Article

Multi-omics profiling of mouse gastrulation at single-cell resolution

https://doi.org/10.1038/s41586-019-1825-8

Received: 18 October 2018

Accepted: 22 October 2019

Published online: 11 December 2019

Ricard Argelaguet[1,17], Stephen J. Clark[2,17]*, Hisham Mohammed[2,17], L. Carine Stapel[2,17], Christel Krueger[2], Chantriolnt-Andreas Kapourani[3,4], Ivan Imaz-Rosshandler[5,6], Tim Lohoff[2,5], Yunlong Xiang[7,8], Courtney W. Hanna[2,9], Sebastien Smallwood[2], Ximena Ibarra-Soria[10], Florian Buettner[11], Guido Sanguinetti[3], Wei Xie[7,8], Felix Krueger[12], Berthold Göttgens[5,6], Peter J. Rugg-Gunn[2,5,6,9], Gavin Kelsey[2,9], Wendy Dean[13], Jennifer Nichols[5], Oliver Stegle[1,14,15]*, John C. Marioni[1,10,16]* & Wolf Reik[2,9,16]*

Formation of the three primary germ layers during gastrulation is an essential step in the establishment of the vertebrate body plan and is associated with major transcriptional changes[1–5]. Global epigenetic reprogramming accompanies these changes[6–8], but the role of the epigenome in regulating early cell-fate choice remains unresolved, and the coordination between different molecular layers is unclear. Here we describe a single-cell multi-omics map of chromatin accessibility, DNA methylation and RNA expression during the onset of gastrulation in mouse embryos. The initial exit from pluripotency coincides with the establishment of a global repressive epigenetic landscape, followed by the emergence of lineage-specific epigenetic patterns during gastrulation. Notably, cells committed to mesoderm and endoderm undergo widespread coordinated epigenetic rearrangements at enhancer marks, driven by ten-eleven translocation (TET)-mediated demethylation and a concomitant increase of accessibility. By contrast, the methylation and accessibility landscape of ectodermal cells is already established in the early epiblast. Hence, regulatory elements associated with each germ layer are either epigenetically primed or remodelled before cell-fate decisions, providing the molecular framework for a hierarchical emergence of the primary germ layers.

Recent technological advances have enabled the profiling of multiple molecular layers at single-cell resolution[9–13], providing novel opportunities to study the relationship between the transcriptome and epigenome during cell-fate decisions. We applied single-cell nucleosome, methylome and transcriptome sequencing[12] (scNMT-seq) to profile 1,105 single cells isolated from mouse embryos at four developmental stages (embryonic day (E)4.5, E5.5, E6.5 and E7.5) representing the exit from pluripotency and primary germ-layer specification (Fig. 1a–d, Extended Data Fig. 1). Cells were assigned to a specific lineage by mapping their RNA-expression profiles to a comprehensive single-cell atlas[4] from the same stages when available or using marker genes (Extended Data Fig. 2). Using dimensionality reduction, we show that all three molecular layers contain sufficient information to separate cells by stage (Fig. 1b–d) and lineage identity (Extended Data Figs. 2, 3).

Epigenome dynamics at pluripotency exit

We characterized the changes in DNA methylation and chromatin accessibility during each stage transition. Globally, methylation levels increase from approximately 25% to approximately 75% in embryonic tissues and to about 50% in extra-embryonic tissues, driven mainly by a wave of de novo methylation from E4.5 to E5.5 that preferentially targets CpG-poor genomic loci[6,8,14] (Fig. 1e, Extended Data Fig. 3). By contrast, we observed a more gradual decline in global chromatin accessibility from around 38% at E4.5 to around 30% at E7.5 (Fig. 1f), with no differences between embryonic and extra-embryonic tissues (Extended Data Fig. 3). To relate epigenetic changes to the transcriptional dynamics across stages, we calculated—for each gene and across all embryonic cells—the correlation between RNA expression and the corresponding DNA methylation or chromatin accessibility at the promoter. Out of

[1]European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [2]Epigenetics Programme, Babraham Institute, Cambridge, UK. [3]School of Informatics, University of Edinburgh, Edinburgh, UK. [4]MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [5]Wellcome–MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK. [6]Department of Haematology, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK. [7]Center for Stem Cell Biology and Regenerative Medicine, MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China. [8]THU-PKU Center for Life Sciences, Tsinghua University, Beijing, China. [9]Centre for Trophoblast Research, University of Cambridge, Cambridge, UK. [10]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [11]Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany. [12]Bioinformatics Group, Babraham Institute, Cambridge, UK. [13]Department of Biochemistry and Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada. [14]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [15]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [16]Wellcome Sanger Institute, Cambridge, UK. [17]These authors contributed equally: Ricard Argelaguet, Stephen J. Clark, Hisham Mohammed, L. Carine Stapel. *e-mail: stephen.clark@babraham.ac.uk; o.stegle@dkfz.de; john.marioni@cruk.cam.ac.uk; wolf.reik@babraham.ac.uk

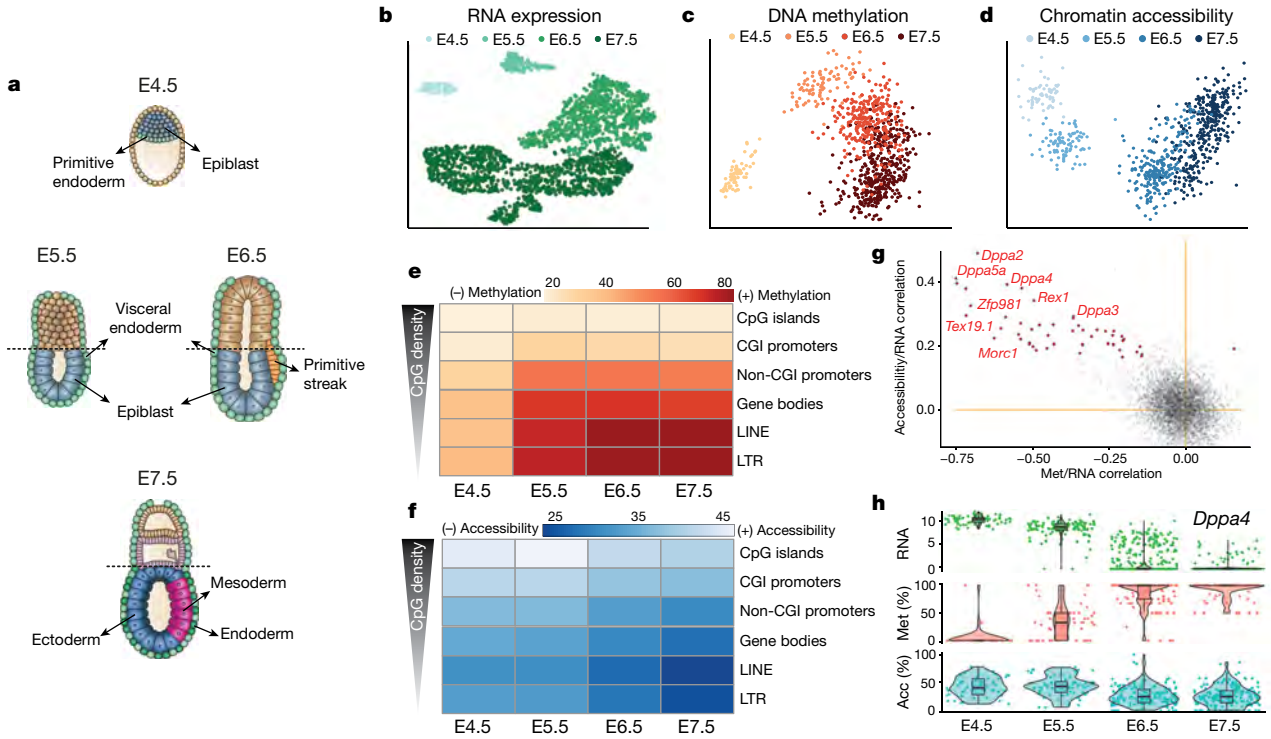


Fig. 1 | Single-cell multi-omics profiling of mouse gastrulation. a, Schematic of the developing mouse embryo, with stages and lineages considered in this study labelled. **b**, Dimensionality reduction of RNA-expression data using UMAP. Cells are coloured by stage. There are 1,061 cells included from 28 embryos sequenced using scNMT-seq and 1,419 cells from 26 embryos sequenced using scRNA-seq. **c**, **d**, Dimensionality reduction of DNA methylation data (**c**) and chromatin accessibility data (**d**) from scNMT-seq using factor analysis (Methods). Cells are coloured by stage. There are 986 cells included for DNA methylation data and 864 cells for chromatin accessibility data. **e**, **f**, Heat map of per cent DNA methylation levels (**e**) and per cent chromatin accessibility levels (**f**) by stage and genomic context. **g**, Scatter plot of Pearson correlation coefficients of promoter methylation (Met) versus RNA expression (*x* axis) and promoter accessibility versus RNA expression (*y* axis). Each dot corresponds to one gene ($n = 4{,}927$). Red dots depict significant associations for both correlation types ($n = 39$, false discovery rate (FDR) < 10%). Examples of early pluripotency and germ cell markers among the significant hits are labelled. **h**, Illustrative example of epigenetic repression of *Dppa4*. Box and violin plots show the distribution of RNA expression (log normalized counts, green), promoter methylation (red) and accessibility (Acc) (blue) per stage. Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. Each dot corresponds to one cell.

5,000 genes tested, we identified 125 genes the expression of which shows significant correlation with promoter DNA methylation and 52 with expression significantly correlated with chromatin accessibility (Fig. 1g, Extended Data Fig. 4, Supplementary Tables 1, 2). These loci largely comprise markers of early pluripotency and germ cells, such as *Dppa4*, *Zfp42*, *Tex19.1* and *Pou3f1* (Fig. 1g, h, Extended Data Fig. 4), which are repressed, coinciding with the global increase in methylation and decrease in accessibility. In addition, this analysis identified genes, including *Trap1a* and *Zfp981*, that may have unknown roles in development. Notably, of the genes that are upregulated after E4.5, only 39 and 9 show a significant correlation between RNA expression and promoter methylation or accessibility, respectively (Extended Data Fig. 4). This suggests that the upregulation of these genes is probably controlled by other regulatory elements.

Characterizing germ-layer epigenomes

To understand the relationships between all three molecular layers during germ-layer commitment we next applied multi-omics factor analysis (MOFA)[15] to cells collected at E7.5. MOFA performs unsupervised dimensionality reduction simultaneously across multiple data modalities, thereby capturing the global sources of cell-to-cell variability via a small number of inferred factors. Notably, the model makes use of multimodal measurements from the same cells, thereby detecting coordinated changes between the different data modalities.

As input to the model we used RNA-sequencing (RNA-seq) data across protein-coding genes and DNA methylation and chromatin accessibility data across putative regulatory elements. This includes promoters and germ-layer-specific chromatin immunoprecipitation with DNA sequencing (ChIP–seq) peaks for distal H3K27ac (enhancers) and H3K4me3 (transcription start sites)[16] (Extended Data Fig. 5). MOFA identified six factors, with the top two (sorted by variance explained) capturing the emergence of the three germ layers (Fig. 2a, b). Notably, MOFA links the variation at the gene-expression level to concerted methylation and accessibility changes at lineage-specific enhancer marks (Fig. 2c). By contrast, epigenetic changes at promoters or at H3K4me3-marked regions show much weaker associations with germ-layer formation (Fig. 2a, Extended Data Fig. 6, Supplementary Tables 3, 4). This supports other studies that have identified distal enhancers as lineage-driving regulatory regions[8,17–19]. Inspection of gene–enhancer associations identified enhancers linked to key germ-layer markers including *Lefty2* and *Mesp2* (mesoderm), *Foxa2* and *Bmp2* (endoderm), and *Bcl11a* and *Sp8* (ectoderm) (Fig. 2c, Extended Data Fig. 7). Notably, ectoderm-specific enhancers display fewer associations than their mesoderm and endoderm counterparts, a finding that is explored further below.

The four remaining factors correspond to additional transcriptional and epigenetic signatures related to anterior–posterior axial patterning (factor 3), notochord formation (factor 4), mesoderm patterning (factor 5) and cell cycle (factor 6) (Extended Data Fig. 8).

Finally, we sought to identify transcription factors that could drive or respond to epigenetic changes in germ-layer commitment. Integrating differential-expression information with motif enrichment at differentially accessible loci revealed that lineage-specific enhancers were

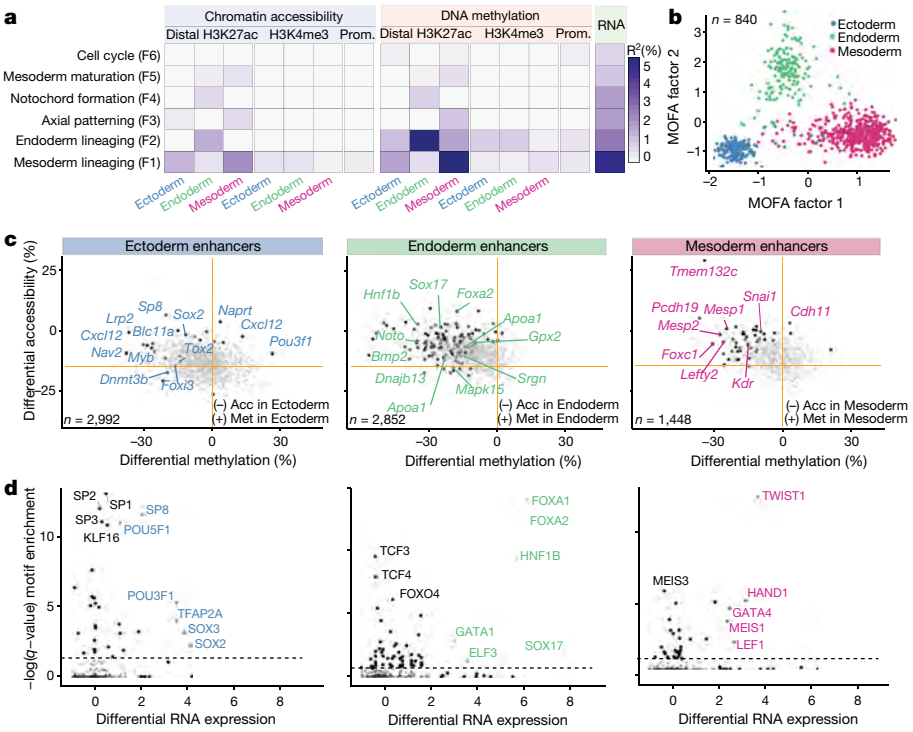


Fig. 2 | Multi-omics factor analysis reveals coordinated epigenetic and transcriptomic variation at enhancer elements during germ-layer commitment. a, Percentage of variance explained (R^2) by each MOFA factor (rows) across data modalities (columns). **b**, Scatter plot of MOFA factor 1 (*x* axis) and MOFA factor 2 (*y* axis). Cells are coloured according to their lineage assignment ($n = 840$). **c**, Scatter plots showing differential DNA methylation (*x* axis) and chromatin accessibility (*y* axis) at lineage-specific enhancers at E7.5. Ectoderm versus non-ectoderm cells (left, $n = 2,992$), endoderm versus non-endoderm cells (middle, $n = 2,852$) and mesoderm versus non-mesoderm cells (right, $n = 1,448$) are shown. Black dots depict gene–enhancer pairs with significant changes in RNA expression and methylation or accessibility (Pearson's χ^2 test, FDR <10%). **d**, Transcription factor motif enrichment at lineage-defining enhancers. Motif enrichment (Fisher's exact test, −log(*q* value), *y* axis, $n = 746$ motifs) versus differential RNA expression (log fold change, *x* axis) of the corresponding transcription factor is shown. The analysis is performed separately for ectoderm- (left), endoderm- (middle) and mesoderm- (right) defining enhancers. Transcription factors with significant motif enrichment (FDR <1%) and differential RNA expression (edgeR quasi-likelihood test, FDR <1%) are labelled.

enriched for binding sites associated with key developmental transcription factors, including POU3F1, SOX2 and SP8 for ectoderm, SOX17, HNF1B, and FOXA2 for endoderm, and GATA4, HAND1 and TWIST1 for mesoderm (Fig. 2d).

Time resolution of the enhancer epigenome

We next investigated how the epigenomic patterns associated with germ-layer specification arise during development. DNA methylation levels in endoderm- and mesoderm-defining enhancers follow the genome-wide dynamics, increasing from an average of 25% to 80% in all cell types (Fig. 3, Extended Data Fig. 9). Upon lineage specification, they undergo concerted demethylation to about 50% in a cell-type-specific manner. The opposite pattern is observed for chromatin accessibility; accessibility of mesoderm- and endoderm-defining enhancers initially decreases from approximately 40% to 30% (following the genome-wide dynamics) before becoming more accessible (approximately 45%) upon lineage specification. The general dynamics of demethylation and chromatin opening of enhancers during embryogenesis are therefore apparently conserved in zebrafish, *Xenopus* and mouse[19]. Consistent with these data, when quantifying the H3K27ac levels of lineage-defining enhancers in more-differentiated tissues (E10.5 midbrain, E12.5 intestine and E10.5 heart)[20,21], we observe that a substantial number of enhancers remain marked by H3K27ac (Extended Data Fig. 5). This indicates that the enhancers established at E7.5 are, to a large extent, maintained later in development.

In contrast to the mesoderm and endoderm enhancers, the ectoderm enhancers are open and demethylated as early as E4.5 in the epiblast (Fig. 3, Extended Data Fig. 9). Only in cells committed to mesendoderm fate do the ectoderm enhancers become partially repressed. Consistently, when measuring the accessibility dynamics at sites containing motifs for ectoderm-defining transcription factors (SOX2 and SP8), we find that these motifs are already accessible in the epiblast and lose accessibility specifically upon mesendoderm commitment. Conversely, motifs associated with endoderm- and mesoderm-defining transcription factors become accessible in their respective lineages only at E7.5 (Extended Data Fig. 9).

These observations can be explained by either priming of an ectodermal signature in the epiblast or the maintenance of a pluripotency signature in the ectoderm. To investigate this, we overlapped the E7.5 enhancer annotations with published H3K27ac ChIP–seq data from embryonic stem cells (ES cells) and E10.5 midbrain[21,22]. The E7.5 ectoderm enhancers display almost-exclusively pluripotent or neural signatures with notably different DNA methylation and chromatin accessibility dynamics (Extended Data Fig. 10). Pluripotency enhancers show an increase in methylation and a decrease in accessibility over time, suggesting a repression of these enhancers with similar dynamics to promoters of pluripotency genes (Fig. 1g, h). By contrast, neuroectoderm enhancers remain hypomethylated and accessible from E4.5 (Extended Data Fig. 10).

Finally, to infer temporal dependencies of enhancer activation, we used the RNA-expression profiles to order cells across two trajectories corresponding to mesoderm and endoderm commitment (Extended Data Fig. 11). By plotting the average DNA methylation and chromatin accessibility for each class of lineage-defining enhancer, we find that the methylation gain (and accessibility loss) of ectoderm enhancers precedes the demethylation (and accessibility gain) of mesoderm and

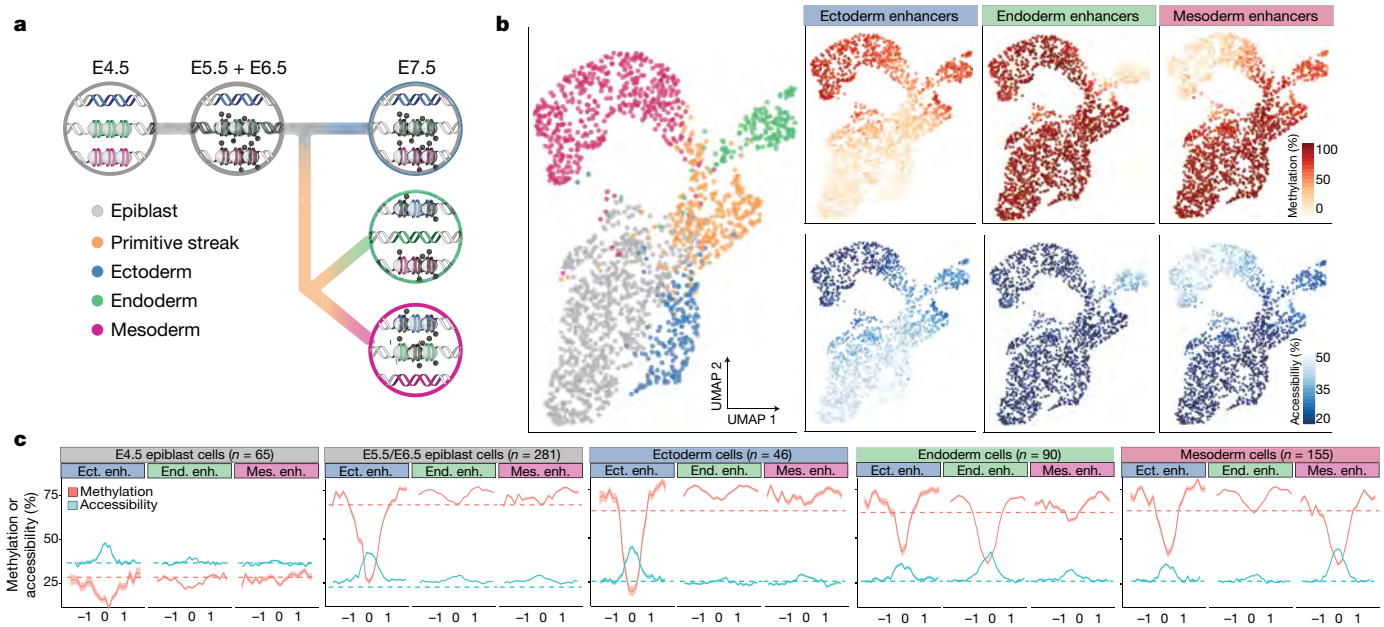


Fig. 3 | DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers across development. a, Illustration of the hierarchical model of enhancer epigenetic dynamics associated with germ-layer commitment. **b**, UMAP projection based on the MOFA factors inferred using all embryonic cells (n = 1,928). Main plot, cells are coloured by lineage. Smaller plots, cells are coloured by average methylation (top) or accessibility (bottom) at lineage-defining enhancers. For cells with RNA-expression data only, the MOFA factors were used to estimate the methylation and accessibility levels. **c**, Profiles of methylation (red) and accessibility (blue) at lineage-defining enhancers (enh.) (n = 3,918 for ectoderm, n = 1,930 for endoderm, n = 1,417 for mesoderm) across development. Running averages in 50-bp windows around the centre of the ChIP–seq peaks (2 kb upstream and downstream) are shown. Solid lines show the mean across cells and shaded areas represent the s.d. E5.5 and E6.5 epiblast cells show similar profiles and are combined. Dashed horizontal lines represent genome-wide background levels for methylation (red) and accessibility (blue).

endoderm enhancers. In both cases, changes in methylation and accessibility co-occur, suggesting tight co-regulation of the two epigenetic layers.

TET enzymes drive enhancer demethylation

TET methylcytosine dioxygenase enzymes have been implicated in enhancer demethylation[23,24], and loss-of-function experiments suggest that TET enzymes are vital for gastrulation[25,26]. To test whether TET enzymes drive lineage-specific demethylation, we differentiated both wild-type ES cells and ES cells deficient for all three TET enzymes (*Tet* TKO) into embryoid bodies and analysed the cells using scNMT-seq.

Mapping the RNA-expression profiles to the in vivo gastrulation atlas shows that wild-type embryoid bodies recapitulate the transition from a pluripotent epiblast at day 2 of differentiation to the primitive streak between days 4 and 5 (Fig. 4a, b). At days 6 and 7, we observe the emergence of mature mesoderm structures including haematopoietic cell types (Fig. 4a, b, Extended Data Fig. 12). Expression of marker genes is restricted to the expected lineage and differential expression between lineages agrees with the in vivo results (Extended Data Fig. 12). Moreover, the global dynamics of DNA methylation and chromatin accessibility in wild-type embryoid bodies substantially mirror the in vivo data (Extended Data Fig. 12).

Comparison of wild type with *Tet* TKO differentiation in the epiblast-like cells at day 2 revealed higher DNA methylation in ectoderm enhancers in the *Tet* TKO cells, but no differences in mesoderm or endoderm enhancers (Fig. 4c). Re-analysis of methylation measurements from *Tet* TKO embryos confirms that the same pattern is observed in vivo[25] (Extended Data Fig. 12). Impaired demethylation is also associated with differences in differentiation timing, with *Tet* TKO cells showing an increased proportion of early mesendoderm differentiation at day 4 to 5 (Fig. 4a, b). However, at day 6 to 7 *Tet* TKO cells do not properly demethylate lineage-specific enhancers and do not differentiate into mature mesodermal cell types (Fig. 4c).

These observations indicate that demethylation of lineage-defining enhancers is at least partially driven by TET proteins. Although enhancer demethylation does not seem to be required for early mesoderm commitment, the lack of haematopoietic cells in the *Tet* TKO cells suggests that demethylation may be important for subsequent lineage progression. Consistently, *Tet* TKO embryos are able to initiate gastrulation, but by E8.5 they display defects in mesoderm-derived cell types, including heart or somites[25].

Discussion

Our results show that pluripotent epiblast cells are epigenetically primed for an ectoderm fate as early as E4.5. This finding supports the existence of a 'default' path in Waddington's epigenetic landscape model, providing a potential mechanism for the phenomenon of 'default' differentiation of neurectodermal tissue from ES cells[27,28]. By contrast, endoderm and mesoderm are actively diverted from the default path by demethylation and chromatin opening at the corresponding enhancer elements[17,24,25]. Thus, the germ-layer epigenome is defined during gastrulation by a hierarchical, or asymmetric, epigenetic model (Fig. 3a).

More generally, these results have important implications for the role of the epigenome in defining lineage commitment. We speculate that asymmetric epigenetic priming—whereby early progenitors are epigenetically primed for a default cell type—may be a more general feature of lineage commitment in vivo. In support of this hypothesis, two recent studies have identified default pathways in foregut specification and osteogenesis[29,30]. Future studies that use multi-omics approaches to investigate cell populations have the potential to transform our understanding of cell-fate decisions, with important implications for stem cell biology.

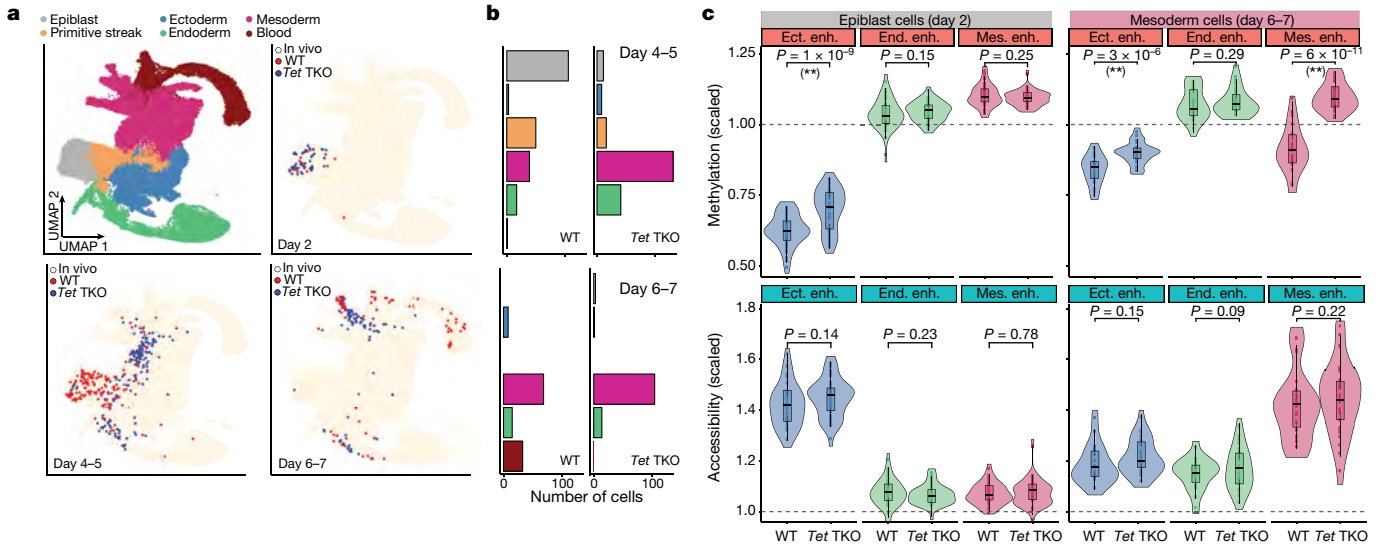


Fig. 4 | TET enzymes are required for efficient demethylation of mesoderm-defining enhancers and subsequent blood differentiation in embryoid bodies. a, UMAP projection of stages E6.5 to E8.5 of the atlas dataset (no extra-embryonic cells). Top left, cells coloured by lineage assignment. The remaining plots show, for different days of embryoid body differentiation, the nearest neighbours that were used to assign cell-type labels to the embryoid body dataset. Wild-type (WT) cells are red ($n = 438$), *Tet* TKO cells are blue ($n = 436$). We grouped days 4–5 and 6–7 together because of the similarity in the cell types recovered. **b**, Bar plots showing the numbers of each cell type for each day of embryoid body differentiation, grouped by genotype ($n = 438$ WT and 436 KO). **c**, Overlaid box and violin plots show the distribution of DNA methylation (top) or chromatin accessibility (bottom) for lineage-defining enhancers in epiblast-like cells at day 2 ($n = 46$ (WT) and $n = 44$ (*Tet* TKO)) and mesoderm-like cells at days 6–7 ($n = 22$ (WT) and $n = 32$ (*Tet* TKO)). The *y* axes show methylation or accessibility scaled to the genome-wide levels. Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. *P* values shown result from comparisons of group means (*t*-test). Asterisks denote significant differences (FDR <10%).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1825-8.

1. Peng, G. et al. Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. *Dev. Cell* **36**, 681–697 (2016).
2. Mohammed, H. et al. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).
3. Wen, J. et al. Single-cell analysis reveals lineage segregation in early post-implantation mouse embryos. *J. Biol. Chem.* **292**, 9840–9854 (2017).
4. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
5. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
6. Auclair, G., Guibert, S., Bender, A. & Weber, M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol.* **15**, 545 (2014).
7. Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710–719 (2014).
8. Zhang, Y. et al. Dynamic epigenomic landscapes during early lineage specification in mouse embryos. *Nat. Genet.* **50**, 96–105 (2018).
9. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
10. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
11. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
12. Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
13. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
14. Smith, Z. D. et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
15. Argelaguet, R. et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
16. Xiang, Y. et al. Epigenomic analysis of gastrulation reveals a unique chromatin state for primed pluripotency. *Nat. Genet.* https://doi.org/10.1038/s41588-019-0545-1 (2019).
17. Cusanovich, D. A. et al. The *cis*-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
18. Daugherty, A. C. et al. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Res.* **27**, 2096–2107 (2017).
19. Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
20. Kazakevych, J., Sayols, S., Messner, B., Krienke, C. & Soshnikova, N. Dynamic changes in chromatin states during specification and differentiation of adult intestinal stem cells. *Nucleic Acids Res.* **45**, 5770–5784 (2017).
21. Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
22. Kim, H. S. et al. Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells. *Nature* **556**, 510–514 (2018).
23. Rasmussen, K. D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. *Genes Dev.* **30**, 733–750 (2016).
24. Sardina, J. L. et al. Transcription factors drive Tet2-mediated enhancer demethylation to reprogram cell fate. *Cell Stem Cell* **23**, 727–741.e9 (2018).
25. Dai, H.-Q. et al. TET-mediated DNA demethylation controls gastrulation by regulating Lefty–Nodal signalling. *Nature* **538**, 528–532 (2016).
26. Li, X. et al. Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. *Proc. Natl Acad. Sci. USA* **113**, E8267–E8276 (2016).
27. Tropepe, V. et al. Direct neural fate specification from embryonic stem cells: a primitive mammalian neural stem cell stage acquired through a default mechanism. *Neuron* **30**, 65–78 (2001).
28. Muñoz-Sanjuán, I. & Brivanlou, A. H. Neural induction, the default model and embryonic stem cells. *Nat. Rev. Neurosci.* **3**, 271–280 (2002).
29. Rauch, A. et al. Osteogenesis depends on commissioning of a network of stem cell transcription factors that act as repressors of adipogenesis. *Nat. Genet.* **51**, 716–727 (2019).
30. Banerjee, K. K. et al. Enhancer, transcriptional, and cell fate plasticity precedes intestinal determination during endoderm development. *Genes Dev.* **32**, 1430–1442 (2018).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Embryos and single cell isolation

All mice used in this study were C57BL/6Babr and were bred and maintained in the Babraham Institute Biological Support Unit. Ambient temperature was about 19–21 °C and relative humidity was 52%. Lighting was provided on a 12 h:12 h light:dark cycle, including 15 min 'dawn' and 'dusk' periods of subdued lighting. After weaning, mice were transferred to individually ventilated cages with 1–5 mice per cage. Mice were fed CRM (P) VP diet (Special Diet Services) ad libitum and received seeds (for example, sunflower or millet) at the time of cage-cleaning as part of their environmental enrichment. All mouse experimentation was approved by the Babraham Institute Animal Welfare and Ethical Review Body. Animal husbandry and experimentation complied with existing European Union and United Kingdom Home Office legislation and local standards. Sample sizes were determined to obtain at least 50 cells for each germ layer. No randomization or blinding was performed. Sex of embryos was not known at the time of collection. Single-cells from E4.5 to E5.5 embryos were collected as previously described[2]. E6.5 and E7.5 embryos were dissected to remove extra-embryonic tissues and dissociated in TryplE for 10 min at room temperature. Undigested portions were physically removed and the remainder filtered through a 30-µm filter before isolation using flow cytometry.

Tet TKO cell culture

Tet1$^{-/-}$*Tet2*$^{-/-}$*Tet3*$^{-/-}$ (C57BL6/129/FVB) and matching wild-type mouse ES cells[31] were cultured in 2i+LIF medium (50/50 DMEM-F12 (Gibco, 31330-038) and Neurobasal medium (Gibco, 21103-49) with serum-free N2B27 (0.5% N2 and 1% B27; Gibco), 0.1 mM 2-mercaptoethanol (Life Technologies, 31350-010) and 2 mM L-glutamine (Life Technologies, 25030-024) supplemented with LIF, MEK inhibitor PD0325901 (1 µM) and GSK3 inhibitor CHIR99021 (3 µM), all from Department of Biochemistry, University of Cambridge). ES cells were cultured on tissue culture plastic pre-coated with 0.1% gelatine in H_2O and were passaged when approaching confluence (every 2–3 days).

For the embryoid body differentiation assay, 2×10^4 ES cells were collected in medium consisting of DMEM (Life Technologies, 10566-016), 15% fetal bovine serum (Gibco, 10270106), 1× non-essential amino acids (NEAA) (Life Technologies, 11140050), 0.1 mM 2-mercaptoethanol (Life Technologies, 31350-010), 2 mM L-glutamine (Life Technologies, 25030-024) in ultra-low attachment 96-well plates (Sigma-Aldrich, CLS7007). All cells were cultured in a humidified incubator at 37 °C in 5% CO_2 and 20% O_2. Embryoid bodies were collected 2, 4, 5, 6 and 7 days after induction of differentiation and dissociated into single cells using accutase before flow sorting. Cell lines were subject to routine mycoplasma testing using the MycoAlert testing kit (Lonza) and tested negative. Cell lines were not authenticated.

scNMT-seq library preparation

Single cells were flow-sorted (E6.5 and E7.5 stages, using a BD Influx or BD Aria III) or manually picked when cell numbers were too low (E4.5, E5.5). Cells were isolated into 96-well PCR plates containing 2.5µl of methylase reaction buffer (1 × M.CviPI Reaction buffer (NEB), 2 U M.CviPI (NEB), 160 µM S-adenosylmethionine (NEB), 1 U µl^{-1} RNasein (Promega), 0.1% IGEPAL CA-630 (Sigma)). Samples were incubated for 15 min at 37 °C to methylate accessible chromatin before the reaction was stopped with the addition of RLT plus buffer (Qiagen) and samples frozen down and stored at −80 °C before processing. Poly-A RNA was captured on oligo-dT conjugated to magnetic beads and amplified cDNA was prepared according to the G&T-seq[32] and Smartseq2 protocols[33]. The lysate containing gDNA was purified on AMPureXP beads before bisulfite-sequencing (BS-seq) libraries were prepared according to the scBS-seq protocol[34].

A subset of embryo cells were processed for scRNA-seq only (1,419 cells after QC). These followed the same protocol but we discarded the gDNA after separation.

A full step-by-step protocol for scNMT-seq is available at https://doi.org/10.17504/protocols.io.6jnhcme.

Sequencing

All sequencing was carried out on a NextSeq500 instrument. BS-seq libraries were sequenced in 48-plex pools using 75-bp paired-end reads in high-output mode. RNA-seq libraries were pooled as either 384 plexes and sequenced using 75-bp paired-end reads in high-output mode or 192 plexes and sequenced using 75-bp paired-end reads in mid-output mode. This yielded a mean raw sequencing depth of 8.5 million (BS-seq) and 1 million (RNA-seq) paired-end reads per cell.

RNA-seq alignment and quantification

RNA-seq libraries were aligned to the GRCm38 mouse genome build using HiSat2[35] (v.2.1.0) using options –dta–sp. 1000,1000–no-mixed–no-discordant, yielding a mean of 681,000 aligned reads per cell. Subsequently, gene expression counts were quantified from the mapped reads using featureCounts[36] with the Ensembl gene annotation[37] (v.87). Only protein-coding genes matching canonical chromosomes were considered. The read counts were log-transformed and size-factor adjusted[38].

BS-seq alignment and methylation/accessibility quantification

BS-seq libraries were aligned to the bisulfite converted GRCm38 mouse genome using Bismark[39] (v.0.19.1) in single-end nondirectional mode. Following the removal of PCR duplicates, we retained a mean of 1.6 million reads per cell. Methylation calling and separation of endogenous methylation (from A-C-G and T-C-G trinucleotides) and chromatin accessibility (G-C-A, G-C-C and G-C-T trinucleotides) was performed with Bismark using the–NOMe option of the coverage2cytosine script.

Following a previous approach[40], individual CpG or GpC sites in each cell were modelled using a binomial distribution in which the number of successes is the number of reads that support methylation and the number of trials is the total number of reads. A CpG methylation or GpC accessibility rate for each site and cell was calculated by maximum likelihood. The rates were subsequently rounded to the nearest integer (0 or 1).

When aggregating over genomic features, CpG methylation and GpC accessibility rates were computed assuming a binomial model, with the number of trials being the number of observed CpG sites and the number of successes being the number of methylated CpGs. Notably, this implies that DNA methylation and chromatin accessibility is quantified as a rate (or a percentage). We avoid binarizing DNA methylation and chromatin accessibility values into low and high states, as this is not a good representation of the continuous nature of the data (Extended Data Fig. 3).

ChIP–seq data processing

ChIP–seq data were obtained from the Gene Expression Omnibus accession code GSE125318. Reads were trimmed using Trim Galore (v.0.4.5, cutadapt 1.15, single end mode) and mapped to *Mus musculus* GRCm38 using Bowtie2[41] (v.2.3.2). Read 2 was excluded from the analysis for paired-end samples because of low-quality scores (Phred <25). All analyses were performed using SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). For quantification, read length was extended to 300 bp and regions of coverage outliers and extreme strand bias were excluded as these were assumed to be alignment artefacts. Comparison of datasets with different read lengths did not reveal major mapping differences, and thus mapped, extended reads were merged for samples that were sequenced across more than one lane.

Samples were similar overall regarding total mapped read numbers, distribution of reads and ChIP enrichment.

To best represent the underlying ChIP–seq signal, different methods to define enriched genomic regions were used for H3K4me3 and H3K27ac marks. For H3K4me3, a SeqMonk implementation of MACS[42] with the local rescoring step omitted was used ($P < 10^{-15}$, fragment size 300 bp), and enriched regions closer than 100 bp were merged. Peaks were called separately for each lineage. For H3K27ac, reads were quantitated per 500-bp tiles correcting per million total reads and excluding duplicate reads. Smoothing subtraction quantification was used to identify local maxima (value >1), and peaks closer than 500 bp apart were merged. Lineage-specific peak annotations exclude peaks that are also present in one of the other lineages, and only peaks present in both replicates were considered (Extended Data Fig. 5).

Publicly available ChIP–seq libraries for H3K27ac[20–22] were processed with Trim Galore and Bowtie2 (see above), and analysed in Seqmonk. Read counts were determined for 1-kb non-overlapping tiles and, separately, for lineage-specific enhancers (average length 1.2 kb). The genomic tiles were used to determine the distribution of H3K27ac across the genome. Enhancers were classified as marked if their read counts were within the top 5% of the distribution.

scRNA-seq and scBS-seq quality control

For RNA expression, cells with less than 100,000 mapped reads and with less than 500 expressed genes were excluded. For DNA methylation and chromatin accessibility, cells with less than 50,000 CpG sites and 500,000 GpC sites covered, respectively, were discarded (Extended Data Fig. 1).

Lineage assignment using RNA expression

Lineages were assigned by mapping the RNA-expression profiles to a comprehensive single-cell atlas from the same stages[4], when available (stages E6.5 and E7.5), or by SC3[43] otherwise (stages E4.5 and E5.5) (Extended Data Fig. 2). Extra-embryonic cells were identified by these methods and excluded from further analyses.

The mapping was performed by matching mutual nearest neighbours[44]. First, count matrices from both experiments were concatenated and normalized together. Highly variable genes were selected[38] from the resulting expression matrix and were used as input for principal components analysis. Subsequently, batch correction was applied to remove the technical variability between the two experiments and a *k*-nearest neighbours graph was computed between them. For each scNMT-seq cell, the cell type was selected as the mode from a Dirichlet distribution given by the cell type distribution of the top 30 nearest neighbours in the atlas (that is, majority voting).

Correlation analysis

To identify genes with an association between the mRNA expression and promoter epigenetic status, we calculated the correlation coefficient for each gene across all cells between the RNA expression and the corresponding DNA methylation or chromatin accessibility levels at the gene's promoter ±2 kb around the transcription start site (TSS).

As a filtering criterion, we required, for each genomic feature, a minimum number of 1 CpG (methylation) or 5 GpC (accessibility) measurements in at least 50 cells. Additionally, the top-5,000 most variable genes (across all cells) were selected, according to the rationale of independent filtering[45]. Two-tailed Student's *t*-tests were performed to test for evidence against the null hypothesis of no correlation, and *P* values were adjusted for multiple testing using the Benjamini–Hochberg procedure[46].

Differential DNA methylation and chromatin accessibility analysis

Differential analysis of DNA methylation and chromatin accessibility was performed using a Fisher exact test independently for each genomic element. Cells were aggregated into two exclusive groups and, for a given genomic element, we created a contingency table by aggregating (across cells) the number of methylated and unmethylated nucleotides. Multiple testing correction was applied using the Benjamini–Hochberg procedure. As a filtering criteria, we required 1 CpG (methylation) and 5 GpC (accessibility) observations in at least 10 cells per group. Non-variable regions were filtered out before differential testing.

Motif enrichment

To find transcription factor motifs enriched in lineage-associated sites, we used H3K27ac sites that were identified as differentially accessible between lineages as explained above. We tested for enrichment over a background of all H3K27ac sites using ame (meme suite[47] v.4.10.1) with parameters –method fisher –scoring avg. Position frequency matrices were downloaded from the Jaspar core vertebrates database[48]. This is a curated list of experimentally derived binding motifs and not an exhaustive set, which means that some important transcription factors will not be analysed, owing to absence of their motifs.

Differential RNA-expression analysis

Differential RNA-expression analysis between prespecified groups of interest was performed using the genewise negative binomial generalized linear model with quasi-likelihood test from edgeR[49]. Significant hits were called with a 1% FDR (Benjamini–Hochberg procedure) and a minimum $\log_2$ fold change of 1. Genes with low expression (mean $\log_2$ counts <0.5) were filtered out before differential testing[45].

Dimensionality reduction for DNA methylation and chromatin accessibility data using Bayesian factor analysis

To handle the large number of missing values in DNA methylation and chromatin accessibility data, we used a linear Bayesian factor analysis model[15]. The linearity assumption renders the model output directly interpretable, and more robust to changes in hyperparameters than nonlinear methods, particularly with small numbers of cells. We trained every model using the top-5,000 most variable features and we constrained the latent space to two latent factors, which were used for visualization (Fig. 1c, d, Extended Data Fig. 3). Variance-explained estimates were computed using the coefficient of determination as previously described[15].

MOFA

The input to MOFA is a list of matrices, in which each matrix represents a different data modality. RNA-expression measurements were defined as one data modality. For DNA methylation and chromatin accessibility, we defined separate matrices for promoters, distal H3K27ac sites (enhancers) and H3K4me3 (TSS). Promoters were defined as a bidirectional 2-kb window around the TSS of protein-coding genes. For each genomic context, we created a DNA methylation matrix and a chromatin accessibility matrix by quantifying *M*-values for each cell and genomic element.

As a filtering criterion, genomic features were required to have a minimum of 1 CpG (methylation) or 5 GpC (accessibility) observed in at least 25 cells. Genes were required to have a minimum cellular detection rate of 25%. In addition, to reduce computational complexity, the top 1,000 most variable features were selected per view. Similarly, the top 2,500 most variable genes were selected for RNA expression.

Similar to most latent dimensionality reduction methods, the optimization procedure of MOFA is not guaranteed to find a global optimum. Following ref. [15], model selection was performed by selecting the model with the highest evidence lower bound out of ten trials.

The number of factors was calculated by requiring a minimum of 1% variance explained in the RNA. The robustness of factors across trials was assessed by calculating the correlation coefficients between every

pair of factors across the ten trials. All inferred factors were consistently found in all model instances.

The downstream characterization of the model output included several analyses. (1) Variance decomposition: quantification of the fraction of variance explained (R^2) by each factor in each view, using a coefficient of determination[15]. (2) Visualization of weights/loadings: the model learns a weight for every feature in each factor, which can be interpreted as a measure of feature importance. Features with large weights (in absolute value) are highly correlated with the factor values. (3) Visualization of factors: each MOFA factor captures a different dimension of cellular heterogeneity. All together, they define a latent space that maximizes the variance explained in the data (under some important sparsity assumptions[15]). The cells can be visualized in the latent space by plotting scatter plots of combinations of factors. (4) Gene set enrichment analysis: when inspecting the weights for a given factor, multiple features can be combined into a gene set-based annotation. For a given gene set G, we evaluate its significance via a parametric t-test (two-sided), whereby we compare the mean of the weights of the foreground set (features that belong to the set G) with the mean of the weights in the background set (features that do not belong to the set G). Resulting P values are adjusted for multiple testing using the Benjamini–Hochberg procedure from which significant pathways are called (FDR <10%).

Reporting summary

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

Data availability

Raw sequencing data together with processed files (RNA counts, CpG methylation reports, GpC accessibility reports) are available in the Gene Expression Omnibus under accession number GSE121708. Processed data can be downloaded from ftp://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation.

Code availability

All code used for analysis is available at https://github.com/rargelaguet/scnmt_gastrulation.

31. Hu, X. et al. Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell* **14**, 512–522 (2014).
32. Macaulay, I. C. et al. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat. Protoc.* **11**, 2081–2103 (2016).
33. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
34. Clark, S. J. et al. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12**, 534–547 (2017).
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
36. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
37. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44** (D1), D710–D716 (2016).
38. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000 Res.* **5**, 2122 (2016).
39. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
40. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
42. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
44. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
45. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* **107**, 9546–9551 (2010).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).
47. McLeay, R. C. & Bailey, T. L. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
48. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46** (D1), D260–D266 (2018).
49. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
50. Ohnishi, Y. et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16**, 27–37 (2014).
51. Yeom, Y. I. et al. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development* **122**, 881–894 (1996).
52. Kalantry, S. et al. The amnionless gene, essential for mouse gastrulation, encodes a visceral-endoderm-specific protein with an extracellular cysteine-rich domain. *Nat. Genet.* **27**, 412–416 (2001).
53. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
54. Liang, G. et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl Acad. Sci. USA* **101**, 7357–7362 (2004).
55. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
56. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).

Acknowledgements R.A. is a member of Robinson College at the University of Cambridge. We thank K. Tabbada, C. Murnane and N. Forrester of the Babraham Next Generation Sequencing Facility for assistance with Illumina sequencing; members of the Babraham Flow Cytometry Core Facility for cell sorting and the Babraham Biological Support Unit for animal work; Y. Zhang for help in processing the ChIP–seq data. L.C.S. was supported by an EMBO postdoctoral fellowship (ALTF 417-2018) and is currently a Marie Sklodowska-Curie fellow funded by the European Commission under the H2020 Programme. J.C.M. is supported by core funding from EMBL and CRUK. R.A. is supported by the EMBL International Predoc Programme. X.I.-S. is supported by Wellcome Trust Grant 108438/E/15/Z. F.B. is supported by the UK Medical Research Council (Career Development Award MR/M01536X/1). B.G. and J.N. are supported by core funding by the MRC and Wellcome Trust to the Wellcome–MRC Cambridge Stem Cell Institute. W.R. is supported by Wellcome (105031/Z/14/Z; 210754/Z/18/Z) and BBSRC (BBS/E/B/000C0422). O.S. is supported by core funding from EMBL and DKFZ and the EU (ERC project DECODE 810296).

Author contributions H.M., W.D. and W.R. conceived the project. S.S. and H.M. designed the study and generated pilot data. W.D., J.N. and L.C.S. performed embryo dissections and single-cell isolation. L.C.S and T.L. performed in vitro differentiation experiments. S.J.C. and H.M. performed scNMT-seq library preparation. F.K. processed and managed sequencing data. C.K. analysed ChIP–seq datasets with assistance from Y.X. and C.W.H. R.A. and S.J.C. performed pre-processing and quality control of scNMT-seq data. R.A. and I.I.-R. mapped cells to the scRNA-seq atlas. R.A., S.J.C., F.B., L.C.S., X.I.-S., C.-A.K. and C.K. performed computational analysis. R.A. generated figures. R.A. S.J.C., L.C.S., O.S., J.C.M. and W.R. interpreted results and drafted the manuscript. G.S., P.J.R-G., W.X., G.K., O.S., B.G., J.C.M. and W.R. supervised the project. All authors read and approved the final manuscript.

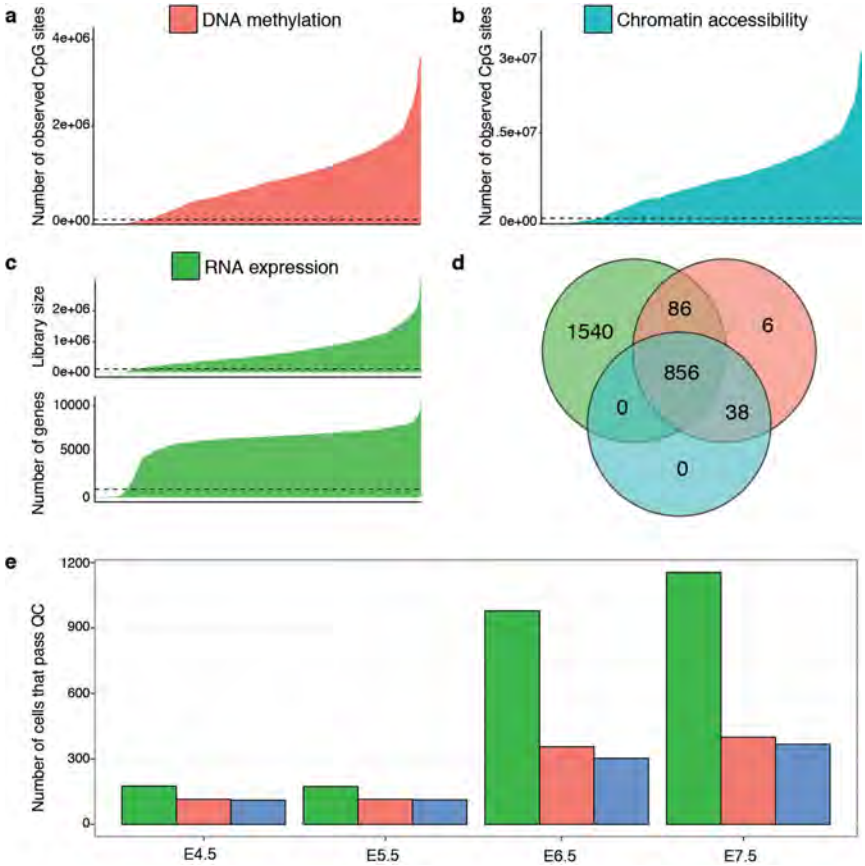
Competing interests W.R. is a consultant and shareholder of Cambridge Epigenetix. The remaining authors declare no competing interests.

Additional information

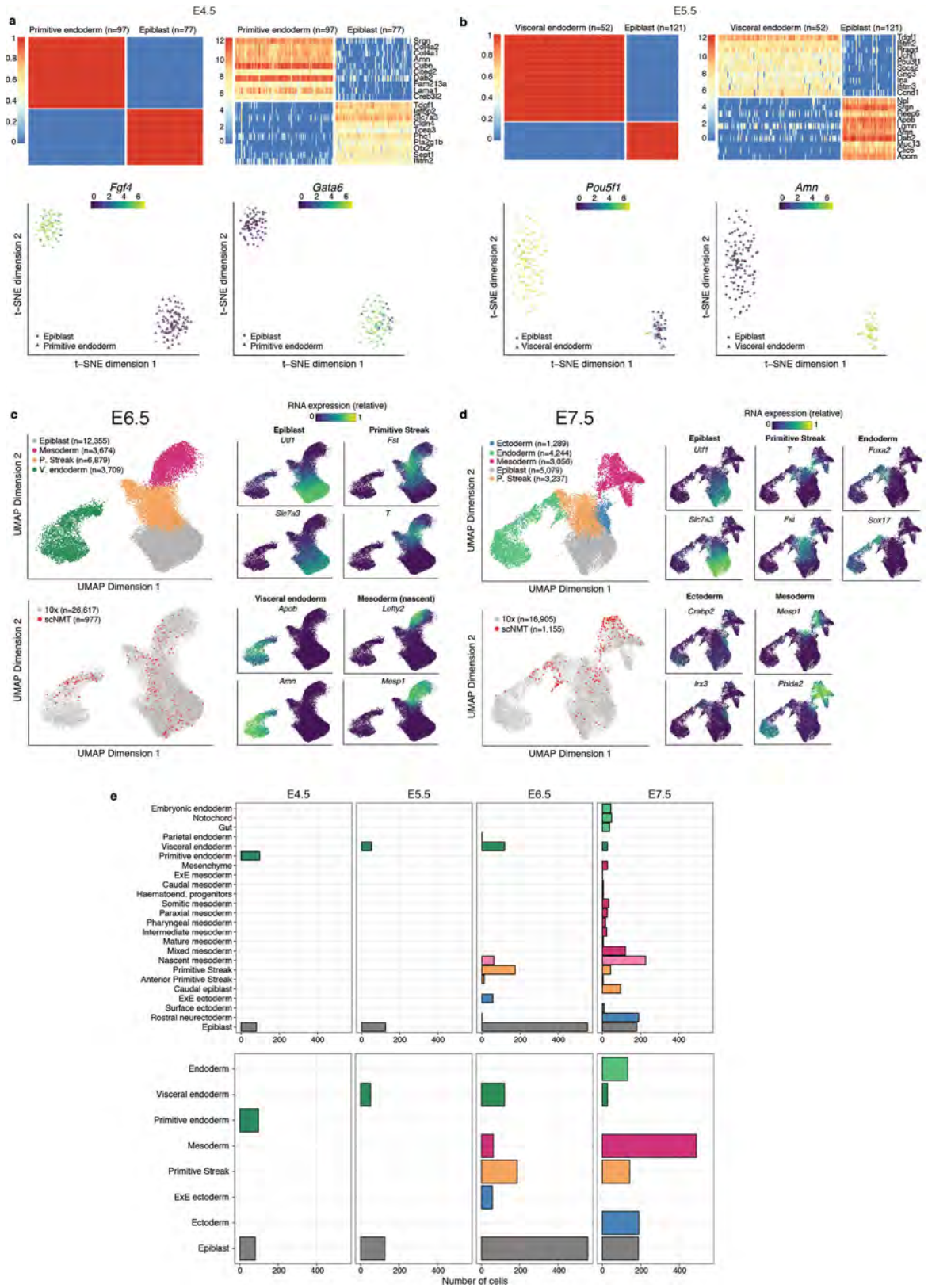
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1825-8.

Correspondence and requests for materials should be addressed to S.J.C., O.S., J.C.M. or W.R.

Peer review information *Nature* thanks Andrew Adey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

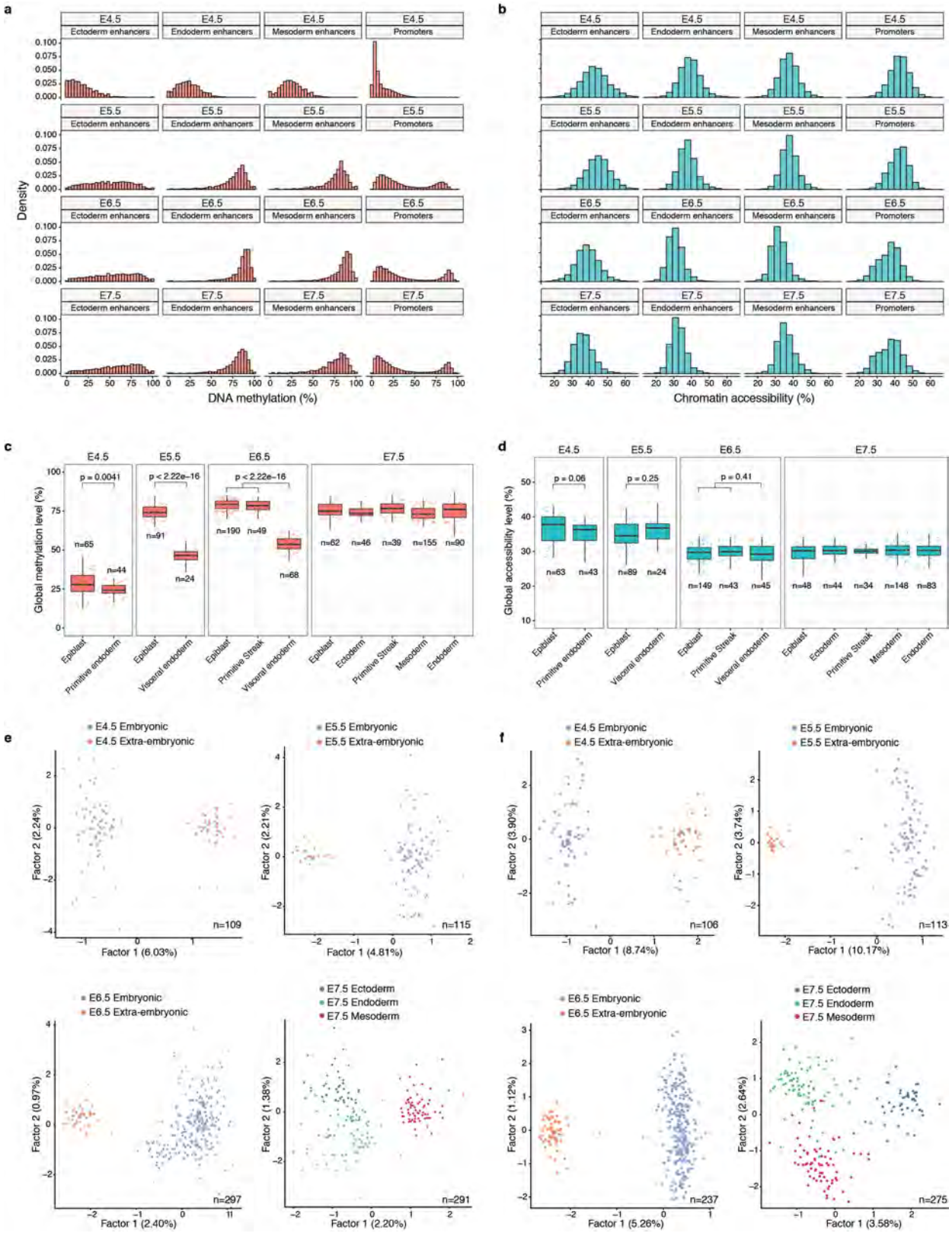


Extended Data Fig. 1 | scNMT-seq quality controls. a, **b**, Number of observed cytosines in CpG (red; **a**) or GpC (blue; **b**) contexts respectively. Each bar corresponds to one cell. Cells are sorted by total number of CpG or GpC sites. Cells below the dashed line were discarded on the basis of poor coverage (n = 1,105). **c**, RNA-library size per cell. Top, total number of reads. Bottom, number of expressed genes (read counts >0). Cells below the dashed line were discarded on the basis of poor coverage (n = 2,524). **d**, Venn diagram displaying the number of cells that pass quality control for RNA expression (green), DNA methylation (red) and chromatin accessibility (blue). **e**, Number of cells that pass quality control for each molecular layer, grouped by stage. For 1,419 out of 2,524 total cells, only the RNA expression was sequenced.



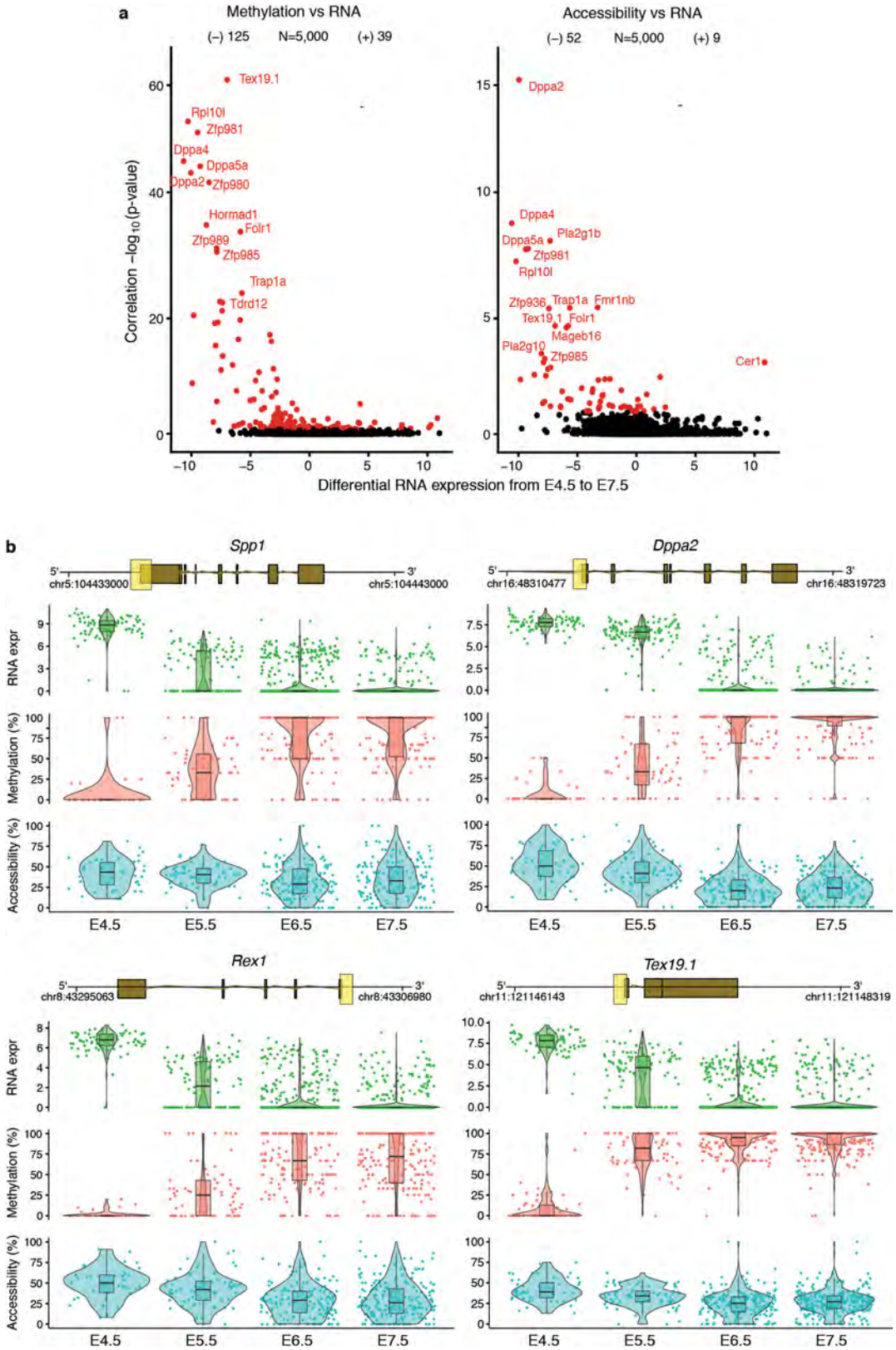
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Cell-type assignments based on RNA expression. a, **b**, Lineage assignment of E4.5 cells (**a**; $n = 175$) and E5.5 cells (**b**; $n = 173$). Top left, SC3 consensus plots representing the similarity between cells on the basis of averaging of clustering results from multiple combinations of clustering parameters. Top right, heat map showing the RNA expression (log normalized counts) of the ten most informative gene markers for each cluster. Bottom left, *t*-distributed stochastic neighbour embedding (*t*-SNE) representation of the RNA-expression data coloured by the expression of *Fgf4* and *Pou5f1*, known E4.5 and E5.5 epiblast markers[50,51], respectively. Bottom right, *t*-SNE representation of the RNA-expression data coloured by the expression of *Gata6* and *Amn*, known E4.5 primitive endoderm and E5.5 visceral endoderm markers[52]. **c**, **d**, Lineage assignment of E6.5 cells (**c**; $n = 977$) and E7.5 cells (**d**; $n = 1{,}155$). Left, UMAP projection of the atlas dataset (stages E6.5 to E7.0 to assign E6.5 cells and E7.0 to E8.0 to assign E7.5 cells). In the top-left panel, cells are coloured by lineage assignment. In the bottom-left panel, the cells coloured in red are the nearest neighbours that were used to transfer labels to the scNMT-seq dataset. In right panels, cells are coloured by the relative RNA expression of lineage-marker genes. **e**, Top, number of cells per lineage, using the maximally resolved cell types reported in ref. [4]. Bottom, number of cells per lineage after aggregation of cell types belonging to the same germ layer or extra-embryonic tissue type, as used in this study.

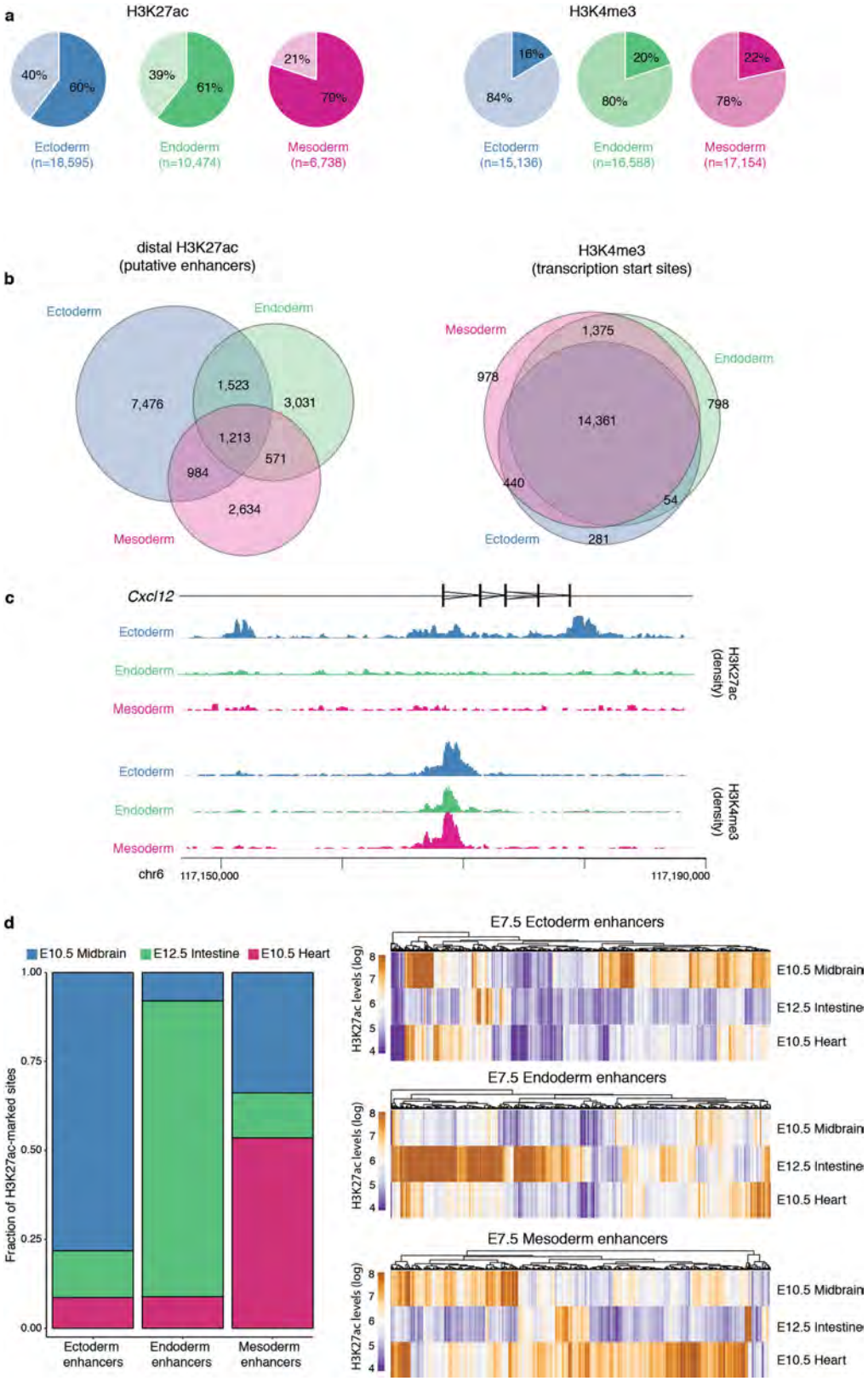


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Global methylation and chromatin accessibility dynamics. a, **b**, Distribution of DNA methylation (**a**) and chromatin accessibility levels (**b**) per stage and genomic context. When aggregating over genomic features, CpG methylation and GpC accessibility levels (%) are computed assuming a binomial model, with the number of trials being the total number of observed CpG (or GpC) sites and the number of successes being the number of methylated CpG (or GpC) sites (Methods). Notably, this implies that DNA methylation and chromatin accessibility are quantified as a percentage and are not binarized into low or high states. As this figure shows, the distribution of DNA methylation and chromatin accessibility across loci (after aggregating measurements across all cells per stage) is largely continuous and does not show bimodality. Hence, a binary approach similar to that sometimes used for differentiated cell types would not provide a good representation of the data. **c**, **d**, Box plots showing the distribution of genome-wide CpG methylation levels (**c**) or GpC accessibility levels (**d**) per stage and lineage. Each dot represents a single cell. Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. At a significance threshold of 0.01 (*t*-test, two-sided), the global DNA methylation levels differ between embryonic and extra-embryonic lineages, but the global chromatin accessibility levels do not. **e**, **f**, Dimensionality reduction of DNA methylation (**e**) and chromatin accessibility (**f**) data. To perform dimensionality reduction while handling the large amount of missing values, we used a Bayesian factor analysis model (Methods). Scatter plots of the first two latent factors (sorted by variance explained) for models trained with cells from the indicated stages are shown. From E4.5 to E6.5, cells are coloured by embryonic and extra-embryonic origin. At E7.5, cells are coloured by the primary germ layer. All lineage assignments were made using the cells' corresponding RNA-expression levels (Extended Data Fig. 2). The fraction of variance explained by each factor is displayed in parentheses. The input data were *M*-values quantified over DNase I hypersensitive sites profiled in ES cells ($n = 175,231$, subset to the top 5,000 most variable sites to fit the model).

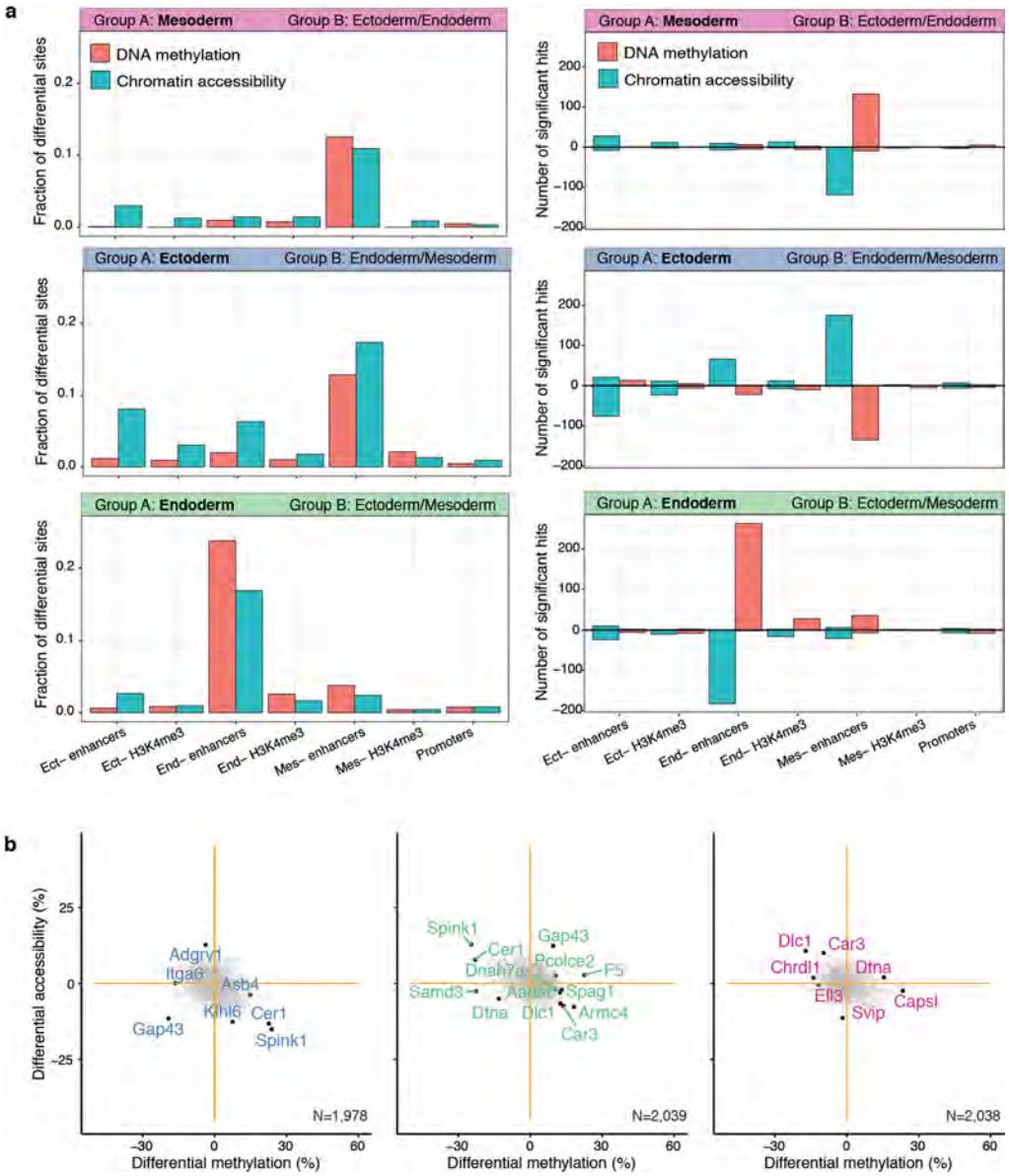


Extended Data Fig. 4 | DNA methylation and chromatin accessibility changes in promoters are associated with repression of early pluripotency and germ cell markers. a, Volcano plots display differential RNA-expression levels between E4.5 and E7.5 cells (in $\log_2$ counts, *x* axis) versus adjusted correlation *P* values (FDR <10% in red, Benjamini–Hochberg correction, $n = 5{,}000$ genes). Left, DNA methylation versus RNA-expression correlations; right, chromatin accessibility versus RNA expression. Negative values for differential RNA expression indicate higher expression in E4.5, whereas positive values indicate higher expression in E7.5. **b**, Illustrative examples of epigenetic repression of early pluripotency and germ cell markers. Box and violin plots show the distribution of RNA expression ($\log_2$ counts, green), DNA methylation (red) and chromatin accessibility (blue) levels per stage. Box plots show median coverage and the first and third quartile, whiskers show 1.5× the interquartile range. Each dot corresponds to one cell. For each gene a genomic track is shown on top, and the promoter region that is used to quantify DNA methylation and chromatin accessibility levels is highlighted in yellow.

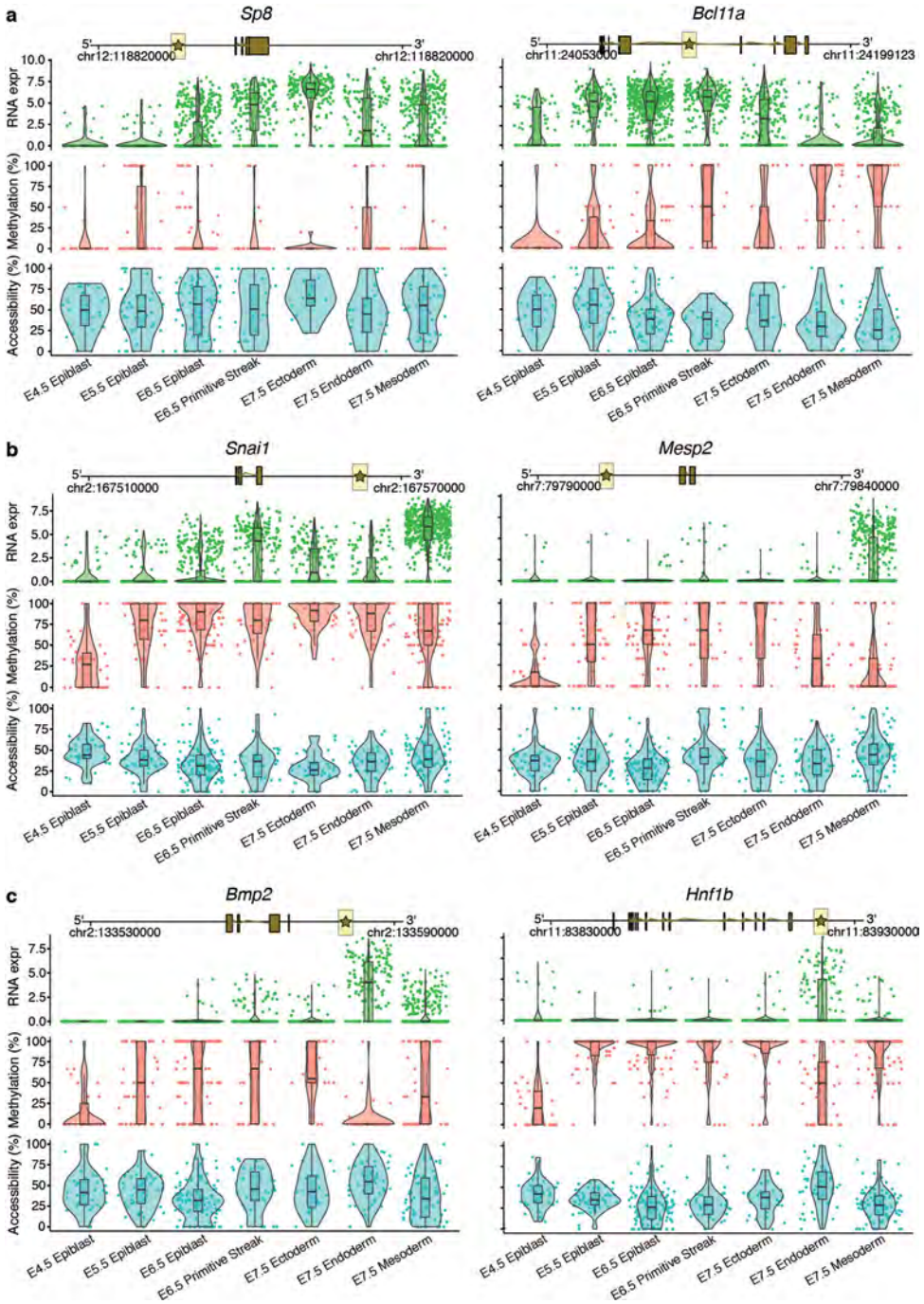


Extended Data Fig. 5 | See next page for caption.

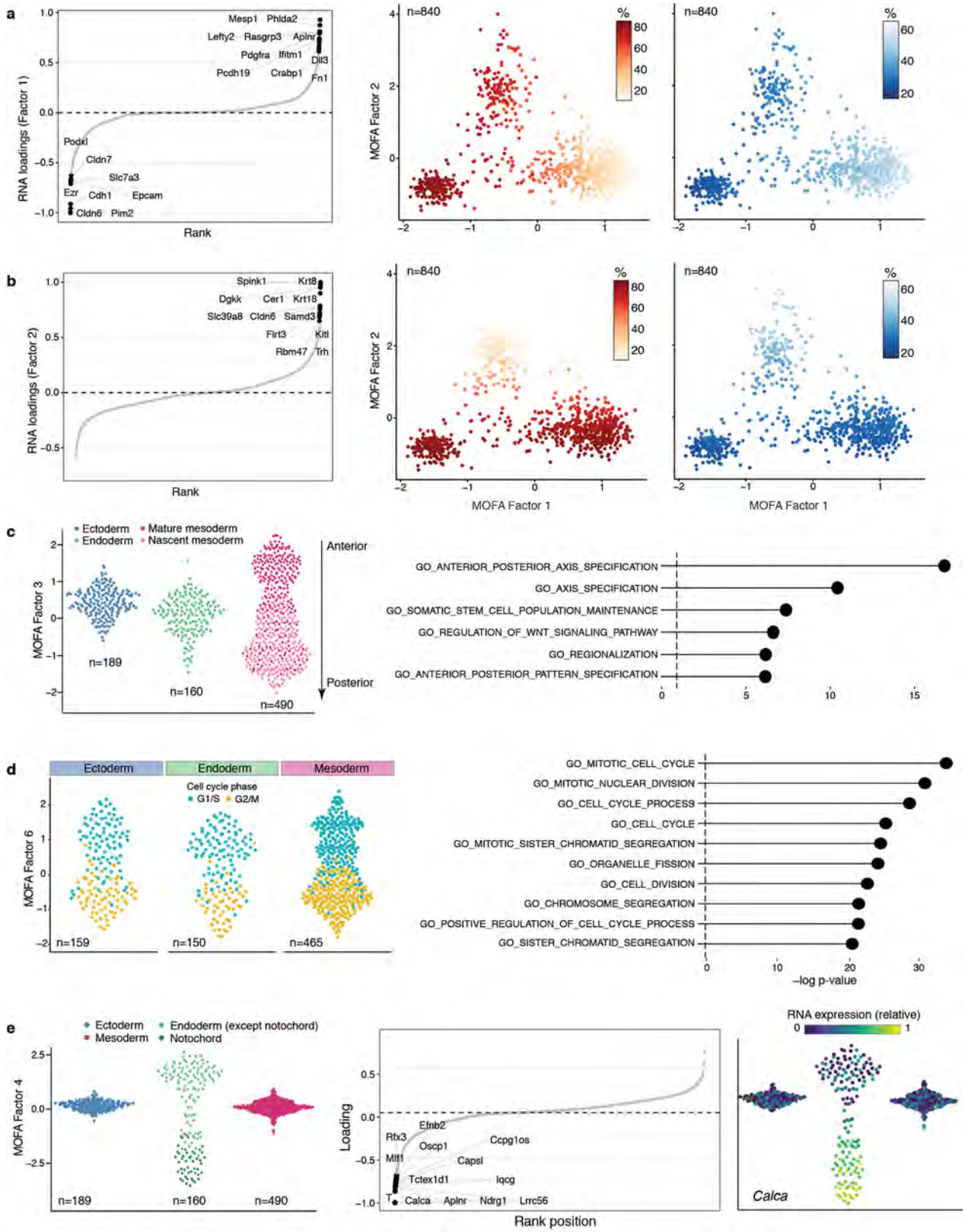
Extended Data Fig. 5 | Characterization of lineage-specific H3K27ac and H3K4me3 ChIP–seq data. a, Percentage of peaks overlapping promoters (±500 bp of TSS of annotated mRNAs (Ensembl v.87); lighter colour) and not overlapping promoters (distal peaks, darker colour). H3K27ac peaks tend to be distal from the promoters, marking putative enhancer elements[53]. H3K4me3 peaks tend to overlap promoter regions, marking TSS[54]. **b**, Venn diagrams showing overlap of peaks for each lineage, for distal H3K27ac (left) and H3K4me3 (right). This shows that H3K27ac peaks tend to be lineage-specific, whereas H3K4me3 peaks tend to be shared between lineages. **c**, Illustrative example of the ChIP–seq profile for the ectoderm marker *Cxcl12*. The top tracks show wiggle plots of ChIP–seq read density (normalized by total read count) for lineage-specific H3K27ac and H3K4me3. The coding sequence is shown in black. The bottom tracks show the lineage-specific peak calls (Methods). H3K27ac peaks are split into distal (putative enhancers) and proximal to the promoter. **d**, Left, bar plot of the fraction of E7.5 lineage-specific enhancers (n = 691 for ectoderm, 618 for endoderm and 340 for mesoderm) that are uniquely marked by H3K27ac in either E10.5 midbrain, E12.5 gut or E10.5 heart. Right, heat map displaying H3K27ac levels at individual lineage-specific enhancers (n = 2,039 for ectoderm, 1,124 for endoderm and 631 for mesoderm) in more differentiated tissues. E7.5 enhancers are predominantly marked in their differentiated-tissue counterparts (midbrain for ectoderm, gut for endoderm and heart for mesoderm).



Extended Data Fig. 6 | Differential DNA methylation and chromatin accessibility analysis at E7.5 for different genomic contexts. a, Bar plots showing the fraction (left) or the total number (right) of differentially methylated (red) or accessible (blue) loci (FDR <10%, *y* axis) per genomic context (*x* axis). Each subplot corresponds to the comparison of one cell type (group A) against cells comprising the other cell types present at E7.5 (group B). In the graphs on the right, positive values indicate an increase in DNA methylation or chromatin accessibility in group A, whereas negative values indicate a decrease in DNA methylation or chromatin accessibility. Differential analysis of DNA methylation and chromatin accessibility was performed independently for each genomic element using a two-sided Fisher's exact test of equal proportions (Methods). **b**, Scatter plots showing differential DNA methylation (*x* axis) versus chromatin accessibility (*y* axis) analysis at promoters. Ectoderm versus non-ectoderm cells (left), endoderm versus non-endoderm cells (middle) and mesoderm versus non-mesoderm cells (right) are shown. Each dot corresponds to a gene ($n = 2{,}038$). Labelled black dots highlight genes with lineage-specific RNA expression that show significant differential methylation or accessibility in their promoters (FDR <10%).

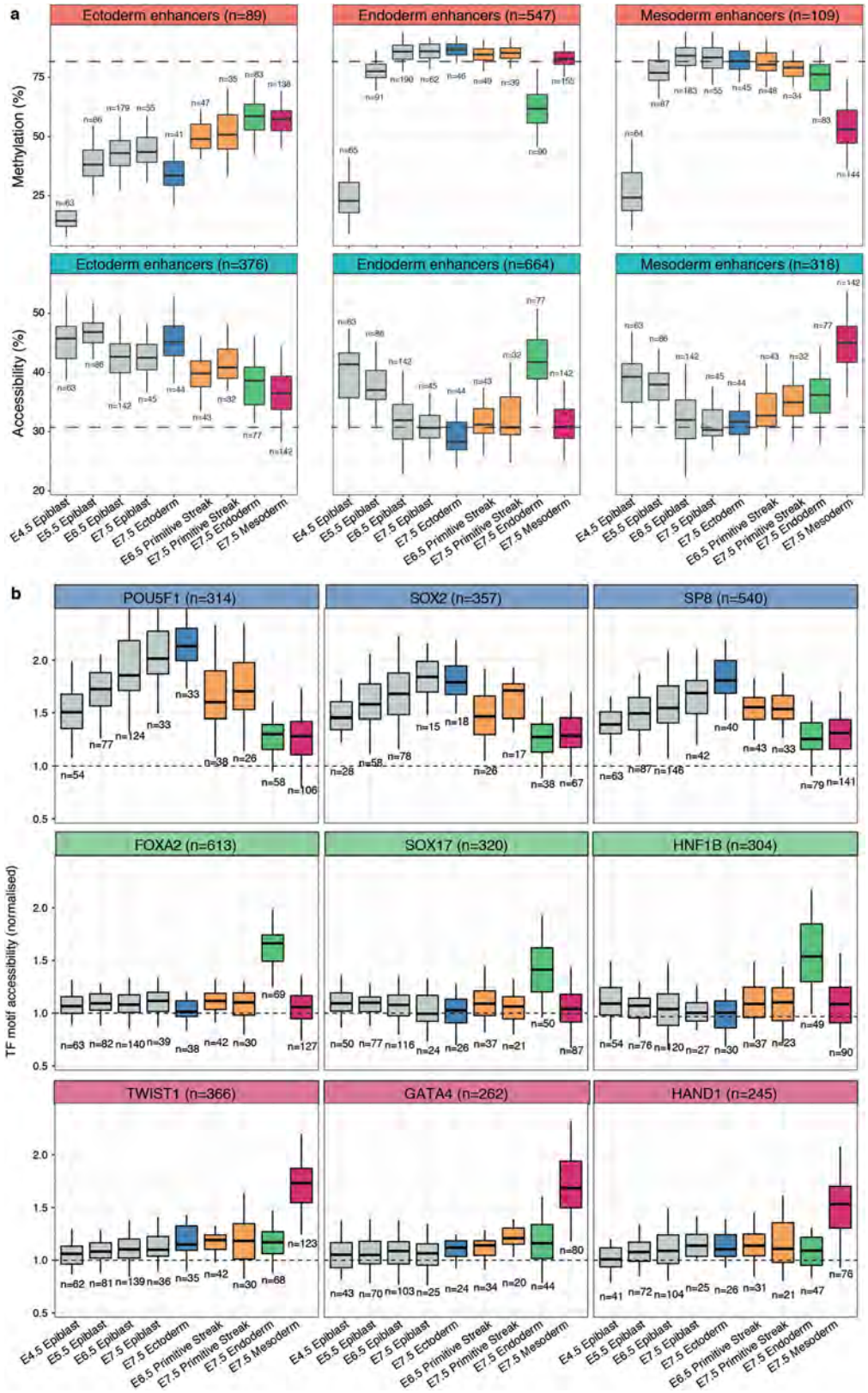


Extended Data Fig. 7 | Illustrative examples of putative epigenetic regulation in enhancer elements during germ-layer commitment. a–**c**, Box and violin plots showing the distribution of RNA expression (log_2 counts, green), enhancer DNA methylation (red) and chromatin accessibility (blue) levels for key germ-layer markers per stage and cell type. Marker genes for ectoderm (**a**), mesoderm (**b**) and endoderm (**c**) are shown. Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. Each dot corresponds to a single cell. For each gene, a genomic track is shown on the top. The enhancer region that is used to quantify DNA methylation and chromatin accessibility levels is represented with a star and highlighted in yellow. Genes were linked to putative enhancers by overlapping genomic coordinates with a maximum distance of 50 kb.

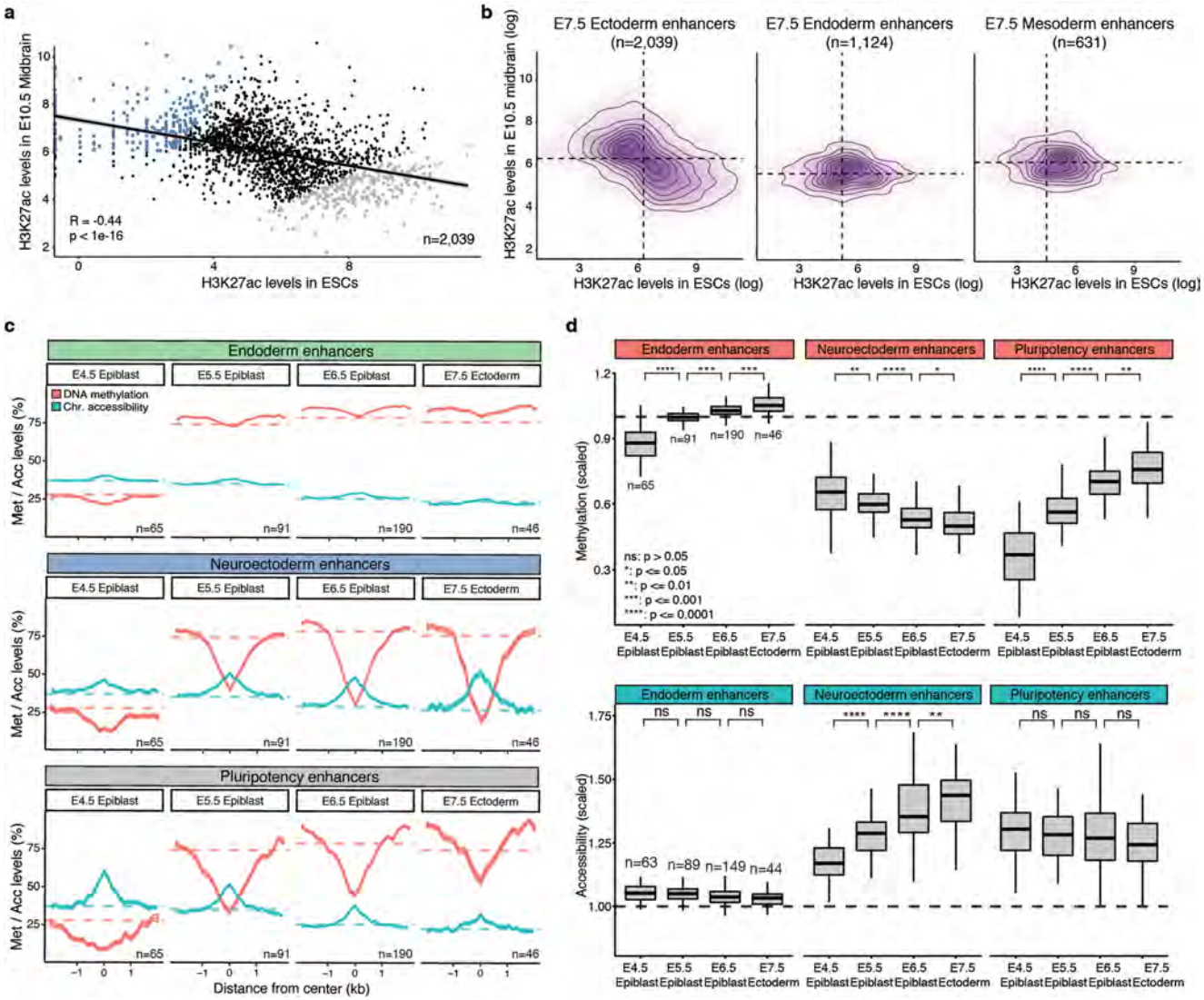


Extended Data Fig. 8 | See next page for caption.

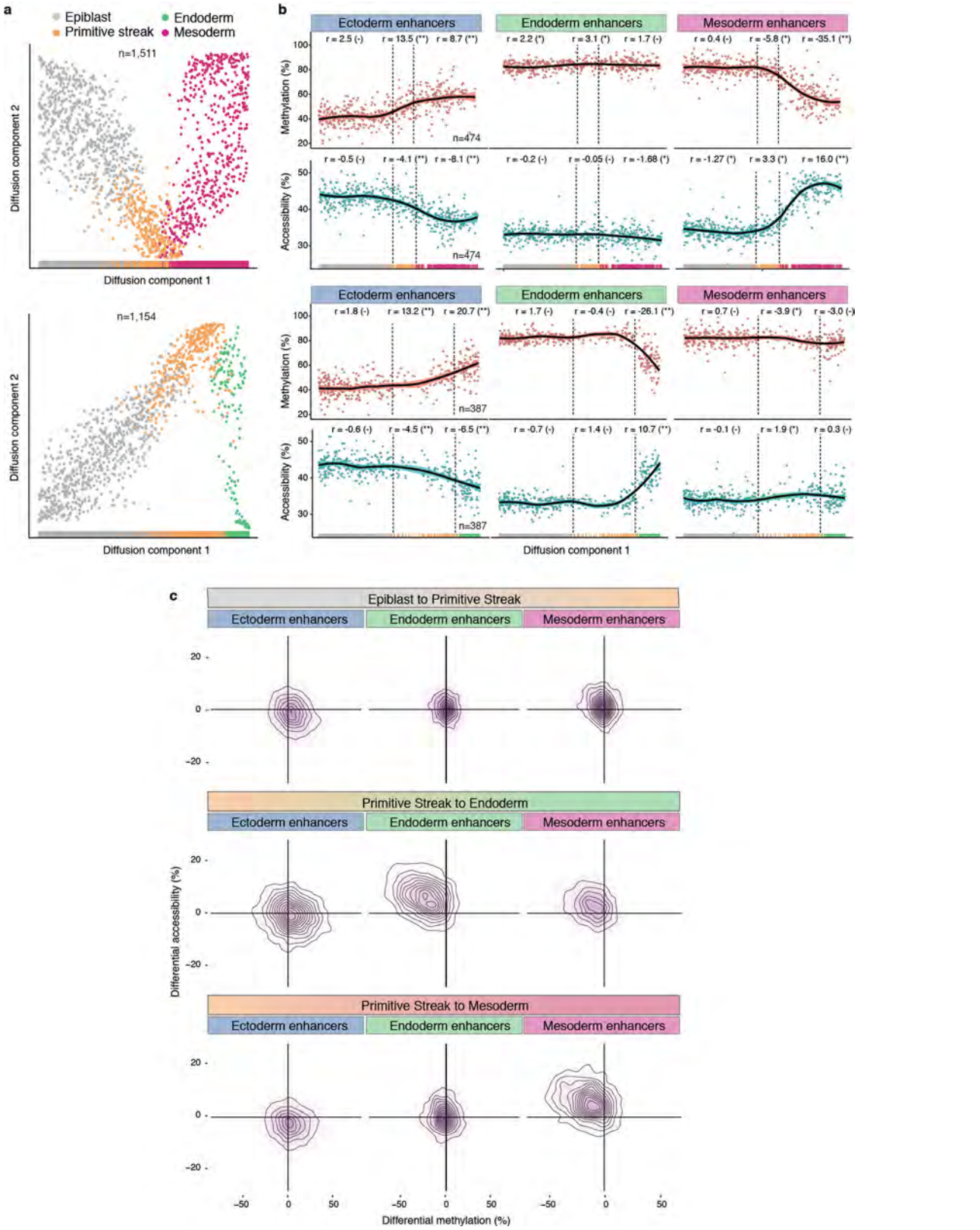
Extended Data Fig. 8 | Characterization of MOFA factors. a, Factor 1 as mesoderm commitment factor. Left, RNA-expression loadings for factor 1. Genes with large positive loadings increase expression in the positive factor values (mesoderm cells). Middle, scatter plot of factor 1 (*x* axis) and factor 2 (*y* axis) values. Each dot corresponds to a single cell, coloured by the average methylation levels of the top 100 enhancers with highest loading. Right, as the middle panel, except cells are coloured by the average accessibility levels. **b**, Factor 2 as the endoderm commitment factor. Left, RNA-expression loadings for factor 2. Genes with large positive loadings increase expression in the positive factor values (endoderm cells). Middle, scatter plot of factor 1 (*x* axis) and factor 2 (*y* axis) values. Each dot corresponds to a single cell, coloured by the average methylation levels (%) of the top 100 enhancers with highest loading. Right, as the middle panel, but cells are coloured by the average accessibility levels. **c**, Characterization of MOFA factor 3 as anteroposterior axial patterning and mesoderm maturation. Left, bee swarm plot of factor 3 values, grouped and coloured by cell type. The mesoderm cells are subclassified into nascent and mature mesoderm (Extended Data Fig. 2). Right, gene set enrichment analysis of the gene loadings of factor 3. The top most significant pathways from MSigDB C2[55] (Methods) are shown. **d**, Characterization of MOFA Factor 6 as cell cycle. Left, bee swarm plot of factor 6 values, grouped by cell type and coloured by inferred cell-cycle state using cyclone[56] (G1/2, cyan; G2/M, yellow). Right, gene set enrichment analysis of the gene loadings of factor 6. The top most significant pathways from MSigDB C2[55] are shown. **e**, Characterization of MOFA factor 4 as notochord formation. Left, bee swarm plot of factor 4 values, grouped and coloured by cell type. The endoderm cells are subclassified into notochord (dark green) and not notochord (green) (Extended Data Fig. 2). Middle, RNA-expression loadings for factor 4. Genes with large negative loadings increase expression in the negative factor values (notochord cells). Right, same bee swarm plots as in left but coloured by the relative RNA expression of *Calca* (gene with the highest loading).



Extended Data Fig. 9 | DNA methylation and chromatin accessibility dynamics of E7.5 lineage-specific enhancers and transcription factor motifs across development. a, Box plots showing the distribution of DNA methylation (top) or chromatin accessibility (bottom) levels of E7.5 lineage-defining enhancers, across stages and cell types. Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. The dashed lines represent the global background levels of DNA methylation at E7.5 (Extended Data Fig. 3). **b**, Box plots showing the distribution of chromatin accessibility levels (scaled to the genome-wide background) for 200-bp windows around transcription factor motifs associated with commitment to ectoderm (top), endoderm (middle) and mesoderm (bottom). Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range.

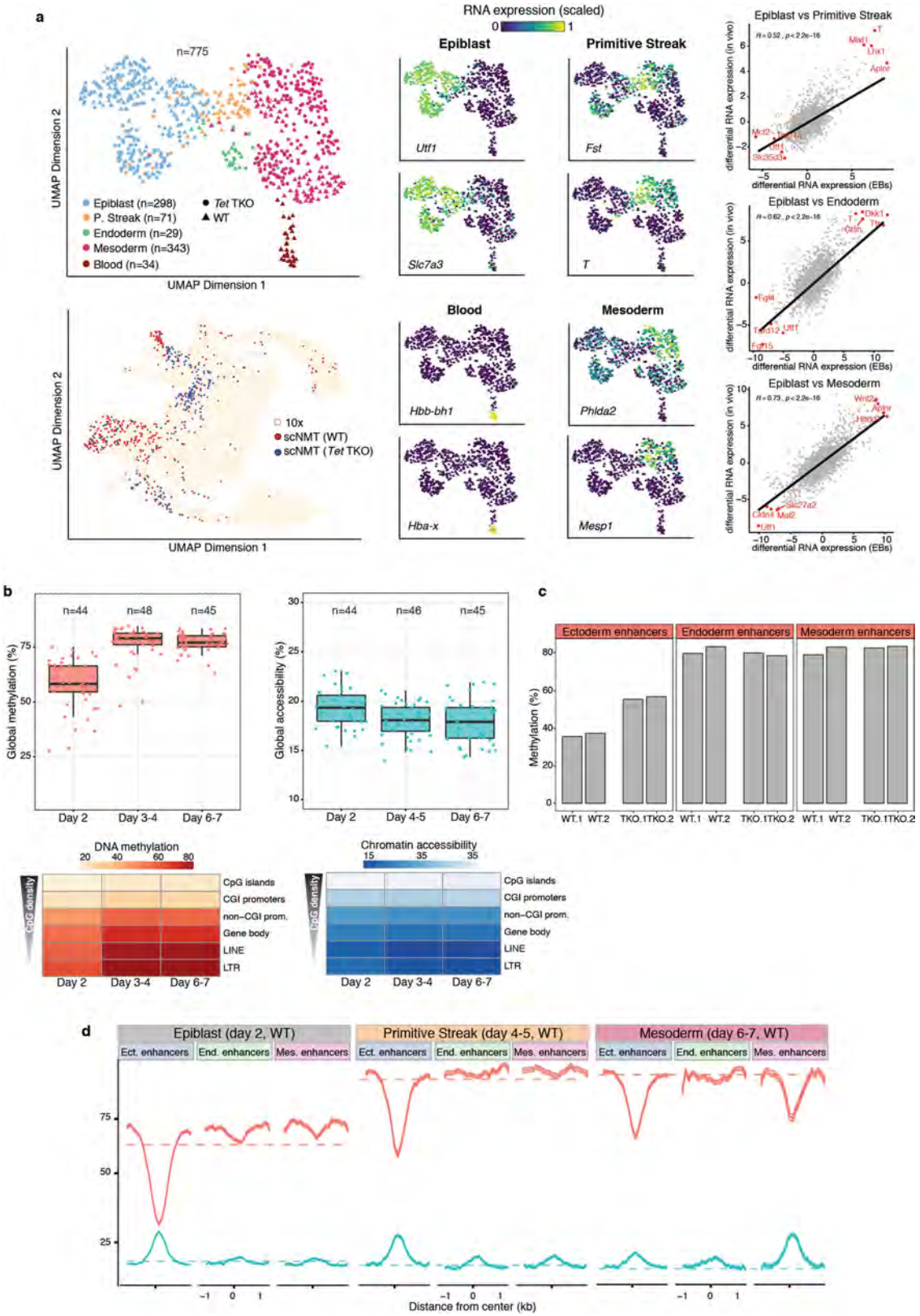


Extended Data Fig. 10 | E7.5 ectoderm enhancers contain a mixture of pluripotency and neural signatures with different epigenetic dynamics. a, Scatter plot showing H3K27ac levels for individual ectoderm enhancers ($n = 2,039$) quantified in serum-grown ES cells (pluripotency enhancers, x axis) versus E10.5 midbrain (neuroectoderm enhancers, y axis). H3K27ac levels in the two lineages are negatively correlated (Pearson's $R = -0.44$), indicating that most enhancers are either marked in ES cells or in the brain. The top 250 enhancers that show the strongest differential H3K27ac levels between midbrain and ES cells (blue for midbrain-specific enhancers and grey for ES cell-specific enhancers) are highlighted. **b**, Density plots of H3K27ac levels in ES cells versus E10.5 midbrain. H3K27ac levels are negatively correlated at E7.5 ectoderm enhancers, but not in E7.5 endoderm ($n = 1,124$) or mesoderm enhancers ($n = 631$). **c**, Profiles of DNA methylation (red) and chromatin accessibility (blue) along the epiblast–ectoderm trajectory. Panels show different genomic contexts: E7.5 ectoderm enhancers that are specifically marked by H3K27ac in the midbrain (middle) or ES cells (bottom) (highlighted populations in **a**). Running averages of 50-bp windows around the centre of the ChIP–seq peaks (2 kb upstream and downstream) are shown. Solid lines display the mean across cells (within a given lineage) and shading displays the s.d. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue). For comparison, we have also incorporated E7.5 endoderm enhancers (top), which follow the genome-wide repressive dynamics. **d**, Box plots of the distribution of DNA methylation (top) and chromatin accessibility (bottom) levels along the epiblast–ectoderm trajectory. Panels show different genomic contexts: E7.5 ectoderm enhancers that are specifically marked by H3K27ac in the midbrain (middle) or ES cells (right) (highlighted populations in **a**). Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. Dashed lines denote background DNA methylation and chromatin accessibility levels at the corresponding stage and lineage. For comparison, we have also incorporated E7.5 endoderm enhancers (left), which follow the genome-wide repressive dynamics.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Silencing of ectoderm enhancers precedes activation of mesoderm and endoderm enhancers. a, Reconstructed mesoderm (top) and endoderm (bottom) commitment trajectories using a diffusion pseudotime method applied to the RNA-expression data (Methods). Scatter plots of the first two diffusion components are shown, with cells coloured according to their lineage assignment (n = 1,154 for endoderm and n = 1,511 for mesoderm). For both cases, ranks along the first diffusion component are selected to order cells according to their differentiation state. **b**, DNA methylation (red) and chromatin accessibility (blue) dynamics of lineage-defining enhancers along the mesoderm (top) and endoderm (bottom) trajectories. Each dot denotes a single cell (n = 387 for endoderm and n = 474 for mesoderm) and black curves represent non-parametric locally estimated scatterplot smoothing regression estimates. In addition, for each scenario we fit a piecewise linear regression model for epiblast, primitive streak and mesoderm or endoderm cells (vertical lines indicate the discretized lineage transitions). For each model fit, the slope (r) and its significance level are displayed in the top (– for nonsignificant, 0.01<*P < 0.1 and **P < 0.01). **c**, Density plots showing differential DNA methylation (x axis) and chromatin accessibility (y axis) at lineage-defining enhancers calculated for each of the lineage transitions.



Extended Data Fig. 12 | See next page for caption.

Extended Data Fig. 12 | Embryoid bodies recapitulate the transcriptional, methylation and accessibility dynamics of the embryo. a, Embryoid bodies show high transcriptional similarity to gastrulation-stage embryos. Top left, UMAP projection of RNA expression for the embryoid body dataset ($n = 775$). Cells are coloured by lineage assignment and shaped by genotype (WT or *Tet* TKO). Bottom left, UMAP projection of stages E6.5 to E8.5 of the atlas dataset (no extra-embryonic cells) with the nearest neighbours that were used to assign cell type labels to the scNMT-seq embryoid body dataset coloured in red (WT) or blue (*Tet* TKO). Middle, UMAP projection of embryoid body cells coloured by the relative RNA expression of marker genes. Right, scatter plot of the differential gene expression ($\log_2$ normalized counts) between different assigned lineages for embryoid bodies (*x* axis) versus embryos (*y* axis). Each dot represents one gene. Pearson correlation coefficient with corresponding *P* value (two-sided) are displayed. Lines show the linear regression fit. The top-four genes with the largest differential expression are highlighted in red. **b**, Global DNA methylation and chromatin accessibility levels during embryoid body differentiation. Top, box plots showing the distribution of genome-wide CpG methylation (left) or GpC accessibility levels (right) per time point and lineage (compare with Extended Data Fig. 3). Each dot represents a single cell (only wild-type cells are used). Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. Bottom, heat map of DNA methylation (left) or chromatin accessibility (right) levels per time point and genomic context (compare with Fig. 1e, f). **c**, Ectoderm enhancers are more methylated in *Tet* TKO compared with wild-type epiblast cells in vivo. Bar plots show the mean (bulk) DNA methylation levels for ectoderm (left), endoderm (middle) and mesoderm (right) enhancers in E6.5 epiblast cells[25]. For each genotype, two replicates are shown. **d**, Profiles of DNA methylation (red) and chromatin accessibility (blue) at lineage-defining enhancers quantified over different lineages across embryoid body differentiation (only wild-type cells). Running averages in 50-bp windows around the centre of the ChIP–seq peaks (2 kb upstream and downstream) are shown. Solid lines display the mean across cells and shading displays the corresponding s.d. Dashed horizontal lines represent genome-wide background levels for methylation (red) and accessibility (blue).

Stephen J Clark
Wolf Reik
John C Marioni
Corresponding author(s): Oliver Stegle

Last updated by author(s): 18 October 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing was performed using an Illumina Nextseq500 instrument running NextSeq Control Software v4.0
Data analysis	All analysis code is available at https://github.com/rargelaguet/scnmt_gastrulation

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data together with processed files (RNA counts, CpG methylation reports, GpC accessibility reports) are available in the Gene Expression Omnibus under accession GSE121708. A link to the processed data is available in the GitHub project.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined in order to obtain at least 50 cells for each germ layer.
Data exclusions	Regions of coverage outliers and extreme strand bias excluded as these were assumed to be alignment artefacts.
Replication	For each developmental stage, we collected cells from at least 3 individual embryos and results were consistent across embryos.
Randomization	This is not relevant since we did not use different experimental groups or conditions in our study.
Blinding	This is not relevant since we did not use different experimental groups or conditions in our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Tet[l-/- ,2 -/- ,3 -/-] (C57BL6/129/FVB) and matching wild-type mouse ES cells (Hu, X. et al. Cell Stem Cell 2014)
Authentication	None
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination with the MycoAlert testing kit (Lonza).
Commonly misidentified lines (See ICLAC register)	None

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mus musculus, C57BL/6Babr. Embryos at 4.5 to 7.5 days post fertilization. Sex was unknown at the time of collection due to early embryonic stage.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not involve field-collected samples.
Ethics oversight	All mouse experiments were approved by the Babraham Institute Animal Welfare and Ethical Review Body.

Note that full information on the approval of the study protocol must also be provided in the manuscript.