. *et al.* Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).

15. Zeller, P. *et al.* T-ChIC: multi-omic detection of histone modifications and full-length transcriptomes in the same single cell. *bioRxiv* 2024.05.09.593364 (2024) doi:10.1101/2024.05.09.593364.

16. Salmen, F. *et al.* High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol.* **40**, 1780–1793 (2022).

17. Fishman, L. *et al.* Single-cell temporal dynamics reveals the relative contributions of transcription and degradation to cell-type specific gene expression in zebrafish embryos.

bioRxiv (2023) doi:10.1101/2023.04.20.537620.

18. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).
19. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
20. Vastenhouw, N. L. *et al.* Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464**, 922–926 (2010).
21. de la Calle Mustienes, E., Gómez-Skarmeta, J. L. & Bogdanović, O. Genome-wide epigenetic cross-talk between DNA methylation and H3K27me3 in zebrafish embryos. *Genom Data* **6**, 7–9 (2015).
22. Bradford, Y. M. *et al.* Zebrafish information network, the knowledgebase for Danio rerio research. *Genetics* **220**, (2022).
23. Persad, S. *et al.* SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01716-9.
24. Murphy, P. J., Wu, S. F., James, C. R., Wike, C. L. & Cairns, B. R. Placeholder Nucleosomes Underlie Germline-to-Embryo DNA Methylation Reprogramming. *Cell* **172**, 993–1006.e13 (2018).
25. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
26. Payumo, A. Y., McQuade, L. E., Walker, W. J., Yamazoe, S. & Chen, J. K. Tbx16 regulates hox gene activation in mesodermal progenitor cells. *Nat. Chem. Biol.* **12**, 694–701 (2016).
27. Dooley, C. M. *et al.* The gene regulatory basis of genetic compensation during neural crest induction. *PLoS Genet.* **15**, e1008213 (2019).
28. Gheldof, A., Hulpiau, P., van Roy, F., De Craene, B. & Berx, G. Evolutionary functional analysis and molecular regulation of the ZEB transcription factors. *Cell. Mol. Life Sci.* **69**, 2527–2541 (2012).
29. Hickey, G. J. *et al.* Establishment of developmental gene silencing by ordered polycomb

complex recruitment in early zebrafish embryos. *Elife* **11**, (2022).

30. Veronezi, G. M. B. & Ramachandran, S. Nucleation and spreading maintain Polycomb domains every cell cycle. *Cell Rep.* **43**, 114090 (2024).
31. San, B. *et al.* Normal formation of a vertebrate body plan and loss of tissue maintenance in the absence of ezh2. *Sci. Rep.* **6**, 24658 (2016).
32. Yette, G. A., Stewart, S. & Stankunas, K. Zebrafish Polycomb repressive complex-2 critical roles are largely Ezh2- over Ezh1-driven and concentrate during early embryogenesis. *bioRxiv* 2020.12.31.424918 (2021) doi:10.1101/2020.12.31.424918.
33. Sedykh, I. *et al.* Zebrafish Rfx4 controls dorsal and ventral midline formation in the neural tube. *Dev. Dyn.* **247**, 650–659 (2018).
34. Esterberg, R. & Fritz, A. dlx3b/4b are required for the formation of the preplacodal region and otic placode through local modulation of BMP activity. *Dev. Biol.* **325**, 189–199 (2009).
35. Kaaij, L. J. T. *et al.* Enhancers reside in a unique epigenetic environment during early zebrafish development. *Genome Biol.* **17**, 146 (2016).
36. Schmid, M., Durussel, T. & Laemmli, U. K. ChIC and ChEC; genomic mapping of chromatin proteins. *Mol. Cell* **16**, 147–157 (2004).
37. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
38. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
39. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
40. Li, H. *et al.* 692 (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2693.
41. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
42. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–5 (2016).

43. Kim, D., Langmead, B. & Salzberg, S. HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes. Preprint at (2017).

44. Bhardwaj, V. & Mourragui, S. User-friendly exploration of epigenomic data in single cells using sincei. *bioRxiv* 2024.07.27.605424 (2024) doi:10.1101/2024.07.27.605424.

45. Akdogan-Ozdilek, B., Duval, K. L., Meng, F. W., Murphy, P. J. & Goll, M. G. Identification of chromatin states during zebrafish gastrulation using CUT&RUN and CUT&Tag. *Dev. Dyn.* **251**, 729–742 (2022).

46. Bhardwaj, V. *et al.* snakePipes: facilitating flexible, scalable and integrative epigenomic analysis. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz436.

47. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

48. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol.* **22**, 258 (2021).

49. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

50. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20**, 665–672 (2023).

51. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

52. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).

53. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

54. Řehůřek, R. & Sojka, P. Software Framework for Topic Modelling with Large Corpora. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* 45–50 (ELRA, Valletta, Malta, 2010).

55. Singhal, A., Buckley, C. & Mitra, M. Pivoted Document Length Normalization. *SIGIR Forum* **51**, 176–184 (2017).

56. Halko, N., Martinsson, P.-G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *arXiv [math.NA]* (2009).

57. Heinig, M. *et al.* histoneHMM: Differential analysis of histone modifications with broad genomic footprints. *BMC Bioinformatics* **16**, 60 (2015).

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

59. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

60. Baranasic, D. *et al.* Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements. *Nat. Genet.* **54**, 1037–1050 (2022).

61. van Heeringen, S. J. & Veenstra, G. J. C. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).

62. Pachkov, M., Erb, I., Molina, N. & van Nimwegen, E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.* **35**, D127–31 (2007).

63. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

64. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).

Fig. 1

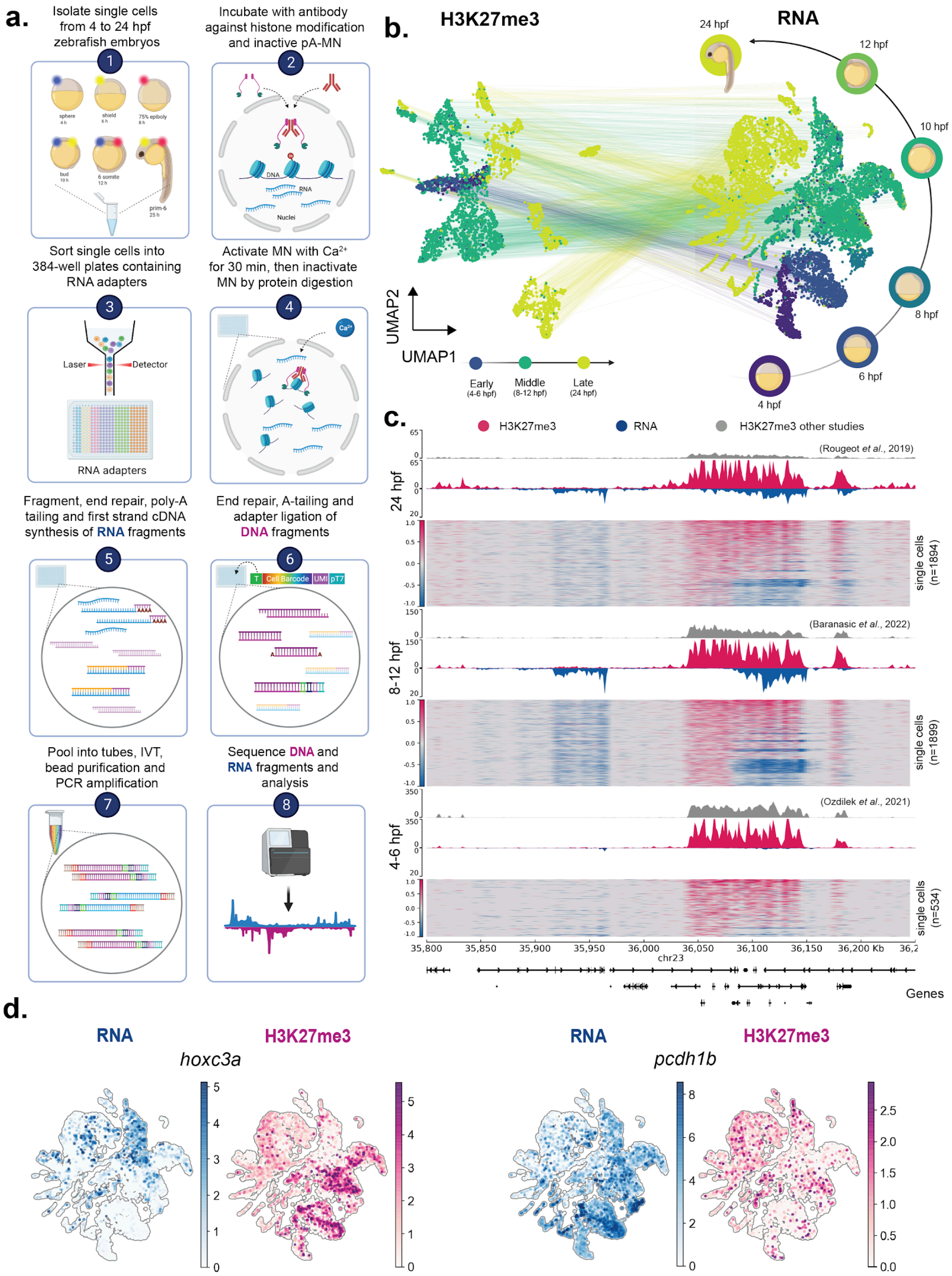

Figure 1: T-ChIC quantifies of full length transcripts and histone modifications from the same single cell

a. T-ChIC experimental workflow: Labeled single cells from different time points are pooled and sorted in 384-well plated to be processed for DNA and RNA fragments, before re-pooling and amplification to produce sequencing libraries. **b. UMAP projections of single cells** using signals from H3K27me3 (left) and transcriptome (right) and colored by timepoints. The 6 sampled timepoints (right) are pooled into 3 groups (early/middle/late) based on the complexity of H3K27me3 signal. **c. Single-cell trackplot** showing signal on the 450-kb region around the hoxc gene cluster. The heatmaps show signals in single cells, while the coverage tracks on top show the pseudo-bulk signal (blue: RNA, pink: H3K27me3). Publicly available bulk H3K27me3 datasets are shown for comparison (grey) **d. UMAP projections** (based on transcriptome signal) showing gene-level normalized signal for RNA (blue) or H3K27me3 (pink) on 2 selected genes.

Fig. 2

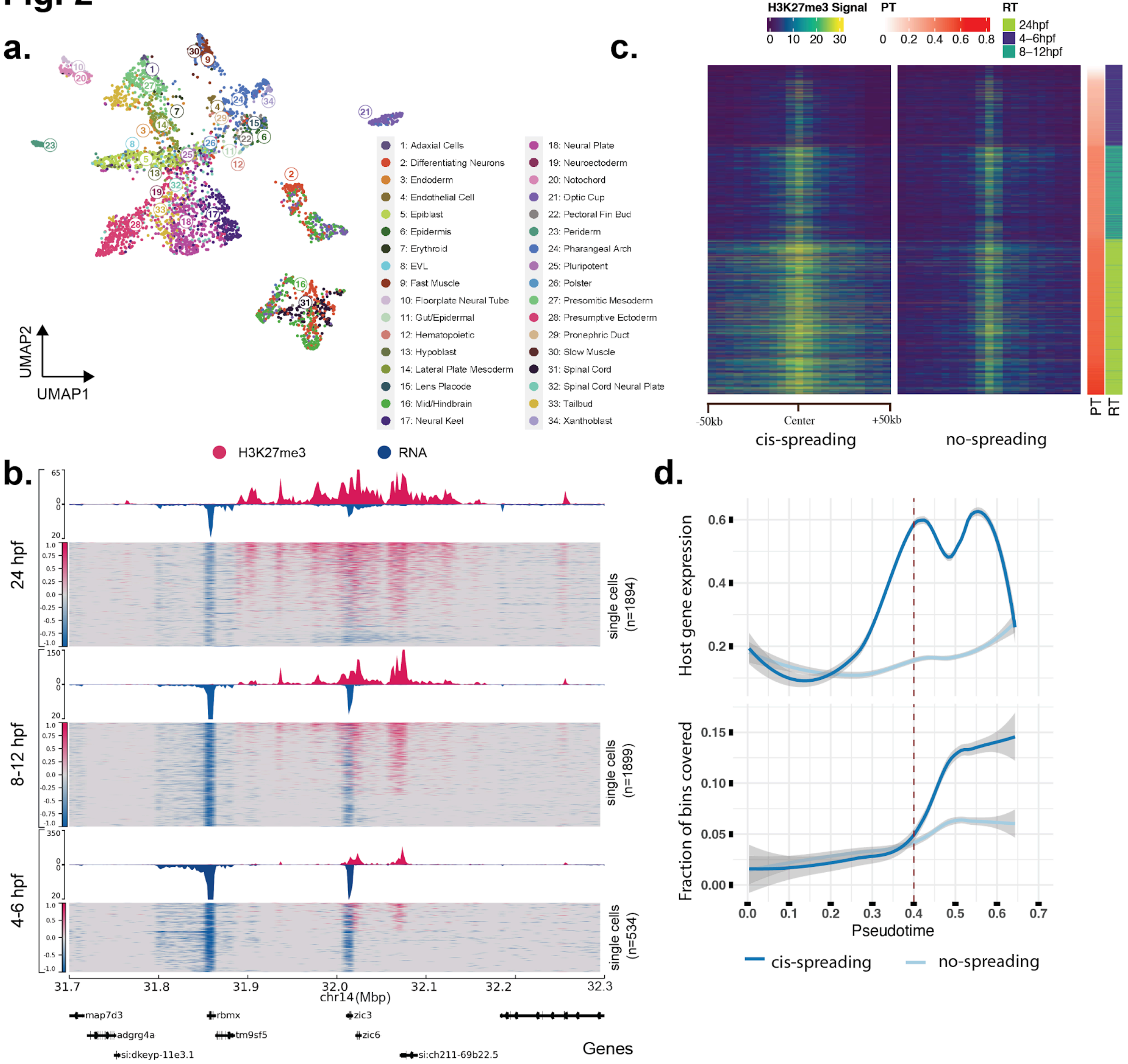

Figure 2: Spatio-temporal spreading of H3K27me3 correlates with gene silencing

a. UMAP projection of cells based on H3K27me3 signal (same as Fig 1B, left) indicating the different cell types annotated using the transcriptome signal. **b. An example genomic locus** that demonstrates the cis-spreading of H3K27me3 signal (pink) around the *zic3* gene with time during development. Note that apart from the *zic3* gene, the spreading also correlates with a downregulation of the expression for the nearby gene (*rbmx*) with time. **c. Single-cell heatmap** (each row is a single cell) showing the average H3K27me3 signal for the top 100 genes detected by the linear model, showing increase in spreading of signal at the 100-kb region surrounding the center bin with pseudotime. Center bin was identified as the bin with non-zero signal in pluripotent cells. **d. Line plots comparing H3K27me3 spreading and gene expression with pseudotime**. The bottom panel shows the average fraction of bins that show H3K27me3 signal on 2 sets of genes (spreading, non-spreading) with time, while the top panel shows the average gene expression of these genes-sets along pseudotime. Spreading is initially de-coupled with gene expression, and silencing is achieved only after a certain fraction of the gene's body has been covered.

Fig. 3

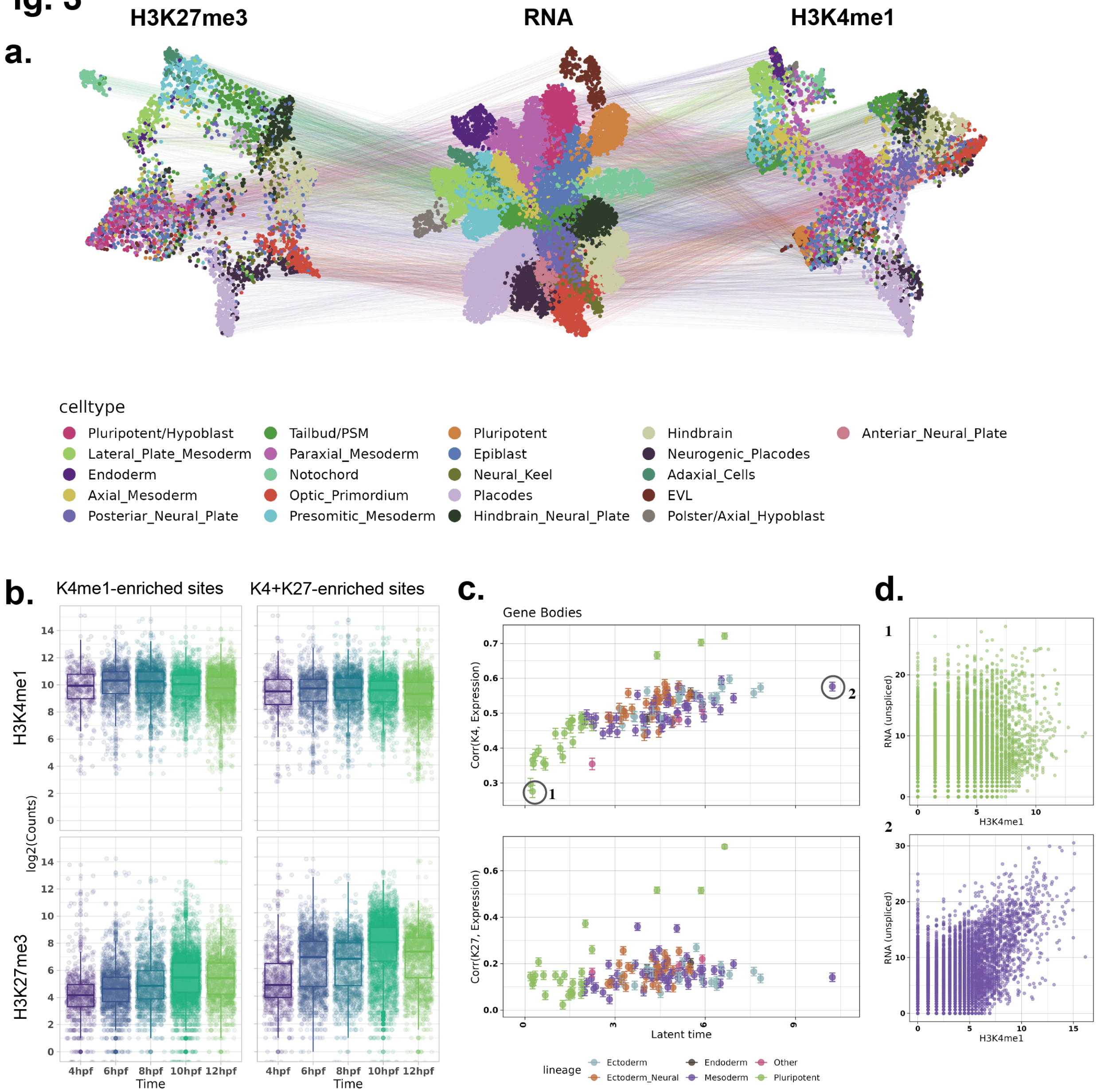

Figure 3: Integrative analysis of H3K27me3, H3K4me1 and transcriptome

a. UMAP projection of the single cells based on the 3 modalities H3K27me3 (left), RNA (center) and H3K4me1 (right) after integration of the two batches and annotation of cells. **b. Total UMI counts per cell on H3K4me1 enriched regions** from the integrated dataset, divided into K4me1-unique regions and regions co-enriched for K27me3. H3K4me1 signal (top panels) on these regions remains unchanged with time, while H3K27me3 signal (bottom panels) increases. **c. Correlation between histone modification and gene expression with latent time**. Top plot shows metacells arranged by pseudotime (Y-axis) and the Pearson correlation coefficient between H3K4me1 and RNA (top) and H3K27me3 and RNA. **d. Scatterplot showing correlation between H3K4me1 and unspliced RNA** for two selected metacells indicated in C (early and late in latent time).

Fig. 4

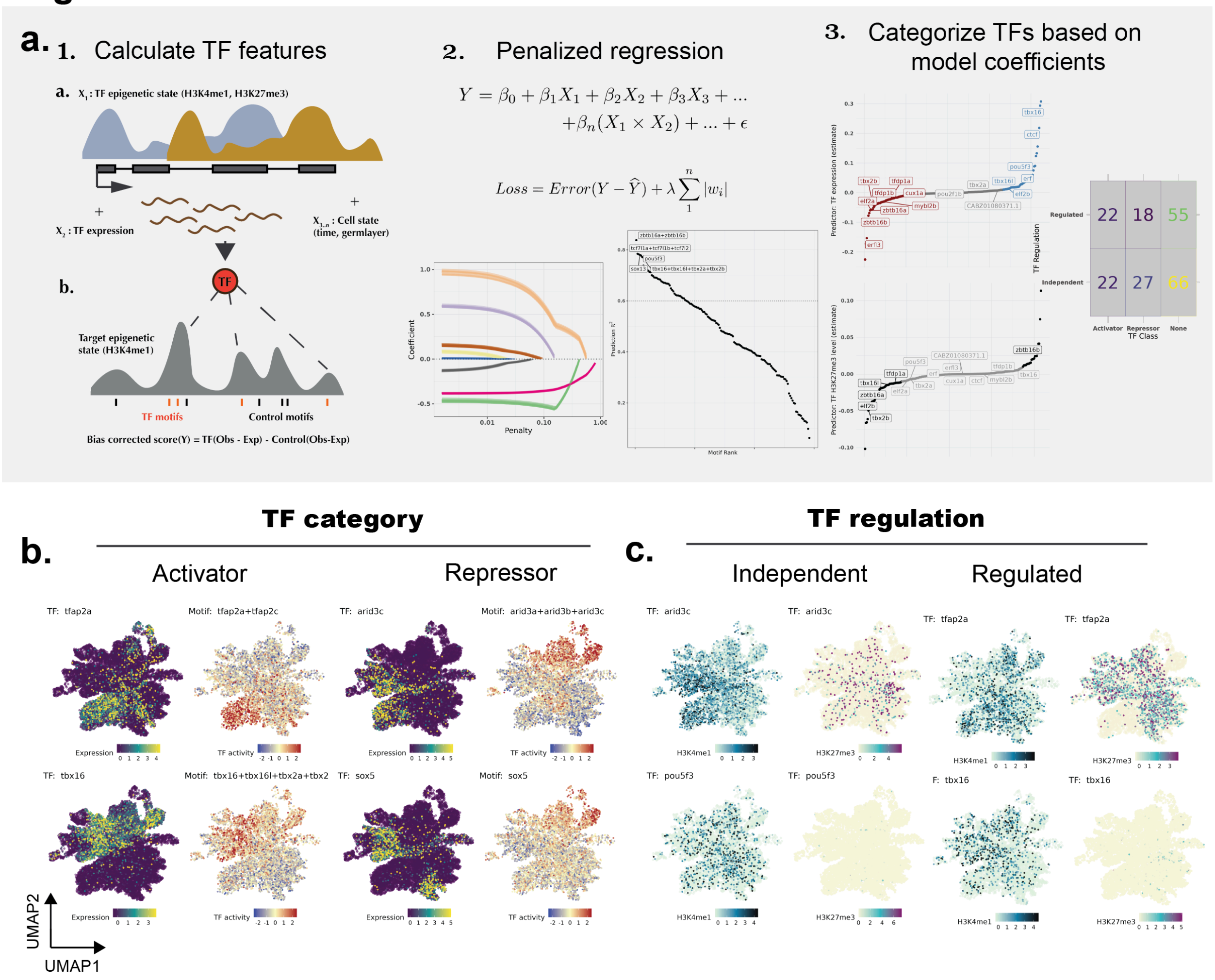

Figure 4: Prediction of TF activity using TF epigenetics and transcription

a. Schematic and outputs of the prediction model. 1) (top) Normalized H3K4me1 and H3K27me3 signal on the TF locus, TF (spliced) RNA signal, and indicators of cell state (pseudotime and lineage) are, used to predict TF activity (bottom) 2) The lasso regression model is used to select the most useful predictors for each TF. Bottom right plot shows the TFs sorted by the R^2 values on the independent test data. 3) Coefficients from the final models are ranked and compared to categorize the TFs. The number of TFs classified as activator/repressor, or regulated/independent are shown in the right plot. **b. UMAPs showing**

the expression and motif activities of TFs classified as “activators” or “repressors” using the above model. TF expression (left) is based on (normalized) spliced RNA and TF activity (right) is based on H3K4me1 signal on TFBS . For activators, the motif activities on TFBS are gained in cells where the TFs are expressed, while for the repressors, the motif activities are lost in the cells expressing the TF. **c. UMAPs show (normalized) H3K4me1 signal and H3K27me3 signal** on TF gene body, for TFs classified as “regulated” or “independent”. The histone modifications on regulated TFs are correlated with their activities.