Decoding gene regulation in the mouse embryo using single-cell multi-omics

Ricard Argelaguet[#,1,10], Tim Lohoff[1,6,11], Jingyu Gavin Li[1,9], Asif Nakhuda[2], Deborah Drage[1,10], Felix Krueger[3,10], Lars Velten[4,5], Stephen J. Clark[#,1,10], Wolf Reik[#,1,6,7,8,10]

1 Epigenetics Programme, Babraham Institute, Cambridge CB22 3AT, UK.
2 Gene Targeting Facility, Babraham Institute, Cambridge CB22 3AT, UK.
3 Bioinformatics Group, Babraham Institute, Cambridge CB22 3AT, UK.
4 Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
5 Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
6 Wellcome Trust – Medical Research Council Stem Cell Institute, University of Cambridge, Jeffrey Cheah Biomedical Centre, Puddicombe Way, Cambridge CB2 0AW, UK
7 Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK
8 Centre for Trophoblast Research, University of Cambridge, Cambridge, UK
9 Chu Kochen Honors College of Zhejiang University, Hangzhou, Zhejiang, China
10 Current address: Altos Labs Cambridge Institute, Granta Park, Cambridge, CB21 6GP, UK.
11 Current address: Forbion Capital Partners, Gärtnerplatz 6, 80469 Munich, Germany

Correspondence should be addressed to rargelaguet@altoslabs.com, sclark@altoslabs.com, wreik@altoslabs.com

Abstract

Following gastrulation, the three primary germ layers develop into the major organs in a process known as organogenesis. Single-cell RNA sequencing has enabled the profiling of the gene expression dynamics of these cell fate decisions, yet a comprehensive map of the interplay between transcription factors and cis-regulatory elements is lacking, as are the underlying gene regulatory networks. Here we generate a multi-omics atlas of mouse early organogenesis by simultaneously profiling gene expression and chromatin accessibility from tens of thousands of single cells. We develop a computational method to leverage the multi-modal readouts to predict transcription factor binding events in cis-regulatory elements, which we then use to infer gene regulatory networks that underpin lineage commitment events. Finally, we show that these models can be used to generate *in silico* predictions of the effect of transcription factor perturbations. We validate this experimentally by showing that Brachyury is essential for the differentiation of neuromesodermal progenitors to somitic mesoderm fate by priming cis-regulatory elements.

Introduction

In mammals, specification of the basic body plan occurs during gastrulation, when the pluripotent epiblast is patterned to give rise to the three primary germ layers. Subsequently, these progenitors generate all major organ systems in a process known as organogenesis (Arnold and Robertson, 2009; Bardot and Hadjantonakis, 2020; Tam and Loebel, 2007). In the mouse, germ layer formation and early organogenesis have been profiled using a variety of genomics technologies, including single-cell RNA-sequencing (scRNA-seq), which led to the annotation of multiple cell types and the characterisation of differentiation trajectories (Cao et al., 2019; Ibarra-Soria et al., 2018; Pijuan-Sala et al., 2019). Some efforts to profile the epigenome during these stages have produced bulk chromatin accessibility using ATAC-seq and histone profiling with ChIP-seq at E7.5 (Xiang et al., 2020), single-nucleus (sn) chromatin accessibility maps at E8.25 with snATAC-seq (Pijuan-Sala et al., 2020) and single-cell transcriptome, nucleosome positioning and DNA methylation up to E7.5 with scNMT-seq (Argelaguet et al., 2019). These data demonstrate the dynamic remodelling that the epigenome undergoes during development. However, a comprehensive characterisation of the epigenome changes and the cis-regulatory elements involved in the transition from gastrulation to early organogenesis is still lacking, as well as an integration of this information with the transcriptome. Furthermore, the genomic positions and the target genes of the various transcription factors (TFs) that control these developmental trajectories have only been explored for a limited set of TFs and using *in vitro* systems. A catalogue of TF binding sites during mouse early organogenesis *in vivo* is lacking.

Single-cell multimodal technologies have huge potential for the study of gene regulation (Chen et al., 2019; Clark et al., 2018; Luo et al., 2022; Ma et al., 2020; Zhu et al., 2019, 2021). In particular, the ability to link epigenomic with transcriptomic changes allows the inference of gene regulatory networks (GRNs)(Aibar et al., 2017; Davidson and Erwin, 2006; Kamimoto et al., 2020; Kartha et al., 2021; Materna and Davidson, 2007). GRNs are able to capture the interplay between TFs, cis-regulatory DNA sequences and the expression of target genes (Garcia-Alonso et al., 2019; Levine and Davidson, 2005; Stadhouders et al., 2018), and can hold predictive power of cell fate transitions and gene perturbations (Kamimoto et al., 2020). Methods that derive GRNs from single-cell genomics data have been developed (Aibar et al., 2017; Fleck et al., 2021; Kamimoto et al., 2020; Kartha et al., 2021) and applied to the developing fly brain (Janssens et al., 2022) but similar analyses of mammalian development are lacking. In addition, GRN inference relies on accurate TF binding data, yet limited knowledge of TF binding exists for early embryonic development due to limitations in experimental methods such as ChIP-seq or CUT&RUN, which require large numbers of cells (Skene and Henikoff, 2017) and faithful antibodies. It is thus unrealistic to profile a large fraction of all TFs even in a single biological context (Lambert et al., 2018; Park, 2009). Instead, TF binding sites are typically inferred from the presence of a sequence motif within accessible chromatin (Castro-Mondragon et al., 2021; Schep et al., 2017; Weirauch et al., 2014). This approach can be successful for some TFs that display non-redundant DNA motifs with high sequence specificity, but the presence of a TF motif does not guarantee the existence of an active binding site (Wang et al., 2012). Moreover, the use of DNA motifs as a proxy for TF binding is not well suited for the study of TFs that share similar DNA motifs, and also for TFs linked to short motifs. Thus, alternative methods for predicting TF binding sites are required.

Recent technological advances have enabled the simultaneous profiling of RNA expression and epigenetic modalities from single cells at high-throughput (Chen et al., 2019; Ma et al., 2020; Zhu et al., 2019). This provides a unique opportunity to systematically decode the TF activities and the GRN structure that underpins cell fate transitions. Here, we perform snATAC-seq and snRNA-seq from the same nuclei from a time course of mouse embryonic development from E7.5 to E8.75. We develop a computational method to leverage the multi-modal readouts to predict TF binding events in cis-regulatory elements, which we then use to build GRNs that underlie cell fate transitions. Finally, we show that these models can be used to generate *in silico* predictions of the effect of TF perturbations.

Results

Simultaneous profiling of RNA expression and chromatin accessibility during mouse early organogenesis at single-cell resolution

We employed the 10x Multiome technology to profile RNA expression and chromatin accessibility from single nuclei collected between E7.5 and E8.75 (**Figure 1a**). A total of 61,781 cells passed quality control for both data modalities, with a median detection of 4,322 genes expressed per cell and a median of 29,710 ATAC fragments per cell (**Figure S1**). Cell type assignments were made by mapping the RNA expression profiles to a reference atlas from similar stages (Pijuan-Sala et al., 2019) (**Figure 1b-c**, **Figure S2**). To evaluate the cell type assignments we performed multi-modal dimensionality reduction with MOFA+ (Argelaguet et al., 2020), revealing that both molecular layers contain sufficient information to distinguish cell type identities (**Figure 1c**). Similar results are obtained when applying dimensionality reduction to single data modalities. To further validate the measurements obtained from both data modalities, we compared the RNA expression and chromatin accessibility profiles with published data sets profiled with scRNA-seq (E7.5 to E8.5 embryos)(Pijuan-Sala et al., 2019) and snATAC-seq (E8.25 embryos)(Pijuan-Sala et al., 2020). Despite differences in the technology and in the molecular input (i.e. whole cell versus single nuclei in the case of RNA expression) we observe close agreements in both gene expression (**Figure S3**) and gene accessibility measurements (**Figure S4**).

A catalogue of cis-regulatory elements

To define open chromatin regions that represent putative cis-regulatory elements we performed peak calling on the snATAC-seq data using the ArchR pipeline (Granja et al., 2021). Briefly, peaks are defined by an iterative overlapping strategy where cells are aggregated by cell type into pseudo-bulk replicates. This approach has been shown to optimally preserve cell type-specific peaks (Granja et al., 2021). We obtained a total of 192,251 ATAC peaks, which we classified into four groups depending on their genomic location: Promoter (16.92%), Exonic (5.77%), Intronic (41.57%) and Intergenic (35.75%) (**Figure S5**). 81% of peaks display differential accessibility in at least one cell type comparison (**Methods**). 69% of peaks were assigned to genes based on genomic proximity (less than 50kb from the gene body), with an average of ~20 peaks linked to a gene and an average of ~2.3 genes associated to a peak. ~35% of peak-to-gene associations displayed significant positive correlation with the RNA

expression levels of at least one of the proximal genes, whereas ~11% displayed a negative correlation (**Figure S5**).

Molecular characterisation of lineage-specific cis-regulatory elements

Next, we sought to characterise the transcriptomic and epigenetic variability of lineage-defining genes. We used the pairwise differential RNA expression results between cell types to define cell type-specific upregulated marker genes (**Figure 1d left, Methods**). Then, we quantified the average RNA expression and chromatin accessibility (at promoter regions) for each class of marker genes and each cell type (**Figure 1e right, Figure S6**). As a positive control, we performed the same quantification for a set of canonical housekeeping genes, which are constitutively expressed and have an open chromatin profile. As a negative control, we included a set of olfactory receptors genes, which are not expressed until later in development and display a closed chromatin profile (**Figure 1e left, Figure S6**). In marker genes, we observe the highest levels of expression and chromatin accessibility in the cell types that they mark, as expected. In all other cell types expression of these marker genes is still detected but at reduced levels. Promoter accessibility is also lower for marker genes in the cell types that they mark, however the differences are much less pronounced than for gene expression (**Figure 1e**). This suggests that promoter accessibility may have a limited function in driving differences in gene expression across cell types. Then, we asked whether cis-regulatory elements that are distal to promoter regions (Intronic and Intergenic peak sets) also display the same behaviour. We defined cell type-specific marker peaks by performing pairwise differential accessibility analysis (**Figure 1d right, Methods**), and then compared the average chromatin accessibility at promoter regions of marker genes versus marker peaks (**Figure 1f**). We find distal cis-regulatory elements to be more dynamic, with accessibility levels similar to promoters in the cell types where they become active, but much lower accessibility in the cell types where they are not active (**Figure 1f**). Consistent with previous reports (Argelaguet et al., 2019; Cusanovich et al., 2018), our results indicate a more prominent role of distal regulatory regions in cell fate decisions. A representative example is the *Gata6* locus shown in (**Figure 1g**). This gene encodes a zinc finger transcription factor that is active in multiple cell types derived from lateral mesoderm (Morrisey et al., 1996). Consistently, this gene is expressed in multiple late mesodermal cell types, including Cardiomyocytes, Pharyngeal mesoderm and Allantois. However, the promoter region is homogeneously open across all cell types, whereas three regulatory regions located within 50 kilobases of the gene body gain accessibility exclusively in the cell types where *Gata6* is expressed. Other representative examples are shown in **Figure S6**.

An *in silico* ChIP-seq library for mouse organogenesis

Cell fate decisions are molecularly driven by changes in gene regulatory networks (GRN) orchestrated by the interaction between transcription factors (TFs) and their target genes (Levine and Davidson, 2005). Nevertheless, limited knowledge of TF binding exists for early embryonic development. First, experimental methods such as ChIP-seq or CUT&RUN require large numbers of cells to accurately profile TF binding events making it challenging to apply to embryos (Skene and Henikoff, 2017). Second, the success of the experiments depend on properties of available antibodies and on the properties of the TF itself, making it unrealistic to profile even a fraction of all transcription factors in the genome(Lambert et al., 2018; Park,

2009). Current methods for ATAC-seq data analysis link TFs to regulatory regions by the presence of TF motifs (Castro-Mondragon et al., 2021; Schep et al., 2017; Weirauch et al., 2014). This approach can be successful for some TFs that display non-redundant DNA motifs with high sequence specificity, but it has important shortcomings. First, the presence of a TF motif does not guarantee the existence of an active binding site (Wang et al., 2012). Second, a large fraction of TFs belong to families that share the same motif, even when having different functions and expression patterns. Representative examples are the GATA, HOX and the FOX family of transcription factors (**Figure S7**). As a result of these issues, it is extremely challenging to link TFs to regulatory elements when exclusively using a combination of genomic and epigenomic information. This can be illustrated by the large number of TF motifs that are contained within each ATAC peak (**Figure S7**). Here, we developed a novel computational approach that integrates genomic, epigenomic and transcriptomic information to predict functional TF binding events.

Intuitively, we consider an ATAC peak *i* to be a putative binding site for TF *j* if it contains the *j* motif and its chromatin accessibility is correlated with the RNA expression of the TF (**Figure 2a**). We combine three metrics (motif score, average chromatin accessibility and correlation) to devise a quantitative *in silico* binding score for each combination of TF and ATAC peak (**Methods**). Note that our approach is unsupervised and does not require ChIP-seq data as input. This stands in contrast with other approaches that have been proposed to predict TF binding from multi-omics data, which employ supervised models that require labelled training data from ChIP-seq experiments (Avsec et al., 2021; Karimzadeh and Hoffman, 2019). We will refer to this approach as *in silico* ChIP-seq. As expected, the number of predicted binding sites for each TF is a function of the minimum score threshold, which ranges from 0 to 1 after scaling (**Figure 2b**). Notably, the incorporation of RNA expression massively reduces the amount of predicted binding sites for each TF as well as the amount of TFs that can be linked to each regulatory element (**Figure 2c-d**).

To validate the *in silico* ChIP-seq library, we used publicly available ChIP-seq experiments for a set of TFs that are known to play key roles during mouse gastrulation and early organogenesis, and defined this as the ground truth for TF binding events. Due to the limited availability of *in vivo* ChIP-seq datasets, we had to rely on *in vitro* models that more closely resemble the gastrulating embryo (**Supplementary Table 1**). Yet, we observe remarkable agreement between the *in silico* TF binding scores and the observed ChIP-seq signal (**Figure 2d-e**). Worse agreement is obtained when excluding the transcriptomic information from the model (**Figure 2d**). Representative examples of TF binding predictions are shown alongside ChIP-seq data in **Figure 2f-g**. Interestingly, for all TFs we benchmarked, the consistency with ChIP-seq measurements exclusively holds true for ATAC peaks that are positively correlated with TF expression (**Figure S8**), which is consistent with these TFs acting as chromatin activators. Our approach also predicts repressive interactions with chromatin (not to be confused with transcriptional repression of target genes, as we will discuss below). Chromatin repressors are known to be important for gene regulation, and they generally involve the recruitment of chromatin remodelers, including histone modifiers, to turn chromatin from an open to a closed state (Berest et al., 2019; Gaston and Jayaraman, 2003; Iurlaro et al., 2021; Janssens et al., 2022; Lambert et al., 2018). However, insufficient ChIP-seq data exists for chromatin repressors in the context of embryonic development, thus limiting our benchmark. In consequence, we only consider activatory links between TFs and regulatory regions for downstream analyses.

Generation of a catalogue of cell type-specific transcription factor activities

The most popular method to quantify TF activities per cell using snATAC-seq data is chromVAR (Schep et al., 2017). Briefly, this method computes, for each cell (or pseudo-bulk cell type) and TF motif, a z-score that measures the difference between the total number of fragments that map to motif-containing peaks and the expected number of fragments.
While useful when only having access to chromatin accessibility data, chromVAR scores are often not representative of true TF activities, mainly because accessible DNA motifs are not always good proxies for actual TF binding events. This can be illustrated with the Fox family of transcription factors, which all share a similar DNA motif, but nevertheless have different roles during mouse gastrulation: whereas Foxb1 is a pioneer TF in the ectodermal lineage (Labosky et al., 1997), Foxc1 is active in the mesodermal lineage (Wilm et al., 2004). Their distinct roles are evidenced by the RNA expression profiles in our data set and their mapping to different spatial locations in the embryo (Lohoff et al., 2022) (**Methods, Figure 3a**). Yet, due to their motif similarity, the chromVAR scores of these two TFs are indistinguishable (**Figure 3a**). Here, we modified the chromVAR algorithm to use the predicted TF binding sites from the *in silico* ChIP-seq library (instead of all ATAC peaks that contain the TF motif). We find that this yields TF activity scores that are more consistent with the RNA expression patterns of the corresponding TFs (**Figure 3b**). For clarity, we will refer to this approach as chromVAR-Multiome.

Next, we used the chromVAR-Multiome scores to perform pairwise differential analysis between cell types and parse the results to quantify TF activities for each combination of TF and cell type (**Methods**) (**Figure 3c-d**). Reassuringly, using this approach we recover canonical TF markers for a variety of cell types (**Figure 3e**), including Foxa2 and Sox17 for endodermal cell types; Mesp11/2 and Mixl1 for the Primitive Streak and mesodermal cell types; Sox2 and Rfx4 for ectodermal cell types; Tbx5 and Nkx2-5 for Cardiomyocytes; Runx1 and Tal1 for Blood progenitors and Erythroids. Notably, the resolution of the data enables us to provide quantifications of TF activities for cell types that are challenging to study due the low cell numbers and difficult cell isolation, including Primordial Germ Cells (PGCs) and Neural crest cells (**Figure 3f**). For the Neural crest, we recover many TFs that have been previously associated with Neural crest identity in different species: Pax7, Foxd3, Tfap2a, Tfap2b, Sox10, Sox5, Ets1, Nr2f1 and Mef2c (**Figure S9**, **Supplementary Table 2**). For example, Tfap2a has been shown to be essential for Neural crest specification in Xenopus embryos (de Crozé et al., 2011). In mice, disruption of the *Tfap2a* gene results in craniofacial malformations and embryonic lethality(Schorle et al., 1996). In humans, missense mutations in the corresponding orthologous gene results in branchio-oculo-facial syndrome, which is also characterised by craniofacial abnormalities (Milunsky et al., 2008). For PGCs we also recover TFs described to be important for PGC specification in mice, including Prdm1 (also called Blimp-1), Esrrb and Pou5f1 (also called Oct4) (**Figure S9**, **Supplementary Table 2**). For example, Blimp-1 has been shown to be essential for the repression of the somatic programme upon PGC specification(Ohinata et al., 2005). In addition, we also predict several TFs with unknown roles in PGC formation that could be suitable candidates for further characterisation, including Ybx2, Bbx and Klf8 (**Figure S9**).

Interestingly, visualisation of TF activities across all cell types reveals that (1) cell types are defined by a combinatorial activity of multiple TFs and (2) most TFs are active across multiple

cell types (**Figure 3g**). The first observation can be illustrated with the Neural crest: Of the canonical TFs shown in Figure 3f, none are uniquely active in the Neural crest, with the exception of Dlx2 and Sox10 (**Figure 3h**). The second observation sometimes arises from the hierarchical nature of lineage specification (such as Pax7 being active in multiple ectodermal-derived cell types, Foxa2 in all endodermal-derived cell types and Tal1 in all cell types that are linked to blood formation). However, in other cases we observe the same TF active in cell types from different germ layer origins, thus suggesting widespread pleiotropic activity where TFs define cellular identities via combinatorial context-dependent activity (Reiter et al., 2017; Spitz and Furlong, 2012). Representative examples are Sox9, active in the Neural crest, Brain, Definitive endoderm, and Notochord; Tfap2c, active in the Neural crest, ExE ectoderm and PGCs; and Ets1, active in Neural crest, Endothelium and Blood Progenitors (**Figure 3h**).

Mapping the transcription factor regulatory network that underlies differentiation of neuromesodermal progenitors

In the previous section, we used the *in silico* ChIP-seq library and the chromVAR-Multiome algorithm to generate a catalogue of TF activities linked to discrete cellular identities. For simplicity, we ignored interactions between TFs. Next, we sought to quantify interactions between TFs by inferring gene regulatory networks (GRNs) and connecting them to continuous cell fate transitions.

We employed a multi-step algorithm to infer GRNs (**Methods, Figure S10**). First, we subset cells of interest and infer metacells (Persad et al., 2022), with the goal of achieving a resolution that retains the cellular heterogeneity while overcoming the sparsity issues of single-cell data. Second, we used the *in silico* ChIP-seq library to link TFs to cis-regulatory elements (ATAC peaks). Third, we linked cis-regulatory elements to potential target genes by genomic proximity (here a maximum distance of 50kb), which is a reasonable approximation in the absence of 3D chromatin contact information (Janssens et al., 2022; Kamal et al., 2021). This results in a directed network where each parent node corresponds to a TF, and each child node corresponds to a target gene. Finally, following the approach of (Kamimoto et al., 2020), we estimated the weights of the edges by fitting a linear regression model of the expression of a target gene as a function of the parent TF's expression. Importantly, while our benchmark of the *in silico* ChIP-seq does not support the inclusion of repressive links between TFs and ATAC peaks, evidence exists that TFs can repress the expression of target genes (Gaston and Jayaraman, 2003; Liang et al., 2017). Thus, in the GRN model we allowed for negative associations between TF expression and target gene expression.

We applied this methodology to study the complex gene regulatory network that determines differentiation of Neuromesodermal progenitors (NMPs)(Gouti et al., 2017). Briefly, NMPs are a population of bipotent stem cells that fuel axial elongation by simultaneously giving rise to Spinal cord cells, an ectodermal cell type, as well as posterior somites, a mesodermal cell type (Sambasivan and Steventon, 2020) (**Figure 4a**). Notably, these cell fate transitions occur when most of the cells in the embryo are already committed to one of the germ layers(Sambasivan and Steventon, 2020). Molecularly, NMPs are characterised by the counterbalanced co-expression of the mesodermal factor Brachyury and the neural factor Sox2 (Henrique et al., 2015), but studies have suggested that these are just two players of a

complex regulatory landscape, thus calling for more complex GRN models (Gouti et al., 2017). Here we applied the GRN methodology described above to study the NMP differentiation trajectory (**Figure 4a, Methods**). For ease of interpretation and visualisation, we restricted the target genes to be other TFs. Besides Sox2 and Brachyury, the network consisted of 40 additional TFs with with 379 activatory associations and 48 repressive associations (**Figure 4b**), Notably, we find that Cdx and Hox TFs display the highest centrality of the network by establishing an activatory self-regulatory loop that sustains NMP identity (**Figure 4c,d**). This observation agrees with studies that showed that all three Cdx genes contribute additively to axial elongation and the development of posterior embryonic structures, with the most important one being Cdx2 (Chawengsaksophak et al., 2004; van Rooijen et al., 2012). To further validate the predicted interaction between Cdx and Hox genes, we used ChIP-seq data for Cdx2 profiled in Epiblast Stem Cells exposed to Wnt and Fgf signalling, which induces posterior axis elongation and generates cells that resemble NMPs (Amin et al., 2016). Consistent with our inferred GRN, we find widespread binding of Cdx2 within the Hoxb cluster of genes (**Figure 4e**). Notably, this interaction between Cdx and Hox genes also agrees with *in vitro* studies that described the upregulation of posterior Hox genes in NMP-like cells upon induction of Cdx factors (Amin et al., 2016; Neijts et al., 2016). In addition to its role as transcriptional activator of Hox genes, we also find that Cdx2 displays a pleiotropic role by repressing TFs that direct the transition to Somitic mesoderm (Foxc2, Brachyury, Meox1) and Spinal cord (Pax6) (**Figure 4f and Figure S11**).

Brachyury controls the differentiation of Neuromesodermal progenitors to Somitic mesoderm by priming cis-regulatory elements

Our results above indicate that Cdx and Hox factors sustain the bipotent NMP identity but can also act as a fate switch by repressing TFs that direct the transition to Somitic mesoderm, most notably Brachyury. Similar to Cdx2 null mutants, Brachyury null embryos generate the first set of anterior somites, but axial elongation is impaired (Martin, 2016). Previous studies have shown that Brachyury is required for the transition from NMPs to posterior Somitic mesoderm (Guibentif et al., 2021). To validate whether this role of Brachyury is captured by our models trained on the reference data set, we first visualised the chromatin accessibility dynamics of cis-regulatory elements that are putative Brachyury binding sites, as inferred from the *in silico* ChIP-seq library applied to NMP differentiation. We find that these cis-regulatory elements increase in accessibility when transitioning from NMP to Somitic mesoderm, but not to Spinal cord. Interestingly, some of these binding sites are linked to mesodermal genes that become expressed in the Somitic mesoderm, including Tbx6, Mesp1 and Fgf4 (**Figure S12**). This behaviour is suggestive of epigenetic priming, whereby the chromatin of cis-regulatory elements becomes accessible before transcription of the target gene (Argelaguet et al., 2019; Ma et al., 2020). Next, we employed the GRN of NMP differentiation to perform an *in silico* knock out of Brachyury using the CellOracle framework (Kamimoto et al., 2020) *(***Methods***)*. We find that the *in silico* knock-out of Brachyury disrupts the transition from NMP to Somitic mesoderm (**Figure 4g**). Although this result was expected based on previous findings (Guibentif et al., 2021), it demonstrates how GRNs inferred from unperturbed single-cell multi-omics data have the potential to provide functional insights into cell fate transitions (Kartha et al., 2021).

To validate our predictions, we generated Brachyury KO embryos by direct delivery of CRISPR/cas9 as a ribonucleoprotein (RNP) complex via electroporation, targeting exon 3 of the *Brachyury* (*T*) gene in zygotes at one-cell stage (**Methods, Figure 5a**). Control embryos received Cas9 protein only. Embryos were transferred into pseudopregnant females and collected at E8.5 for 10x Multiome sequencing. In total, we obtained 6,797 cells from 3 embryos at E8.5 with a wildtype (WT) background and 6,572 cells from 7 embryos with a *Brachyury* KO background. Cell types were again annotated by mapping the RNA expression to the transcriptomic atlas (**Figure 5b**). Consistent with our predictions and the results of (Gouti et al., 2017; Guibentif et al., 2021), we observe a relative underrepresentation of (posterior) somitic mesoderm and allantois cells in the *Brachyury* KO embryos, together with a relative overrepresentation of NMP cells (**Figure 5c**). No significant difference is observed in the abundance of Spinal cord cells, suggesting that the neural differentiation capacity of NMPs is not affected in the absence of Brachyury. Notably, we also observe defects in the Erythropoiesis trajectory (**Figure 5c**), suggesting pleiotropic effects of Brachyury across multiple developmental trajectories(Bruce and Winklbauer, 2020). To further explore the effect of the *Brachyury* KO in NMPs, we mapped the cells onto the NMP differentiation trajectory reconstructed from the transcriptomic reference atlas (Pijuan-Sala et al., 2019) (**Figure 5d**). Again, we find that WT cells map across the entire trajectory, but *Brachyury* KO cells map only onto the transition between NMP and Spinal cord. Additionally, RNA velocity analysis of these cells shows that WT NMP cells transition towards both Spinal cord and Somitic mesoderm fates, whereas in the *Brachyury* KO only the Spinal cord displays a coherent differentiation trajectory (**Figure 5e**).

Next, we performed differential accessibility analysis between WT and Brachyury KO NMP cells (**Methods**) finding that most of the differentially accessible (DA) peaks are more closed in Brachyury KO cells (**Figure 5f**). This set of DA peaks display enrichment for the T-box motif and a higher *in silico* TF binding score for Brachyury than non DA peaks (**Figure 5g-h**), hence suggesting that this set of regulatory regions are directly regulated by Brachyury. Consistent with our predictions of Brachyury target sites above, we find that the set of DA peaks in NMPs are markers of Somitic mesoderm in the reference atlas (**Figure 5i**), again consistent with a potential role of a Brachyury-driven epigenetic priming in NMPs. More generally, our results hint that the dysregulation of individual cis-regulatory elements can be predicted, to some extent, using only the reference data set. A representative example is shown in **Figure 5j**, which shows a cis-regulatory region that corresponds to a Brachyury binding site located upstream of *Mesp1*, a gene that is not expressed in NMPs but becomes expressed in the Somitic mesoderm (**Figure S12**). This cis-regulatory element becomes partially open in WT NMP cells, but not in *Brachyury* KO NMP cells, and attains its highest accessibility levels in WT Somitic mesoderm cells, while becoming closed in Spinal cord cells. Similar patterns can be observed for the cis-regulatory elements linked to *Tbx6* and *Fgf4* (**Figure S13)**.

Conclusion

We have generated a single-cell multi-omic atlas of mouse early organogenesis by simultaneously profiling RNA expression and chromatin accessibility between E7.5 and E8.75, spanning late gastrulation and early organogenesis. Taking advantage of the simultaneous profiling of TF expression and cognate motif accessibility, we developed a novel tool to quantitatively predict TF binding events in cis-regulatory elements, which we used to quantify celltype-specific TF activities and infer gene regulatory networks that underlie cell fate transitions. We show that these computational models trained on unperturbed data can be

used to predict the effect of transcription factor perturbations. We validate this experimentally by showing that Brachyury is essential for the differentiation of neuromesodermal progenitors to somitic mesoderm fate by priming cis-regulatory elements.

Author contributions

R.A., T.L., S.J.C. and W.R. conceived the project
T.L. and D.D performed embryo dissections
T.L. and S.J.C performed nuclear extractions
F.K. processed and managed sequencing data
A.N. performed gene targeting
R.A. and L.V. conceived and implemented the *in silico* ChIP-seq method.
R.A. and S.J.C. performed pre-processing and quality control
R.A. and G.L. performed the computational analysis
R.A. generated the figures
R.A. and S.J.C. interpreted results and drafted the manuscript.
W.R., S.J.C. and L.V. supervised the project.
All authors read and approved the final manuscript.

Code availability

Code to reproduce the analysis is available at https://github.com/rargelaguet/mouse_organogenesis_10x_multiome_publication

Data availability

Raw sequencing data together is available in the Gene Expression Omnibus under accession GSE205117 (reviewer token token gzcxaugylriplkn). Links to processed objects as well as to a Shiny app for interactive data analysis are available in the github repository.

Acknowledgments

We thank Paula Kokko-Gonzales, Nicole Forrester and Amelia Edwards of the Babraham Institute Sequencing Facility for assistance with 10x Genomics library preparation, Katarzyna Kania and members of the CRUK-CI Genomics Core for 10x Genomics library preparation and Illumina sequencing and the Babraham Biological Support Unit for animal work. We thank Bart Theeuwes and Brendan Terry for comments on the manuscript. We thank all members of the Reik lab for their discussions and support.

The following sources of funding are gratefully acknowledged. This work was supported by the Wellcome Trust (awards 210754/Z/18/Z and 220379/Z/20/Z) and the BBSRC (award BBS/E/B/000C0421). T.L. was funded by the Wellcome Trust 4-Year PhD Programme in Stem Cell Biology and Medicine and the University of Cambridge, UK (203813/Z/16/A and 203813/Z/16/Z). This research was funded in whole or in part by the Wellcome Trust. R.A. was supported by the Wellcome for a Collaborative Award in Science (award 220379/Z/20/Z). The funding sources mentioned above had no role in the study design, in the collection, analysis and interpretation of data, in the writing of the manuscript and in the decision to submit the manuscript for publication.

Conflict of interest statement

W.R. is a consultant and shareholder of Cambridge Epigenetix. R.A., D.D., F.K., S.J.C. and W.R. are employees of Altos Labs. The remaining authors declare no competing financial interests.

Methods

RNA data processing

Raw sequencing files were processed with CellRanger arc 2.0.0 using default arguments. Reads were mapped to the mm10-2020-A-2.0.0 genome and counted with GRCm38.92 annotation. Low-quality cells were filtered based on the distribution of QC metrics. Cells were required to have a minimum of 2000 UMIs per cell, a maximum of 40% mitochondrial reads and a maximum of 20% ribosomal reads. The resulting count matrix was stored using a SingleCellExperiment (Amezquita et al., 2019) (v 1.14.1) object. Normalisation and log transformation was performed using scran (Lun et al., 2016) (v1.20.1) and scuttle (McCarthy et al., 2017)(v1.2.1). Doublet detection was performed using the hybrid approach in the scds (v1.8.0) package.

ATAC data processing

We used the ArchR package (Granja et al., 2021)(v1.0.1) for preprocessing of ATAC data. Briefly, arrow files were created from the ATAC fragment files. Cells were required to have a minimum of 3500 fragments per cell, a minimum TSS enrichment of 9, and a maximum blacklist ratio of 0.05. Pseudo-bulk replicates were obtained per cell type and peak calling was performed using macs2 (Zhang et al., 2008) (v2.2.7.1) using the cell type identified from the RNA expression as a group. A consensus peak set was obtained by an iterative overlapping strategy which is better at preserving cell type-specific peaks. Motif annotations were extracted from the CISBP (Weirauch et al., 2014) (v2) and JASPAR 2000 database (Castro-Mondragon et al., 2021). Motif matches for each peak were obtained using motifmatchr (v1.14.1), with a minimum motif width of 7 and a maximum q-value of 1e-4. Bigwig files were exported for each cell type for visualisation on the IGV browser (Robinson et al., 2011) (v2.11.0).

Velocity analysis

Spliced and unspliced count matrices were extracted using velocyto (La Manno et al., 2018)(v0.17.17). Velocity analysis was performed using scVelo (Bergen et al., 2020) (v0.2.1) in dynamical mode.

Metacell inference

When exploring continuous trajectories we summarised the data into metacells with the goal of achieving a resolution that retains the heterogeneity while overcoming the sparsity issues of single-cell data. We identified metacells (i.e. groups of cells that represent singular cell-states from single-cell data) using SEACells (Persad et al., 2022). Following the method guidelines, metacells were computed separately for each sample using approximately one metacell for every seventy-five cells. Following metacell identification, we regenerated gene

expression and chromatin accessibility count matrices summarised at the metacell level. Sample-specific count matrices were then concatenated and normalised using log-transformed counts per million.

490

***In silico* ChIP-seq library**

The in silico ChIP-seq library is a computational approach to link TFs to cis-regulatory elements in the form of ATAC peaks. Intuitively, we consider an ATAC peak *i* to be a putative binding site for TF *j* if *i* contains the j motif and its chromatin accessibility correlates with the
495 RNA expression of *j*. Formally, we calculate the in silico TF binding score for ATAC peak *i* and TF *j* with the following equation:

$$x_{ij} = \boldsymbol{\sigma}_{ij}\, minmax(\boldsymbol{\theta}_{ij}\boldsymbol{\pi}_i)$$

where $\boldsymbol{\sigma}_{ij}$ is the correlation between the chromatin accessibility of peak *i* and the RNA expression of TF *j*. $\boldsymbol{\theta}_{ij}$ is the motif score for TF *i* in peak *j*, and $\boldsymbol{\pi}_i$ is the maximum chromatin
500 accessibility of peak *j* (across cell types). Note that the TF binding score ranges from -1 to 1 due to the *minmax* normalisation, which is applied across all TFs and peaks. A negative *in silico* TF binding score value denotes a repressive event, where the chromatin accessibility of ATAC peak *i* is negatively regulated by TF *j*. In contrast, a positive value denotes an activatory event, where the chromatin accessibility of peak i is positively regulated by TF *j*. Although the
505 TF *in silico* score is continuous, some analysis require a binarised association between TFs and cis-regulatory elements, including GRN inference. In this case the *in silico* TF binding score can be modulated as a hyperparameter. Small values will lead to many predicted TF binding events, a high false positive rate and a low true positive rate. Large values will lead to fewer predicted TF binding events, but a low false positive rate and a high true positive rate.
510 We performed grid search and found that values between 0.10 and 0.30 provide reasonable trade-offs between the number of predicted TF binding events and the accuracy of the predictions in our benchmark.

Quantification of transcription factor activities per cell type using chromVAR-Multiome

515 TF activities were calculated using the chromVAR algorithm (Schep et al., 2017). The method takes as input the ATAC peak matrix and a set of position-specific weight matrices (PWMs) encoding TF sequence affinities. Here we used the JASPAR (2022)(Castro-Mondragon et al., 2021) and CISBP (v2.0)(Weirauch et al., 2014) databases. Briefly, for each TF motif contained within an ATAC peak and each cell (or cell type, when calculated at the pseudo-bulk level),
520 chromVAR calculates a z-score that measures the difference between the total number of fragments that map to motif-containing peaks and the expected number of fragments (based on the average of all cells). Importantly, the normalisation and scaling that chromVAR applies is aimed at mitigating technical biases between cells (Tn5 tagmentation efficiency, PCR amplification, etc.) and features (GC content, mean accessibility, etc.). While useful when only
525 having access to scATA-seq data, chromVAR z-scores are often not representative of true TF activities, mainly because DNA motifs are not always good proxies for actual TF binding. Here we modified the input to the chromVAR algorithm: instead of using all ATAC peaks with the presence of the TF motif, we selected putative binding sites with an *in silico* TF binding score higher than 0.15.

530

Dimensionality reduction using MOFA

We generated a multi-modal latent embedding using MOFA+ (Argelaguet et al., 2020). Briefly, the method takes as input multiple data modalities and performs multi-view matrix factorisation

to generate a set of latent factors that can be used for a variety of downstream tasks. Here we used as input to MOFA the RNA expression and ATAC peak matrix. Feature selection was performed to enrich for highly variable features (3,000 genes and 25,000 ATAC peaks). Optionally, one can also use as input latent variables that result from linear dimensionality reduction (Principal Component Analysis in the case of the RNA expression and Latent Semantic Indexing in the case of ATAC peaks). This leads to a significant increase in speed and also mitigates challenges linked to class imbalance (i.e. the two views having many different features). We ran MOFA with a fixed set of 30 factors, which we subsequently used as input to the UMAP algorithm(McInnes et al., 2018) to generate a (non-linear) two-dimensional embedding that is suitable for visualisation.

TF marker scores

We used the chromVAR-Multiome values to define TF marker scores for each combination of cell type and TF. We adopted a similar algorithm as used for the definition of marker genes in Seurat (*FindMarkers* function) and scran (*findMarkers* function). First, we performed differential analysis between each pair of cell types using a t-test. Then, for each TF *i* and cell type *j* we counted the number of significant differential comparisons between cell type *j* and all other cell types different from *j*. Instead of aggregating the p-values and fold changes, as done in Seurat and scran, we adopt a more intuitive metric and define the TF marker score as the fraction of differential comparisons. Intuitively, the higher the score of TF *i* in cell type *j* the more active that TF *i* is in cell type *j* when comparing the chromVAR-Multiome values to the other cell types. The maximum TF marker score value is 1, when all differential comparisons are significant. When defining the catalogue of TF activities per cell type (Figure 3f), we set a minimum TF marker score of 0.75.

Gene accessibility scores

Here we quantified promoter accessibility by adding all reads that map to the region that is 500bp upstream and 100bp downstream of the transcription start site (TSS). TSS annotations are obtained from the BioMart database using the Bioconductor GenomicFeatures package (v1.48.1). Note that here we disabled ArchR's default gene accessibility model, which incorporates information from cis-regulatory elements that are located near the TSS. Although this approach is more predictive of changes in gene expression, it is problematic when applied to genomic regions with high gene density, as cis-regulatory elements cannot be confidently linked to genes.

Pooling cells from the same cell type into pseudo-bulk replicates

The sparsity of the single-cell data limits the statistical analysis, the visualisation strategies and overall the biological insights that can be extracted from the data (Squair et al., 2021). For some analysis that involve cell type comparisons (including differential analysis, peak calling or *in silico* ChIP-seq inference), we create "pseudo-bulk" replicates by aggregating reads from all cells that belong to the same cell type. The pseudo-bulk strategy is particularly important for snATAC-seq data, as ATAC peaks typically have very few reads per cell. For differential analysis between cell types, we follow the approach suggested in(Crowell et al., 2020) and create the same number of replicates per cell type by bootstrapping cells assigned to the same cell type. Besides reducing sparsity, this approach also helps address the problem of having a different number of samples per group when doing differential analysis at single-cell resolution, which often leads to p-values being systematically different depending on the number of samples per group.

Genome Browser visualisation

We use the getGroupBW function in ArchR to group, summarise and export a bigwig file for each cell type. Briefly, the function calculates normalised accessibility values along the genome using 100bp tiles. We visualise the ATAC bigwig files as separate tracks in the IGV Browser (v2.11.0)(Thorvaldsdottir et al., 2013)

Differential RNA expression and chromatin accessibility

Following the guidelines from previous studies (Squair et al., 2021), we performed differential analysis using pseudo-bulk replicates for each cell type (and genotype, in the Brachyury KO study). For each group we derived 5 replicates by bootstrapping different subsets of cells at random. Each pseudo-bulk replicate contained 30% of the total number cells, with at least 25 cells per replicate. Subsequently, read counts were aggregated for each group, followed by normalisation with log-transformed counts per million (CPMs). Note that this "pseudo-bulk-with-replicates" approach yields the same number of samples per group, which facilitates differential analysis comparisons. Differential analysis was performed using the negative binomial model with a quasi-likelihood test implemented in edgeR (Robinson et al., 2010). Significant hits were called with a 1% FDR (Benjamini–Hochberg procedure) and a minimum log2 fold change of 1. Hits with small average expression values (log normalised counts <=2) were ignored, as this can lead to artificially large fold change values.

Identification of marker genes in the reference atlas

Cell type-specific marker genes and peaks were identified using the reference cells (i.e. the cells form the Brachyury KO study were not included). First, we performed differential analysis between each pair of cell types using the strategy outlined above. Then, for each cell type, we labelled as marker genes or as marker peaks those hits that are differentially expressed/accessible and upregulated in the cell type of interest in more than 85% of the comparisons.

Mapping to a reference atlas and cell type assignment

Cell types were assigned by mapping the RNA expression profiles to a reference atlas from the same stages (Pijuan-Sala et al., 2019). The mapping was performed by matching mutual nearest neighbours with the fastMNN algorithm (*batchelor* R package v1.8.1)(Haghverdi et al., 2018). First, count matrices from both experiments were concatenated and normalised together using scran (v1.20.1). Highly variable genes were selected(Lun et al., 2016) from the resulting expression matrix and were used as input for Principal Component Analysis. A first round of batch correction was applied within the atlas cells to remove technical variability between samples. A second round of batch correction was applied to integrate query and atlas cells within a joint PCA space. Then, for each query cell we used the *queryKNN* function in BiocNeighbors to identify the 25 nearest neighbours from the atlas. Finally, a cell type was inferred for each query cell by majority voting among the atlas neighbour cells.

Mapping to the spatial atlas and imputation of spatially-resolved ChromVAR-Multiome scores

Mapping of the 10x Multiome cells to the spatially-resolved transcriptomic atlas was done using the same approach described above for the scRNA-seq reference atlas. This integration is however more challenging due to the sparsity of the seqFISH data set and the different nature of the size factors. Here we followed the strategy outlined in(Lohoff et al., 2022) and applied cosine normalisation on the log-normalised counts. For simplicity, we used as reference a single z-slice from a representative E8.5 embryo.

630 Finally, we used the mapping to impute spatially-resolved TF activities. We transferred the chromVAR-Multiome scores from the 10x Multiome cells onto the nearest neighbours of spatial atlas. Due to the noisy estimates in single-cell data and the presence of outliers, we performed kNN denoising before visualisation.

635 **Inference of the TF regulatory network underlying differentiation of Neuromesodermal progenitors.**

First, we selected metacells of the NMP differentiation trajectory. Note that we discourage the use of data at single cell resolution, as the sparsity of snATAC-seq makes it challenging to obtain reliable associations between the RNA expression of TFs (which are typically lowly 640 expressed genes) and chromatin accessibility of cis-regulatory regions. Second, we used the *in silico* ChIP-seq methodology to link TFs with cis-regulatory elements. Third, we linked cis-regulatory regions to nearby genes via a maximum genomic distance of 50kb. Note that this step results in a many-to-many mapping, where each gene can be linked to multiple cis-regulatory regions, and each cis-regulatory region can be linked to many genes. Fourth, we 645 built a linear regression model of target gene RNA expression as a function of the TF's RNA expression. Finally, we visualise the GRN as a directed graph where nodes correspond to TFs and target genes (which can also be other TFs), where the edge width is given by the slope of the linear regression models.

650 ***In silico* TF perturbation with CellOracle**

Briefly, CellOracle leverages a gene regulatory network and a differentiation trajectory to predict shifts in cellular identities by simulating the effects of TF perturbations on the GRN configuration. It simulates gene expression values upon TF perturbation, which are then compared with the gene expression of local neighbourhoods to estimate transition probabilities 655 between cell states. Finally, CellOracle creates a transition trajectory graph to project the predicted identity of these cells upon TF perturbation. Here we used the GRN inferred from the NMP differentiation trajectory as input, where target genes are constrained to also be TFs. Given the improved signal-to-noise ratio in the metacell representation, we disable the default kNN denoising step.

660

Embryos and nuclear isolation

C57BL/6Babr mice were bred and maintained by the Babraham Institute Biological Support Unit. All mouse experimentation was approved by the Babraham Institute Animal Welfare and Ethical Review Body. Animal husbandry and experimentation complied with existing European 665 Union and United Kingdom Home Office legislation and local standards.

Following dissection, embryos from the same stages were pooled to give sufficient cell numbers. Embryos were dissociated into single-cells using 200µl of TriplE Express for 10 minutes at 37°C on a shaking incubator. 1ml of ice-cold 10% FBS in PBS was added to quench 670 and cells were filtered using a 40µM Flowmi cell strainer. Following centrifugation at 300g for 5 minutes, the supernatant was discarded and cells were resuspended in 50µl of PBS containing 0.04% BSA. Cells were counted and viability assessed using trypan blue staining on a Countess II instrument (Invitrogen). >95% of cells were negative for trypan blue indicating high sample quality.

675

Nuclear isolation was carried out according to the low-cell input version of the 10X protocol for cell lines and PBMCs

(https://assets.ctfassets.net/an68im79xiti/6t5iwATCRaHB4VWOJm2Vgc/bdfd23cdc1d0a321487c8b231a448103/CG000365_DemonstratedProtocol_NucleiIsolation_ATAC_GEX_Sequencing_RevB.pdf). Specifically, the 50µl cell suspension was transferred to a 0.2ml PCR tube and centrifuged at 300g for 5 minutes. After removing the supernatant, cells were resuspended in 50µl ice cold nuclear extraction (NE) buffer (10mM Tris pH 7.5, 10mM NaCl, 3mM MgCl2, 1% BSA, 0.1% Tween, 1mM DTT, 1U/ul RNaseIn (Promega), 0.1% NP40, 0.01% Digitonin) and incubated on ice for 4 minutes. 50µl of wash buffer (identical to NE buffer but lacking NP40 and digitonin) was added and nuclei were centrifuged at 500g for 5 minutes at 4°C. After removing the supernatant, nuclei were washed once in 50µl of diluted nuclei buffer (10x Genomics), span down and finally resuspended in 7ul of dilute nuclei buffer (10x Genomics). 1µl was used to assess quality using a microscope and count nuclei using a Countess II instrument. >99% of nuclei stained positive for trypan blue and the nuclei were found to have the expected morphology. Nuclei were diluted such that a maximum of 16,000 were taken forward for 10x Multiome library preparation.

Brachyury gene targeting

One-cell stage zygotes were obtained from C57BL/6Babr superovulated matings. CRISPR/Cas9 reagent consisted of Cas9 protein (200ng/ul) and a sgRNA targeting exon 3 of the Brachyury gene (120ng/ul, ACTCTCACGATGTGAATCCG), diluted in Opti-MEM I (Thermo Fisher). Control embryos received Cas9 but no gRNA. Super electroporator NEPA21 and platinum plate electrodes 1mm gap (CUY501P1-1.5) were used for electroporation. Four repeats of poring pulses (40V, 3.5ms length and 50ms intervals) and five repeats of transfer pulses (7V, 50ms length, 50ms intervals) were applied to zygotes. Approximately 50 embryos were added to 5-6ul of CRISPR/Cas9mix per electroporation. Embryos were cultured overnight and only 2-cell stage embryos were transferred into pseudo-pregnant recipients, which were later harvested to obtain E8.5 embryos. In total this yielded 3 control embryos and 7 Brachyury KO embryos which were pooled for processing.

For genotyping embryonic yolk sacs were lysed using QuickExtract buffer prior to PCR amplification of a region spanning the predicted cut site (forward: GTAGGCAGTCACAGCTATGA, reverse: GGGTTTAATGGTGTATAGCG). The resulting amplicon was Sanger sequenced and the trace was analysed using Synthego ICE analysis producing a KO score of 93% (https://www.synthego.com/products/bioinformatics/crispr-analysis)(Conant et al., 2022).

10x Multiome library preparation and sequencing

Libraries were prepared using the 10x Genomics Chromium and sequenced on a Novaseq 6000 instrument (Illumina) using the recommended read-lengths. This yielded medians of 720 million RNA-seq reads and 481 million ATAC reads per sample. We recovered a median of 7,700 cells per sample prior to quality control.

ChIP-seq data processing

ChIP-seq data for TFs Cdx2, Foxa2, Gata1, Gata4, Tal1 and Tbx5 was obtained from the Gene Expression Omnibus. Due to the limited availability of *in vivo* ChIP-seq datasets, we used *in vitro* models that more closely resemble the gastrulating embryo (**Supplementary Table 1**). Reads were trimmed using Trim Galore (v0.4.5) and mapped to *M. musculus* GRCm38 using Bowtie2 (Langmead and Salzberg, 2012) (v2.3.2). Bigwig files were generated

for genome browser visualisation using samtools (v1.13)(Li et al., 2009) and bamCoverage (v3.5.1)(Ramírez et al., 2016). Peak calling was performed using macs2 (v2.2.7.1)(Zhang et al., 2008) with the "--broad and --broad-cutff 0.1" arguments.

References

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086. .

Amezquita, R.A., Lun, A.T.L., Becht, E., Carey, V.J., Carpp, L.N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., et al. (2019). Orchestrating single-cell analysis with Bioconductor. Nat. Methods *17*, 137–145. .

Amin, S., Neijts, R., Simmini, S., van Rooijen, C., Tan, S.C., Kester, L., van Oudenaarden, A., Creyghton, M.P., and Deschamps, J. (2016). Cdx and T Brachyury Co-activate Growth Signaling in the Embryonic Axial Progenitor Niche. Cell Rep. *17*, 3165–3177. .

Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C.W., et al. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature *576*, 487–491. .

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. *21*, 111. .

Arnold, S.J., and Robertson, E.J. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. Nat. Rev. Mol. Cell Biol. *10*, 91–103. .

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366. .

Bardot, E.S., and Hadjantonakis, A.-K. (2020). Mouse gastrulation: Coordination of tissue patterning, specification and diversification of cell fate. Mech. Dev. *163*, 103617. .

Berest, I., Arnold, C., Reyes-Palomares, A., Palla, G., Rasmussen, K.D., Giles, H., Bruch, P.-M., Huber, W., Dietrich, S., Helin, K., et al. (2019). Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. Cell Rep. *29*, 3147–3159.e12. .

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

Bruce, A.E.E., and Winklbauer, R. (2020). Brachyury in the gastrula of basal vertebrates. Mech. Dev. *163*, 103625. .

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502. .

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173. .

Chawengsaksophak, K., de Graaff, W., Rossant, J., Deschamps, J., and Beck, F. (2004). Cdx2 is essential for axial elongation in mouse development. Proc. Natl. Acad. Sci. U. S. A. *101*, 7641–7645. .

Chen, S., Lake, B.B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat. Biotechnol. *37*, 1452–1457. .

Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. *9*, 781. .

Conant, D., Hsiau, T., Rossi, N., Oki, J., Maures, T., Waite, K., Yang, J., Joshi, S., Kelso, R., Holden, K., et al. (2022). Inference of CRISPR Edits from Sanger Trace Data. CRISPR J *5*, 123–130. .

Crowell, H.L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M.D. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. Nat. Commun. *11*, 6077. .

de Crozé, N., Maczkowiak, F., and Monsoro-Burq, A.H. (2011). Reiterative AP2a activity controls sequential steps in the neural crest gene regulatory network. Proc. Natl. Acad. Sci. U. S. A. *108*, 155–160. .

Cusanovich, D.A., Reddington, J.P., Garfield, D.A., Daza, R.M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H.A., Christiansen, L., Qiu, X., Steemers, F.J., et al. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature *555*, 538–542. .

Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796–800. .

Fleck, J.S., Jansen, S.M.J., Wollny, D., Seimiya, M., Zenk, F., Santel, M., He, Z., Gray Camp, J., and Treutlein, B. (2021). Inferring and perturbing cell fate regulomes in human cerebral organoids.

Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. *29*, 1363–1375. .

Gaston, K., and Jayaraman, P.S. (2003). Transcriptional repression in eukaryotes: repressors and repression mechanisms. Cell. Mol. Life Sci. *60*, 721–741. .

Gouti, M., Delile, J., Stamataki, D., Wymeersch, F.J., Huang, Y., Kleinjung, J., Wilson, V., and Briscoe, J. (2017). A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development. Dev. Cell *41*, 243–261.e7. .

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. *53*, 403–411. .

Guibentif, C., Griffiths, J.A., Imaz-Rosshandler, I., Ghazanfar, S., Nichols, J., Wilson, V., Göttgens, B., and Marioni, J.C. (2021). Diverse Routes toward Early Somites in the Mouse Embryo. Dev. Cell *56*, 141–153.e6. .

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. *36*, 421–427. .

Henrique, D., Abranches, E., Verrier, L., and Storey, K.G. (2015). Neuromesodermal progenitors and the making of the spinal cord. Development *142*, 2864–2875. .

Ibarra-Soria, X., Jawaid, W., Pijuan-Sala, B., Ladopoulos, V., Scialdone, A., Jörg, D.J., Tyser, R.C.V., Calero-Nieto, F.J., Mulas, C., Nichols, J., et al. (2018). Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. Nat. Cell Biol. *20*, 127–134. .

Iurlaro, M., Stadler, M.B., Masoni, F., Jagani, Z., Galli, G.G., and Schübeler, D. (2021). Mammalian SWI/SNF continuously restores local accessibility to chromatin. Nat. Genet. *53*, 279–287. .

Janssens, J., Aibar, S., Taskiran, I.I., Ismail, J.N., Gomez, A.E., Aughey, G., Spanier, K.I., De Rop, F.V., González-Blas, C.B., Dionne, M., et al. (2022). Decoding gene regulation in the fly brain. Nature *601*, 630–636. .

Kamal, A., Arnold, C., Claringbould, A., Moussa, R., Daga, N., Nogina, D., Kholmatov, M., Servaas, N., Mueller-Dott, S., Reyes-Palomares, A., et al. (2021). GRaNIE and GRaNPA: Inference and evaluation of enhancer-mediated gene regulatory networks applied to study macrophages.

Kamimoto, K., Hoffmann, C.M., and Morris, S.A. (2020). CellOracle: Dissecting cell identity via network inference and in silico gene perturbation.

Karimzadeh, M., and Hoffman, M.M. (2019). Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome.

Kartha, V.K., Duarte, F.M., Hu, Y., Ma, S., Chew, J.G., Lareau, C.A., Earl, A., Burkett, Z.D., Kohlway, A.S., Lebofsky, R., et al. (2021). Functional Inference of Gene Regulation using Single-Cell Multi-Omics.

Labosky, P.A., Winnier, G.E., Jetton, T.L., Hargett, L., Ryan, A.K., Rosenfeld, M.G., Parlow, A.F., and Hogan, B.L. (1997). The winged helix gene, Mf3, is required for normal development of the diencephalon and midbrain, postnatal growth and the milk-ejection reflex. Development *124*, 1263–1274. .

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498. .

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *175*, 598–599. .

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

Levine, M., and Davidson, E.H. (2005). Gene regulatory networks for development. Proc. Natl. Acad. Sci. U. S. A. *102*, 4936–4942. .

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. .

Liang, Z., Brown, K.E., Carroll, T., Taylor, B., Vidal, I.F., Hendrich, B., Rueda, D., Fisher, A.G., and Merkenschlager, M. (2017). A high-resolution map of transcriptional repression. Elife *6*, e22767. .

Lohoff, T., Ghazanfar, S., Missarova, A., Koulena, N., Pierson, N., Griffiths, J.A., Bardot, E.S., Eng, C.-H.L., Tyser, R.C.V., Argelaguet, R., et al. (2022). Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nat. Biotechnol. *40*, 74–85. .

Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. *5*, 2122. .

Luo, C., Liu, H., Xie, F., Armand, E.J., Siletti, K., Bakken, T.E., Fang, R., Doyle, W.I., Stuart, T., Hodge, R.D., et al. (2022). Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. Cell Genomics *2*, 100107. .

Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. Cell *183*, 1103–1116.e20. .

Martin, B.L. (2016). Factors that coordinate mesoderm specification from neuromesodermal progenitors with segmentation during vertebrate axial extension. Semin. Cell Dev. Biol. *49*, 59–67. .

Materna, S.C., and Davidson, E.H. (2007). Logic of gene regulatory networks. Curr. Opin. Biotechnol. *18*, 351–354. .

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics *33*, 1179–1186. .

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Milunsky, J.M., Maher, T.A., Zhao, G., Roberts, A.E., Stalker, H.J., Zori, R.T., Burch, M.N., Clemens, M., Mulliken, J.B., Smith, R., et al. (2008). TFAP2A mutations result in branchio-oculo-facial syndrome. Am. J. Hum. Genet. *82*, 1171–1177. .

Morrisey, E.E., Ip, H.S., Lu, M.M., and Parmacek, M.S. (1996). GATA-6: a zinc finger transcription factor that is expressed in multiple cell lineages derived from lateral mesoderm. Dev. Biol. *177*, 309–322. .

Neijts, R., Amin, S., van Rooijen, C., Tan, S., Creyghton, M.P., de Laat, W., and Deschamps, J. (2016). Polarized regulatory landscape and Wnt responsiveness underlie Hox activation in embryos. Genes Dev. https://doi.org/10.1101/gad.285767.116.

Ohinata, Y., Payer, B., O'Carroll, D., Ancelin, K., Ono, Y., Sano, M., Barton, S.C., Obukhanych, T., Nussenzweig, M., Tarakhovsky, A., et al. (2005). Blimp1 is a critical determinant of the germ cell lineage in mice. Nature *436*, 207–213. .

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680. .

Persad, S., Choo, Z.-N., Dien, C., Masilionis, I., Chaligné, R., Nawy, T., Brown, C.C., Pe'er, I., Setty, M., and Pe'er, D. (2022). SEACells: Inference of transcriptional and epigenomic cellular states from single-cell genomics data.

Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. Nature *566*, 490–495. .

Pijuan-Sala, B., Wilson, N.K., Xia, J., Hou, X., Hannah, R.L., Kinston, S., Calero-Nieto, F.J., Poirion, O., Preissl, S., Liu, F., et al. (2020). Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse

organogenesis. Nat. Cell Biol. *22*, 487–497. .

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. .

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81. .

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. .

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

van Rooijen, C., Simmini, S., Bialecka, M., Neijts, R., van de Ven, C., Beck, F., and Deschamps, J. (2012). Evolutionarily conserved requirement of Cdx for post-occipital tissue emergence. Development *139*, 2576–2583. .

Sambasivan, R., and Steventon, B. (2020). Neuromesodermal Progenitors: A Basis for Robust Axial Patterning in Development and Evolution. Front Cell Dev Biol *8*, 607516. .

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

Schorle, H., Meier, P., Buchert, M., Jaenisch, R., and Mitchell, P.J. (1996). Transcription factor AP-2 essential for cranial closure and craniofacial development. Nature *381*, 235–238. https://doi.org/10.1038/381235a0.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife *6*, e21856. .

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

Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692. .

Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., et al. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. Nat. Genet. *50*, 238–249. .

Tam, P.P.L., and Loebel, D.A.F. (2007). Gene function in mouse embryogenesis: get set for gastrulation. Nat. Rev. Genet. *8*, 368–381. .

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics *14*, 178–192. https://doi.org/10.1093/bib/bbs017.

Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., et al. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. *22*, 1798–1812. .

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443. .

Wilm, B., James, R.G., Schultheiss, T.M., and Hogan, B.L.M. (2004). The forkhead genes, Foxc1 and Foxc2, regulate paraxial versus intermediate mesoderm cell fate. Dev. Biol. *271*, 176–189. .

Xiang, Y., Zhang, Y., Xu, Q., Zhou, C., Liu, B., Du, Z., Zhang, K., Zhang, B., Wang, X., Gayen, S., et al. (2020). Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency. Nat. Genet. *52*, 95–105. .

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. .

Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M.M., Hu, M., and Ren, B. (2019). An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Nat. Struct. Mol. Biol. *26*, 1063–1070. .

Zhu, C., Zhang, Y., Li, Y.E., Lucero, J., Behrens, M.M., and Ren, B. (2021). Joint profiling of histone modifications and transcriptome in single cells from mouse brain. Nat. Methods *18*, 283–292. .

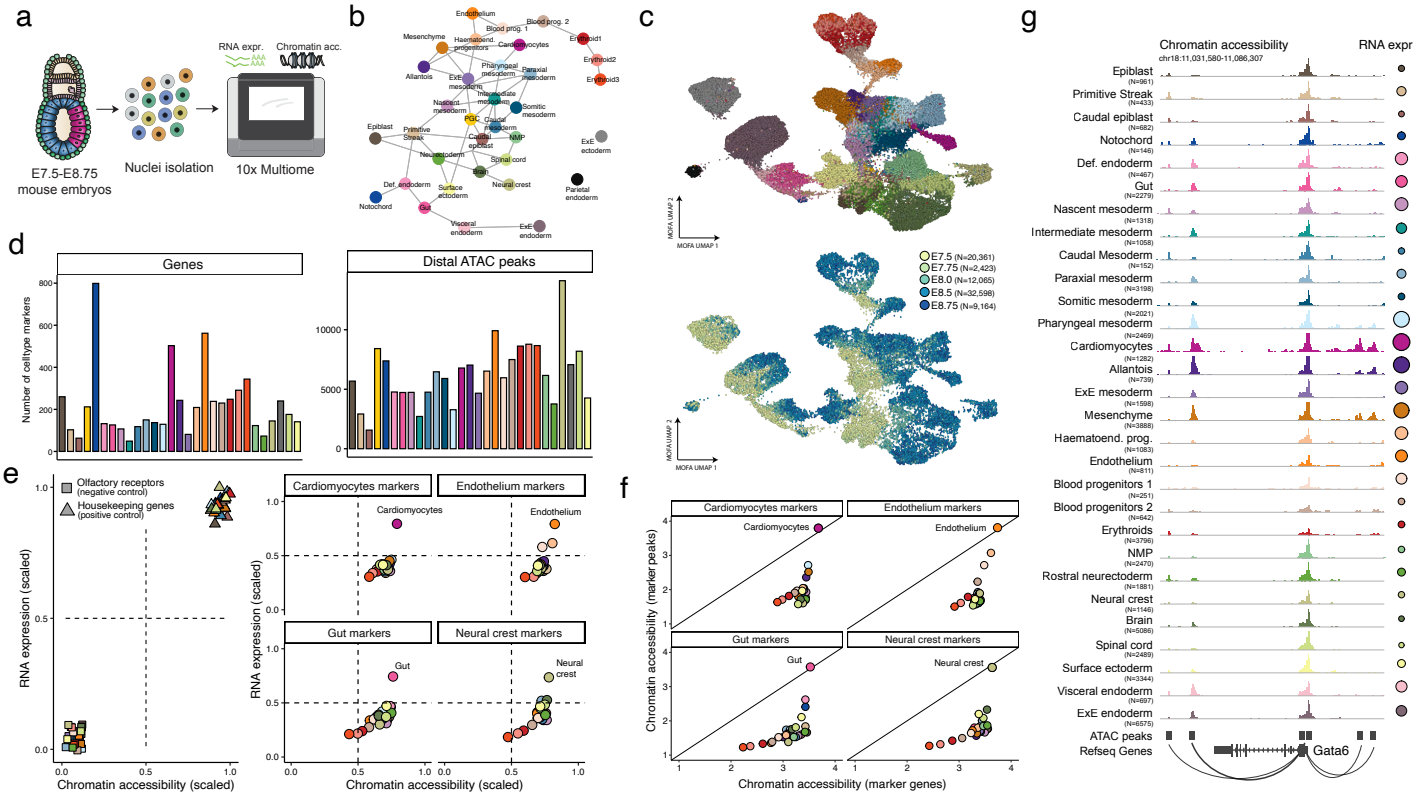

Figure 1: Simultaneous profiling of RNA expression and chromatin accessibility from single cells during mouse early organogenesis

(a) Schematic display of the experimental design. Mouse embryos are dissociated into single cells then lysed to extract nuclei which are processed for simultaneous snATAC and snRNA-seq from the same cell using the 10x Multiome protocol.

(b) Partition-based graph abstraction (PAGA) (Wolf et al., 2019) of the reference atlas (Pijuan-Sala et al., 2019), where each node corresponds to a different cell type. Cell types are coloured as per (Pijuan-Sala et al., 2019).

(c) Multi-modal dimensionality reduction using MOFA, followed by UMAP (Argelaguet et al., 2020). Cells are coloured by cell type (top, see (b) for key) and stage (bottom).

(d) Number of marker genes (left) and marker peaks (right) per cell type. See (b) for cell type colour key.

(e) RNA expression and promoter chromatin accessibility values of different gene sets quantified separately for each cell type. The left panel shows olfactory receptors (negative control, non-expressed genes with closed chromatin) and housekeeping genes (positive control, highly expressed genes with open chromatin). The right panel shows different gene sets of cell type marker genes. Each dot corresponds to a pseudobulk cell type, coloured as in (b). Note that RNA expression and chromatin accessibility values are quantified as an average across all genes from each gene set.

(f) Chromatin accessibility values of cell type marker genes (x-axis) and marker distal ATAC peaks from the same cell type (y-axis). Each panel shows gene and peak sets for different cell types. The diagonal line shows the values where both promoter and peak chromatin accessibility values are identical. Quantification of chromatin accessibility is done as in (e).

(g) Genome browser snapshot of the *Gata6* locus. Each track displays pseudobulk ATAC-seq signal for a given cell type. Note the dynamic patterns of distal regulatory regions both upstream and downstream of the gene, compared to the uniformly open promoter region.

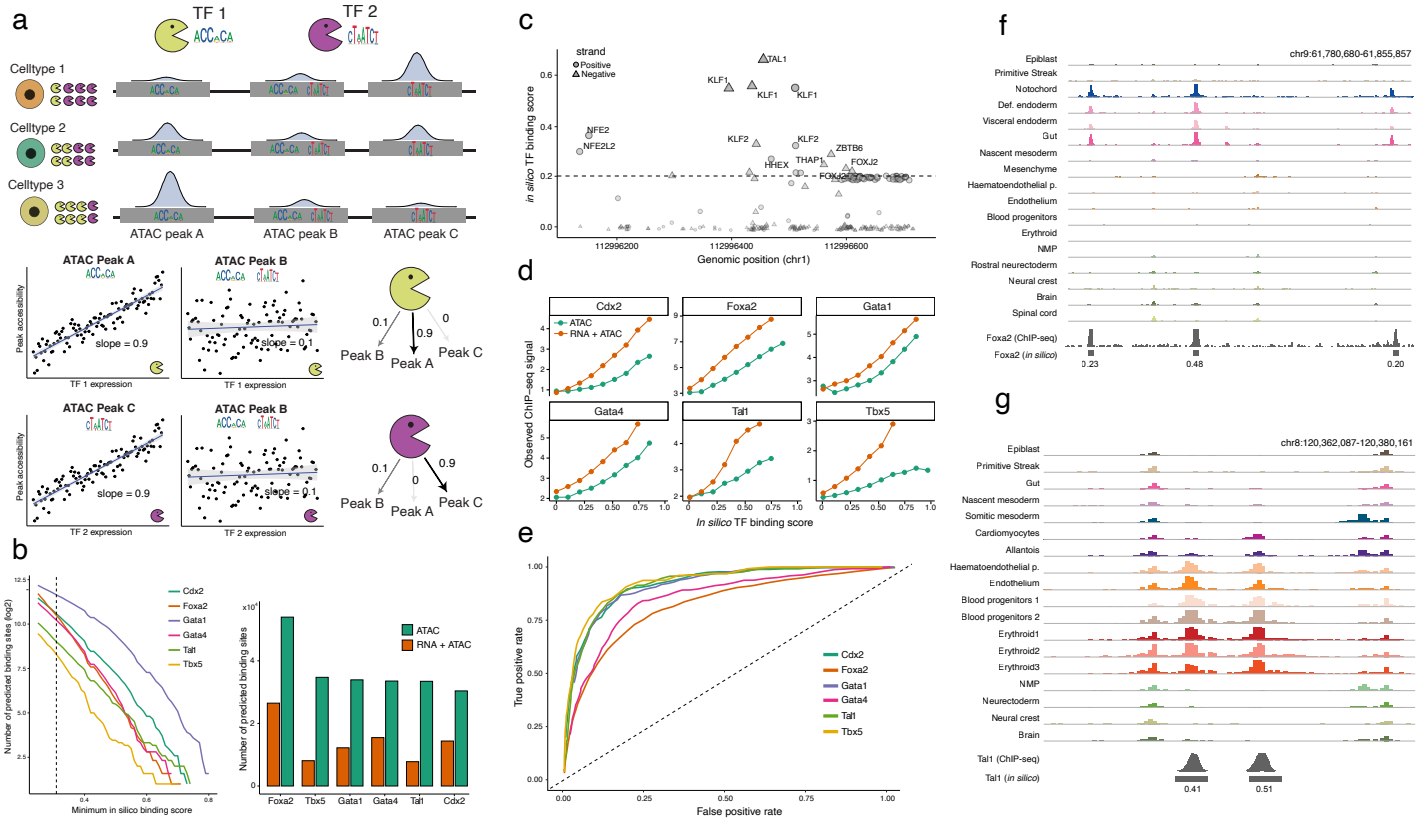

Figure 2: *In silico* ChIP-seq: leveraging multi-modal information to perform accurate prediction of transcription factor binding sites.

(a) Schematic of the *in silico* ChIP-seq methodology. Consider two different TFs (Pacmans), each one with different DNA binding preferences encoded in the form of different position-specific weight matrices; and three cis-regulatory elements represented as ATAC peaks (grey boxes), each one containing different instances of the TF motifs. Each row displays a different cell (or metacell or cell type, depending on the level of data aggregation). Each cell type is associated with different values of TF RNA expression (see changes in Pacman abundance) and chromatin accessibility of the cis-regulatory elements (see changes in the density histogram). The *in silico* ChIP-seq model exploits the correlation between TF RNA expression and the chromatin accessibility of the ATAC peaks that contain at least one instance of its TF motif to derive a quantitative TF binding score. In the schematic Peak A contains the TF 1 motif, and its accessibility correlates with the RNA expression of TF 1, thus leading to a high TF binding score. Peak B also contains the TF 1 motif, but its accessibility correlates poorly with the TF's RNA expression, which leads to a non-zero but low TF binding score. Peak C does not contain the TF 1 motif, which leads to a zero TF binding score.

(b) Left: the number of predicted binding sites for 6 representative TFs as a function of the minimum *in silico* TF binding score. Dashed line indicates the minimum score used in subsequent analyses. Right: Bar plots showing the number of predicted binding sites in the *in silico* ChIP-seq model when incorporating the RNA expression (orange) versus just using ATAC information (green).

(c) A representative instance of an ATAC peak highlighting the large number of TF motifs contained within a 600bp locus. Shown are the positions of all TF motifs within the ATAC peak (x-axis) against the *in silico* ChIP-seq score (y-axis). Note that only a subset of TF motif instances display high *in silico* ChIP-seq score. The dashed line indicates the cutoff used to determine a putative binding site, as in (b).

(d) Comparison of *in silico* TF binding scores (x-axis) versus experimental ChIP-seq signal (y-axis), using the same 6 TFs as in (b). Orange line displays scores derived from the *in silico* ChIP-seq model, whereas the green line displays scores derived when just using ATAC-seq information (i.e. omitting the TF RNA expression from the model). Scores were binned from 0 to 1 in intervals of 0.1,and each dot corresponds to the average value across all cis-regulatory regions from the interval. ChIP-seq datasets are all derived from publicly available data sets that most closely resemble mouse embryos at the gastrulation and organogenesis stage (Supplementary Table 1).

(e) Receiver Operating Characteristic (ROC) curves comparing the predicted TF binding sites vs the real TF binding sites (inferred from peak calling on the experimental ChIP-seq data).

(f) Genome browser snapshot of a locus containing (f) Foxa2 or (g) Tal1 predicted binding sites. Each track displays pseudobulk ATAC-seq signal for a given celltype. The experimental ChIP-seq values are shown in the bottom, together with the *in silico* TF binding scores for the ATAC peaks that have a TF binding score higher than 0.20 (same threshold as in (b)).

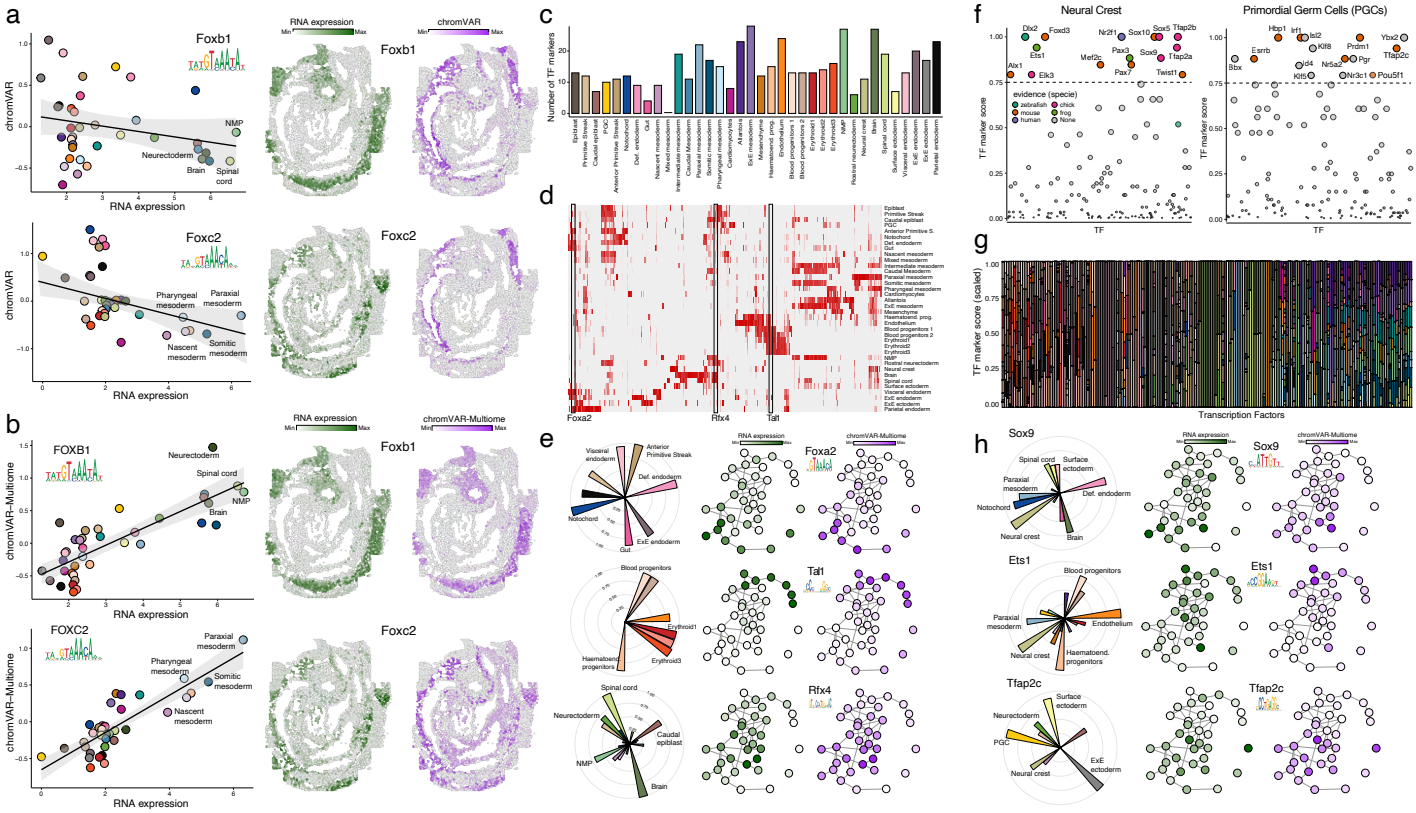

Figure 3: A catalogue of cell type-specific transcription factor activities reveals widespread pleiotropic activity.

(a) Comparison of chromVAR and chromVAR-Multiome for quantification of transcription factor activity. (Left) Scatter plots show the correlation between the TF's RNA expression and the chromatin activity of target regions, quantified at the pseudobulk level using chromVAR. Each dot corresponds to a different cell type. (Right) Spatially-resolved TF RNA expression (imputed values from Lohoff et al, 2021, coloured in green) and TF chromatin activity (coloured in purple). Note that spatially-resolved TF chromatin activity values are inferred by mapping the 10x Multiome cells onto the spatial transcriptomic data (Methods).

(b) Same as (a), but TF activities quantified using chromVAR-Multiome.

(c) Barplot displaying the number of TF markers per cell type. TF markers are inferred using the TF activity scores, which results from performing differential analysis with the chromVAR-Multiome values (Methods). The higher the score for TF i in celltype j, the more active this TF is predicted to be in cell type j, with a minimum score of 0 and a maximum score of 1.

(d) Heatmap displaying TF activity scores for each celltype (rows) and each TF (column).

(e) Left: polar plots displaying the celltype TF activity scores for three different TFs: Foxa2 (top), Tal1 (middle) and Rfx4 (bottom). Right: PAGA representation of the transcriptomic atlas as in Figure 1b for the three same TFs, with each node coloured by the RNA expression of the TF (green) and the corresponding chromVAR-Multiome score (purple).

(f) Dot plots displaying the TF activity scores for all TFs in Neural Crest cells (left) and PGCs (right). TFs with the highest TF activity score are labelled and coloured to indicate whether a known function has been reported and in which species the evidence was obtained (Supplementary Table 1).

(g) Stacked bar plots displaying the TF activity scores for all TFs across all cell types. Each column corresponds to a TF.

(h) as (e) but for TFs that display a pleiotropic effect (i.e. they are markers of very distinct cell types).

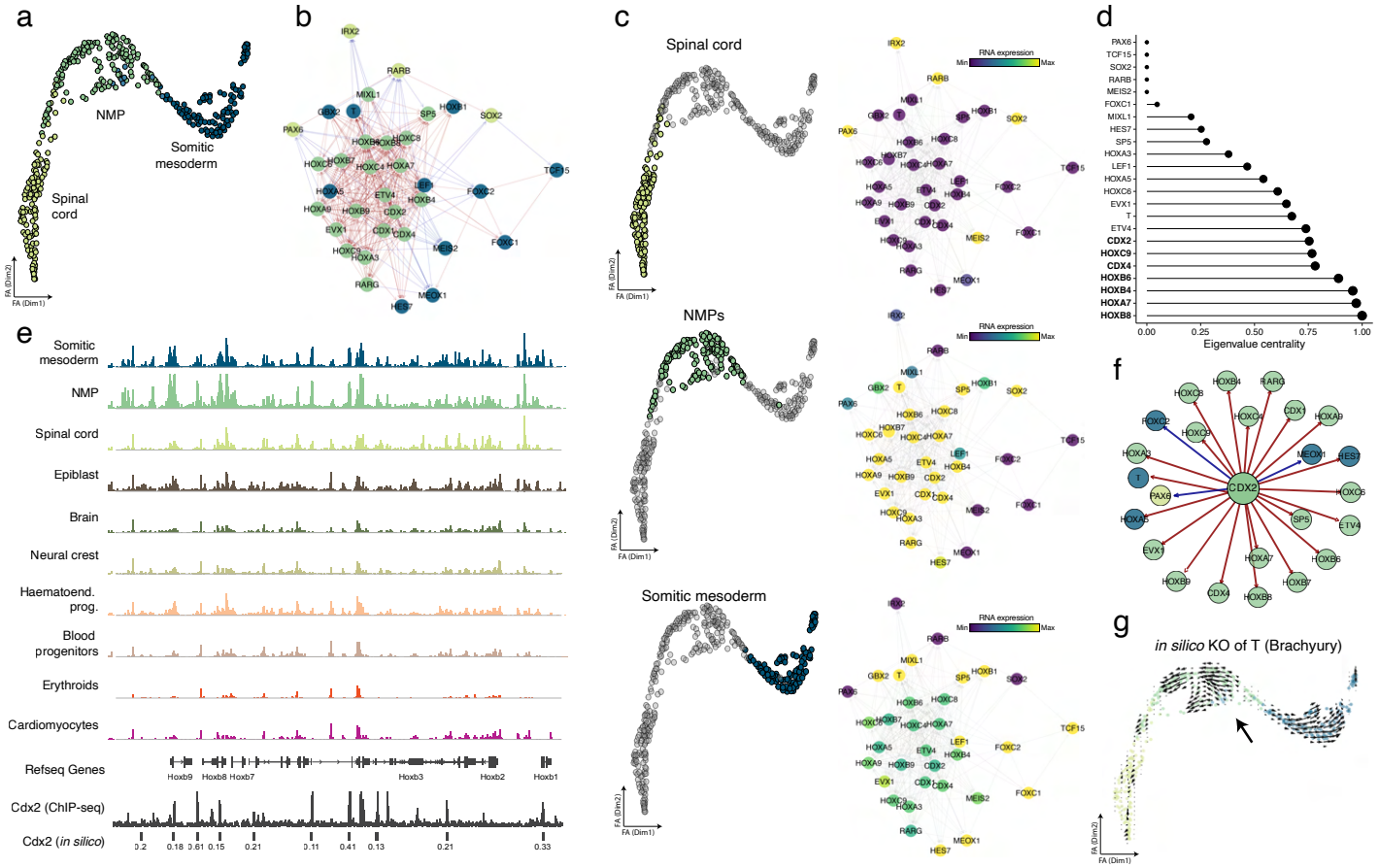

Figure 4: Characterisation of the Transcription factor regulatory networks underlying Neuromesodermal progenitors.

(a) Force-atlas layout of the NMP differentiation trajectory. Each dot corresponds to a metacell, coloured by cell type identity.

(b) TF regulatory network inferred using the NMP trajectory. Each node corresponds to a TF, coloured by the cell type where the TF displays the highest expression. Edges denote regulatory relationships. Red edges denote activatory relationships (the expression of the parent node is positively correlated with the expression of the child node), whereas blue edges denote repressive relationships (the expression of the parent node is negatively correlated with the expression of the child node). We refer the reader to Figure S10 for a schematic of the GRN inference procedure.

(c) Left: same force-atlas as (a) but highlighting each of the three cell types of the trajectory: Spinal cord (top), NMP (middle) or Somitic mesoderm (bottom). Right: Same TF regulatory network as in (b), but nodes are coloured based on the average expression of the TF in each of the three cell types of the trajectory: Spinal cord (top), NMP (middle) or Somitic mesoderm (bottom). For clarity, we increased the transparency of edges.

(d) Genome browser snapshot of the Hoxb loci. Each track displays pseudobulk ATAC-seq signal for a given celltype. Shown in the bottom is the *in silico* ChIP-seq predictions for Cdx2 and the experimental ChIP-seq signal for Cdx2 profiled in NMP-like cells.

(e) Eigenvalue centrality for each TF in the network.

(f) Regulatory connections between Cdx2 and downstream TFs. As in (b), nodes are coloured by the cell type where the TF displays the highest expression.

(g) *in silico* knock-out of Brachyury using CellOracle (Kamimoto et al, 2021). Shown is the force-atlas layout of the NMP differentiation trajectory. Each dot corresponds to a metacell, coloured by cell type identity. Arrows display the predicted changes in cell fate for different parts of the trajectory when knocking out Brachyury and propagating the signal through the GRN (Methods).

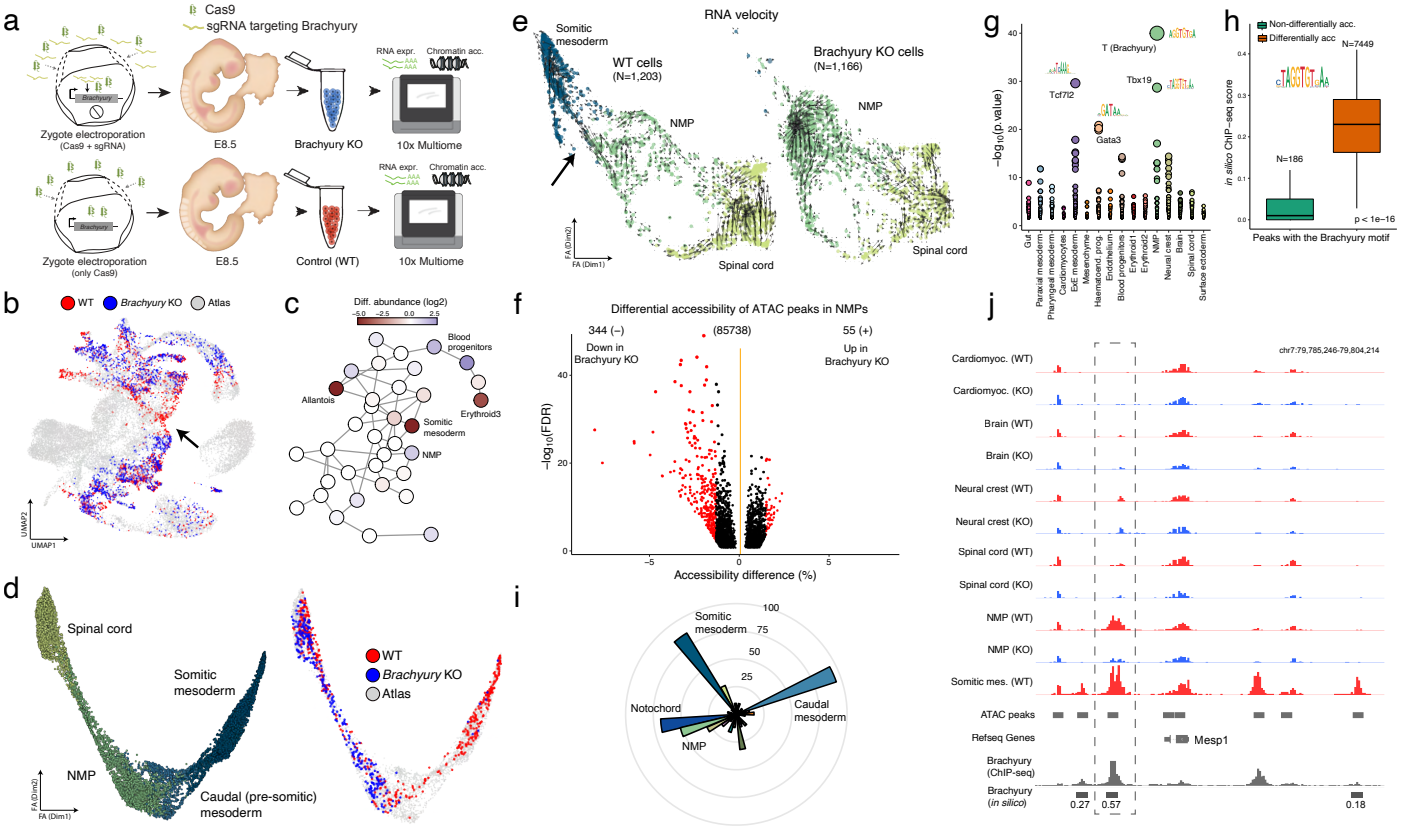

Figure 5: Brachyury controls the transition from neuromesodermal progenitors to posterior somitic mesoderm by priming cis-regulatory elements.

(a) Schematic showing the experimental design. We generated Brachyury KO embryos by electroporation of Cas9 protein and a single guide RNA (sgRNA) targeting the Brachyury (T) gene into one-cell stage zygote (Methods). Control embryos received Cas9 but no sgRNA. Embryos were transferred into pseudopregnant females and collected at E8.5 for 10x Multiome sequencing.

(b) Mapping cells to the reference atlas (Pijuan-Sala et al., 2019). Highlighted are cells in the reference dataset that are nearest neighbours to wildtype cells (red) or Brachyury KO cells (blue) in this experiment.

(c) PAGA representation of the reference atlas (Pijuan-Sala et al. 2019), where each node corresponds to a cell type. Nodes are coloured by differences in cell type abundance between Brachyury KO and WT control cells. Positive values indicate more abundance in the Brachyury KO, negative values indicate less abundance in the Brachyury KO.

(d) Force-directed layout of the trajectory that connects Neuromesodermal Progenitor (NMP) cells to either Spinal cord or Somitic mesoderm, inferred using the reference atlas (Pijuan-Sala et al., 2019). Left: each cell is coloured by cell type. Right: mapping cells to the reference NMP trajectory. Highlighted are cells in the reference trajectory that are nearest neighbours to wildtype cells (red) or Brachyury KO cells (blue) in this experiment.

(e) RNA velocity analysis of the NMP trajectory using scVelo (Bergen et al. 2020) on the 10x Multiome cells. Shown are WT cells (left) and Brachyury KO cells (right). The arrow highlights the trajectory from NMP to Somitic mesoderm that is present in WT cells but absent in Brachyury KO cells.

(f) Volcano plot displaying differential accessibility analysis of ATAC peaks between WT and Brachyury KO NMP cells. Coloured in red are ATAC peaks that pass significance threshold (Methods).

(g) Polar plots display which cell types are marked by the differentially accessible ATAC peaks in NMP cells. Each observation corresponds to a differentially accessible ATAC peak. Most of these peaks are markers of posterior mesodermal cell types (Somitic mesoderm and Caudal mesoderm).

(h) Shown is the *in silico* TF binding scores for Brachyury within ATAC peaks that contain the Brachyury motif. ATAC peaks are split based on their differentially accessibility significance when comparing WT and Brachyury KO NMP cells. Note that the *in silico* ChIP-seq is inferred using metacells from the NMP trajectory, instead of using all cells from the 10x Multiome reference.

(i) TF motif enrichment analysis in differentially accessible peaks per cell type (x-axis). The y-axis displays the FDR-adjusted p-values of a Fisher exact test. Each dot corresponds to a different TF motif, coloured by the cell type where the differential accessibility analysis is performed.

(j) Genome browser plot highlighting a Brachyury binding site within a differentially accessible ATAC peak between WT and Brachyury KO NMP cells. The highlighted ATAC peak displays high ChIP-seq signal as well as high *in silico* TF binding score. Note that the cis-regulatory region is located proximal to Mesp1, a gene that becomes expressed in Somitic mesoderm cells but not in NMP cells. This is suggestive of a role of Brachyury in epigenetic priming of somitic mesoderm fate in NMP cells.

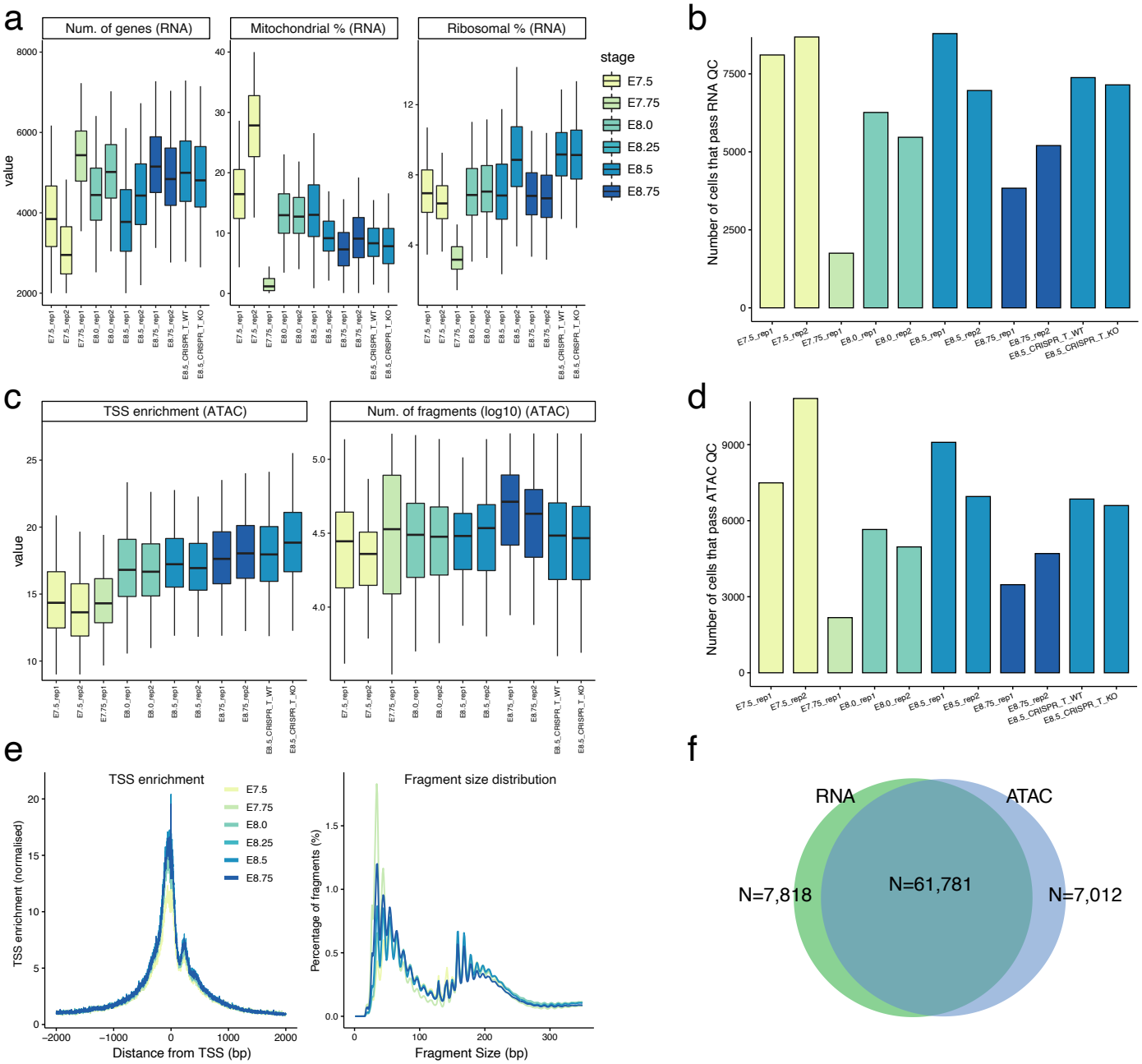

Figure S1: Quality control statistics per sample.

(a) Boxplots displaying RNA-seq quality control (QC) metrics per cell: the number of expressed genes (left), the percentage of mitochondrial reads (middle) and the percentage of Ribosomal reads (right). Each box is a sample, coloured by embryonic stage.

(b) Barplots displaying the number of cells that pass RNA-seq QC for each sample. Bars are coloured by embryonic stage.

(c) Boxplots displaying ATAC-seq quality metrics per cell: the enrichment of reads in the Transcription Start Site (TSS) (Granja et al., 2021) (left) and the number of fragments (right). Each box is a sample, coloured by embryonic stage.

(d) Barplots displaying the number of cells that pass ATAC-seq QC for each sample. Bars are coloured by embryonic stage.

(e) Histograms of QC statistics for ATAC-seq per sample. The left plot shows the number of fragments as a function of the distance from the nearest gene's TSS. The right plot shows the insert size distribution of ATAC-seq fragments.

(f) Venn diagram showing the overlap between cells that pass QC for the two modalities.

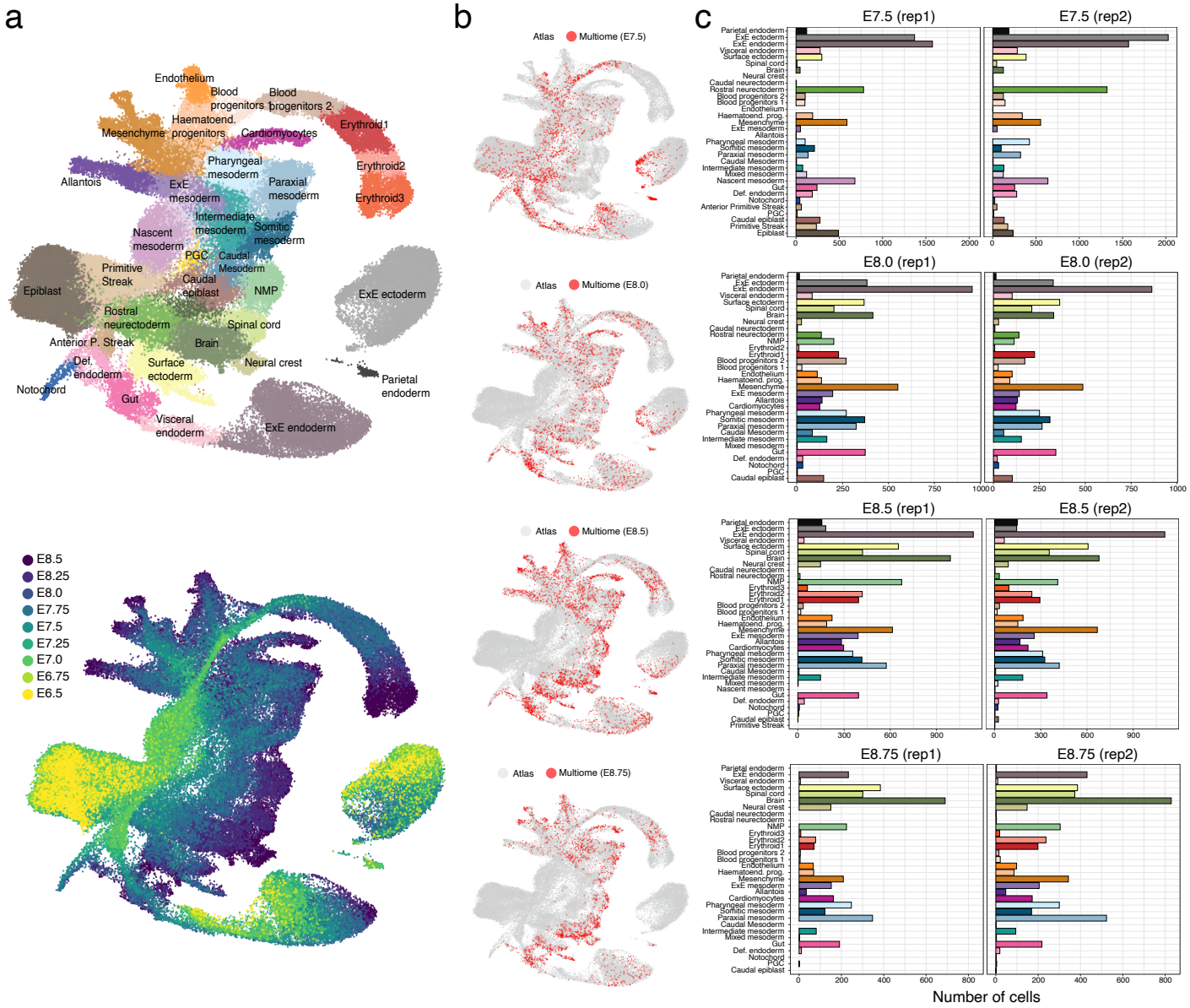

Figure S2: Mapping to the reference atlas and cell type annotation.

(a) UMAP plot from the reference atlas (Pijuan-Sala et al., 2019). Dots are coloured by cell type (top) or embryonic stage (bottom).

(b) Same UMAP plot as in (a). Coloured in red are cells that represent matching nearest neighbours to cells from this study (Methods). f corresponds to different stages.

(c) Bar plots displaying the number of cells for each cell type and sample. Each row corresponds to different stages.

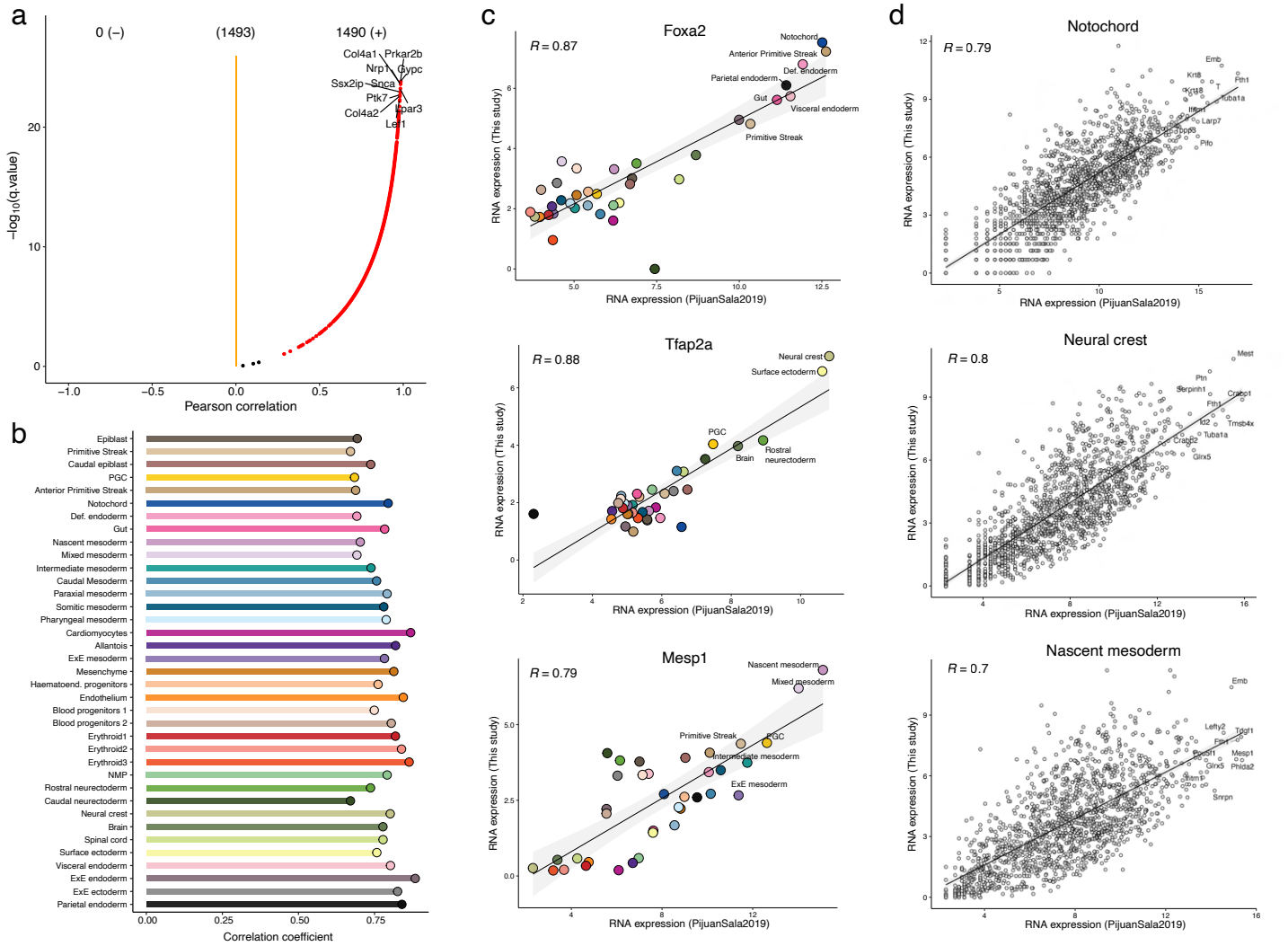

Figure S3: Comparison of the snRNA expression profiles from the 10x Multiome with existing scRNA-seq data from overlapping stages.

(a) Volcano plot displays the results of correlation tests per gene (across cell types) between the reference dataset (Pijuan-Sala et al., 2019) and this study. Correlations were computed at the cell type level after pseudobulk (i.e. each observation corresponds to a different cell type). Only cell type marker genes (N=1493) were considered for this analysis.

(b) Bar plots display the results of correlation tests per cell type (across genes) between the reference data set (Pijuan-Sala et al., 2019) and this study. As in (a), marker genes were considered for this analysis.

(c) Scatter plots show the RNA expression levels for three representative genes between the reference dataset (x-axis) and this study (y-axis). Each dot corresponds to a different cell type. Line represents the linear regression fit. Shown in the top left corner is the Pearson correlation coefficient.

(d) Scatter plots show the RNA expression levels for three representative cell types between the reference dataset (x-axis) and this study (y-axis). Each dot corresponds to a different gene. Line represents the linear regression fit. Shown in the top left corner is the Pearson correlation coefficient.

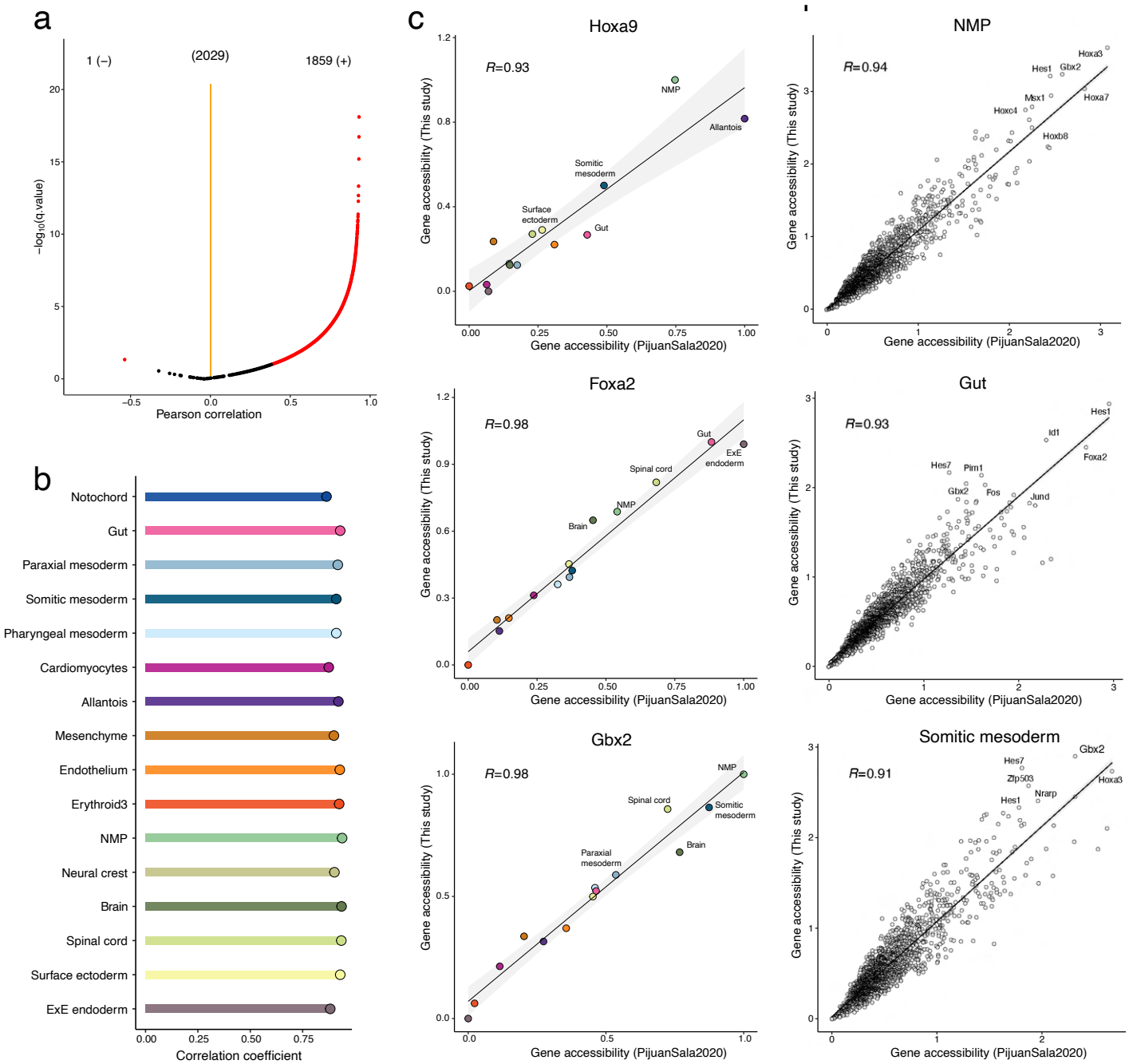

Figure S4: Comparison of the 10x Multiome chromatin accessibility data with existing scATAC-seq data from E8.25.

(a) Volcano plot displays the results of correlation tests per gene (across cell types) between the reference dataset (Pijuan-Sala et al., 2020) and this study. Correlations were computed at the cell type level after pseudobulk (i.e. each observation corresponds to a different cell type). Chromatin accessibility gene scores for marker genes were considered for this analysis (Methods).

(b) Bar plots display the results of correlation tests per cell type (across genes) between the reference data set (Pijuan-Sala et al., 2020) and this study. As in (a), marker genes were considered for this analysis.

(c) Scatter plots show the chromatin accessibility levels for three representative genes between the reference dataset (x-axis) and this study (y-axis). Each dot corresponds to a different cell type. Line represents the linear regression fit. Shown in the top left corner is the Pearson correlation coefficient.

(d) Scatter plots show the chromatin accessibility levels for three representative cell types between the reference dataset (x-axis) and this study (y-axis). Each dot corresponds to a different gene. Line represents the linear regression fit. Shown in the top left corner is the Pearson correlation coefficient.

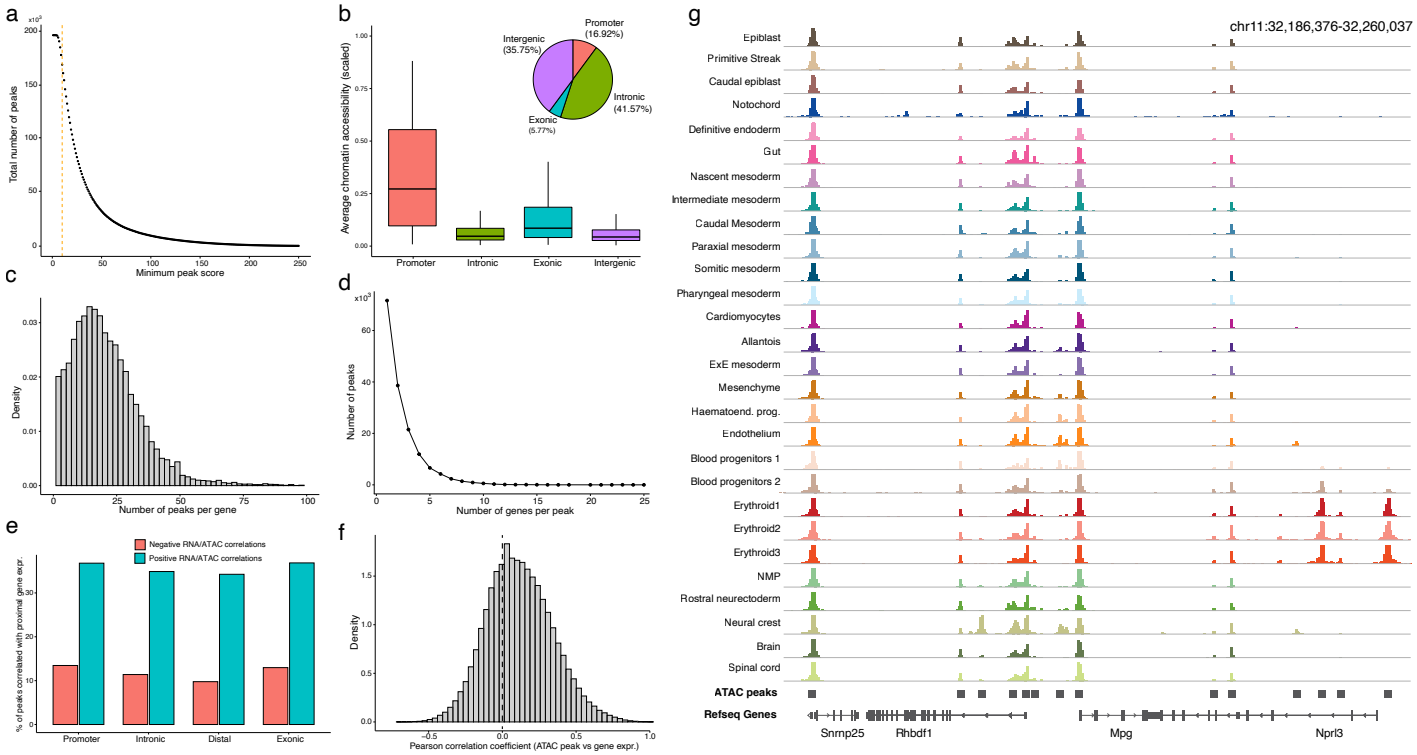

Figure S5: Identification of cis-regulatory elements.

(a) Scatter plot showing the relationship between the ATAC peak score cutoff (x-axis) and the corresponding number of ATAC peak calls (y-axis). Dashed line indicates the cutoff used in subsequent analyses.

(b) Boxplots showing the mean chromatin accessibility across all cells for peaks overlapping different genomic contexts. Inset: pie chart showing the percentage of peaks overlapping each genomic context.

(c) Histogram showing the number of ATAC peaks linked to each gene (maximum genomic distance of 50kb).

(d) Line plot showing the the number of genes linked to each peak.

(e) Barplot showing the percentage of ATAC peaks whose accessibility correlates with expression of at least one linked gene (q-value¡0.01 and a minimum absolute correlation of 0.25). Positive correlates are coloured in blue whereas negative correlations are coloured in red.

(f) Histogram displaying the distribution of Pearson correlation coefficients between ATAC peak accessibility and RNA expression (quantified at the pseudobulk level across cell types).

(g) Genome browser plot of a representative genomic locus that contains ATAC peaks that display variability in chromatin accessibility across cell types as well as peaks that are relatively homogeneous across cell types. Note that highly variable ATAC typically map to intergenic or intronic regions, whereas homogeneous ATAC peaks are found in promoter regions.

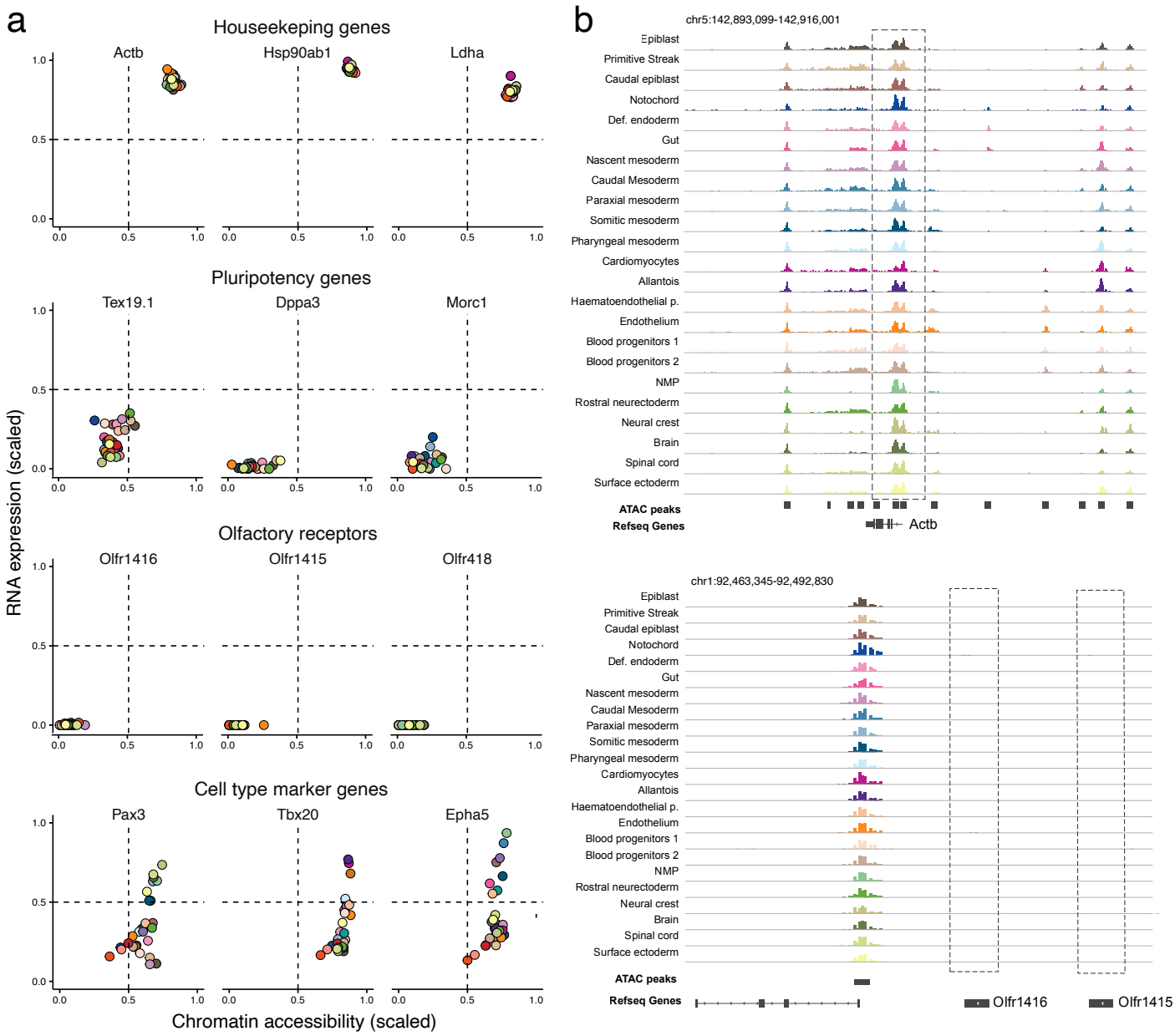

Figure S6: Representative examples of RNA expression and chromatin accessibility values for different gene sets.

(a) RNA expression and promoter chromatin accessibility values of different genes quantified for each cell type. The first row shows examples of housekeeping genes (positive control, highly expressed genes with open chromatin). The second row shows examples of naive pluripotency genes. The third row shows examples of olfactory receptors (negative control, non-expressed genes with closed chromatin). The fourth row shows examples of cell type marker genes.

(b) Genome browser snapshots displaying the *Actb* loci (housekeeping gene) and the *Olfr1416* loci (olfactory receptor). Each track displays pseudobulk ATAC-seq signal for a given celltype.

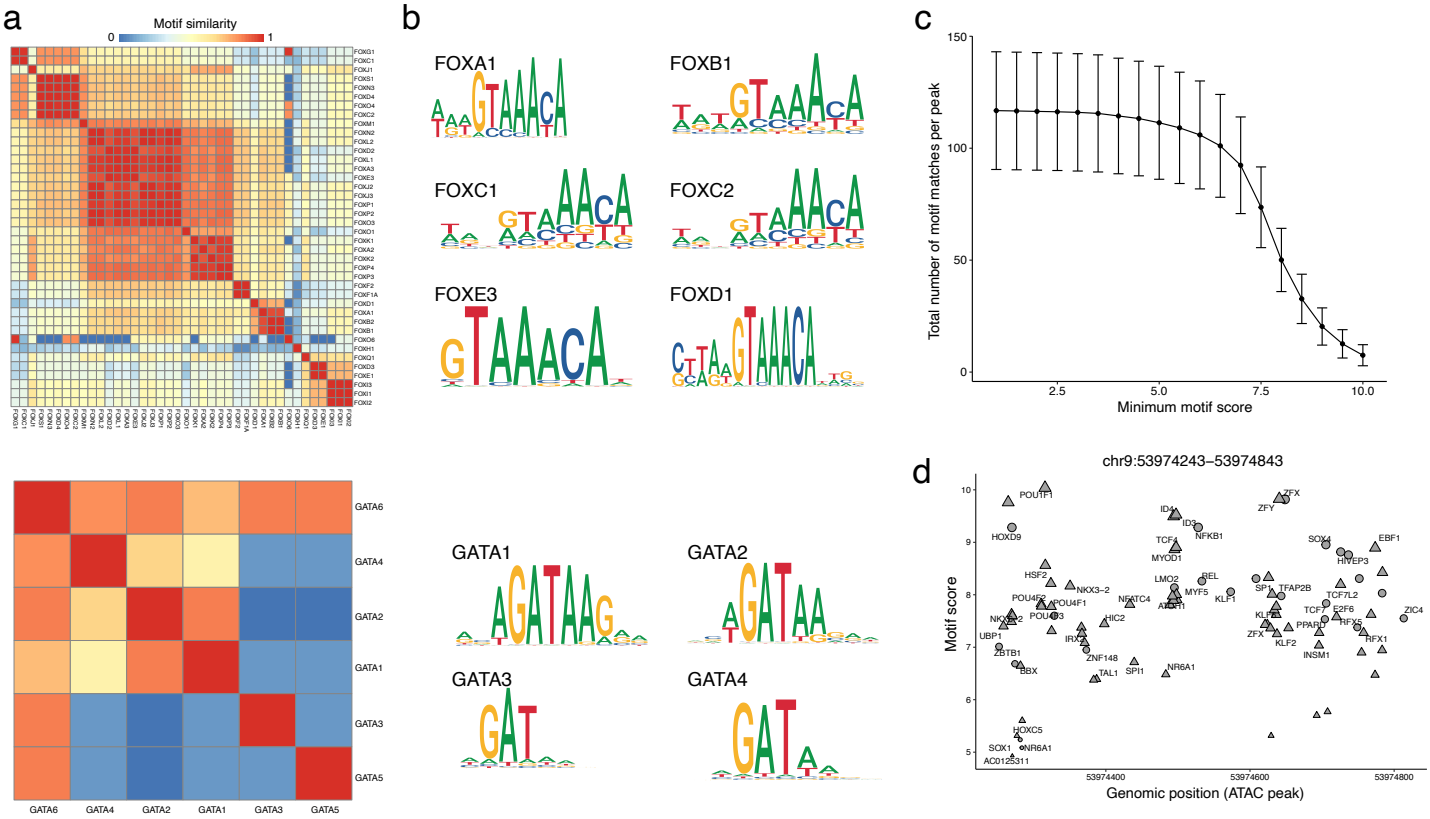

Figure S7: Transcription Factor motif similarities poses challenges for studying gene regulation.

(a) Heatmap showing the similarity between motif sequences for FOX (top) and GATA motifs (bottom). The similarity score is a normalised version of the sum of column correlations proposed in (Pietrokovski 1996). A score of 1 is expected for two identical motifs, whereas a score of 0 is expected for unrelated motifs.

(b) Representative examples of FOX (top) and GATA (bottom) transcription factor motifs to illustrate the similarity between motifs from the same TF family.

(c) Number of TF motifs per peak as a function of the minimum TF motif score cutoff.

(d) Location of motifs within a representative ATAC peak. Each dot represents a TF motif match within the genomic sequence. The x-axis displays the position of the match and the y-axis displays the TF motif score.

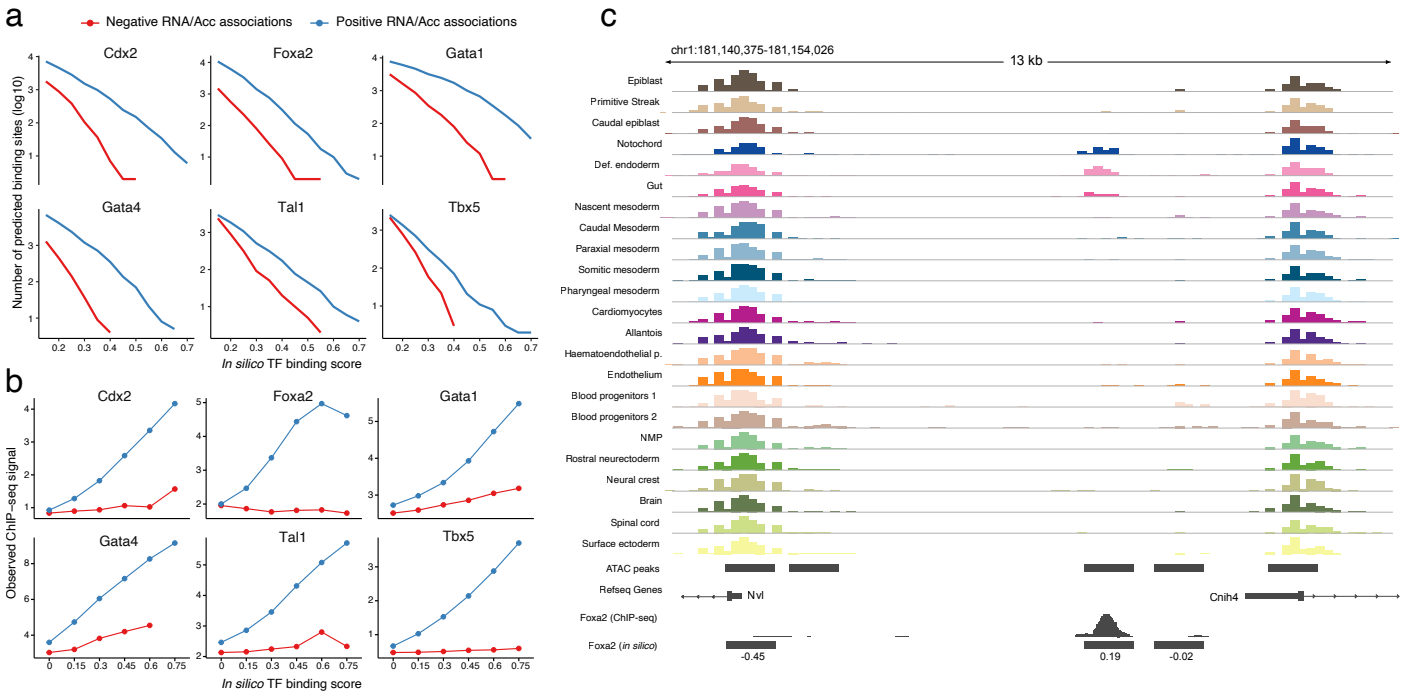

Figure S8: Negative *in silico* TF binding scores that represent repressive interactions with chromatin display low consistency with ChIP-seq data.

(a) Number of predicted binding sites (y-axis) as a function of the *in silico* TF binding score (x-axis), split by the sign of the correlation between RNA expression and ATAC peak accessibility.

(b) Experimental ChIP-seq signal (y-axis) as a function of the *in silico* TF binding score (x-axis), split by the sign of the correlation between RNA expression and ATAC peak accessibility.

(c) Representative genome browser snapshot showing a locus with predicted Foxa2 binding sites. Each track displays pseudobulk ATAC-seq signal for a given cell type. The bottom tracks display ChIP-seq for Foxa2 binding (as a validation) and the *in silico* predicted binding sites for Foxa2.

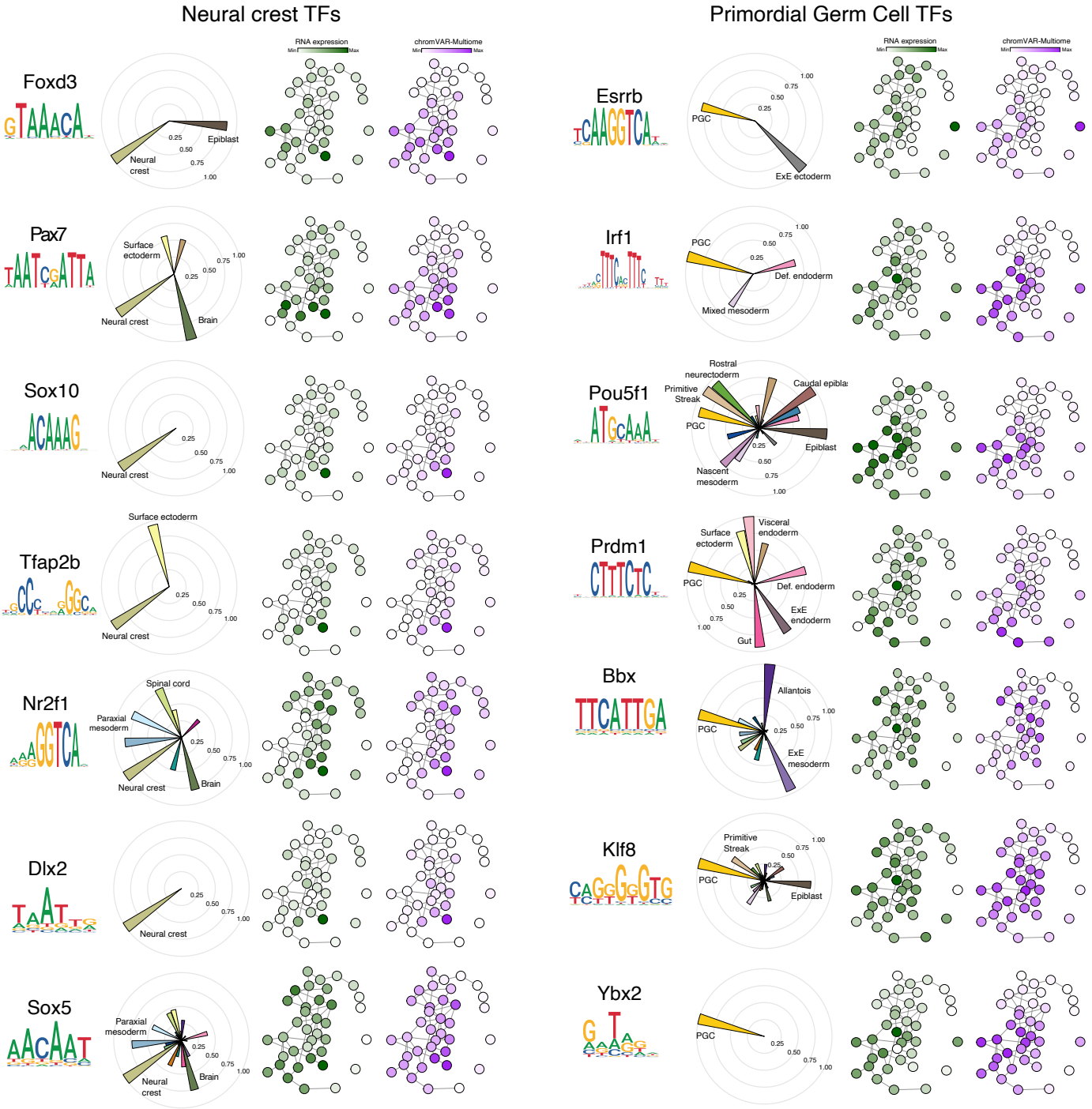

Figure S9: Overview of Transcription Factor activities for Neural Crest and Primordial Germ Cells markers.
Each row shows the TF activities that result from performing differential analysis of the chromVAR-Multiome values (Methods). The higher the score for TF *i* in celltype *j*, the more active TF *i* is predicted to be in cell type *j*, with a minimum score of 0 and a maximum score of 1. Each panel shows: Transcription Factor (TF) of interest, alongside its DNA motif (left). Polar plots displaying the TF activity scores for each cell type (Middle). PAGA representation of cell types (as in Figure 1b) with each node coloured by the gene expression level (green) and chromVAR-Multiome score (purple) (Right). TF markers for Neural Crest are shown in the first column and TF markers for Primordial Germ cells (PGCs) are shown in the second column.

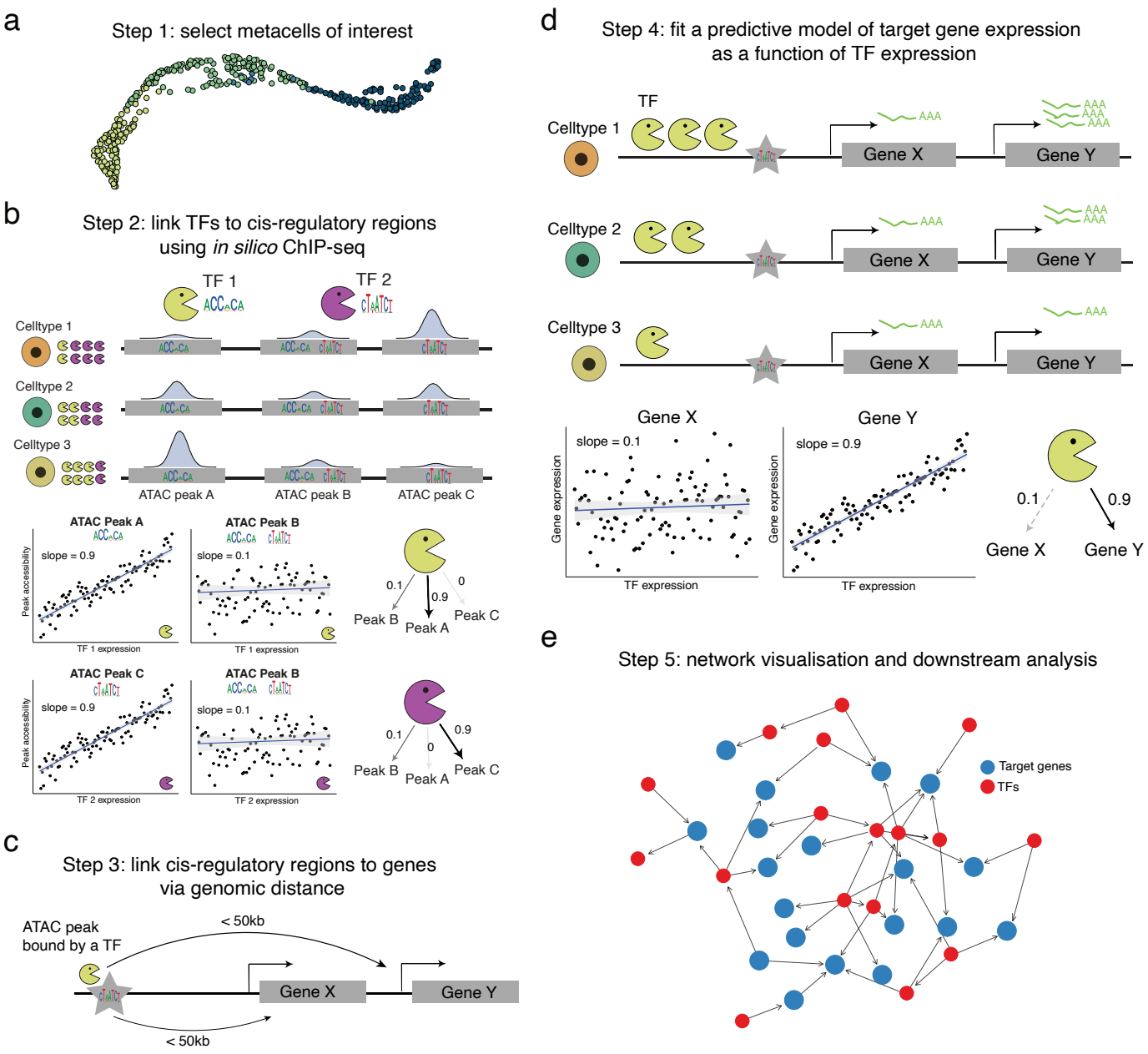

Figure S10: Schematic of the methodology for Gene Regulatory Network (GRN) inference.

(a) The first step is to select metacells of interest. We discourage the use of single cell resolution, as the sparsity of scATAC-seq makes it virtually impossible to obtain reliable association estimates between the RNA expression of Transcription Factors (TFs, which are typically lowly expressed genes) and chromatin accessibility of cis-regulatory regions.

(b) The second step is to use the *in silico* ChIP-seq methodology to link TFs with cis-regulatory elements. This is the same diagram as shown in Figure 2a. Note that the *in silico* ChIP-seq results will vary depending on the metacells that are used as input.

(c) The third step is to link cis-regulatory regions that are predicted to be bound by TFs to nearby genes via genomic distance. Note that this is a many-to-many mapping, where each gene can be linked to multiple cis-regulatory regions, and each cis-regulatory region can be linked to many genes.

(d) The fourth step is to build a predictive model of target gene RNA expression as a function of the TF's RNA expression. Although some GRN inference methods have used non-linear regression models, here we use linear regression models, as they provide more stable, interpretable and generalisable estimates.

(e) The final step is to visualise the GRN as directed graph and perform quantitative analysis on the network. Red notes represent TFs, whereas blue nodes represent target genes. The edge width is given by the slope of the linear models.

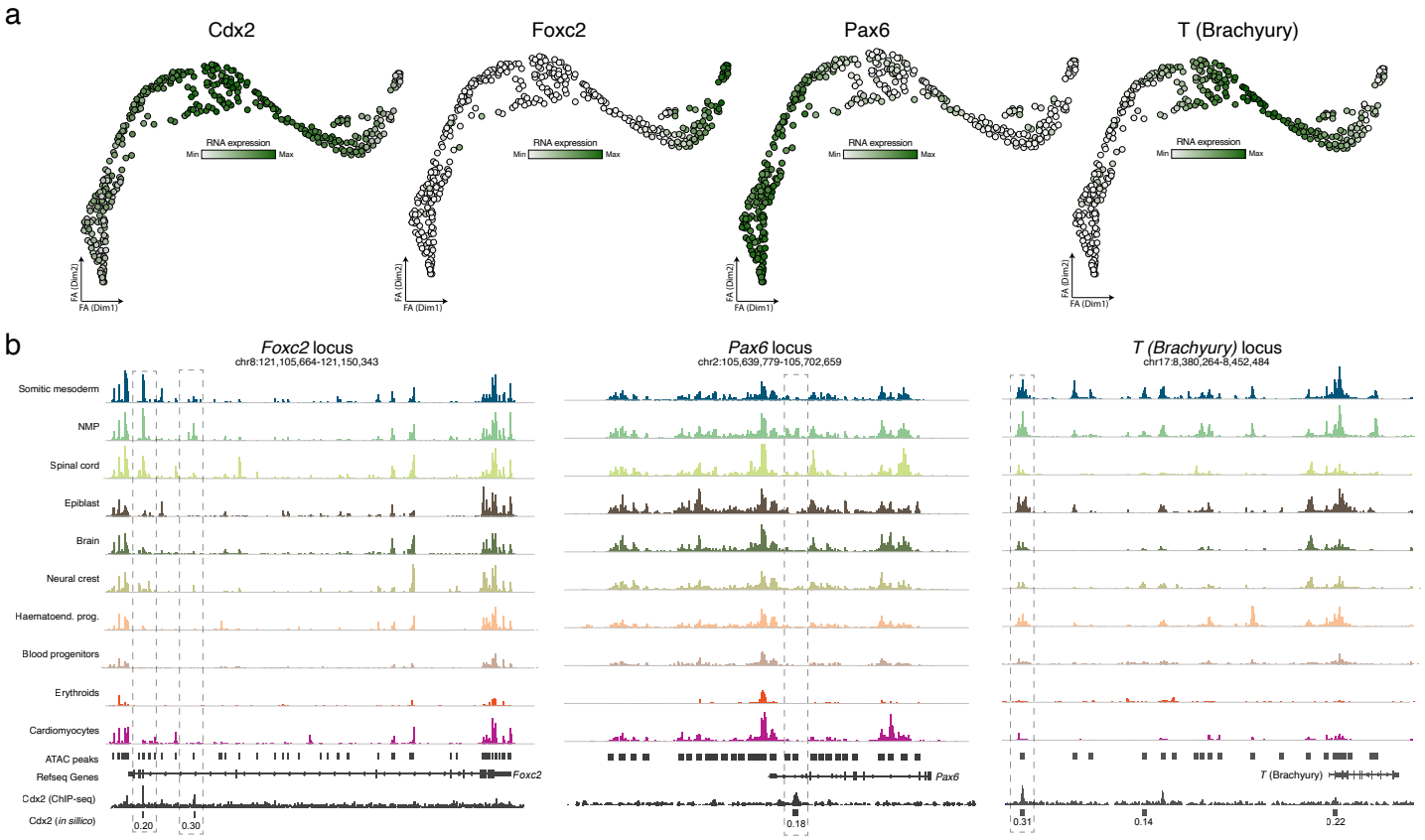

Figure S11: Examples of repressive interactions between Cdx2 and TFs that specify Spinal cord and Somitic mesoderm fate.

(a) Force-atlas layout of the NMP differentiation trajectory. Each dot corresponds to a metacell, coloured by the RNA expression of Cdx2, Foxc2, Pax6 and T, respectively.

(b) Genome browser snapshot of different loci that code for genes associated with Somitic mesoderm fate (Foxc2 and T) and Spinal cord fate (Pax6). Each track displays pseudobulk ATAC-seq signal for a given celltype. Shown in the bottom is the *in silico* ChIP-seq predictions for Cdx2 and the experimental ChIP-seq signal for Cdx2 profiled in NMP-like cells (Amin et al. 2016). Highlighted are Cdx2 binding sites near *Foxc2*.

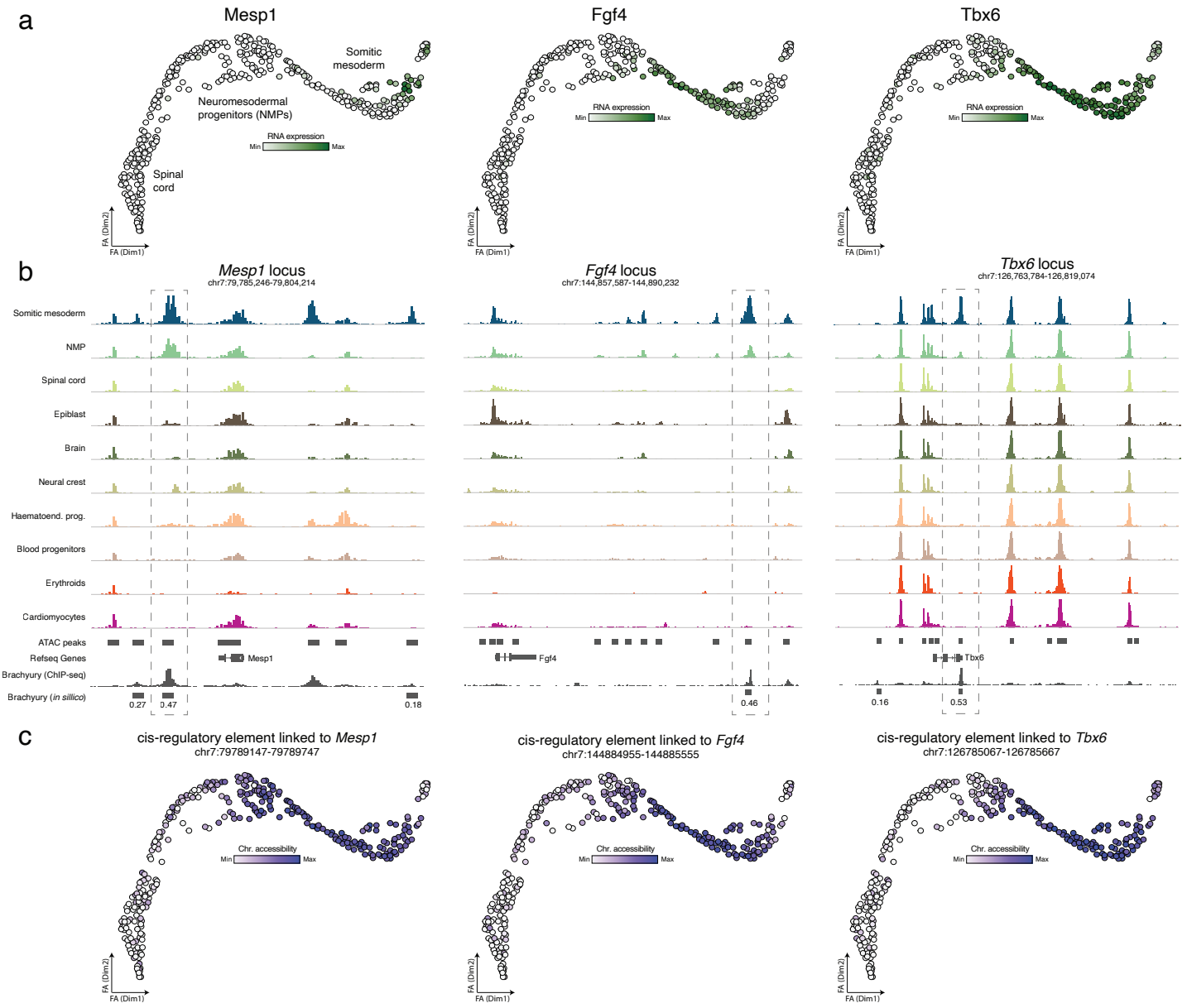

Figure S12: Representative examples of cis-regulatory elements targeted by Brachyury that prime NMP cells towards Somitic mesoderm.

(a) Force-atlas layout of the NMP differentiation trajectory. Each dot corresponds to a metacell, coloured by the RNA expression of Mesp1, Fgf4 and Tbx6, respectively.

(b) Genome browser snapshot of different loci that code for genes associated with Somitic mesoderm fate (Foxc2 and T) and Spinal cord (Pax6) fate. Each track displays pseudobulk ATAC-seq signal for a given celltype. Shown in the bottom are the *in silico* ChIP-seq predictions for Cdx2 and the experimental ChIP-seq signal for Cdx2 profiled in NMP-like cells (Amin et al. 2016). Highlighted are Cdx2 binding sites near *Foxc2*.

(c) Force-atlas layout of the NMP differentiation trajectory. Each dot corresponds to a metacell, coloured by the chromatin accessibility of cis-regulatory regions linked to *Mesp1*, *Fgf4*, *Pax6* and *Tbx6*, respectively.

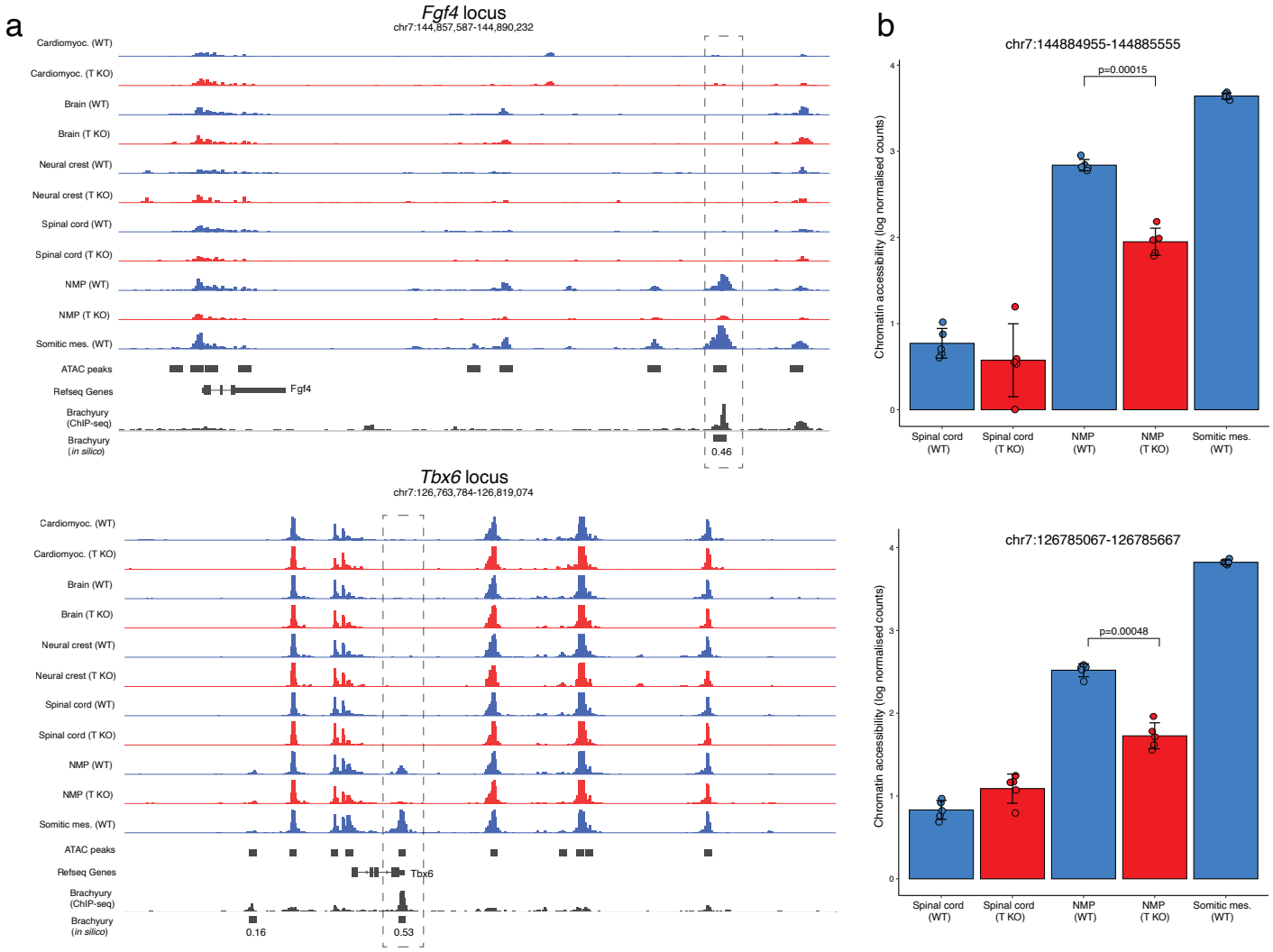

Figure S13: Representative examples of cis-regulatory elements that display impaired epigenetic priming in Brachyury KO NMP cells.

(a) Genome browser snapshot of two loci that code for genes associated with Somitic mesoderm fate: Fgf4(top) and Tbx6 (bottom). Each track displays pseudobulk ATAC-seq signal for a given celltype. Shown in the bottom are the *in silico* ChIP-seq predictions for Brachyury and the experimental ChIP-seq signal for Brachyury profiled in Embryoid Bodies (Tosic et al 2019). Highlighted are cis-regulatory elements bound by Brachyury that are differentially accessible in WT and KO NMP cells.

(b) Bar plots display the chromatin accessibility of cis-regulatory regions per cell type and genotype, quantified at the pseudobulk level with replicates (Methods). Each dot corresponds to a pseudobulk replicate. Error bars display the standard deviation across replicates. Shown on top of the NMP bar plots is the p-value of a t-test comparing the mean accessibility between WT and KO NMP pseudobulk replicates.