Spring 2023 – Epigenetics and Systems Biology
Discussion Session (Epigenetics and Development)
Michael K. Skinner – Biol 476/576
Week 9 (March 9)

Epigenetics of Cell and Developmental Biology

Primary Papers

1. Schworer, et al., (2016) Nature 540:428. (PMID: 27919074)
2. Argelaguet, et al. (2019) Nature 576(7787):487-491. (PMID: 31827285)
3. Lyko F, et al., (2010) PLoS Biol. 2;8(11):e1000506. (PMID: 21072239)

Discussion

Student 22 – Ref #1 above

- What is the epigenetic aging effect observed?
- What stem cell effect was observed?
- How do epigenetics and genetics cooperate in this process?

Student 23 – Ref #2 above

- What was the experimental design to investigate gastrulation?
- What technology was used to examine epigenetics?
- What observations regarding gastrulation DNA methylation and transcriptome were made?

Student 24 – Ref #3 above

- What are the cast systems in the bee?
- How does epigenetics influence the development of the bee?
- What is the environmental factor that alters the epigenetic programming?

LETTER

doi:10.1038/nature20603

Epigenetic stress responses induce muscle stem-cell ageing by *Hoxa9* developmental signals

Simon Schwörer[1], Friedrich Becker[1], Christian Feller[2], Ali H. Baig[1], Ute Köber[1], Henriette Henze[1], Johann M. Kraus[3], Beibei Xin[4], André Lechel[5], Daniel B. Lipka[6], Christy S. Varghese[1], Manuel Schmidt[1], Remo Rohs[4], Ruedi Aebersold[2,7], Kay L. Medina[8], Hans A. Kestler[1,3], Francesco Neri[1], Julia von Maltzahn[1]§, Stefan Tümpel[1]§ & K. Lenhard Rudolph[1,9]§

The functionality of stem cells declines during ageing, and this decline contributes to ageing-associated impairments in tissue regeneration and function[1]. Alterations in developmental pathways have been associated with declines in stem-cell function during ageing[2–6], but the nature of this process remains poorly understood. Hox genes are key regulators of stem cells and tissue patterning during embryogenesis with an unknown role in ageing[7,8]. Here we show that the epigenetic stress response in muscle stem cells (also known as satellite cells) differs between aged and young mice. The alteration includes aberrant global and site-specific induction of active chromatin marks in activated satellite cells from aged mice, resulting in the specific induction of *Hoxa9* but not other Hox genes. *Hoxa9* in turn activates several developmental pathways and represents a decisive factor that separates satellite cell gene expression in aged mice from that in young mice. The activated pathways include most of the currently known inhibitors of satellite cell function in ageing muscle, including Wnt, TGFβ, JAK/STAT and senescence signalling[2–4,6]. Inhibition of aberrant chromatin activation or deletion of *Hoxa9* improves satellite cell function and muscle regeneration in aged mice, whereas overexpression of *Hoxa9* mimics ageing-associated defects in satellite cells from young mice, which can be rescued by the inhibition of Hoxa9-targeted developmental pathways. Together, these data delineate an altered epigenetic stress response in activated satellite cells from aged mice, which limits satellite cell function and muscle regeneration by Hoxa9-dependent activation of developmental pathways.

Age-dependent declines in the number and function of Pax7$^+$ satellite cells (SCs) impair the regenerative capacity of skeletal muscle[2,4,9]. Genes and pathways that contribute to this process[2–6] often also have a role in regulating embryonic development[10–13]. Despite these parallels, the function of the master regulators of development, Hox genes, has not been determined in SC ageing. An analysis of freshly isolated, *in vivo* activated SCs from young adult and aged mice (Extended Data Fig. 1a–e) revealed a specific upregulation of *Hoxa9* in SCs from aged mice, both at the mRNA (Fig. 1a, Extended Data Fig. 2a, b) and protein level (Fig. 1b, c). Similar results were obtained by immunofluorescence staining of SCs (Extended Data Fig. 2c) and myofibre-associated SCs (Fig. 1d, e, Extended Data Fig. 2d) that were activated in culture (Extended Data Fig. 1f, g).

Ageing reduces the proliferative and self-renewal capacity of SCs in wild-type mice[2,9,14,15] (*Hoxa9*$^{+/+}$; Extended Data Fig. 3). Homozygous deletion of *Hoxa9* (*Hoxa9*$^{-/-}$) did not affect the colony-forming capacity of SCs from young adult mice but ameliorated ageing-associated impairment in colony formation of single-cell-sorted SCs in culture (Fig. 2a). *Hoxa9* deletion also increased the self-renewal of myofibre-associated SCs from aged mice in culture but had no effect on SCs from young adult mice under these conditions (Extended Data Fig. 4a–c). Similar results were obtained by short interfering RNA (siRNA)-mediated knockdown of *Hoxa9* in myofibre-associated SC cultures derived from aged mice (Extended Data Fig. 4d–h). The number of SCs decreases in resting tibialis anterior muscle of ageing wild-type mice[2,4,9]; this phenotype was not affected by *Hoxa9* gene status (Extended Data Fig. 5a). However, homozygous deletion or siRNA-mediated knockdown of *Hoxa9* increased the total number of Pax7$^+$ SCs (Fig. 2b, Extended Data Fig. 5b–e) and improved myofibre regeneration

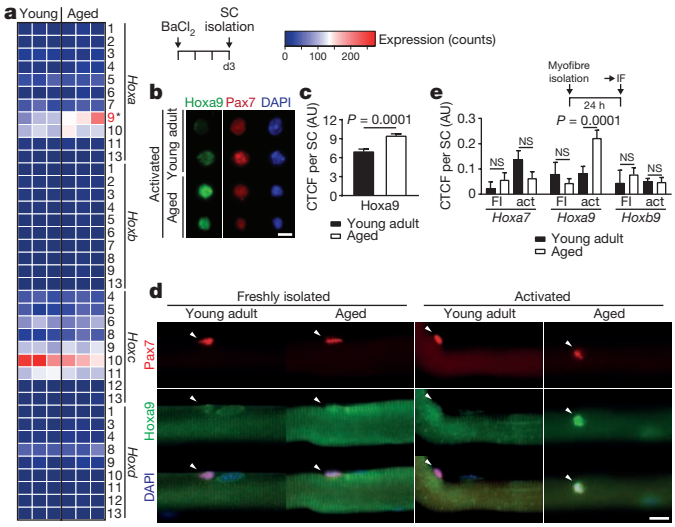


Figure 1 | Upregulation of *Hoxa9* in aged activated SCs. a–c, Analysis of freshly isolated, *in vivo* activated SCs (3 days after muscle injury with $BaCl_2$) from young adult and aged mice. **a**, Heatmap showing the mRNA expression of all *Hox* genes as determined by RNA-sequencing analysis. **b**, Representative immunofluorescence staining for Hoxa9 and Pax7. Nuclei were counterstained with 4′,6-diamidino-2-phenylindole (DAPI). **c**, Corrected total cell fluorescence (CTCF) for Hoxa9 per SC as shown in **b**. AU, arbitrary units. **d**, **e**, Immunofluorescence (IF) staining for Hoxa9 and Pax7 in myofibre-associated SCs that were quiescent (freshly isolated (FI) myofibres) or activated (act; 24 h culture of myofibres). **d**, Representative images with arrowheads denoting Pax7$^+$ cells. **e**, CTCF for indicated *Hox* genes. Note the specific induction of *Hoxa9* in activated SCs isolated from aged mice. Scale bars, 5 μm (**b**) and 20 μm (**d**). *P* values were calculated by two-sided Mann–Whitney *U*-test (**c**) or two-way analysis of variance (ANOVA) (**e**). NS, not significant. $n = 3$ mice in **a**; $n = 134$ nuclei (young), $n = 181$ nuclei (aged) from 3 mice in **c**; $n = 12/13/17/56$ nuclei (*Hoxa7*), $n = 9/42/102/62$ nuclei (*Hoxa9*), $n = 7/35/34/25$ nuclei (*Hoxb9*) from 2 young and 4 aged mice in **e**.

[1]Leibniz-Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena, Germany. [2]Department of Biology, Institute of Molecular Systems Biology, ETH Zürich, Auguste-Piccard-Hof 1, 8093 Zürich, Switzerland. [3]Institute of Medical Systems Biology, Ulm University, James-Franck-Ring, 89081 Ulm, Germany. [4]Molecular and Computational Biology Program, University of Southern California, 1050 Childs Way, Los Angeles, California 90089, USA. [5]Department of Internal Medicine I, Ulm University, Albert-Einstein-Allee 23, 89081 Ulm, Germany. [6]Division of Epigenomics and Cancer Risk Factors, DKFZ, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. [7]Faculty of Science, University of Zürich, Zürich, Switzerland. [8]Department of Immunology, Mayo Clinic, 200 First Street SW, Rochester, Minnesotta 55905, USA. [9]Faculty of Medicine, Friedrich-Schiller-University, Jena, Germany.
§These authors jointly supervised this work.

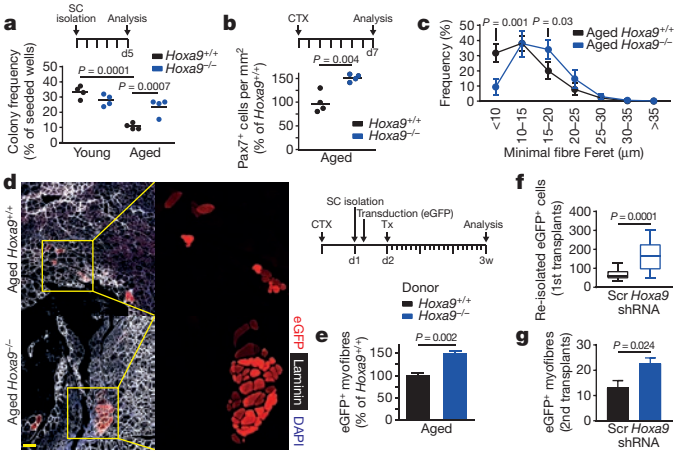


Figure 2 | *Hoxa9* deficiency improves muscle regeneration in aged mice. **a**, Frequency of myogenic colonies derived from single-cell-sorted SCs from young adult or aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice after 5 days (d5) of culture. **b**, **c**, Quantification of Pax7^{+} cells per area (**b**) and frequency distribution of minimal Feret's diameter (**c**) of tibialis anterior muscle fibres from aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice, 7 days after muscle injury with cardiotoxin (CTX). **d**, **e**, Transplantation (Tx) of enhanced green fluorescent protein (eGFP)-labelled SCs from aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice. **d**, Representative immunofluorescence staining for eGFP, laminin and DAPI in engrafted tibialis anterior muscles. Scale bar, 50 μm. **e**, Quantification of donor-derived (eGFP^{+}) myofibres in **d**. **f**, Quantification of donor-derived (eGFP^{+}) SCs re-isolated from primary recipients. Scr, scrambled control shRNA. **g**, Quantification of donor-derived (eGFP^{+}) myofibres from secondary recipients. Data in **f** represent median with 50% confidence interval box and 95% confidence interval whiskers. *P* values were calculated by two-way ANOVA (**a**, **c**), two-sided Student's *t*-test (**b**, **e**, **g**), or two-sided Mann–Whitney *U*-test (**f**). $n = 4$ mice in **a**; $n = 4$ mice in **b**, **c**; $n = 8$ recipient mice in **e**; $n = 20$ recipient mice in **f**; $n = 5$ recipient mice in **g**.

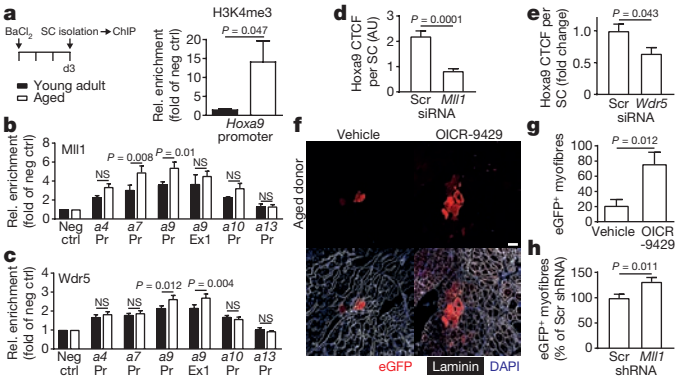


Figure 3 | Mll1 complex-dependent chromatin modification induces *Hoxa9* and limits muscle regeneration in aged mice. **a**–**c**, ChIP–quantitative PCR (qPCR) analysis of the indicated *Hox* promoters (Pr) and exons (Ex) in activated SCs from young adult and aged mice using antibodies against H3K4me3 (**a**), Mll1 (**b**), or Wdr5 (**c**). **d**, **e**, CTCF for Hoxa9 per SC after *Mll1* siRNA (**d**) or *Wdr5* siRNA transfection (**e**) of freshly isolated myofibre-associated SCs from aged mice. **f**–**h**, Transplantation of eGFP-labelled SCs from aged mice. **f**, Representative immunofluorescence staining for eGFP, laminin and DAPI in engrafted tibialis anterior muscles after transplantation of OICR-9429 treated SCs. Scale bar, 50 μm. **g**, **h**, Quantification of donor-derived (eGFP^{+}) myofibres after transplantation of OICR-9429-treated (**g**) or shRNA-treated (**h**) SCs. *P* values were calculated by two-way ANOVA (**b**, **c**), two-sided Student's *t*-test (**a**, **g**, **h**) or two-sided Mann–Whitney *U*-test (**d**, **e**). $n = 6$ mice in **a**; $n = 7$ mice (young), $n = 10$ mice (aged) in **b**, **c**; $n = 109$ nuclei (Scr siRNA), $n = 110$ nuclei (*Mll1* siRNA) from 3 mice in **d**; $n = 116$ nuclei (Scr siRNA), $n = 65$ nuclei (*Wdr5* siRNA) from 3 mice in **e**; $n = 5$ recipient mice in **g**; $n = 6$ recipient mice in **h**.

in injured muscle of aged mice almost to the levels in young adult mice (Fig. 2c, Extended Data Fig. 5f), albeit without affecting overall SC proliferation rates seven days after muscle injury (Extended Data Fig. 5g, h). *Hoxa9* gene deletion also improved the cell-autonomous, *in vivo* regenerative capacity of transplanted SCs derived from aged donor mice but did not affect the capacity of SCs derived from young adult donors (Fig. 2d, e, Extended Data Fig. 6a). Similarly, *Hoxa9* downregulation by short hairpin RNA (shRNA) infection rescued the regenerative capacity and the engraftment of transplanted SCs derived from aged mice almost to the level of SCs from young adult mice (Extended Data Fig. 6b–h). When transduced at similar infection efficiency (Extended Data Fig. 6i), *Hoxa9* shRNA compared to scrambled shRNA improved the self-renewal of serially transplanted SCs from aged mice in primary recipients (Fig. 2f, Extended Data Fig. 6j) as well as the regenerative capacity of 500 re-isolated SCs from primary donors that were transplanted for a second round into the injured tibialis anterior muscle of secondary recipients (Fig. 2g, Extended Data Fig. 6k). Together, these results demonstrate that the induction of *Hoxa9* limits SC self-renewal and muscle regeneration in aged mice, and that the deletion of *Hoxa9* is sufficient to revert these ageing-associated deficiencies.

The expression of *Hoxa9* in development and leukaemia is actively maintained by Mll1-dependent tri-methylation at lysine 4 of histone 3 (H3K4me3)[16–18]. Chromatin immunoprecipitation (ChIP) revealed that H3K4me3 is strongly enriched at the promoter and first exon of *Hoxa9* in activated SCs from aged compared to young adult mice, which was not detected to the same extent for other *Hoxa* genes (Fig. 3a, Extended Data Fig. 7a). ChIP analyses for Mll1 and Wdr5 (a scaffold protein of the Mll1 complex) revealed increased recruitment of these factors to the *Hoxa* cluster with Wdr5 enrichment being confined to the *Hoxa9* locus (Fig. 3b, c). Although no changes were observed for Mll1, both H3K4me3 and Wdr5 showed significantly increased levels in nuclei of myofibre-associated SCs from aged versus young adult mice upon activation (Extended Data Fig. 7b–e). Of note, knockdown of either *Mll1* (also known as *Kmt2a*) or *Wdr5* reduced H3K4me3 levels as well as Mll1 recruitment to the *Hoxa9* locus and ameliorated Hoxa9 induction in activated myofibre-associated SCs from aged mice (Fig. 3d, e, Extended Data Fig. 7f–i). Similar results were obtained by treatment of aged myofibre-associated SCs with OICR-9429, an inhibitor of the Mll1–Wdr5 interaction[19] (Extended Data Fig. 7j, k). Moreover, both *Mll1* knockdown and OICR-9429 treatment increased the self-renewal and lowered the myogenic commitment of myofibre-associated SCs from aged mice (Extended Data Fig. 7l–q), resulting in increased SC numbers in cultures of purified SCs or myofibre-associated SCs derived from aged mice (Extended Data Fig. 7r, s). Notably, Mll1 inhibition by either stable shRNA knockdown (Extended Data Fig. 7t) or OICR-9429 treatment improved the regenerative capacity of SCs from aged mice when transplanted into injured muscle of recipient mice (Fig. 3f–h). Taken together, these experiments demonstrate that the Mll1 complex contributes to *Hoxa9* induction in activated SCs from aged mice, resulting in impairment in SC function and muscle regeneration. Pax7 expression was downregulated in activated SCs of aged mice (Extended Data Fig. 7u–w) and did not correlate with Hoxa9 expression (Extended Data Fig. 7x, y), indicating that Mll1-dependent regulation of Pax7 target genes[20] was not involved in the Mll1-dependent induction of *Hoxa9* in activated SCs from aged mice.

Next, a global analysis of histone post-translational modifications was carried out on freshly isolated SCs obtained before muscle injury (quiescent state) or two, three and five days after *in vivo* SC activation mediated by muscle injury (Fig. 4a, b, Extended Data Fig. 8a). Using a recently developed mass-spectrometry-based proteomic strategy[21], 46 histone H3 and H4 lysine acetylation and methylation motifs were quantified. Quiescent SCs from aged mice compared to young adult mice showed increased levels of repressive marks (H3K9me2 and

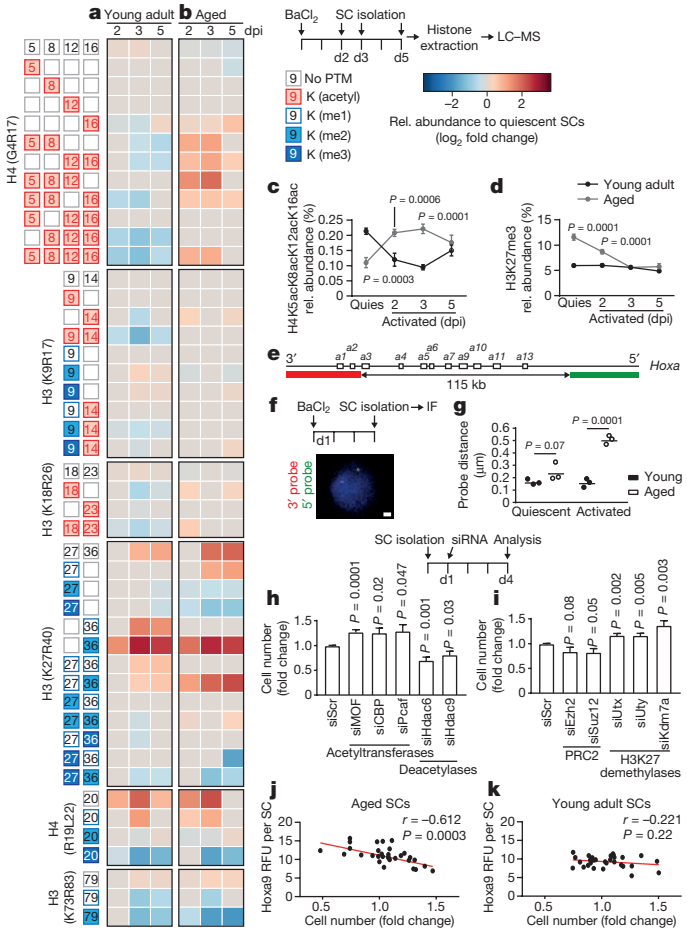


Figure 4 | Altered epigenetic stress response in aged SCs. a, **b**, Heatmap of mass spectrometry (LC–MS) analysis displaying significant ($P < 0.05$) relative changes in abundance of the indicated histone modifications (measured at the indicated peptides) at the indicated days post injury (dpi). **c**, **d**, Trajectory time-course plots showing relative abundance of H4K5acK8acK12acK16ac (**c**) or H3K27me3 (**d**) in freshly isolated quiescent (quies) or *in vivo* activated SCs purified at indicated time points post muscle injury. **e**–**g**, Fluorescence *in situ* hybridization of freshly isolated quiescent or *in vivo* activated SCs with the indicated probes spanning the *Hoxa* cluster (**e**); an exemplary image (**f**); and the average probe distance (**g**). Scale bar, 1 μm. **h**, **i**, Relative changes in SC number 4 days after transfection of freshly isolated SCs from aged mice with the indicated siRNAs. **j**, **k**, Pearson correlation of relative cell number and Hoxa9 immunofluorescence signal of SCs from young adult and aged mice 4 days after transfection with a selection of siRNAs targeting different classes of chromatin modifiers. RFU, relative fluorescence units. *P* values were calculated by two-way ANOVA (**c**, **d**, **g**), two-sided Student's *t*-test (**a**, **b**, **h**, **i**), or Pearson correlation (**j**, **k**). $n = 4$ mice in **a**–**d**; $n = 3$ mice with 50 nuclei per replicate in **g**; $n = 7$ mice (*Ezh2* siRNA), 8 mice (all others) in **h**, **i**; $n = 6$ mice (aged), $n = 3$ mice (young) in **j**, **k**.

H3K27me3; Extended Data Fig. 8a; consistent with ref. 22), and lower amounts of histone modifications typically enriched on active genes (for example, various H4 acetylation motifs, H3K14ac, H3K18ac and H3K36me2; Extended Data Fig. 8a). A time-dependent shift towards a heterochromatic state occurred during SC activation in young adult mice, whereas activation in aged SCs generated the opposite response (Fig. 4a, b). Although selective active marks such as H3 and H4 acetylation motifs declined in SCs from young adult mice during activation, there was a substantial increase in these marks in aged SCs (Fig. 4a–c). Conversely, repressive marks (for example, H3K27me3) decreased in SCs from aged mice but remained stable in SCs from young adult mice during activation (Fig. 4a, b, d). The observed shift of the chromatin towards a more permissive state after SC activation appeared to also affect the *Hoxa* cluster as this locus displayed an increased chromatin decompaction after SC activation in aged mice but not in young adult mice (Fig. 4e–g).

To analyse the functional contribution of different types of chromatin modifications in activated SCs from aged mice, a set of genetic and pharmacological experiments was conducted. The expression of key enzymes involved in chromatin modifications detected by RNA-sequencing analysis was similar in activated SCs from young adult and aged mice (Extended Data Fig. 8b). However, knockdown of the acetyltransferases *MOF* (also known as *Kat8*), *CBP* (*Crebbp*) or *Pcaf* (*Kat2b*) improved the proliferative capacity of SCs from aged mice in bulk culture, whereas knockdown of histone deacetylases led to a reduction (Fig. 4h). Furthermore, knockdown of the H3K27 demethylases *Utx* (also known as *Kdm6a*), *Uty* or *Kdm7a* promoted the proliferation of aged SCs, which was instead inhibited by knockdown of *Suz12* and *Ezh2* (Fig. 4i), members of the PRC2 protein complex responsible for H3K27me3. Multi-acetylation motifs, as observed in activated SC from aged mice (Fig. 4b, c), are preferred binding sites for bromodomain-containing proteins[23]. Eight out of eleven non-toxic bromodomain inhibitors available from the Structural Genomics Consortium exhibited positive effects on the proliferative capacity of SCs from aged mice (Extended Data Fig. 8c, d, $P = 4.2 \times 10^{-4}$). Targeting major classes of chromatin modifiers by a selected set of siRNAs (Supplementary Table 1) revealed a significant inverse correlation ($r = -0.612$) between siRNA-mediated changes in Hoxa9 protein expression and the proliferative capacity of SCs from aged mice, with no such effects observed in SCs from young adult mice (Fig. 4j, k). Similarly, siRNAs against *MOF* and *Utx* as well as bromodomain inhibitors led to significant decreases in the Hoxa9 protein level in activated myofibre-associated SCs from aged mice (Extended Data Fig. 8e–g). In summary, activated SCs from aged mice exhibit site-specific and global aberrations in the epigenetic stress response, resulting in *Hoxa9* activation and profound negative effects on SC function, which are ameliorated by targeting the respective enzymes underlying these alterations.

By analysing the downstream effects of *Hoxa9* induction through lentiviral-mediated *Hoxa9* overexpression, we found a strong reduction in the colony forming and proliferative capacity of SCs from young adult mice (Extended Data Fig. 9a–c). The overexpression of other Hox genes exerted similar effects (Extended Data Fig. 9d) but the *Hoxa9* results are probably most relevant for physiological ageing because only *Hoxa9* was upregulated in activated SCs from aged mice (Fig. 1). The impaired myogenic capacity of SCs in response to *Hoxa9* overexpression was associated with increased rates of apoptosis and decreased cell proliferation (Extended Data Fig. 9e–h). Furthermore, *Hoxa9* induction associated with the suppression of several cell cycle regulators and induction of cell cycle inhibitors and senescence-inducing genes (Extended Data Fig. 9i) as well as with increased staining for senescence-associated β-galactosidase (Extended Data Fig. 9j, k). Microarray expression analysis of *Hoxa9*-overexpressing SCs compared to controls revealed that among the top 12 pathways regulated by *Hoxa9* were several major developmental pathways that have previously been shown to impair SC function and muscle regeneration in the context of ageing[2,3,5,6,9,24,25] (Fig. 5a, Extended Data Fig. 9l–o). ChIP analysis of putative Hoxa9-binding sites (Supplementary Table 1) in *Hoxa9*-overexpressing primary myoblasts indicated that a high number of these genes are probably direct targets of Hoxa9 (Extended Data Fig. 9p; cumulative *P* value over tested genes: $P = 1 \times 10^{-7}$). *Hoxa9* strongly induced downstream targets of the Wnt, TGFβ and JAK/STAT pathways, but targeted activation of each one of these pathways alone only led to slight changes in the expression of target genes of the other two pathways (Extended Data Fig. 9q–s), suggesting that *Hoxa9* acts as a central hub required for the parallel induction of these pathways in aged SCs. Of note, the inhibition of *Stat3*, *Bmp4* or *Ctnnb1* (encoding β-catenin) by shRNAs as well as pharmacological inhibition of the Wnt, TGFβ or JAK/STAT pathway was sufficient to improve the myogenic colony forming capacity of SCs overexpressing *Hoxa9* (Fig. 5b,

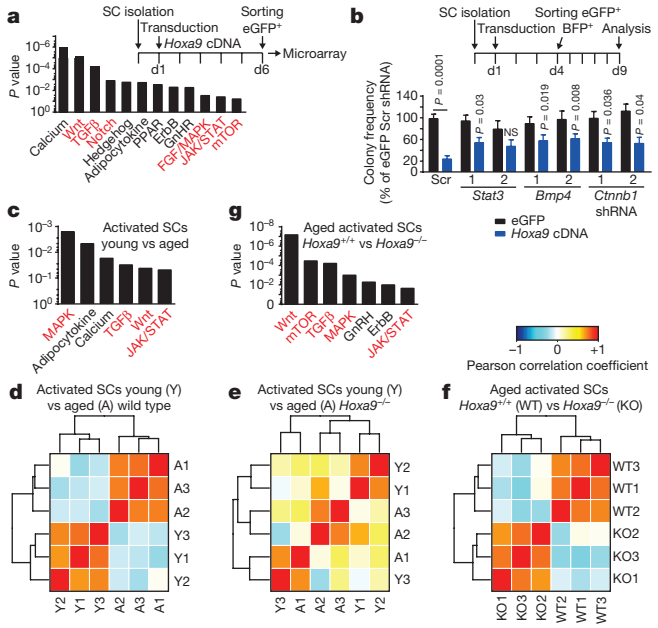


Figure 5 | Activation of *Hoxa9* induces developmental pathways. **a**, KEGG pathway analysis of differentially expressed genes (DEGs) of SCs overexpressing *Hoxa9* compared to eGFP. Red-highlighted pathways were previously shown to impair the function of SCs in aged mice. **b**, Colony formation of single-cell-sorted SCs derived from young adult mice that were co-infected with a *Hoxa9* cDNA and the indicated shRNAs; comparison to *Hoxa9*/scrambled shRNA co-infected cells. **c**, **g**, KEGG analysis of DEGs from indicated transcriptomes. **d**–**f**, Heatmaps displaying Pearson correlation analysis of indicated transcriptomes. *P* values were calculated by two-way ANOVA (**b**). $n = 4$ pools of 3 mice in **a**; $n = 6$ mice (*Stat3* shRNA1/2), $n = 7$ mice (all others) in **b**; $n = 3$ mice per group in **c**–**f**.

Extended Data Fig. 10a, b). In line with previous results, knockdown of *Stat3* also increased the total number and lowered early differentiation of myofibre-associated SCs from aged mice, and in addition, increased the regenerative capacity of transplanted SCs from aged mice to a similar extent as *Hoxa9* knockdown (Extended Data Fig. 10c–g).

Differentially expressed genes were determined using RNA-sequencing data of freshly isolated, *in vivo* activated SCs from young adult and aged wild-type mice as well as from aged *Hoxa9*$^{-/-}$ mice. There was a highly significant overlap between genes induced by *Hoxa9* overexpression in SCs from young adult mice with those genes that were dysregulated in *in vivo* activated SCs from aged compared to young adult mice ($P = 2.2 \times 10^{-19}$; Extended Data Fig. 10h). Pathways that are currently known to be associated with SC ageing were again among the highest ranked pathways differentially expressed in activated SCs from aged compared to young adult mice including MAPK, TGFβ, Wnt and JAK/STAT signalling (Fig. 5c). Of note, *Hoxa9* deletion abrogated the separate clustering of gene expression profiles of activated SCs from aged compared to young adult mice (Fig. 5d, e). Comparing transcriptomes of activated SC from aged *Hoxa9*$^{+/+}$ to aged *Hoxa9*$^{-/-}$ mice re-established the separate clustering (Fig. 5f) characterized by enrichment of the same set of developmental pathways that associate with SC ageing in wild-type mice (Fig. 5g, compare to Fig. 5c).

Taken together, the current study provides experimental evidence that an aberrant epigenetic stress response impairs the functionality of SCs from aged mice by *Hoxa9*-dependent activation of developmental signals (Extended Data Fig. 10i). Notably, a proof of concept is provided that key enzymes that promote global and site-specific alterations in the epigenetic stress response of aged SCs are druggable, and that the inhibition of these targets leads to improvement in SC function and muscle regeneration during ageing. These findings provide experimental support for the recent hypothesis that a 'shadowed' dysregulation of developmental pathways represents a driving force of stem-cell and tissue ageing[26,27].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 16 October 2015; accepted 3 November 2016.
Published online 30 November 2016.

1. Rando, T. A. Stem cells, ageing and the quest for immortality. *Nature* **441,** 1080–1086 (2006).
2. Brack, A. S. *et al.* Increased Wnt signaling during aging alters muscle stem cell fate and increases fibrosis. *Science* **317,** 807–810 (2007).
3. Carlson, M. E. *et al.* Relative roles of TGF-β1 and Wnt in the systemic regulation and aging of satellite cell responses. *Aging Cell* **8,** 676–689 (2009).
4. Sousa-Victor, P. *et al.* Geriatric muscle stem cells switch reversible quiescence into senescence. *Nature* **506,** 316–321 (2014).
5. Conboy, I. M., Conboy, M. J., Smythe, G. M. & Rando, T. A. Notch-mediated restoration of regenerative potential to aged muscle. *Science* **302,** 1575–1577 (2003).
6. Price, F. D. *et al.* Inhibition of JAK-STAT signaling stimulates adult satellite cell function. *Nat. Med.* **20,** 1174–1181 (2014).
7. Krumlauf, R. Hox genes in vertebrate development. *Cell* **78,** 191–201 (1994).
8. Lawrence, H. J., Sauvageau, G., Humphries, R. K. & Largman, C. The role of HOX homeobox genes in normal and leukemic hematopoiesis. *Stem Cells* **14,** 281–291 (1996).
9. Cosgrove, B. D. *et al.* Rejuvenation of the muscle stem cell population restores strength to injured aged muscles. *Nat. Med.* **20,** 255–264 (2014).
10. Artavanis-Tsakonas, S., Rand, M. D. & Lake, R. J. Notch signaling: cell fate control and signal integration in development. *Science* **284,** 770–776 (1999).
11. Lyons, K. M., Pelton, R. W. & Hogan, B. L. Organogenesis and pattern formation in the mouse: RNA distribution patterns suggest a role for bone morphogenetic protein-2A (BMP-2A). *Development* **109,** 833–844 (1990).
12. Muñoz-Espín, D. *et al.* Programmed cell senescence during mammalian embryonic development. *Cell* **155,** 1104–1118 (2013).
13. Sehgal, P. B., Levy, D. E. & Hirano, T. *Signal Transducers And Activators Of Transcription (STATS): Activation And Biology* (Kluwer Academic, 2003).
14. Sinha, M. *et al.* Restoring systemic GDF11 levels reverses age-related dysfunction in mouse skeletal muscle. *Science* **344,** 649–652 (2014).
15. Bernet, J. D. *et al.* p38 MAPK signaling underlies a cell-autonomous loss of stem cell self-renewal in skeletal muscle of aged mice. *Nat. Med.* **20,** 265–271 (2014).
16. Soshnikova, N. & Duboule, D. Epigenetic temporal control of mouse Hox genes *in vivo*. *Science* **324,** 1320–1323 (2009).
17. Ayton, P. M. & Cleary, M. L. Transformation of myeloid progenitors by MLL oncoproteins is dependent on *Hoxa7* and *Hoxa9*. *Genes Dev.* **17,** 2298–2307 (2003).
18. Yu, B. D., Hess, J. L., Horning, S. E., Brown, G. A. & Korsmeyer, S. J. Altered Hox expression and segmental identity in *Mll*-mutant mice. *Nature* **378,** 505–508 (1995).
19. Grebien, F. *et al.* Pharmacological targeting of the Wdr5-MLL interaction in C/EBPα N-terminal leukemia. *Nat. Chem. Biol.* **11,** 571–578 (2015).
20. McKinnell, I. W. *et al.* Pax7 activates myogenic genes by recruitment of a histone methyltransferase complex. *Nat. Cell Biol.* **10,** 77–84 (2008).
21. Feller, C., Forné, I., Imhof, A. & Becker, P. B. Global and specific responses of the histone acetylome to systematic perturbation. *Mol. Cell* **57,** 559–571 (2015).
22. Liu, L. *et al.* Chromatin modifications as determinants of muscle stem cell quiescence and chronological aging. *Cell Reports* **4,** 189–204 (2013).
23. Filippakopoulos, P. *et al.* Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **149,** 214–231 (2012).
24. Tierney, M. T. *et al.* STAT3 signaling controls satellite cell expansion and skeletal muscle repair. *Nat. Med.* **20,** 1182–1186 (2014).
25. Chakkalakal, J. V., Jones, K. M., Basson, M. A. & Brack, A. S. The aged niche disrupts muscle stem cell quiescence. *Nature* **490,** 355–360 (2012).
26. Blagosklonny, M. V. Aging is not programmed: genetic pseudo-program is a shadow of developmental growth. *Cell Cycle* **12,** 3736–3742 (2013).
27. Martin, N., Beach, D. & Gil, J. Ageing as developmental decay: insights from p16^{INK4a}. *Trends Mol. Med.* **20,** 667–674 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Morita and A. Illing for providing guidance regarding FACS analysis. We are thankful to the FLI Core Facilities Functional Genomics (T. Kroll, A. Ploubidou) and DNA Sequencing (M. Groth) for their services. We express our thanks to M. Burkhalter, T. Sperka and A. Illing for discussions and suggestions. We are grateful to B. Wollscheid and S. Goetze for providing support for proteomic measurements. We thank V. Sakk and M. Kettering for mouse husbandry as well as S. Eichwald, K. Tramm and A. Abou Seif for experimental assistance. We are grateful to M. Kessel, M. Kyba, G. Sauvageau and D. Wellik for sharing plasmids with *Hox* cDNAs. We thank the Structural Genomics Consortium and S. Ackloo for providing access to the epigenetic probe library. We further thank M. Cerletti for providing protocols on SC isolation and E. Perdiguero for advice on infection of SCs before

transplantation. Work on this project in K.L.R.'s laboratory was supported by the DGF (RU-745/10, RU-745/12), the ERC (2012-AdG 323136), the state of Thuringia, and intramural funds from the Leibniz association. J.V.M. was supported by a grant from the DFG (MA-3975/2-1). C.F. acknowledges support by the DFG (FE-1544/1-1) and EMBO (long-term postdoctoral fellowship ALTF 55-2015). R.A. was supported by the ERC (AdvGr 670821 (Proteomics 4D)). The funding for the *Hoxa9*$^{-/-}$ mice to K.L.M. was provided by a grant of the NIH (HL096108). R.R. was supported by a grant from the NIH (R01GM106056). This work was further supported by grants to H.A.K. from the DFG (SFB 1074 project Z1), the BMBF (Gerontosys II, Forschungskern SySTaR, project ID 0315894A), and the European Community's Seventh Framework Programme 390 (FP7/2007-2013, grant agreement 602783).

Author Contributions S.S. designed and performed most experiments, analysed data, interpreted results and wrote the manuscript. F.B. designed and performed RNAi, ChIP and FISH experiments on isolated SCs, analysed data, interpreted results and wrote the manuscript. C.F. and R.A. designed and performed LC–MS experiments, analysed data, interpreted results and wrote the manuscript. A.H.B., U.K., H.H., C.S.V. and M.S. performed individual experiments and analysed data. A.L. performed microarray experiments. D.B.L. provided support and suggestions for ChIP experiments. K.L.M. provided *Hoxa9*$^{-/-}$ mice. J.M.K. and H.A.K. performed microarray and pathway analysis, analysed putative Hoxa9-binding sites and provided support for statistical analysis. B.X. and R.R. conducted analysis of putative Hoxa9-binding sites. F.N. analysed RNA-sequencing data and performed correlation analysis. J.V.M. and S.T. conceived the project, designed and performed individual experiments, interpreted results and wrote the manuscript. K.L.R. conceived the project, designed experiments, interpreted results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.V.M. (julia.vonmaltzahn@leibniz-fli.de), S.T. (stefan.tuempel@leibniz-fli.de) or K.L.R. (lenhard.rudolph@leibniz-fli.de).

Reviewer Information *Nature* thanks J. Gil and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to estimate sample size. No randomization was used. No animals were excluded. The evaluator was blinded to the identity of the specific sample as much as the nature of the experiment allowed it.

Mice. We purchased female young adult C57/BL6j mice (3–4 months) and aged C57/BL6J mice (22–28 months) from Janvier (wild-type mice). Female and male *Hoxa9*$^{-/-}$ mice have been described[28] and were obtained together with age- and gender-matched littermate controls from K. L. Medina. Mice were housed in a pathogen-free environment and fed with a standard diet *ad libitum*. Animal experiments were approved by the Thüringer Landesamt für Verbraucherschutz (Germany) under Reg.-Nr. 03-006/13, 03-012/13 and 03-007/15 and by the Regierungspräsidium Tübingen (Germany) under Reg.-Nr. 35/9185.81-3/919.

Muscle injury. Mice were anaesthetized using isoflurane in air and oxygen through a nose cone. For SC activation, muscles were injured by injecting a total volume of 50 µl of 1.2% $BaCl_2$ (Sigma) into approximately 20 sites in the hindlimb muscles. For regeneration and transplantation experiments, tibialis anterior muscle of the right leg was injected with 50 µl cardiotoxin (CTX, 10 µM, Sigma).

SC isolation and FACS. Muscles from hindlimbs from young adult or aged mice were dissected and collected in PBS on ice. Muscles were rinsed with PBS, minced with scissors and incubated in DMEM with Collagenase (0.2%, Biochrom) for 90 min at 37 °C and 70 r.p.m. Digested muscles were washed with 10% FBS in PBS, triturated and incubated in Collagenase (0.0125%) and Dispase (0.4%, Life Technologies) for 30 min at 37 °C and 100 r.p.m. The muscle slurry was diluted with 10% FBS in PBS, filtered through 100-µm cell strainers and spun down at 500*g* for 5 min. Cell pellets were resuspended in FACS buffer (2% FBS in HBSS) and filtered through 40-µm cell strainers and pelleted at 500*g* for 5 min. Pellets were resuspended in FACS buffer and stained with anti-mouse CD45 PE conjugate (30-F11, eBioscience), anti-mouse CD11b PE conjugate (M1/70, eBioscience), anti-mouse Sca-1 PE conjugate (D7, BioLegend), anti-mouse CD31 PE/Cy7 conjugate (390, BioLegend) and anti-mouse α7-integrin Alexa Fluor 647 conjugate (R2F2, AbLab) for 20 min at 4 °C on a rotating wheel. Cells were washed with FACS buffer. Live cells were identified as calcein blue positive (1:1,000, Invitrogen) and propidium iodide negative (PI, 1 µg ml^{-1}, BD Biosciences). SCs were identified as CD45$^-$Sca-1$^-$CD11b$^-$CD31$^-$α7-integrin$^+$. Cell sorting was performed on a FACSAriaIII with Diva Software (BD).

Culture of SCs. SCs and SC-derived primary myoblasts were cultured at 37 °C, 5% CO_2, 3% O_2 and 95% humidity in growth medium on collagen/laminin-coated tissue culture plates for the indicated time periods. Growth medium was comprised of F10 (Life Technologies) with 20% horse serum (GE), 1% penicillin/streptomycin (Life Technologies) and 5 ng ml^{-1} bFGF (Sigma). For coating, tissue culture plates were incubated with 1 mg ml^{-1} collagen (Sigma) and 10 mg ml^{-1} laminin (Life Technologies) in ddH_2O for at least 1 h at 37 °C and allowed to air-dry. For passaging or FACS analysis, cultured cells were incubated with 0.5% trypsin in PBS for 3 min at 37 °C and collected in FACS buffer. Treatment of SCs with noggin (Prepotech) or DKK1 (Prepotech) was done at 100 ng ml^{-1} concentration. SCs and SC-derived primary myoblasts were treated with 1 µM of chemical probes provided by the Structural Genomics Consortium (SGC, http://www.thesgc.org/chemical-probes/epigenetics)[29,30]. OICR-9429 and bromodomain inhibitors were described previously[19,31–39].

Clonal myogenesis assay. Freshly isolated SCs from young adult and aged mice were sorted in growth medium in 96-well plates using the automated cell deposition unit of the FACSAriaIII. After 5 days, wells containing myogenic colonies were counted by brightfield microscopy. For clonal analysis of lentivirus-transduced SCs, infected (eGFP$^+$ and/or BFP$^+$) live (DAPI$^-$) cells were sorted as one cell per well in growth medium and wells containing myogenic colonies were counted by fluorescence microscopy (Axio Observer, Zeiss) after 5 days. A colony was defined by the presence of at least two cells.

Alamar blue assay. SCs or SC-derived primary myoblasts were seeded at 500 cells per well in growth medium into 96-well plates. After 4 days of culture, the viability was measured by adding Alamar Blue (Life Technologies) as 10% of the sample volume. Cells were incubated for 2 h at 37 °C and fluorescence intensity was measured at an excitation/emission wavelength of 560/590 nm.

BrdU assay. SCs were incubated with 5 µM BrdU (Sigma) in growth medium for 2 h. Cells were fixed with 4% PFA, permeabilized with 0.5% Triton X-100 and incubated with 2 N HCl/PBS for 30 min at room temperature. Incorporated BrdU was detected using anti-BrdU (347580, BD Biosciences) and Alexa-594 fluorochrome (Life Technologies) for 1 h at room temperature. Nuclei were counterstained with DAPI in PBS.

TUNEL assay. TUNEL assay was performed using the *In situ* Cell Death Detection Kit, Texas Red (Roche) according to the manufacturer's instructions.

Senescence-associated β-galactosidase assay. SCs were fixed in 4% PFA and stained with staining solution (5 mM potassium ferricyanide, 5 mM potassium ferrocyanide, 2 mM $MgCl_2$, 150 mM NaCl, 1 mg ml^{-1} X-Gal) in citrate/sodium-phosphate buffer (pH 6) overnight at 37 °C. Staining solution was removed by rinsing several times with PBS.

Myofibre isolation and culture. Individual myofibres were isolated from the extensor digitorum longus muscle as described previously[40,41]. Isolated myofibres were cultured in DMEM containing 20% FBS and 1% chicken embryo extract (Biomol) in dishes coated with horse serum. Freshly isolated fibres or fibres cultured for 24–34 h and 72 h were fixed with 2% PFA and subjected to immunofluorescence analysis. Clusters of SCs were counted on at least 10–15 fibres per replicate. A cluster was defined by the presence of at least three adjacent cells. For quantification of immunofluorescence staining of myofibre-associated quiescent and activated SCs, at least 20 fibres were analysed per replicate. Treatment of myofibre-associated SCs with chemical probes provided by the Structural Genomics Consortium (SGC) was done 4 h after isolation at 1 µM concentration.

siRNA transfection. Transfection of SCs was performed in a reverse manner: SCs were seeded in growth medium into individual wells of a 384-well plate pre-filled with transfection mix. For floating cultures of single myofibres, transfections were performed 4 h after isolation in myofibre culture medium. Transfections were done using Lipofectamin RNAiMAX (Life Technologies) according to manufacturer's instructions. For gene knockdown either Silencer Select siRNAs (Life Technologies) or ON-TARGETplus siRNA SMART-pools (Dharmacon) were used. Respective Silencer Select or ON-TARGETplus SMART-pool non-targeting siRNAs were used as negative control. siRNA sequences are listed in Supplementary Table 1. Transfection efficiency was monitored using a Cy3-labelled control siRNA (Life Technologies). After transfection, FACS-sorted SCs or myofibre-associated SCs were cultured for the indicated time periods and fixed in 2% PFA in PBS. *In vivo* knockdown experiments were performed as described earlier[41]. siRNA sequences were modified to the Accell self-delivering format (Dharmacon) and 100 µg Accell siRNA were injected into tibialis anterior muscle 2 days after CTX injury. *In vivo* knockdown was evaluated from SCs isolated from injected tibialis anterior muscle 3 days after transfection. Transfected muscles were collected 5 days after siRNA injection, frozen in 10% sucrose/OCT in liquid nitrogen and stored at −80 °C.

Lentivirus production and transduction. Lentivirus was produced in Lenti-X cells (Clonetech) after co-transfection of 15 µg shRNA or cDNA plasmid, 10 µg psPAX2 helper plasmid and 5 µg pMD2.G according to standard procedures[42]. Virus was concentrated by centrifugation for 2.5 h at 106,800*g* and 4 °C, and virus pellet was resuspended in sterile PBS. Lentiviral transduction was carried out in growth medium supplemented with 8 µg ml^{-1} polybrene (Sigma).

Plasmids. cDNA was inserted into the SF-LV-cDNA-eGFP plasmid[43]. Primers used for cloning of individual *Hox* cDNAs are listed in Supplementary Table 1. shRNA was inserted into the SF-LV-shRNA-eGFP plasmid using mir30 primers (Supplementary Table 1). shRNA sequences are listed in Supplementary Table 1.

SC transplantation. SCs were FACS purified and transduced with a lentivirus on Retronectin (Takara) coated 48-well plates[4]. After 8–10 h, SCs were obtained by resuspension and washed several times with FACS buffer. For each engraftment, 10,000 SCs were resuspended in 0.9% NaCl and immediately transplanted into tibialis anterior muscles of adult immunosuppressed mice that had been injured with CTX 2 days before. Immunosuppression with FK506 (5 mg kg^{-1} body weight, Sigma) was started at the day of injury using osmotic pumps (model 2004, Alzet) and maintained throughout the entire time of engraftment. Engrafted muscles were collected 3 weeks after transplantation and fixed in 4% PFA for 30 min at room temperature followed by incubation in 30% sucrose/PBS overnight at 4 °C. Fixed muscles were frozen in 10% sucrose/OCT in liquid nitrogen and stored at −80 °C.

Immunohistochemistry. Cryosections of 10 µm were cut from frozen muscle using the Microm HM 550. Cryosections were rinsed once with PBS and fixed in 2% PFA in PBS for 5 min at room temperature. Sections were rinsed three times for 5 min with PBS, permeabilized with 0.5% Triton X-100/0.1 M glycine in PBS for 5 min at room temperature followed again by rinsing them three times with PBS. Sections were blocked in PBS supplemented with 5% horse serum and 1:40 mouse on mouse blocking reagent (Vector labs) for 1 h at room temperature. Incubation with primary antibodies was carried out overnight at 4 °C. The next day, sections were rinsed three times with PBS followed by incubation with secondary antibodies for 1 h at room temperature. Sections were rinsed again with PBS and nuclei were counterstained with 1:1,000 DAPI in PBS before mounting with Permafluor (Thermo Scientific). Slides were stored at 4 °C until analysis. The following primary antibodies were used: 1:1,000 chicken anti-GFP (ab6556, AbCam), 1:1,000 rabbit anti-laminin (L9393, Sigma), 1:200 rabbit anti-Ki67 (ab15580, AbCam), undiluted mouse anti-Pax7 (DSHB). The following secondary antibodies were used at 1:1,000: anti-chicken IgG Alexa-Fluor 488, anti-rabbit IgG Alexa-Fluor 488, anti-mouse IgG1 Alexa-Fluor 594 (Life Technologies).

Immunofluorescence. Freshly isolated SCs were allowed to settle on poly-L-lysine-coated diagnostic microscope slides for 30 min at room temperature. All cells and myofibres were fixed with 2% PFA, permeabilized with 0.5% Triton X-100 and blocked with 10% horse serum in PBS for 1 h at room temperature. Cells and fibres were stained with primary antibodies in blocking solution overnight at 4 °C. Samples were washed three times with PBS and incubated with secondary antibodies for 1 h at room temperature. Nuclei were counterstained with DAPI. Cultured cells were kept in PBS; freshly isolated SCs and myofibres were mounted with Permafluor. The following primary antibodies were used: undiluted mouse anti-Pax7 (DSHB), 1:300 rabbit anti-Hoxa9 (07-178, Millipore), 1:500 mouse anti-Mll1 (05-765, Millipore), 1:500 rabbit anti-Wdr5 (A302-429A, Bethyl Laboratories), 1:300 rabbit anti-H3K4me3 (C15410003-50, Diagenode), 1:200 rabbit anti-MyoD (sc-304, Santa Cruz). The following secondary antibodies were used at 1:1,000: anti-rabbit IgG Alexa-Fluor 488, anti-mouse IgG Alexa-Fluor 594, anti-mouse IgG1 Alexa-Fluor 594 (Life Technologies).

Fluorescence *in situ* hybridization (FISH). Chromatin compaction FISH was done as described previously[44]. DNA of the 3′- and 5′ probe (Fosmid clones WIBR1-1312N03 and WIBR1-2209G09, CHORI) was labelled with digoxigenin or biotin by nick-translation (Roche). 100 ng of probe DNA was used per slide, together with 5 μg mouse CotI DNA (Life Technologies) and 5 μg single-stranded DNA (Ambion). Approximately 5,000 freshly sorted SCs were allowed to settle on poly-L-lysine-coated diagnostic microscope slides for 30 min at room temperature and were fixed with 2% PFA for 5 min. After washing three times with PBS, slides were incubated with 0.1 M HCl for 5 min and permeabilized with 0.5% Triton X-100 in 0.5% saponin for 10 min before freeze–thaw in 20% glycerol in PBS. Denaturation was performed in 50% formamide, 1% Tween-20 and 10% dextran sulfate/2× SSC for 5 min at 75 °C before applying the hybridization cocktail. Probes were hybridized overnight at 37 °C in a humified chamber. Slides were rinsed three times with 2× SSC, blocked with 2% BSA in 0.1% Tween-20 in PBS for 1 h at room temperature, and hybridized probes were visualized with anti-digoxigenin-rhodamine (S7165, Millipore) and Streptavidin-Cy2 (016-220-084, IR USA) for 30 min at room temperature. Nuclei were counterstained with DAPI.

Digital image acquisition and processing. Immunofluorescence images of muscle sections, myofibres and freshly isolated SCs were acquired using the upright microscope Axio Imager (Zeiss) with 10×, 20× and 100× objectives and a monochrome camera. Brightfield and immunofluorescence images of cultured SCs were captured using the microscope Axio Observer (Zeiss) with 5×, 10× and 20× objectives and a monochrome camera. Image acquisition and processing was performed using the ZEN 2012 software (Zeiss). Brightness and contrast adjustments were applied to the entire image before the region of interest was selected. For the analysis of muscle sections, several images covering the whole area of the section were acquired in a rasterized manner and assembled in Photoshop CS6 (Adobe) to obtain an image of the entire section. Images were analysed using ImageJ software. The number of $Pax7^+$ cells in regeneration experiments was normalized to the area of the entire muscle section. CTCF was determined for each SC using the calculation: integrated density − (area of selected cell × mean fluorescence of background readings) (ref. 45).

RNA isolation and reverse transcription. Total RNA was isolated from freshly FACS-isolated or cultured SCs by using the MagMAX 96 total RNA Isolation Kit (Ambion) according to the manufacturer's protocol. The GoScript Reverse Transcription System (Promega) was used for cDNA synthesis from total RNA according to manufacturer's instructions.

ChIP. 5×10^4–1×10^5 cells were crosslinked in 1% formaldehyde (Thermo Scientific) for 10 min. Crosslinking was quenched with glycine and cells were washed two times with ice-cold PBS. For ChIP of H3K4me3, cells were lysed in lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.1, 1× Roche cOmplete Protease Inhibitor) and chromatin was sonicated in Snap Cap microTUBEs using a Covaris M220 sonicator to a fragment size of 150–300 bp. Chromatin was cleared for 10 min at 17,000*g*, and one-tenth of the chromatin was removed as input fraction. Chromatin was immunoprecipitated overnight with 20 μl Protein A/G bead mix (1:1, Dynabeads, Invitrogen) pre-coupled with 1 μg antibody (C15410003-50, Diagenode) in ChIP-dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 167 mM NaCl, 16.7 mM Tris-HCl pH 8.1, 1× Roche cOmplete Protease Inhibitor). Beads were washed three times with low-salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 150 mM NaCl, Tris-HCl, pH 8.1) and three times with LiCl buffer (350 mM LiCl, 1% IPEGAL CA630, 1% deoxycholic acid, 1 mM EDTA, 10 mM Tris-HCl, pH 8.1). For ChIP of Mll1, Wdr5 or haemagglutinin (HA)-tagged Hoxa9 cells were resuspended in sonication buffer (0.1% SDS, 1% Triton X-100, 0.1% Na-deoxycholate, 1 mM EDTA, 140 mM NaCl, 50 mM HEPES, pH 7.9), incubated on ice for 10 min and sonicated to a fragment size of 300–600 bp as described above. Chromatin was cleared for 10 min at 17,000*g* and unspecific binding was absorbed with 5 μl of Protein G beads for 1 h. One-tenth (Mll1/Wdr5) or one-twentieth (HA-tag) of the chromatin was removed as input fraction. Chromatin was immunoprecipitated overnight with 2 μg of antibody (Mll1: A300-086A, Wdr5: A302-429A, Bethyl Laboratories; HA-tag: ab9110, Abcam). Chromatin-antibody complexes were captured with 20 μl Protein A/G bead mix (1:1, Dynabeads, Invitrogen) for 2 h. Beads were washed twice with sonication buffer, twice with NaCl buffer (0.1% SDS, 1% Triton X-100, 0.1% Na-deoxycholate, 1 mM EDTA, 500 mM NaCl, 50 mM HEPES, pH 7.9), twice with LiCl buffer and once with TE buffer. Decrosslinking and elution was performed in 50 μl decrosslinking buffer (1% SDS, 100 mM $NaHCO_3$, 250 mM NaCl) for 4 h at 65 °C with continuous shaking and subsequent Proteinase K treatment for 1 h at 45 °C. DNA was purified using Agencourt AMPure XP beads (Beckman Coulter) with a beads:sample ratio of 1.8:1 or MinElute PCR Purification Kit according to manufacturer's protocols.

Quantitative PCR. Quantitative PCR (qPCR) was performed with an ABI 7500 Real-Time PCR System (Applied Biosystems) in technical duplicates from the indicated number of biological replicates. The qPCR was carried out in a volume of 12 μl using the Absolute qPCR Rox Mix (Thermo Scientific) and the Universal Probe Library (Roche). Primer and probe sets for the detection of single genes are listed in Supplementary Table 1. *Gapdh* was detected with rodent *Gapdh* control reagents (Applied Biosystems). Relative expression values were calculated using the ΔC_t method.

$$\Delta C_t = C_t[\text{gene of interest}] - C_t[\textit{Gapdh}]$$
$$\text{Relative expression} = 2^{(-\Delta C_t)}$$

qPCR analysis of ChIP samples was performed using SYBR Green Supermix (Biorad) in a final reaction volume of 10 μl and 0.75 μM final primer concentration. Primers are listed in Supplementary Table 1. HA-tag ChIP signals were calculated as percentage of the input fraction. The $\Delta\Delta C_t$ method was used to calculate fold enrichment of a genomic locus over the ChIP specific background control (*Actb* intergenic region for H3K4me3 or gene desert for *Mll1* and *Wdr5*), both normalized to the signal in the input fraction:

$$\Delta C_t[\text{normalized to input}] = (C_t[\text{ChIP}] - (C_t[\text{input}] - \log_2(\text{input dilution factor})))$$
$$\Delta\Delta C_t = \Delta\, C_t[\text{region of choice normalized to input}] - \Delta C_t[\text{control region normalized to input}]$$
$$\text{Fold enrichment} = 2^{(-\Delta\Delta C_t)}$$

Nanostring analysis. Pellets of freshly isolated SCs were lysed with 3 μl RLT buffer (QIAGEN) and subjected to Nanostring analysis according to manufacturer's instructions using a custom-made Hox gene nCounter Elements TagSet (Nanostring Technologies). Relative expression to the housekeeping genes *Gapdh*, *Hmbs* and *Polr2a* was calculated using nSolver Software (v2.0) after background correction and normalization to hybridized probe signals.

Proteomic analysis of histone modifications. Preparation of histones for mass spectrometry, data acquisition and analysis were essentially performed as described previously[21] with modifications described below. In brief, histones were isolated by acid extraction, derivatised by d6-acetic anhydride (CD_3CO, Aldrich) and digested with sequencing-grade trypsin (Promega) overnight at a trypsin:protein ratio of 1:20. To acetylate free peptide N termini, trypsinised histones were derivatised again for 45 min at 37 °C using 1:20 (v/v) d6-acetic anhydride (CD_3CO, Aldrich) in 50 mM ammonium bicarbonate buffered to pH 8 by ammonium hydroxide solution. After derivatization, peptides were evaporated in a speed-vac at 37 °C to near dryness, resuspended in 50 μl of 0.1 formic acid and purified by a StageTip protocol using two discs of C18 followed by one disc of activated carbon (3 M Empore). After StageTip purification, the samples were evaporated in a speed-vac to near dryness, resuspended in 20 μl of 0.1% formic acid and stored at −20 °C until mass spectrometry acquisition. The histone samples were separated on a reversed-phase liquid chromatography column (75-μm, New Objective) that was packed in-house with a 15-cm stationary phase (ReproSil-Pur C18-AQ, 1.9 μm). The column was connected to a nano-flow HPLC (EASY-nLC 1000; Thermo Scientific) and peptides were electrosprayed in a Q Exactive mass spectrometer (Thermo Fisher Scientific). Buffer A was composed of 0.1% formic acid in HPLC-grade water and buffer B was 0.1% formic acid in ACN. Peptides were eluted in a linear gradient with a flow rate of 300 nl per minute, starting at 3% B and ramping to 35% in 52 min, followed by an increase to 50% B in 4 min, followed by an increase to 98% in 4 min and then holding at 98% B for another 6 min. Mass spectrometry was operated in a combined shotgun-PRM mode targeting positional isomers. Ion chromatograms were extracted with Thermo Xcalibur and Skyline and data summarization and statistical analysis was performed in Excel and R. Relative abundances were calculated from the raw signal reads, according to the formulas described previously[21] without further normalizations.

Microarray and bioinformatics analysis. Gene expression analysis was performed using the Mouse GE 8x60K Microarray Kit (Agilent Technologies, Design ID 028005). 100 ng total RNA isolated from SCs were used for the labelling. Samples were labelled with the Low Input Quick Amp Labelling Kit (Agilent Technologies) according to the manufacturer's instructions. Slides were scanned using a microarray scanner (Agilent Technologies). Expression data were extracted using the Feature Extraction software (Agilent Technologies). Preprocessing of expression data was performed according to Agilent's standard workflow. Using five quality flags (gIsPosAndSignif, gIsFeatNonUnifOL, gIsWellAboveBG, gIsSaturated, and gIsFeatPopnOL) from the Feature Extraction software output, probes were labelled as detected, not detected, or compromised. Gene expression levels were background corrected, and signals for duplicated probes were summarized by geometric mean of non-compromised probes. After $\log_2$ transformation, a percentile shift normalization at the 75% level and a baseline shift to the median baseline of all probes was performed. All computations were performed using the R statistical software framework (http://www.R-project.org). Differentially expressed genes were calculated by the shrinkage T-statistic[46] and controlled for multiple testing by maintaining a false discovery rate (FDR) < 0.05 (ref. 47).

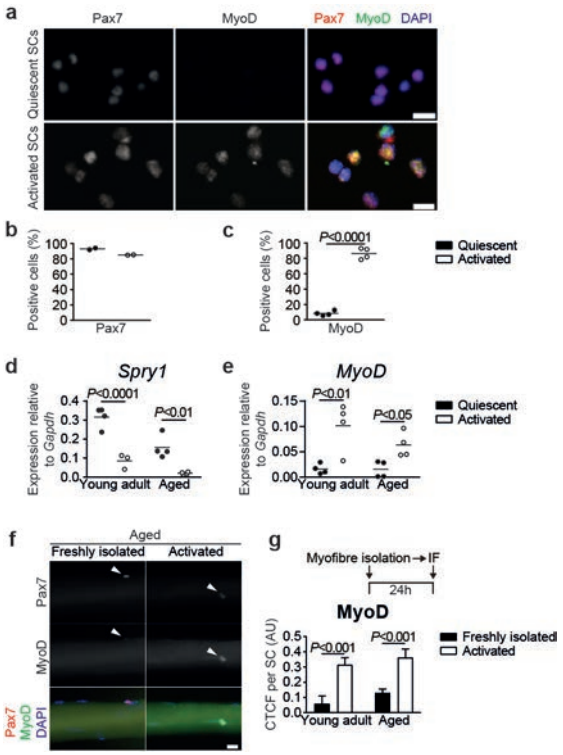
RNA-sequencing analysis. Sequencing reads were filtered out for low quality sequences and trimmed of low quality bases by using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). Mapping to mm9 genome was performed by using TopHat software[48]. Gene quantification was performed by using HT-Seq and differentially expressed genes (DEGs) were estimated by using DESEQ2 (refs 49, 50) within the R statistical software framework (http://www.R-project.org) with $P < 0.01$. Pearson correlation heatmaps were generated by using custom R scripts by selecting genes having more than 10 read counts in all the samples of at least one condition and an interquartile range (IQR) > 0.5. Significance of overlapping DEGs was calculated by normal approximation of hypergeometric probability.

Identification of Hoxa9-binding sites. Transcription start and end sites of putative Hoxa9 target genes were collected from the UCSC Genome Browser[51] with mm8 track. Sequences in gene body regions (from transcription start to end sites), promoter regions ($-2/+1$ kb relative to transcription start sites), and distal intergenic regions ($-50/+50$ kb relative to transcription start sites) of 26 genes were prepared for identification of Hoxa9 binding sites. These sequences were aligned based on the previously reported consensus motifs for Hoxa9-Meis1-Pbx1 (ATGATTTATGGC)[52] and Meis1 (TGTC)[53]. Putative Hoxa9-binding sites were aligned when they contained either no mismatch or one mismatch, and Meis1 motifs were aligned with no mismatch allowed. Hoxa9-binding sites with at least one Meis1-binding site within 300 bp on the same DNA strand were selected for further analysis. Identified Hoxa9-binding sites are listed in Supplementary Table 1.

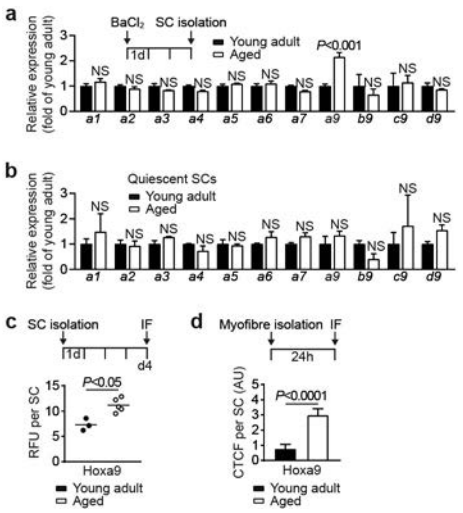
Statistics. If not stated otherwise, results are presented as mean and s.e.m. from the number of samples indicated in the figure legends. Two groups were compared by two-sided Student's *t*-test or two-sided Mann–Whitney *U*-test. For multiple comparisons a two-way ANOVA was performed using a FDR < 0.5 to correct for multiple comparisons. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$. Statistical analysis was done using GraphPad Prism 6 software and R (v3.3.1).

Data availability statement. Microarray and RNA-sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) with the accession code GSE87812. Further data that support the findings of this study are available from the corresponding authors upon reasonable request. Source data for the Figures and Extended Data Figures are provided with the paper.

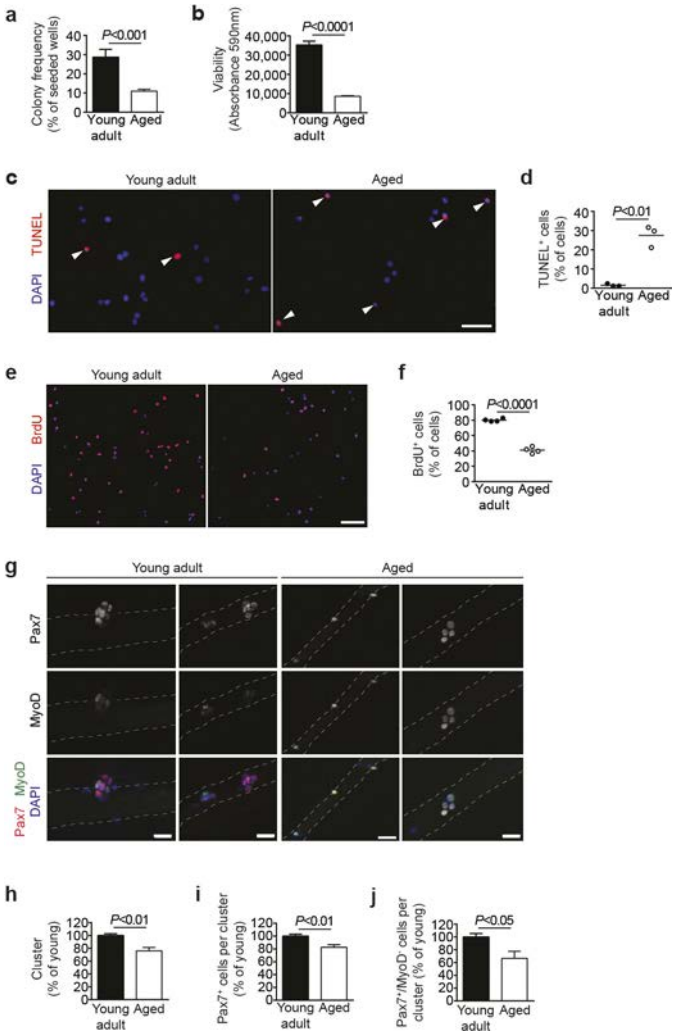
28. Lawrence, H. J. *et al.* Mice bearing a targeted interruption of the homeobox gene *HOXA9* have defects in myeloid, erythroid, and lymphoid hematopoiesis. *Blood* **89,** 1922–1930 (1997).
29. Brown, P. J. & Müller, S. Open access chemical probes for epigenetic targets. *Future Med. Chem.* **7,** 1901–1917 (2015).
30. Barsyte-Lovejoy, D. *et al.* Chemical biology approaches for characterization of epigenetic regulators. *Methods Enzymol.* **574,** 79–103 (2016).
31. Theodoulou, N. H. *et al.* Discovery of I-BRD9, a selective cell active chemical probe for bromodomain containing protein 9 inhibition. *J. Med. Chem.* **59,** 1425–1439 (2016).
32. Picaud, S. *et al.* Generation of a selective small molecule inhibitor of the CBP/p300 bromodomain for leukemia therapy. *Cancer Res.* **75,** 5106–5119 (2015).
33. Picaud, S. *et al.* PFI-1, a highly selective protein interaction inhibitor, targeting BET bromodomains. *Cancer Res.* **73,** 3336–3346 (2013).
34. Martin, L. J. *et al.* Structure-based design of an *in vivo* active selective BRD9 inhibitor. *J. Med. Chem.* **59,** 4462–4475 (2016).
35. Hay, D. A. *et al.* Discovery and optimization of small-molecule ligands for the CBP/p300 bromodomains. *J. Am. Chem. Soc.* **136,** 9308–9319 (2014).
36. Drouin, L. *et al.* Structure enabled design of BAZ2-ICR, a chemical probe targeting the bromodomains of BAZ2A and BAZ2B. *J. Med. Chem.* **58,** 2553–2559 (2015).
37. Clark, P. G. *et al.* LP99: discovery and synthesis of the first selective BRD7/9 bromodomain inhibitor. *Angew. Chem.* **127,** 6315–6319 (2015).
38. Chen, P. *et al.* Discovery and characterization of GSK2801, a selective chemical probe for the bromodomains BAZ2A and BAZ2B. *J. Med. Chem.* **59,** 1410–1424 (2016).
39. Filippakopoulos, P. *et al.* Selective inhibition of BET bromodomains. *Nature* **468,** 1067–1073 (2010).
40. Pasut, A., Jones, A. E. & Rudnicki, M. A. Isolation and culture of individual myofibers and their satellite cells from adult skeletal muscle. *J. Vis. Exp.* **73,** 50074 (2013).
41. Bentzinger, C. F. *et al.* Fibronectin regulates Wnt7a signaling and satellite cell expansion. *Cell Stem Cell* **12,** 75–87 (2013).
42. Schambach, A. *et al.* Lentiviral vectors pseudotyped with murine ecotropic envelope: increased biosafety and convenience in preclinical research. *Exp. Hematol.* **34,** 588–592 (2006).
43. Wang, J. *et al.* A differentiation checkpoint limits hematopoietic stem cell self-renewal in response to DNA damage. *Cell* **148,** 1001–1014 (2012).
44. Chambeyron, S. & Bickmore, W. A. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes Dev.* **18,** 1119–1130 (2004).
45. Burgess, A. *et al.* Loss of human Greatwall results in G2 arrest and multiple mitotic defects due to deregulation of the cyclin B-Cdc2/PP2A balance. *Proc. Natl Acad. Sci. USA* **107,** 12564–12569 (2010).
46. Opgen-Rhein, R. & Strimmer, K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.* **6,** http://dx.doi.org/10.2202/1544-6115.1252 (2007).
47. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B* **57,** 12 (1995).
48. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
49. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
51. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493–D496 (2004).
52. Shen, W. F. *et al.* HOXA9 forms triple complexes with PBX2 and MEIS1 in myeloid cells. *Mol. Cell. Biol.* **19,** 3051–3061 (1999).
53. Huang, Y. *et al.* Identification and characterization of Hoxa9 binding sites in hematopoietic cells. *Blood* **119,** 388–398 (2012).



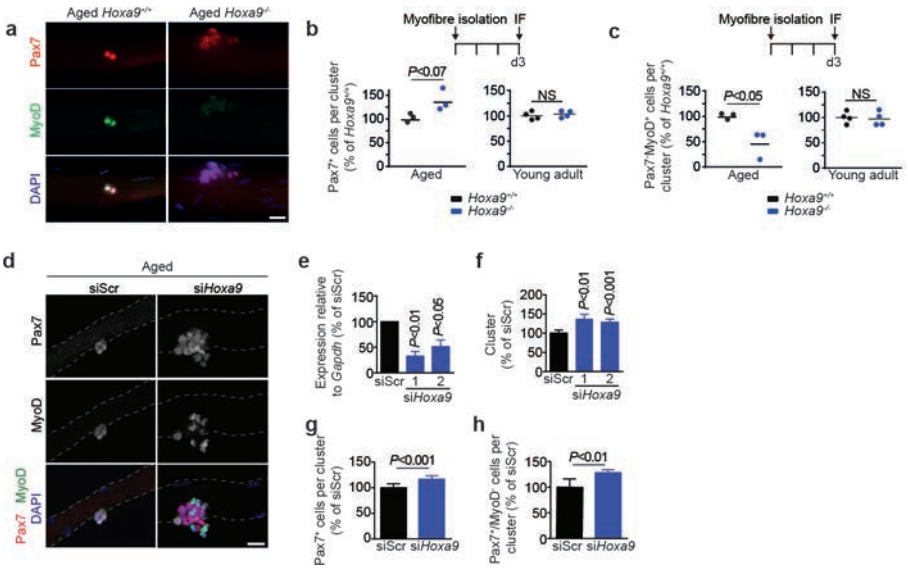
Extended Data Figure 1 | SC activation. a, Immunofluorescence staining for Pax7 and MyoD of freshly isolated SCs from injured (activated SCs) and uninjured muscles (quiescent SCs) from young adult mice. Nuclei were counterstained with DAPI (blue). **b**, **c**, Quantification of $Pax7^+$ cells (**b**) and $MyoD^+$ cells (**c**) in **a**. **d**, **e**, qPCR analysis of *Spry1* (**d**) and *Myod1* (**e**) expression in freshly isolated quiescent and *in vivo* activated SCs of young adult and aged mice. **f**, Immunofluorescence staining for Pax7 and MyoD on freshly isolated and 24-h cultured myofibre-associated SCs from aged mice. Nuclei were counterstained with DAPI (blue). **g**, Corrected total cell fluorescence (CTCF) for MyoD per SC as in **f**. Scale bars, 10 μm (**a**) and 20 μm (**f**). *P* values were calculated by two-sided Student's *t*-test (**b**, **c**) or two-way ANOVA (**d**, **e**, **g**). $n = 2$ mice in **b**; $n = 4$ mice in **c**; $n = 3$ mice (young activated), $n = 4$ mice (all others) in **d**; $n = 4$ mice in **e**; $n = 33/24$ nuclei (young), $n = 35/20$ nuclei (aged) from 3 mice in **g**.



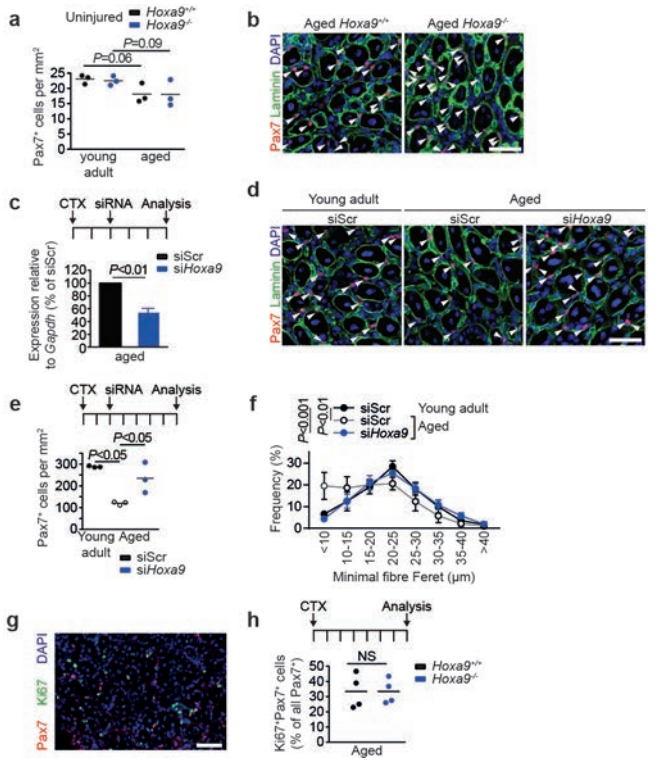
Extended Data Figure 2 | Expression of Hox genes in SCs. a, **b**, Nanostring analysis of mRNA expression of *Hoxa* genes and *Hoxa9* paralogues (*b9-c9-d9*) in *in vivo* activated (**a**) and quiescent (**b**) freshly isolated SCs from young adult and aged mice. **c**, Relative fluorescence units (RFU) for Hoxa9 per SC in 4-day cultured SCs from young adult and aged mice. **d**, Corrected total cell fluorescence (CTFC) for Hoxa9 per activated SC on 24-h cultured myofibres as in Fig. 1d. *P* values were calculated by two-way ANOVA (**a**, **b**) or two-sided Mann–Whitney *U*-test (**c**, **d**). $n = 3$ mice in **a**, **b**; $n = 3$ mice (young), $n = 5$ mice (aged) in **c**; $n = 34$ nuclei (young), $n = 32$ nuclei (aged) from 4 mice in **d**.



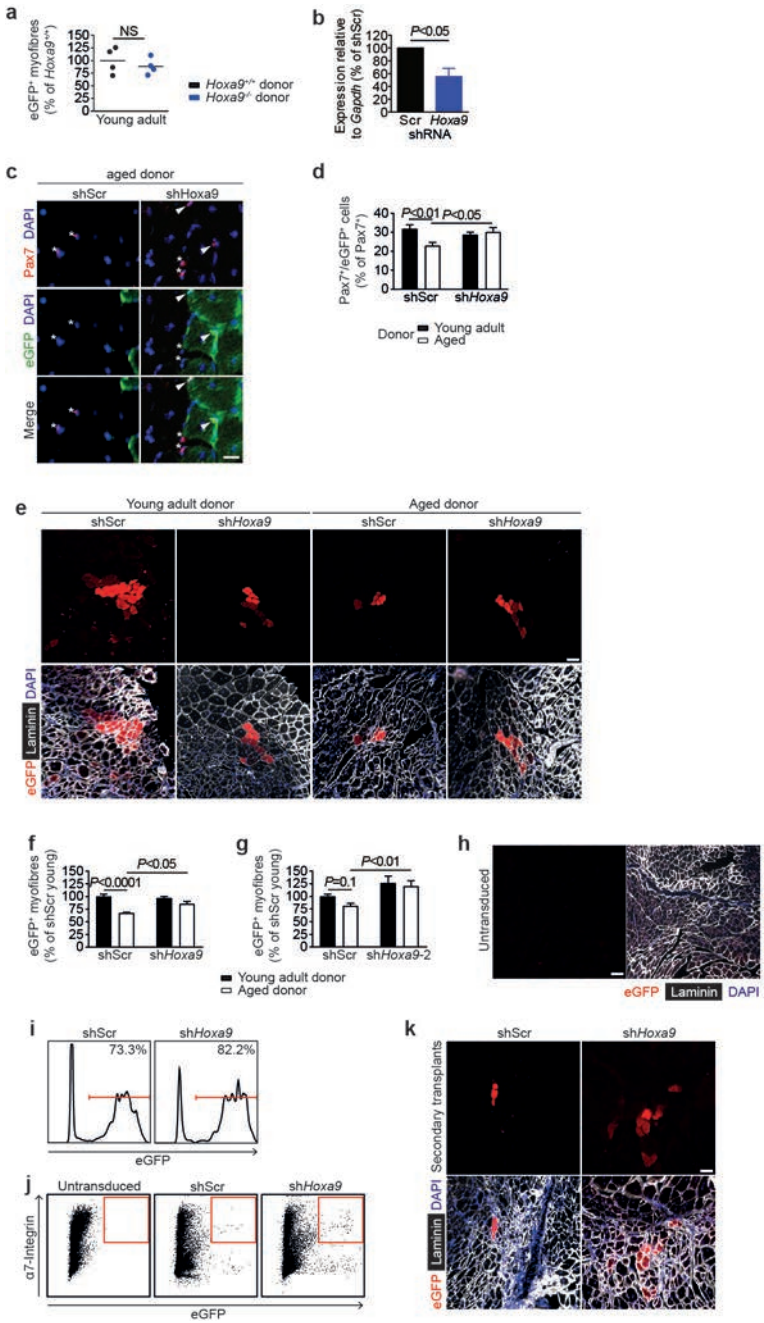
Extended Data Figure 3 | Functional decline in aged SCs. a, SCs from young adult and aged mice were sorted as single cells. After 5 days, the frequency of myogenic colonies was assessed. **b**, Equal numbers of FACS-isolated SCs from young adult and aged mice were cultured for 4 days and Alamar Blue assay was performed. **c**, TUNEL staining of SCs isolated from young adult or aged mice after 4 days of culture. Nuclei were counterstained with DAPI (blue). **d**, Quantification of apoptosis based on TUNEL staining in **c**. **e**, BrdU staining of SCs isolated from young adult or aged mice after 4 days of culture. Nuclei were counterstained with DAPI (blue). **f**, Quantification of proliferation based on BrdU staining in **e**. **g**, Immunofluorescence staining for Pax7 and MyoD on myofibres isolated from young adult and aged mice after 72 h in culture. Nuclei were counterstained with DAPI (blue). **h**–**j**, Quantification of the number of SC-derived clusters with at least 3 adjacent cells (**h**), average number of all $Pax7^+$ cells (**i**), or proportion of $Pax7^+/MyoD^-$ cells (**j**) within clusters as in **g**. Scale bars, 20 μm (**c**, **g**) and 50 μm (**e**). *P* values were calculated by two-sided Student's *t*-test. $n = 8$ mice (young), $n = 10$ mice (aged) in **a**; $n = 7$ mice (young), $n = 5$ mice (aged) in **b**; $n = 3$ mice in **d**; $n = 4$ mice in **f**; $n = 4$ mice (aged) in **j**, $n = 5$ mice (all others) in **h**–**j**.



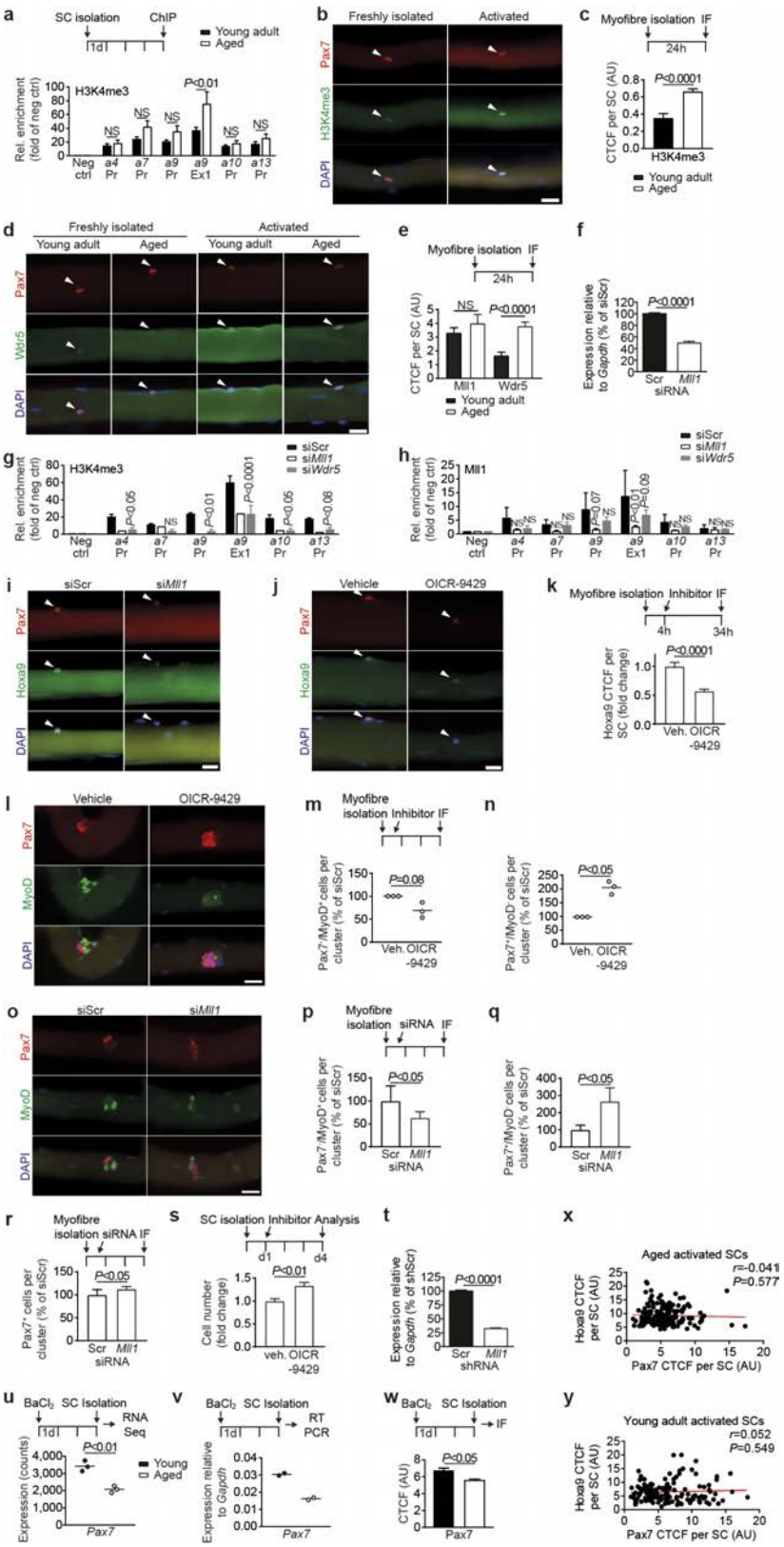
Extended Data Figure 4 | Deletion or knockdown of Hoxa9 improves SC function in myofibre cultures. a, Immunofluorescence staining for Pax7 and MyoD on 72 h cultured myofibre-associated SCs from aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice. **b**, **c**, Average number of all Pax7^{+} cells (**b**) or Pax7^{-}/MyoD^{+} cells (**c**) within clusters from aged or young adult *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice as shown in **a**. **d**, Immunofluorescence staining for Pax7 and MyoD on 72-h cultured myofibres isolated from aged mice transfected with *Hoxa9* or scrambled (Scr) siRNAs. Nuclei were counterstained with DAPI (blue). **e**, qPCR analysis of *Hoxa9* expression in SCs transfected with *Hoxa9* siRNA or scrambled control. Two *Hoxa9* siRNAs with different target sequences (Supplementary Table 1) were used. **f**–**h**, Analysis of 72-h cultured myofibre-associated SCs from **d**. Quantification of the number of SC-derived clusters with at least 3 adjacent cells (**f**), average number of all Pax7^{+} cells (**g**), or proportion of Pax7^{+}/MyoD^{-} cells (**h**) within clusters. Scale bars, 20 μm (**a**, **d**). Dashed lines outline myofibres. *P* values were calculated by two-sided Student's *t*-test. $n = 3$ mice (aged), $n = 4$ mice (young) in **b**, **c**; $n = 3$ mice in **e**; $n = 5$ mice in **f**–**h**.



Extended Data Figure 5 | Inhibition of *Hoxa9* improves muscle regeneration in aged mice. **a**, Quantification of Pax7$^+$ cells per area in uninjured tibialis anterior muscles from young adult and aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice. **b**, Representative immunofluorescence staining for Pax7 and laminin on tibialis anterior muscles from aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice that were collected 7 days after cardiotoxin (CTX) injury. **c**, qPCR analysis of *Hoxa9* expression in SCs isolated from tibialis anterior muscles injected with a self-delivering *Hoxa9* or scrambled siRNA and collected 5 days after muscle injury. **d**, Representative immunofluorescence staining for Pax7 and laminin of injured tibialis anterior muscles from young adult and aged mice that were injected with a self-delivery siRNA and collected 7 days after muscle injury. Nuclei were counterstained with DAPI (blue). Arrowheads denote Pax7$^+$ cells. **e**, Quantification of Pax7$^+$ cells from **d** per area. **f**, Frequency distribution minimal Feret's diameter of muscle fibres from **d**. **g**, Exemplary immunofluorescence staining for Pax7 and Ki67 on tibialis anterior muscles from aged *Hoxa9*$^{+/+}$ and *Hoxa9*$^{-/-}$ mice collected 7 days after muscle injury. Nuclei were counterstained with DAPI (blue). **h**, Quantification of proliferating SCs (Ki67$^+$/Pax7$^+$) as depicted in **g**. Scale bars, 50 µm. *P* values were calculated by two-sided Student's *t*-test (**c**, **h**) or two-way ANOVA (**a**, **e**, **f**). $n = 3$ mice in **a**; $n = 3$ mice in **c**; $n = 3$ mice in **e**, **f**; $n = 4$ mice in **h**.

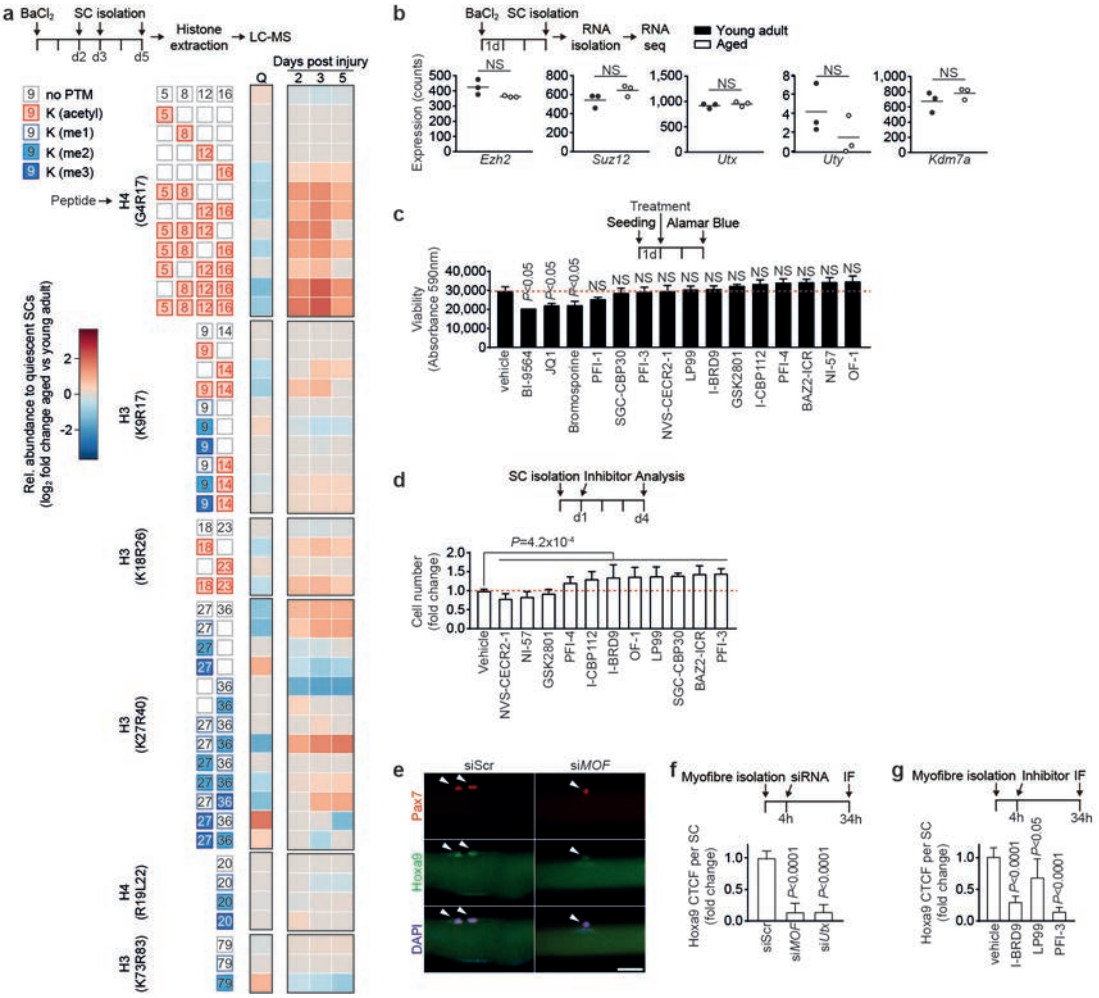


Extended Data Figure 6 | Inhibition of *Hoxa9* improves regenerative capacity of aged SCs. a, Quantification of donor-derived ($eGFP^+$) myofibres from transplantation of SCs from young adult $Hoxa9^{+/+}$ and $Hoxa9^{-/-}$ mice. **b**, qPCR analysis of *Hoxa9* expression in SCs transduced with scrambled control or *Hoxa9* shRNA encoding lentivirus. **c**–**g**, Transplantation of eGFP-labelled SCs from young adult and aged mice that were targeted with shRNAs against *Hoxa9* or a scrambled control. **c**, Representative immunofluorescence staining for Pax7 and eGFP of transplanted muscle sections. Nuclei were counterstained with DAPI (blue). Arrowheads denote $Pax7^+/eGFP^+$ cells, asterisks label $Pax7^+/eGFP^-$ cells. **d**, Quantification of donor-derived ($eGFP^+$) $Pax7^+$ cells in **c**. **e**, Representative immunofluorescence staining for eGFP and laminin of transplanted muscle sections, nuclei were counterstained with DAPI (blue). **f**, **g**, Quantification of donor-derived ($eGFP^+$) myofibres in **e** for two different *Hoxa9* shRNAs in two independent experiments. **h**, Exemplary immunofluorescence staining for eGFP and laminin in tibialis anterior muscles engrafted with untransduced aged SCs. Nuclei were counterstained with DAPI (blue). **i**, Flow cytometric analysis of transduction efficiency of donor SCs used for transplantation in primary recipients analysed in Fig. 2f. **j**, Representative flow cytometry plots for re-isolation of transplanted aged SCs that were untransduced as control or transduced with scrambled control or *Hoxa9* shRNA encoding lentivirus as quantified in Fig. 2f. **k**, Representative immunofluorescence staining for eGFP and laminin in engrafted tibialis anterior muscles from secondary recipients quantified in Fig. 2g. Nuclei were counterstained with DAPI (blue). Scale bars, 20 μm (**c**), 50 μm (**h**) and 100 μm (**e**, **k**). *P* values were calculated by two-sided Student's *t*-test (**a**, **b**) or two-way ANOVA (**d**, **f**, **g**). $n = 4$ recipient mice in **a**; $n = 3$ mice in **b**; $n = 6$ recipient mice (young donors), $n = 4$ recipient mice (aged donors) in **d**, **f**; $n = 5$ recipient mice in **g**.

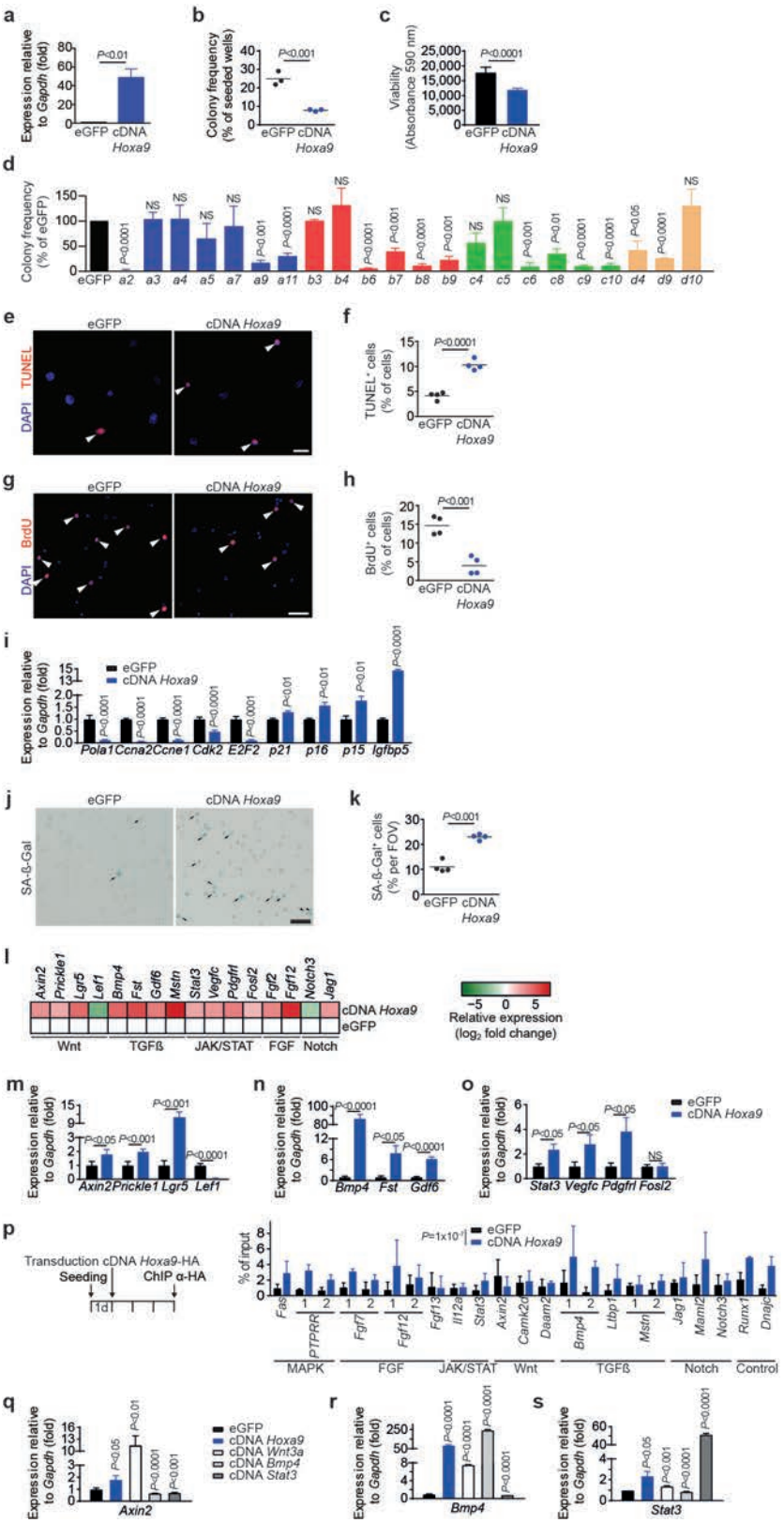


Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Inhibition of Mll1 rescues H3K4me3 induction, *Hoxa9* overexpression, and functional impairment of activated SCs from aged mice. a, ChIP for H3K4me3 at promoters or exons of indicated Hox genes in activated SCs (4 day culture) from young adult and aged mice. **b**, Representative immunofluorescence staining for Pax7 and H3K4me3 on myofibre-associated SCs from aged mice that were freshly isolated or activated by 24-h culture of myofibres. **c**, Corrected total cell fluorescence (CTCF) for H3K4me3 on activated SCs shown in **b**. **d**, Representative immunofluorescence staining for Pax7 and Wdr5 on myofibre-associated SCs from young adult and aged mice that were freshly isolated or activated by 24-h culture of myofibres. **e**, CTCF for Mll1 and Wdr5 per activated SC as shown in **d**. **f**, qPCR analysis of *Mll1* in SCs transfected with *Mll1* siRNA or scrambled control. **g**, **h**, ChIPs for H3K4me3 (**g**) and Mll1 (**h**) in primary myoblasts 3 days after transfection with the indicated siRNAs. **i**, **j**, Immunofluorescence staining for Pax7 and Hoxa9 in myofibres from aged mice after transfection with *Mll1* siRNA or scrambled control (**i**, quantification in Fig. 3d) or after treatment with OICR-9429 or vehicle (**j**). **k**, CTCF for Hoxa9 per SC as shown in **j**. **l**, Representative immunofluorescence staining for Pax7 and MyoD on OICR-9429 treated myofibre-associated SCs from aged mice after 72 h culture. Nuclei were counterstained with DAPI (blue). **m**, **n**, Average number of $Pax7^{-}/MyoD^{+}$ cells (**m**) or $Pax7^{+}/MyoD^{-}$ cells (**n**) within clusters as shown in **l**. **o**, Representative immunofluorescence staining for Pax7 and MyoD on siRNA-treated myofibre-associated SCs from aged mice after 72-h culture. Nuclei were counterstained with DAPI (blue). **p**–**r**, Average number of $Pax7^{-}/MyoD^{+}$ cells (**p**), $Pax7^{+}/MyoD^{-}$ cells (**q**) or $Pax7^{+}$ cells (**r**) within clusters in **o**. **s**, Relative changes in cell number of aged SCs after treatment with OICR-9429 and 4 days of culture, compared to vehicle control. **t**, qPCR analysis of *Mll1* in SCs transduced with *Mll1* shRNA or scrambled control. **u**–**w**, Analysis of Pax7 expression in *in vivo* activated SCs from young adult and aged mice by RNA-sequencing (**u**), qPCR (**v**), or immunofluorescence as depicted in Fig. 1b (**w**). **x**, **y**, Pearson correlation comparing the Hoxa9 immunofluorescence signal (quantification in Fig. 1c) and the Pax7 immunofluorescence signal (quantification in **w**) of activated SCs from aged (**x**) and young adult (**y**) mice. Note, there is no correlation between Hoxa9 expression level and Pax7 expression level in activated SCs from aged mice. Scale bars, 20 μm (**b**, **d**, **i**, **j**, **l**, **o**). *P* values were calculated by two-way ANOVA (**a**, **g**, **h**), two-sided Student's *t*-test (**f**, **m**, **n**, **p**–**v**), two-sided Mann–Whitney *U*-test (**c**, **e**, **k**, **w**) or Pearson correlation (**x**, **y**). $n = 4$ mice (young), $n = 7$ mice (aged) in **a**; $n = 27$ nuclei from 2 mice (young), $n = 27$ nuclei from 4 mice (aged) in **c**; $n = 40/52$ nuclei (Mll1), $n = 44/99$ nuclei (Wdr5) from 3 young/aged mice in **e**; $n = 3$ mice in **f**; $n = 3$ biological replicates (*Wdr5* siRNA), $n = 2$ biological replicates (*Mll1* siRNA) in **g**; $n = 3$ biological replicates in **h**; $n = 173$ nuclei (DMSO), $n = 324$ nuclei (OICR-9429) from 4 mice in **k**; $n = 3$ mice in **m**, **n**; $n = 7$ mice in **p**–**r**; $n = 6$ mice in **s**; $n = 3$ mice in **t**; $n = 3$ mice in **u**; $n = 2$ mice in **v**; $n = 134$ nuclei (young), $n = 181$ nuclei (aged) from 3 mice in **w**–**y**.

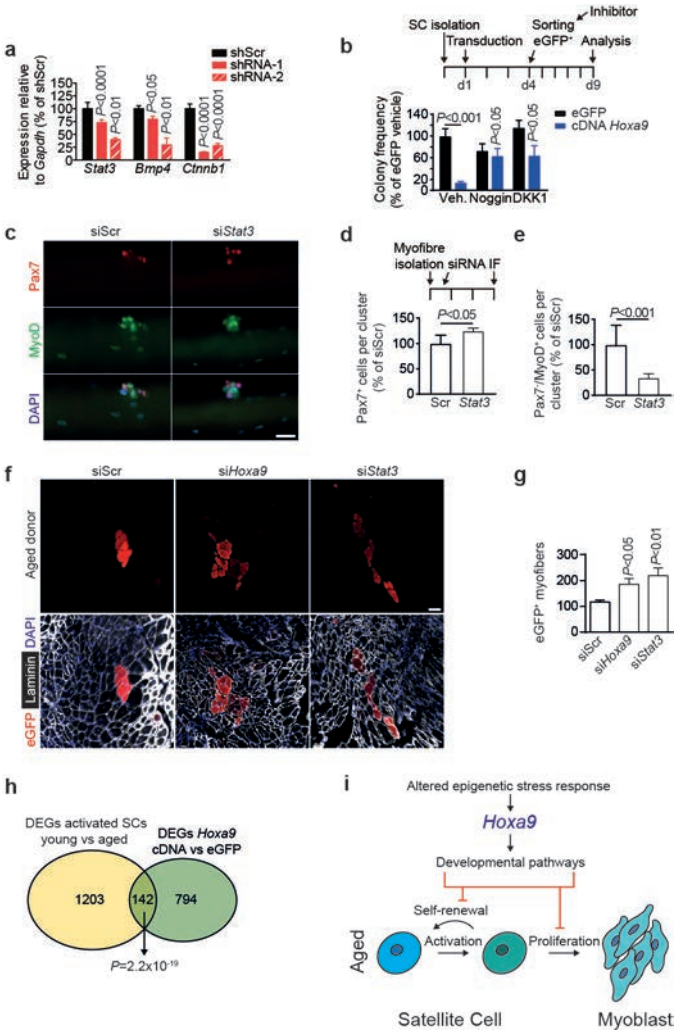


Extended Data Figure 8 | Alterations in the epigenetic stress response of activated SCs from aged mice. a, Heatmap displaying relative changes in abundance of different histone modifications (measured at the indicated peptides) in freshly isolated SCs from aged compared to young adult mice. SCs were analysed in quiescence (Q, derived from uninjured muscle) or at the indicated time points after activation mediated by muscle injury. Relative abundances at indicated days after injury are first normalized to quiescent SCs, and then compared between SCs isolated from aged and young adult mice and $\log_2$ scaled. Only significant changes are shown ($P < 0.05$). **b**, Expression analysis of the indicated genes in freshly isolated *in vivo* activated SCs from young adult and aged mice based on RNA-sequencing. **c**, Viability of primary myoblasts after 48-h treatment with bromodomain inhibitors (1 μM) from the Structural Genomics Consortium probe set, measured by Alamar Blue assay. **d**, Relative changes in cell number of aged SCs after treatment with non-toxic bromodomain inhibitors (1 μM) from **c** and 4 days of culture, compared to vehicle control. A Wilcoxon rank-sum test on the ratio of all cell counts being equal to 1 was performed to test the hypothesis of a general effect of the inhibitors on cell number. **e**, Representative immunofluorescence staining for Pax7 and Hoxa9 in siRNA-treated myofibre-associated SCs from aged mice. Scale bar, 20 μm. **f**, CTCF for Hoxa9 per SC as shown in **e**. **g**, Quantification of immunofluorescence staining for Hoxa9 in $Pax7^+$ cells on myofibre-associated SCs from aged mice treated with bromodomain inhibitors. *P* values were calculated by two-sided Student's *t*-test (**a**–**c**), Wilcoxon rank-sum test (**d**) or two-sided Mann–Whitney *U*-test (**f**, **g**). $n = 4$ mice in **a**; $n = 3$ mice in **b**; $n = 4$ biological replicates in **c**; $n = 6$ mice in **d**; $n = 71$ nuclei (scrambled siRNA), $n = 48$ nuclei (*MOF* siRNA), $n = 98$ nuclei (*Utx* siRNA) from 3 mice in **f**; $n = 60$ nuclei (vehicle), $n = 59$ nuclei (I-BRD9), $n = 38$ nuclei (LP99), $n = 62$ nuclei (PFI-3) from 3 mice in **g**.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Overexpression of *Hox* genes inhibits SC function. a, Expression of *Hoxa9* in SCs transduced with *Hoxa9* cDNA or eGFP as control. **b**, **c**, FACS-isolated SCs from young adult mice were transduced with a lentivirus either containing both eGFP and *Hoxa9* cDNA or only eGFP. Infected ($eGFP^+$) cells were isolated after 3 days. **b**, Frequency of myogenic colonies from single-cell-sorted SCs. **c**, Quantification of cell number based on Alamar Blue assay of bulk cultures. **d**, Frequency of myogenic colonies of SCs overexpressing the indicated Hox genes. **e**, **g**, TUNEL (**e**) or BrdU (**g**) staining of SCs overexpressing *Hoxa9* or eGFP. Infected ($eGFP^+$) cells were isolated 3 days after transduction and analysed 3 days later. Nuclei were counterstained with DAPI (blue). Arrowheads mark TUNEL- or BrdU-positive cells. **f**, **h**, Quantification of apoptosis (**f**) or proliferation (**h**) based on TUNEL or BrdU staining as in **e** or **g**. **i**, qPCR-based expression analysis of various cell-cycle and senescence markers in SCs overexpressing *Hoxa9* compared to eGFP-infected controls, 5 days after infection. **j**, Senescence-associated-β-galactosidase (SA-β-Gal) staining of SCs overexpressing *Hoxa9* or eGFP at day 5 after infection. Arrowheads mark SA-β-Gal-positive cells. **k**, Quantification of senescence per field of view (FOV) based on SA-β-Gal staining in **j**. **l**, Heatmap displaying $\log_2$ fold changes of expression of selected genes from microarray analysis in Fig. 5a. **m**–**o**, qPCR validation of differentially expressed genes annotated to Wnt (**m**), TGFβ (**n**) and JAK/STAT pathways (**o**) as in **l**. **p**, Identification of Hoxa9-binding sites by anti-HA ChIP of primary myoblasts overexpressing HA-tagged *Hoxa9* cDNA or eGFP as control. Shown is the qPCR for 1 or 2 putative Hoxa9-binding sites at the indicated loci. Hoxa9-binding sites at target genes were identified as described in the Methods and are listed in Supplementary Table 1. A two-sided block bootstrap test on the difference of the percentage of bound DNA for all binding sites being equal to 0 was performed to test the hypothesis of a generally increased binding of Hoxa9. **q**–**s**, SCs were infected with lentiviruses expressing *Hoxa9*, *Wnt3a*, *Bmp4* or *Stat3* cDNAs or eGFP. qPCR analysis of expression of the indicated target genes at 5 days after infection: *Axin2* (**q**), *Bmp4* (**r**) and *Stat3* (**s**). Scale bars, 20 μm (**e**, **g**) and 50 μm (**j**). *P* values were calculated by two-sided Student's *t*-test (**a**–**d**, **f**, **h**, **k**, **q**–**s**) or two-way ANOVA (**i**, **m**–**o**). $n = 4$ mice in **a**; $n = 3$ mice in **b**; $n = 7$ mice in **c**; $n = 3$ mice in **d**; $n = 4$ mice in **f**, **h**, **k**; $n = 3$ mice (p15, p21), $n = 6$ mice (p16), $n = 4$ mice (all others) in **i**; $n = 4$ pools of 3 mice in **l**; $n = 4$ mice in **m**–**o**; $n = 3$ biological replicates for **p**; $n = 3$ mice (*Wnt3a*, *Bmp4*, *Stat3*), $n = 4$ mice (eGFP, *Hoxa9*) in **q**–**s**.



Extended Data Figure 10 | Validation of Hoxa9 downstream targets. **a**, Knockdown efficiency of two shRNAs (red bars) for *Stat3*, *Bmp4* and *Ctnnb1*. **b**, SCs from young adult mice were transduced with an *Hoxa9* and *eGFP*-encoding lentivirus. eGFP$^+$ cells were sorted as single cells and cultured in the presence of noggin, DKK1 or 0.1% BSA in PBS as vehicle. Colony frequency was assessed after 5 days and is compared to *Hoxa9* cDNA expressing cells treated with vehicle control. **c**, Representative immunofluorescence staining for Pax7 and MyoD on siRNA-transfected myofibres from aged mice after 72 h of culture. Nuclei were counterstained with DAPI (blue). **d**, **e**, Average number of Pax7$^+$ cells (**d**) or Pax7$^-$/MyoD$^+$ cells (**e**) within clusters in **c**. **f**, Representative immunofluorescence staining for eGFP and laminin in tibialis anterior muscles engrafted with siRNA-transfected SCs isolated from eGFP transgenic aged mice. Nuclei were counterstained with DAPI (blue). **g**, Quantification of donor-derived (eGFP$^+$) myofibres in **f**. **h**, Area-proportional Venn diagram of differentially expressed genes from indicated transcriptomes. **i**, Model for the Hoxa9-mediated impairment of SC function during ageing: quiescent SCs become activated upon muscle injury and proliferate as myoblasts to repair damaged muscle tissue. After activation, aged SCs display global and locus-specific alterations in the epigenetic stress response resulting in overexpression of *Hoxa9*, which in turn induces developmental pathways inhibiting SC function and muscle regeneration in aged mice. Scale bars, 20 μm (**c**), and 100 μm (**f**). *P* values were calculated by two-way ANOVA (**a**, **b**) or two-sided Student's *t*-test (**d**, **e**, **g**). $n = 3$ mice in **a**; $n = 4$ mice in **b**; $n = 5$ mice in **d**, **e**; $n = 5$ recipient mice in **g**; $n = 3$ mice per group (activated SCs), $n = 4$ pools of 3 mice (*Hoxa9* overexpression) in **h**.

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

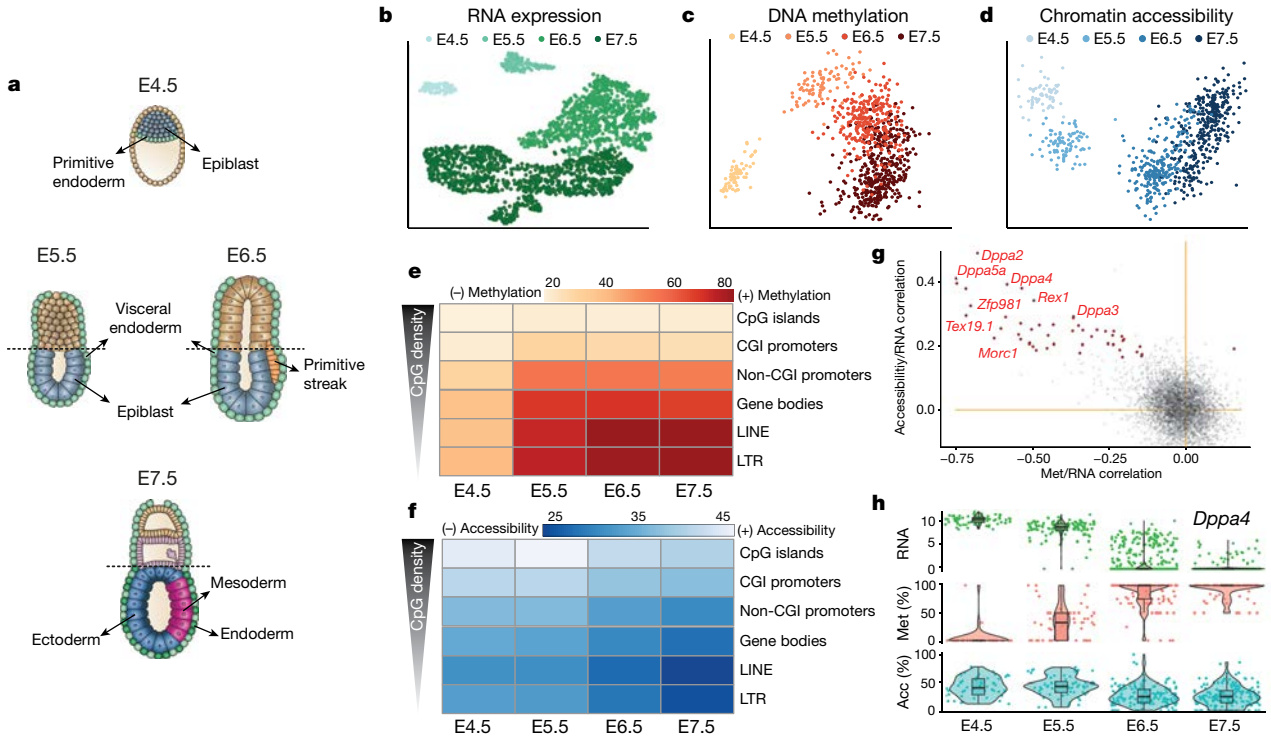


Fig. 1 | Single-cell multi-omics profiling of mouse gastrulation. a, Schematic of the developing mouse embryo, with stages and lineages considered in this study labelled. **b**, Dimensionality reduction of RNA-expression data using UMAP. Cells are coloured by stage. There are 1,061 cells included from 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

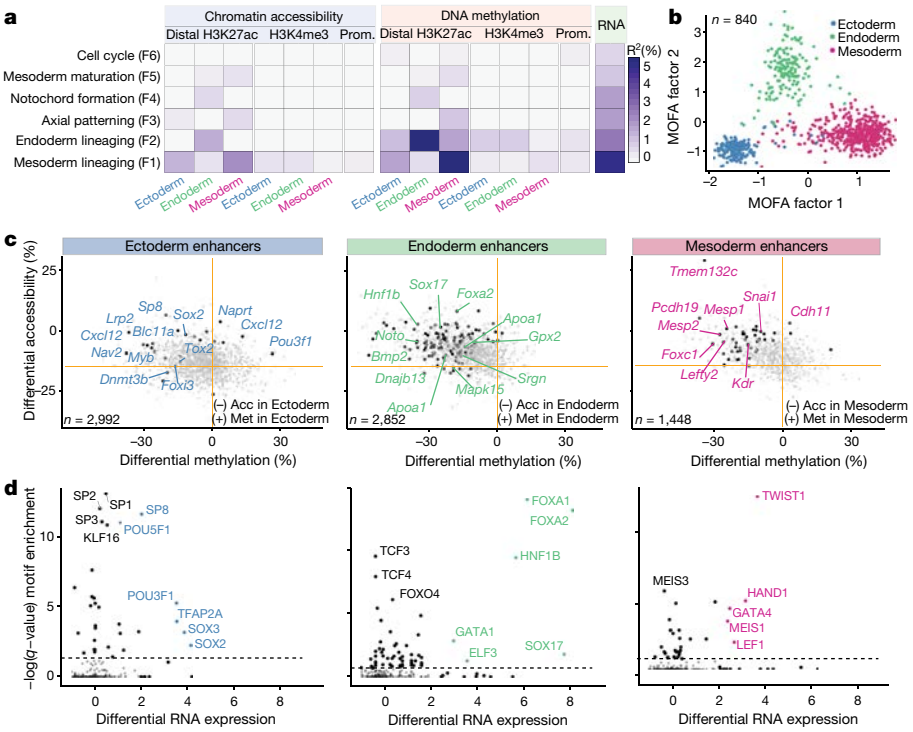


Fig. 2 | Multi-omics factor analysis reveals coordinated epigenetic and transcriptomic variation at enhancer elements during germ-layer commitment. a, Percentage of variance explained (R^2) by each MOFA factor (rows) across data modalities (columns). **b**, Scatter plot of MOFA factor 1 (*x* axis) and 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

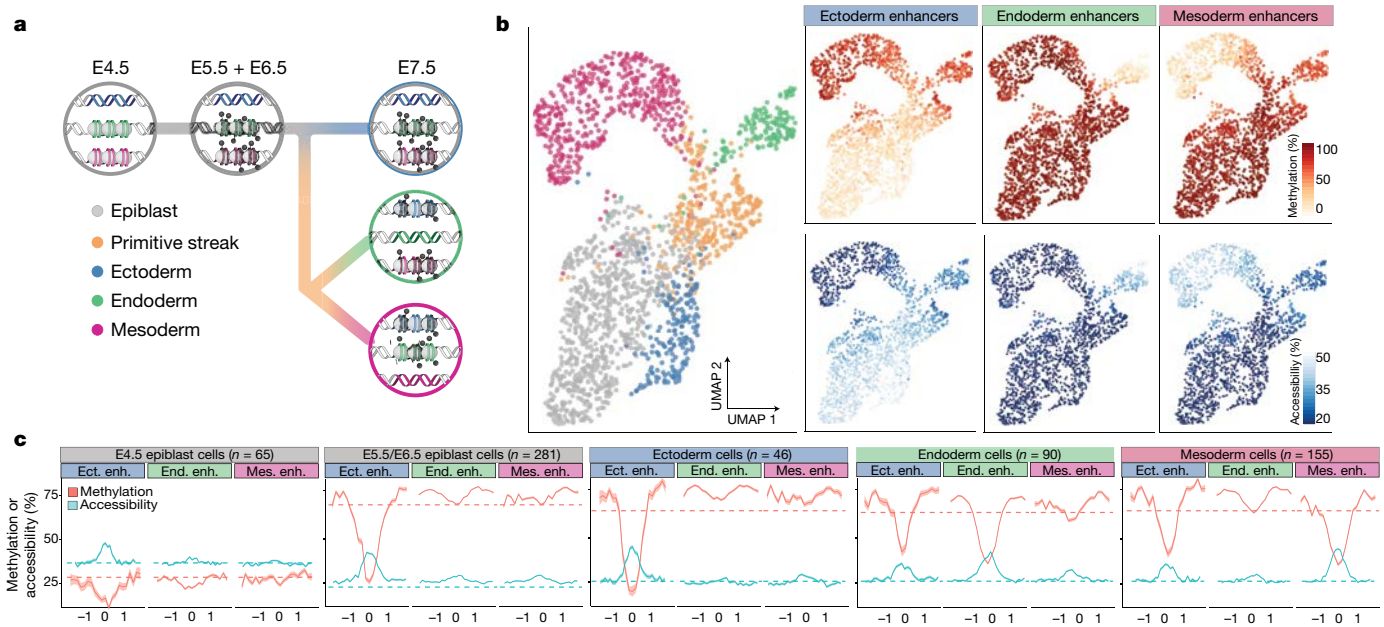


Fig. 3 | DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers across development. a, Illustration of the hierarchical model of enhancer epigenetic dynamics associated with germ-layer commitment. **b**, UMAP projection based on the MOFA factors inferred using all embryonic 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.

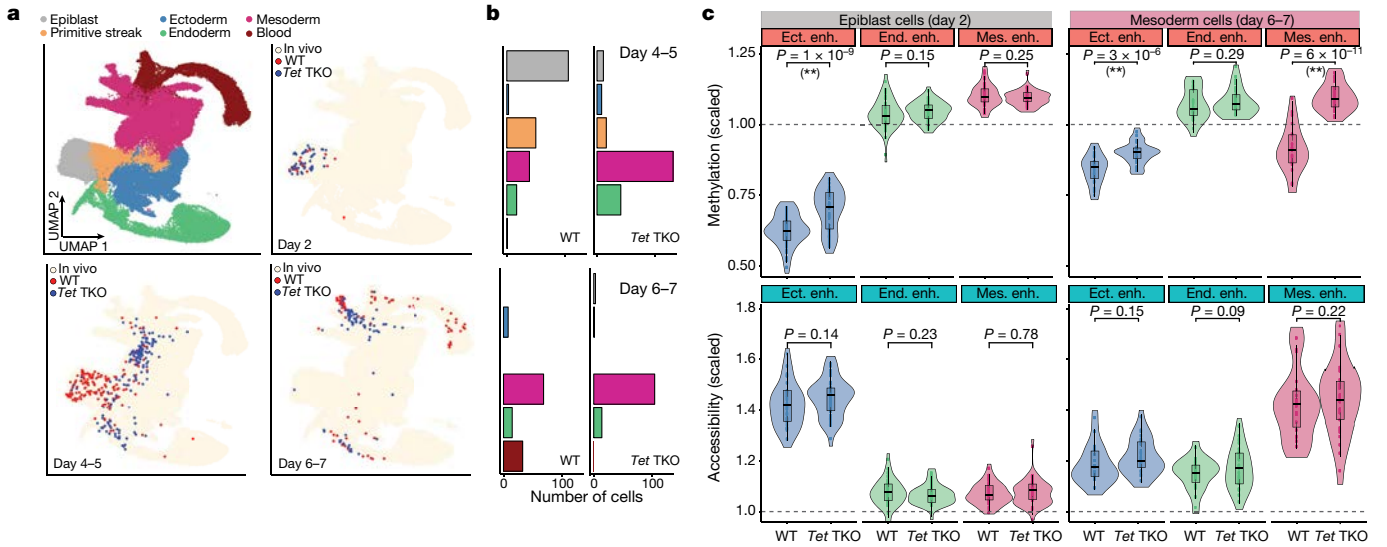


Fig. 4 | TET enzymes are required for efficient demethylation of mesoderm-defining enhancers and subsequent blood differentiation in embryoid bodies. a, UMAP projection of stages E6.5 to E8.5 of the atlas dataset (no extra-embryonic cells). Top left, cells coloured by lineage assignment. The remaining 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.

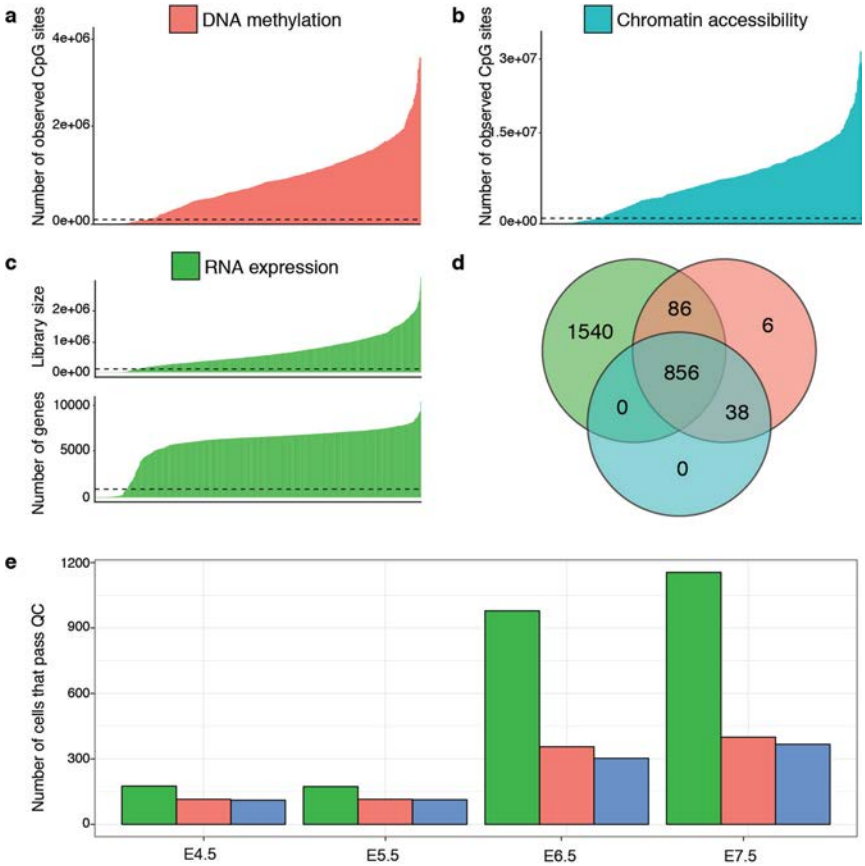
Competing interests W.R. is a consultant and shareholder of Cambridge Epigenetix. The remaining authors declare no competing interests.

Additional information

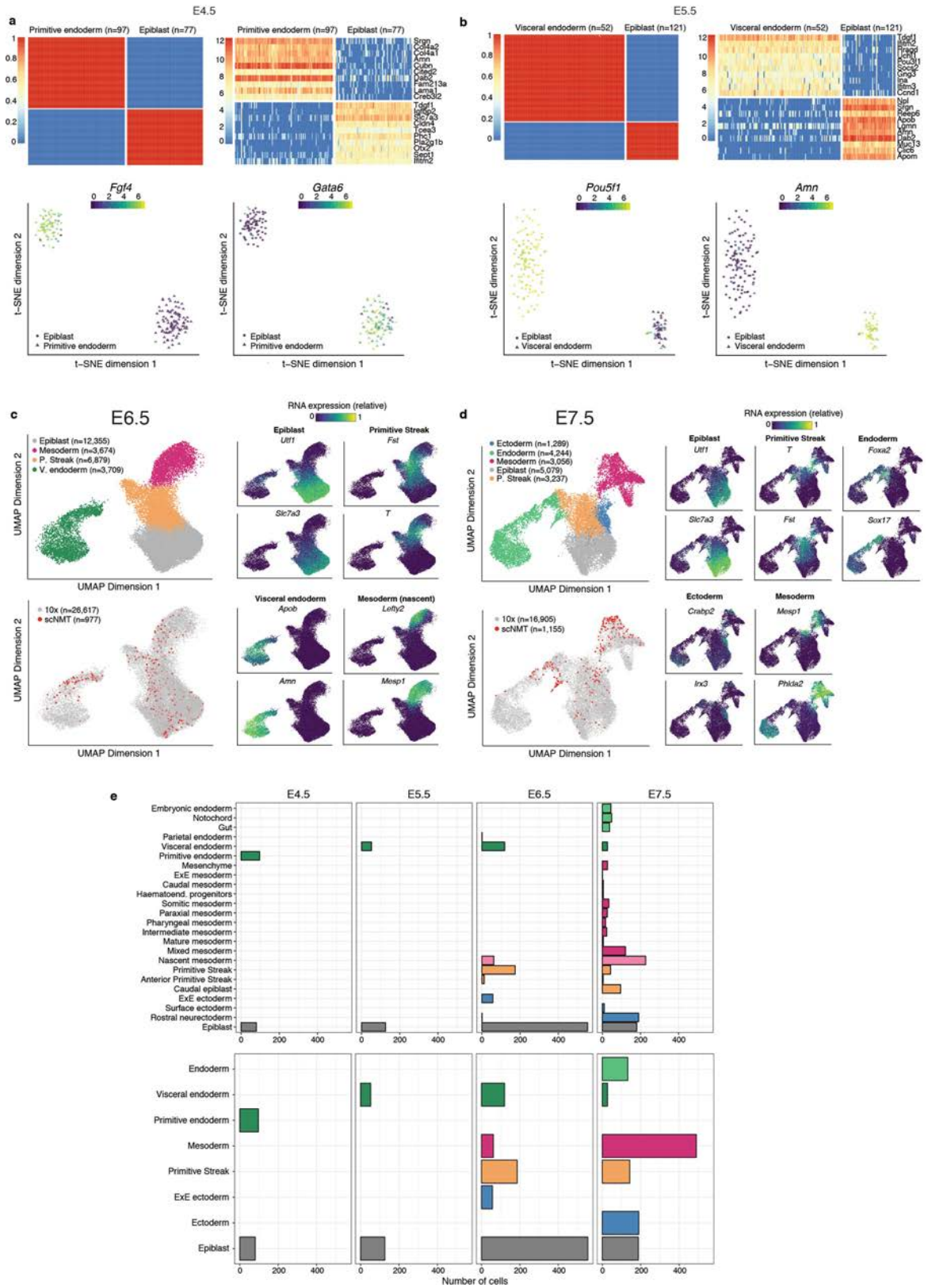
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1825-8.

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.

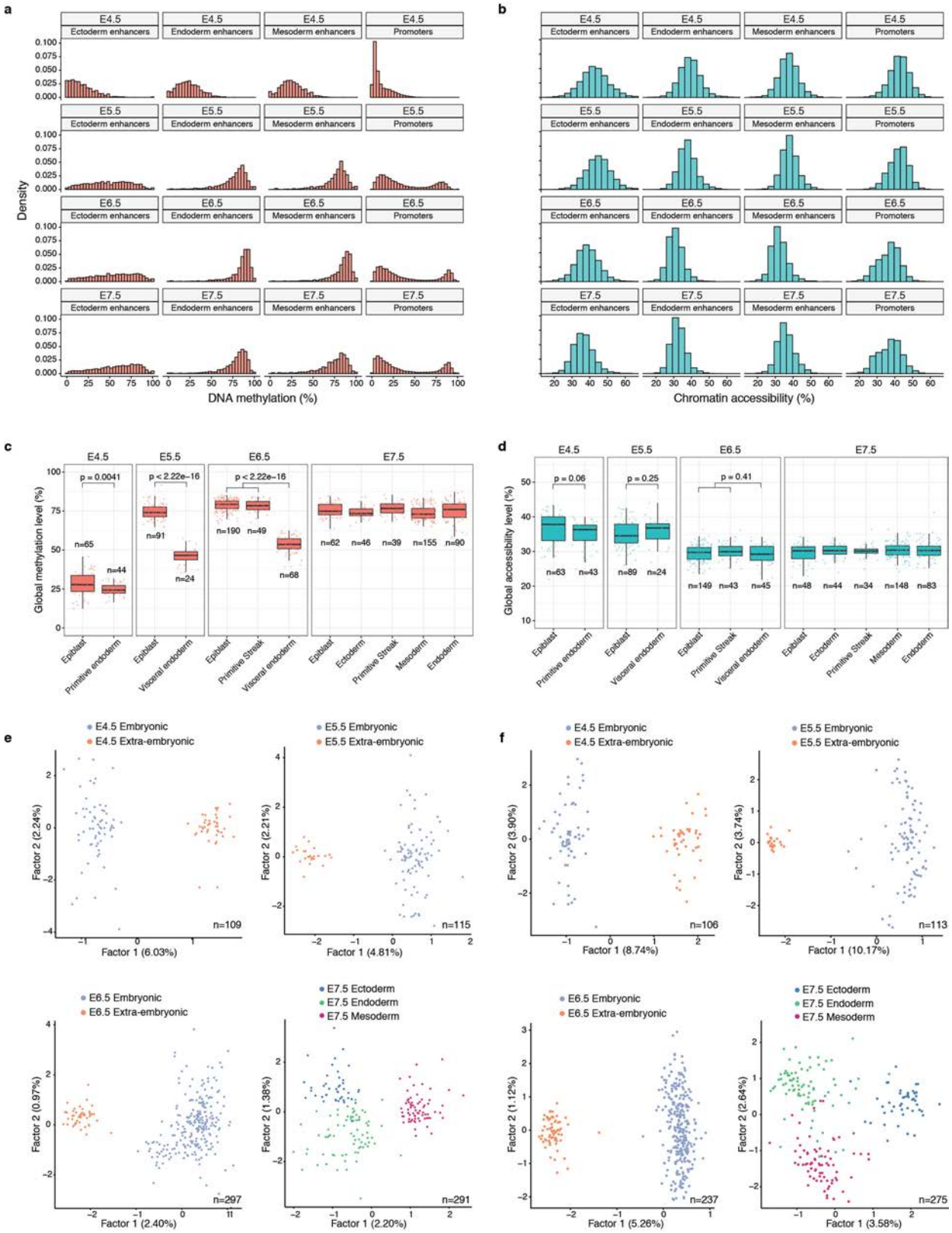


Extended Data Fig. 1 | scNMT-seq quality controls. a, **b**, Number of observed cytosines in CpG (red; **a**) or GpC (blue; **b**) contexts respectively. Each bar corresponds to one cell. Cells are sorted by total number of CpG or GpC sites. Cells below the dashed line were discarded on the basis of 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.



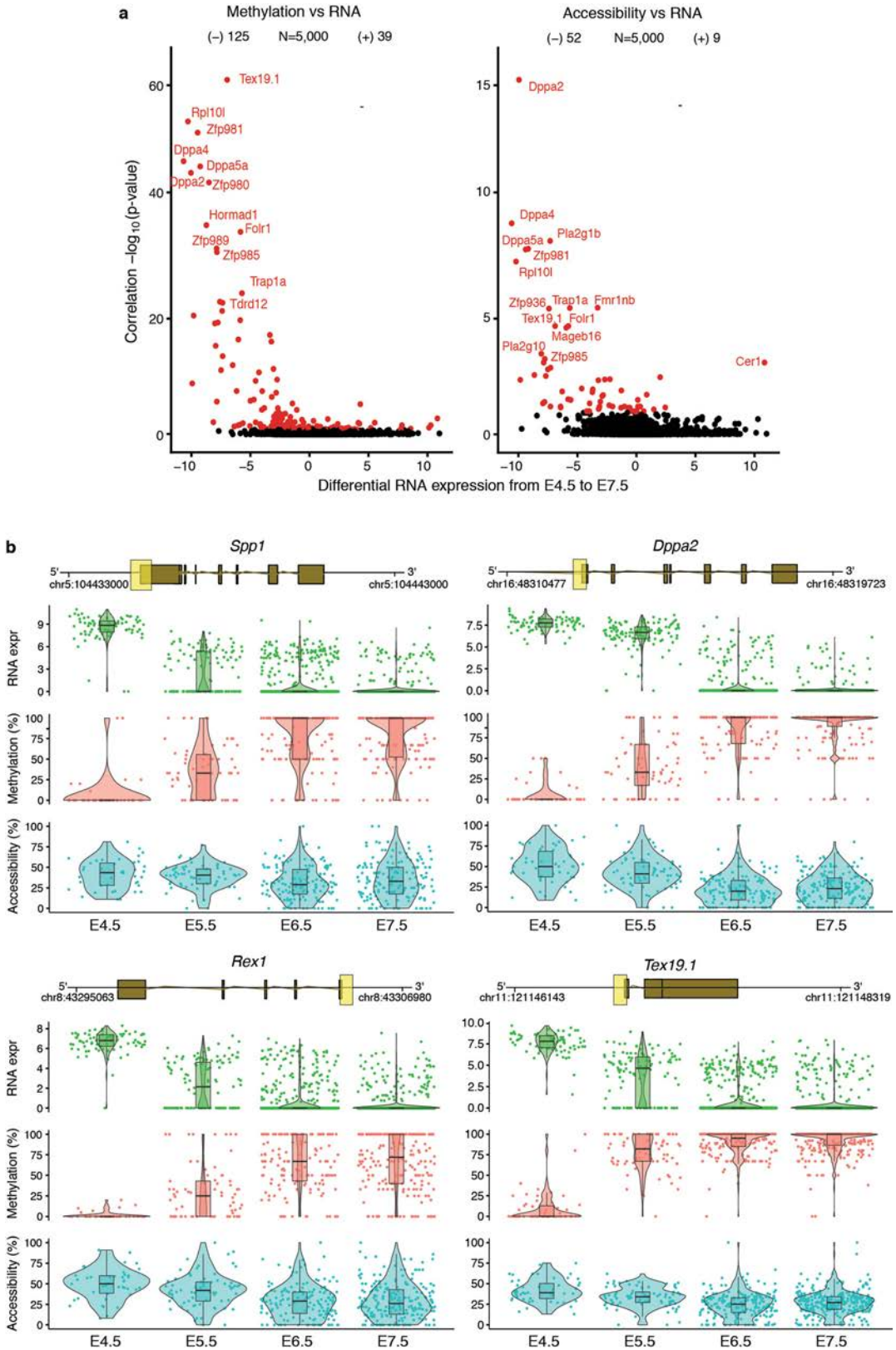
Extended Data Fig. 2 | See next page for caption.

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.

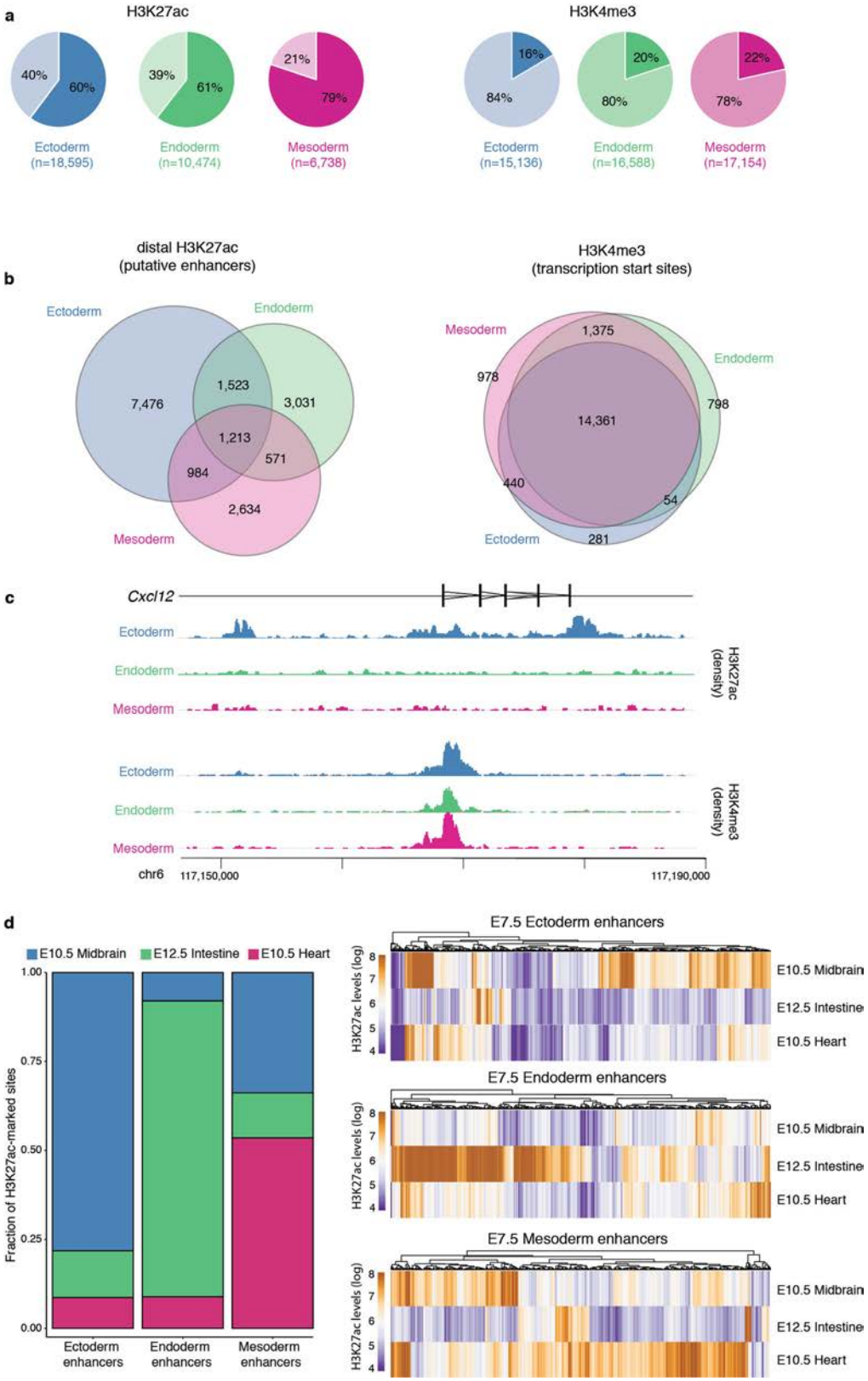


Extended Data Fig. 3 | See next page for caption.

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).

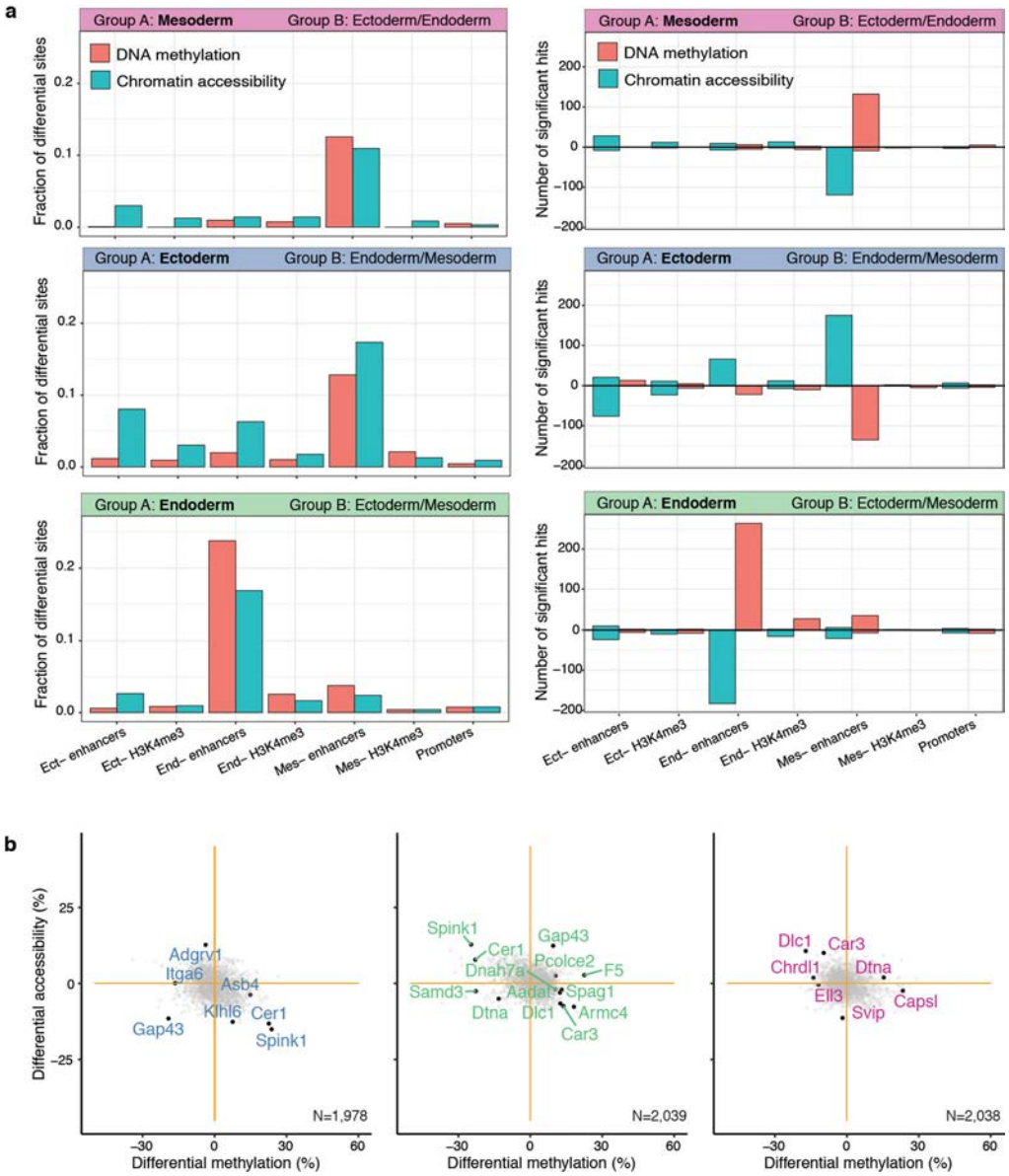


Extended Data Fig. 4 | DNA methylation and chromatin accessibility changes in promoters are associated with repression of early pluripotency and germ cell markers. a, Volcano plots display differential RNA-expression levels between E4.5 and E7.5 cells (in $\log_2$ counts, *x* axis) versus adjusted correlation *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.

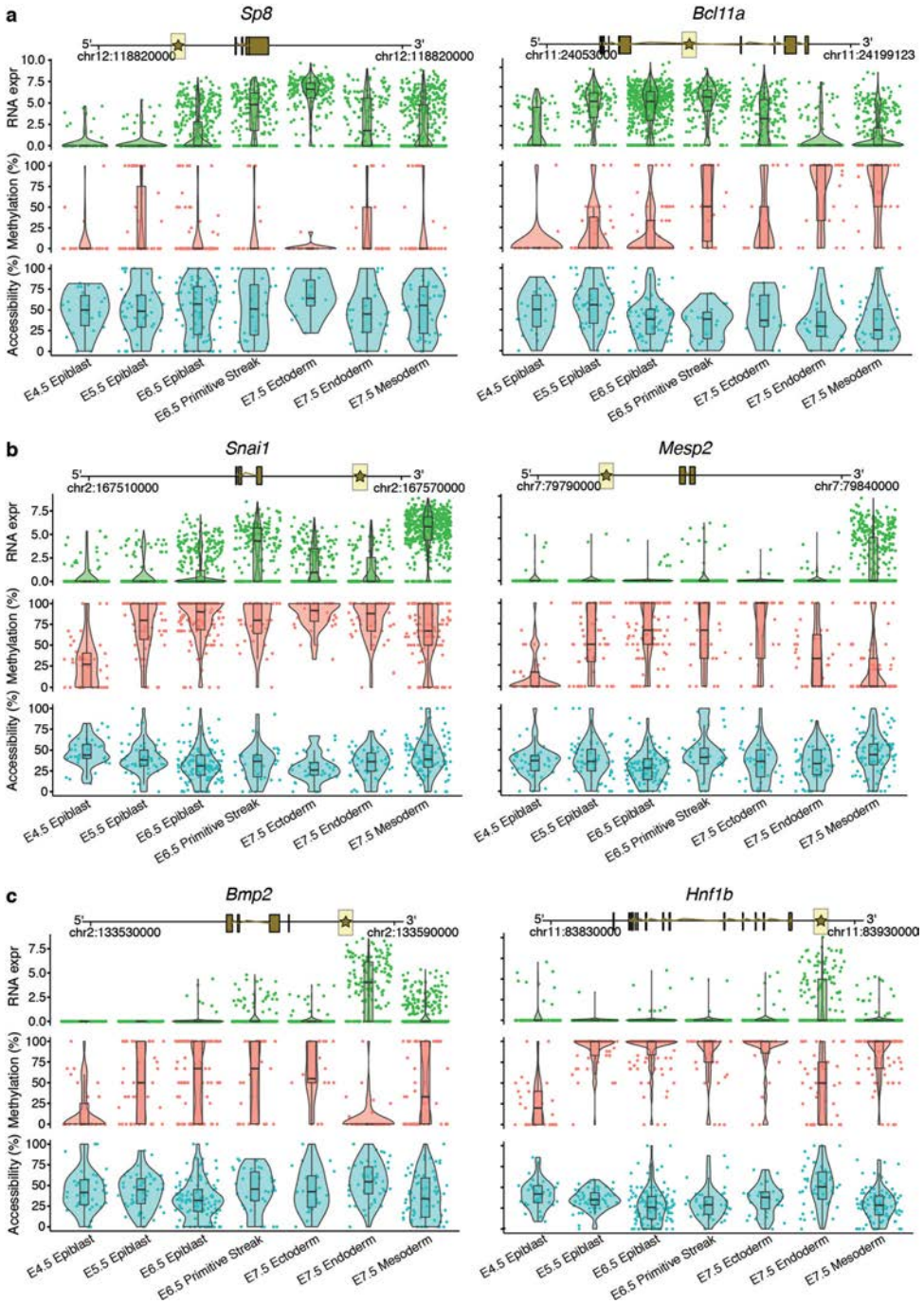


Extended Data Fig. 5 | See next page for caption.

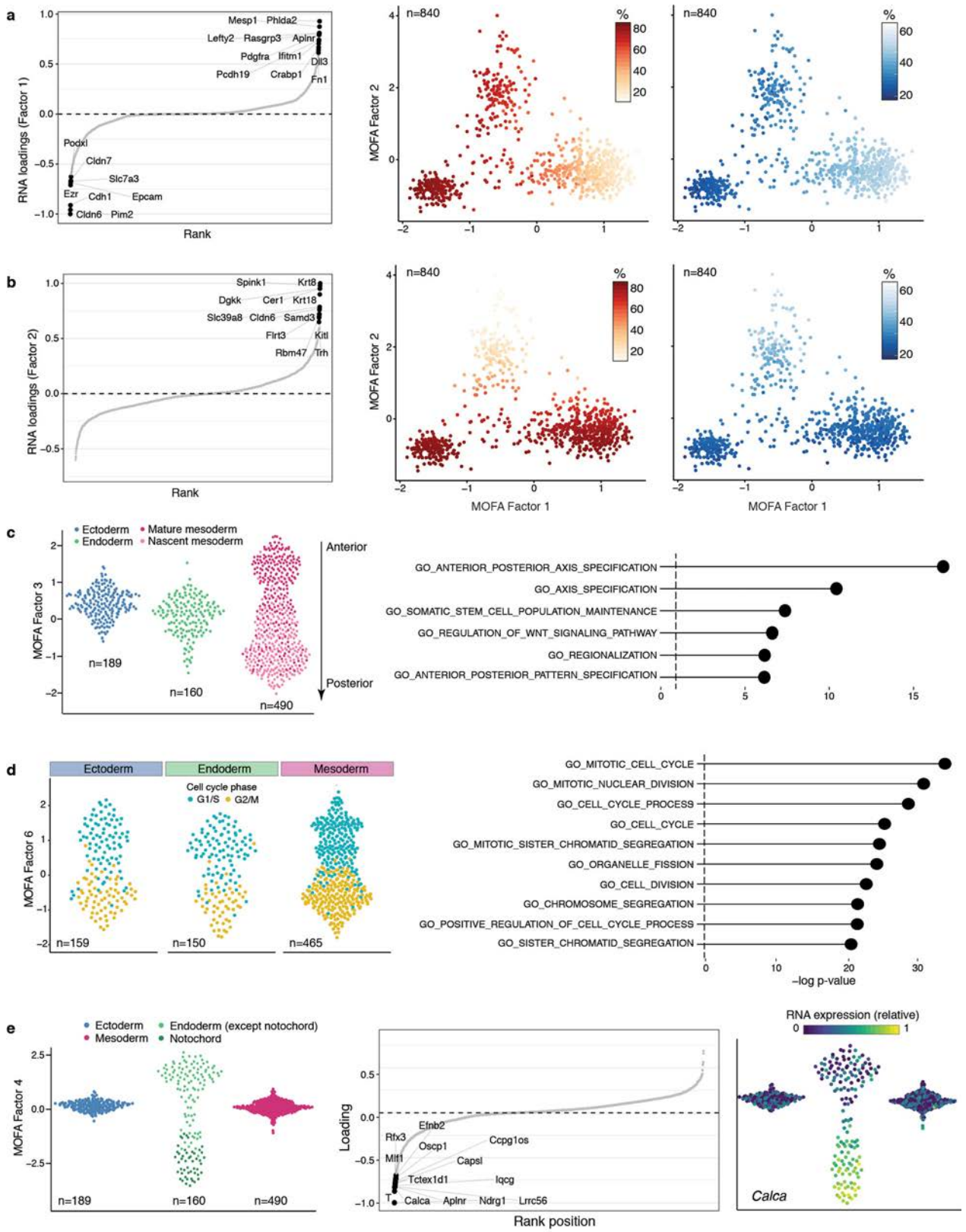
Extended Data Fig. 5 | Characterization of lineage-specific H3K27ac and H3K4me3 ChIP–seq data. a, Percentage of peaks overlapping promoters (±500 bp of TSS of annotated mRNAs (Ensembl v.87); lighter colour) and not overlapping promoters (distal peaks, darker colour). H3K27ac peaks tend to be distal from 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%).

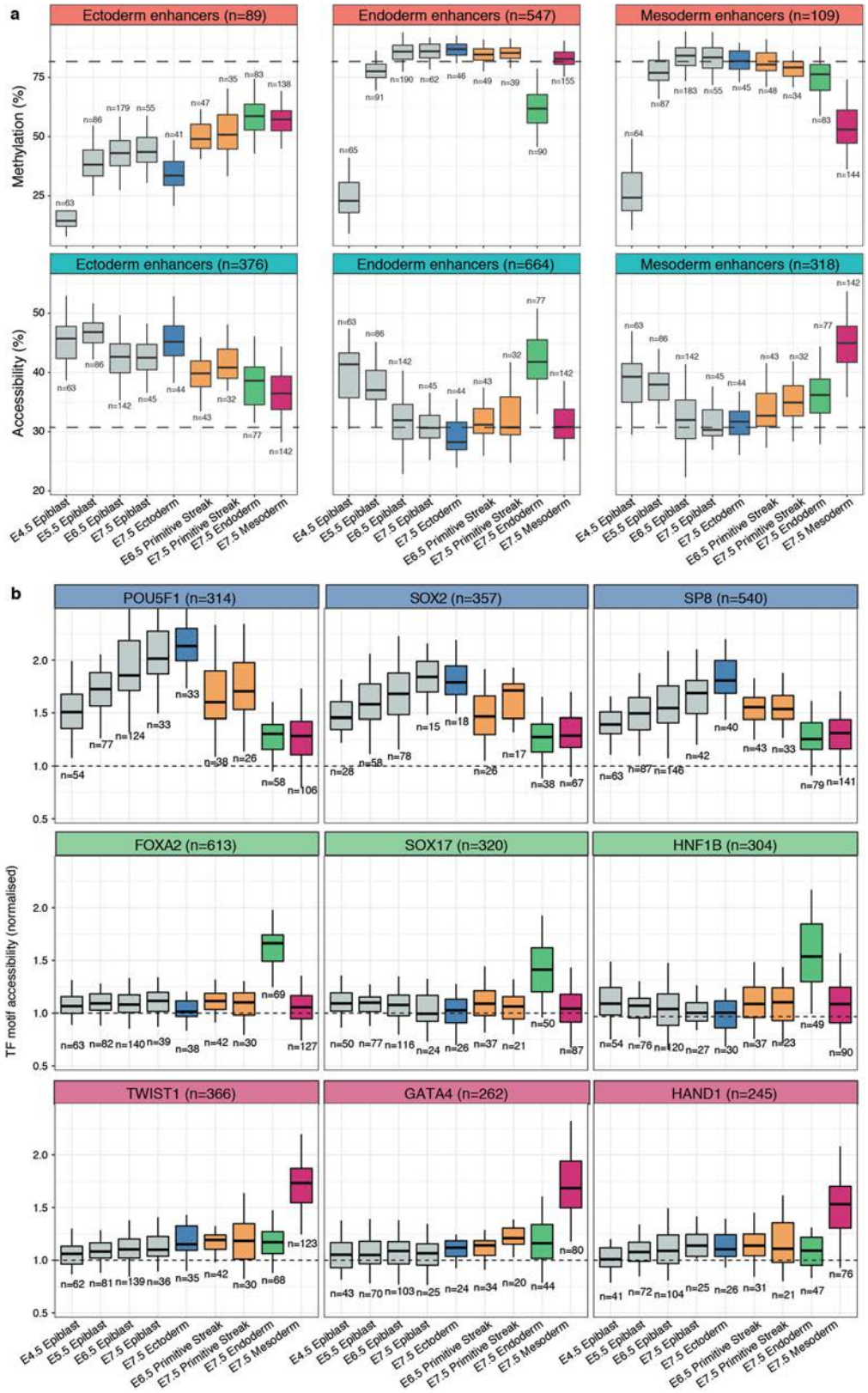


Extended Data Fig. 7 | Illustrative examples of putative epigenetic regulation in enhancer elements during germ-layer commitment. a–**c**, Box and violin plots showing the distribution of RNA expression (log_2 counts, green), enhancer DNA methylation (red) and chromatin accessibility (blue) levels for 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.

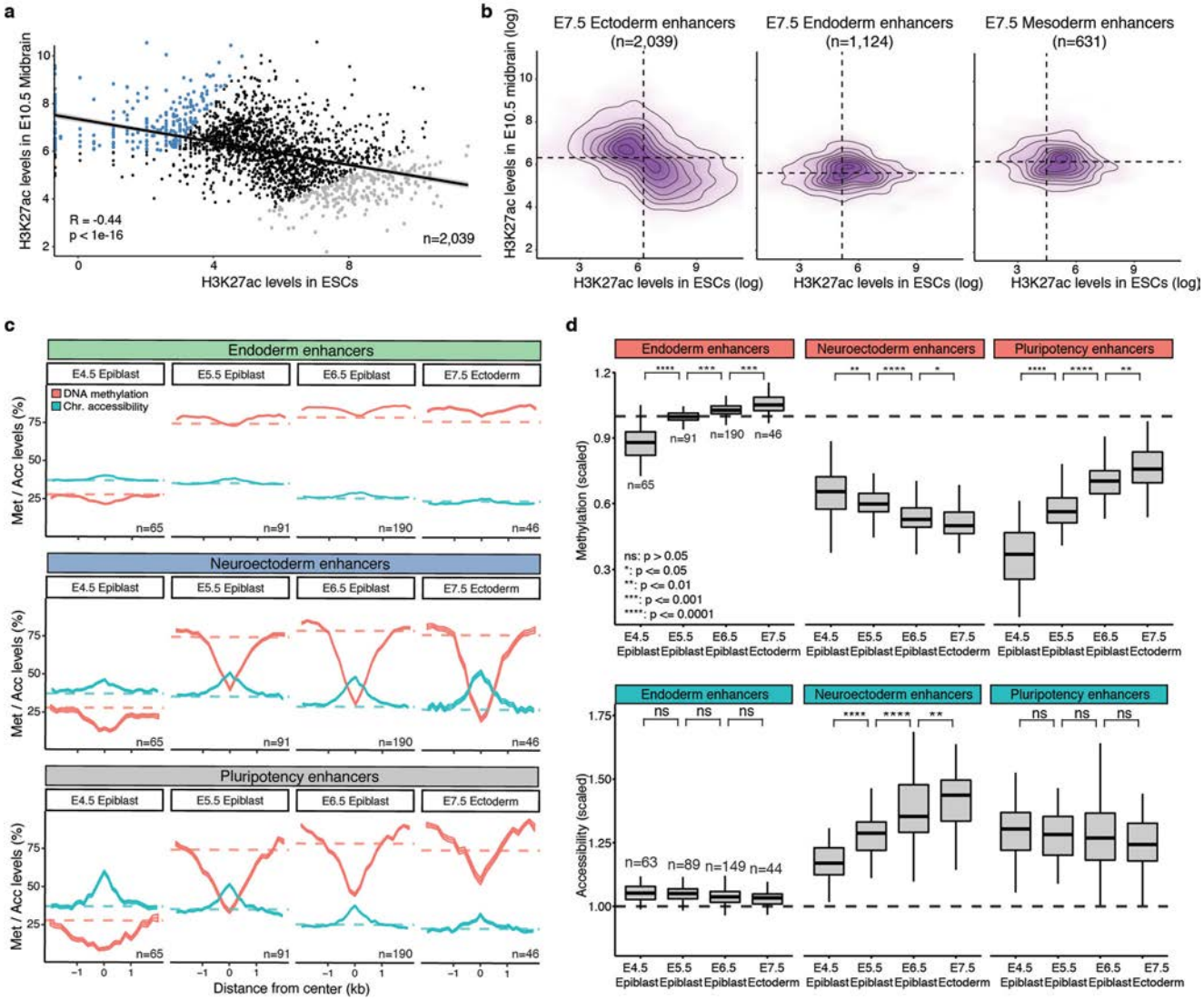


Extended Data Fig. 8 | See next page for caption.

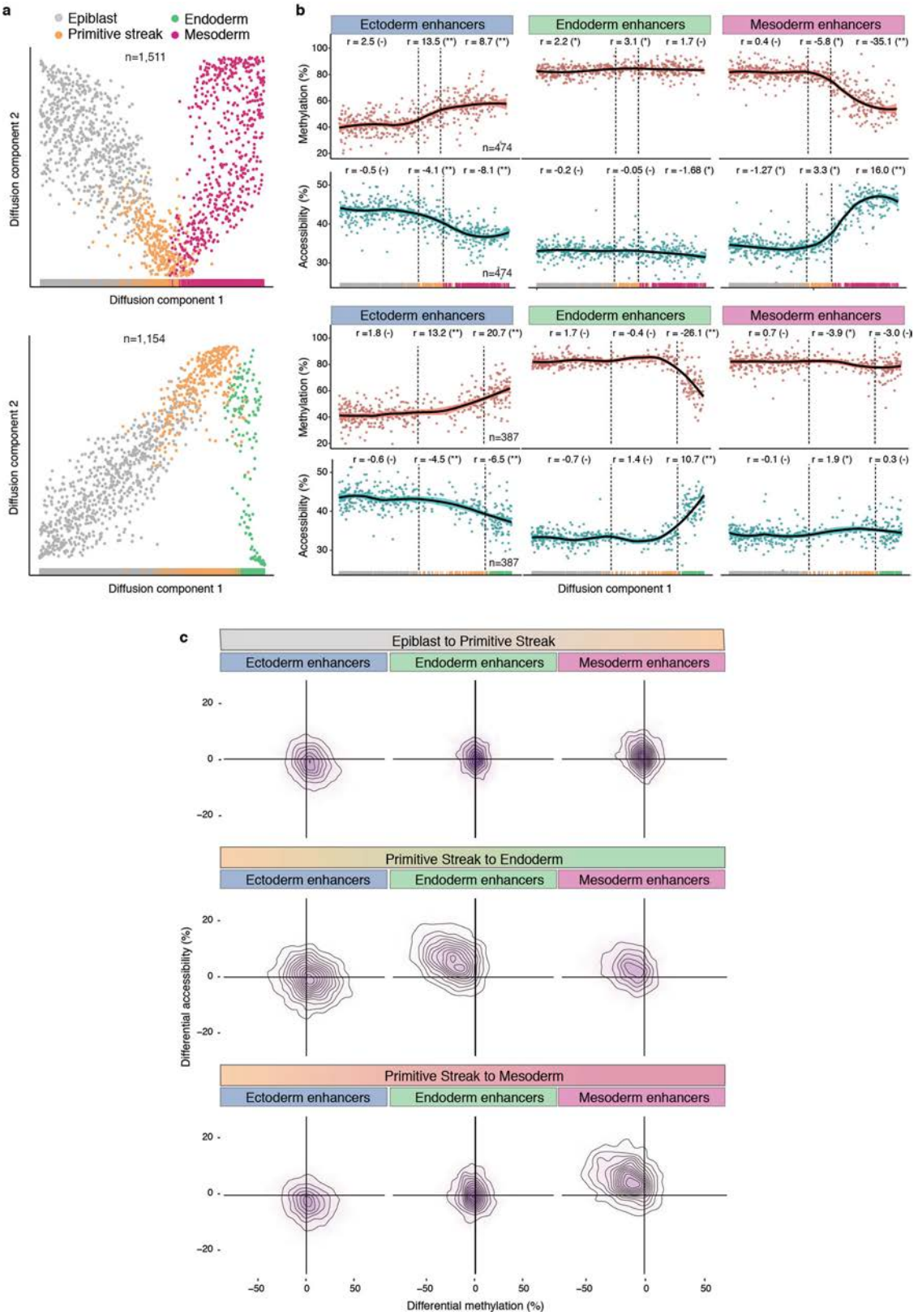
Extended Data Fig. 8 | Characterization of MOFA factors. a, Factor 1 as mesoderm commitment factor. Left, RNA-expression loadings for factor 1. Genes with large positive loadings increase expression in the positive factor values (mesoderm cells). Middle, scatter plot of factor 1 (*x* axis) and factor 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.

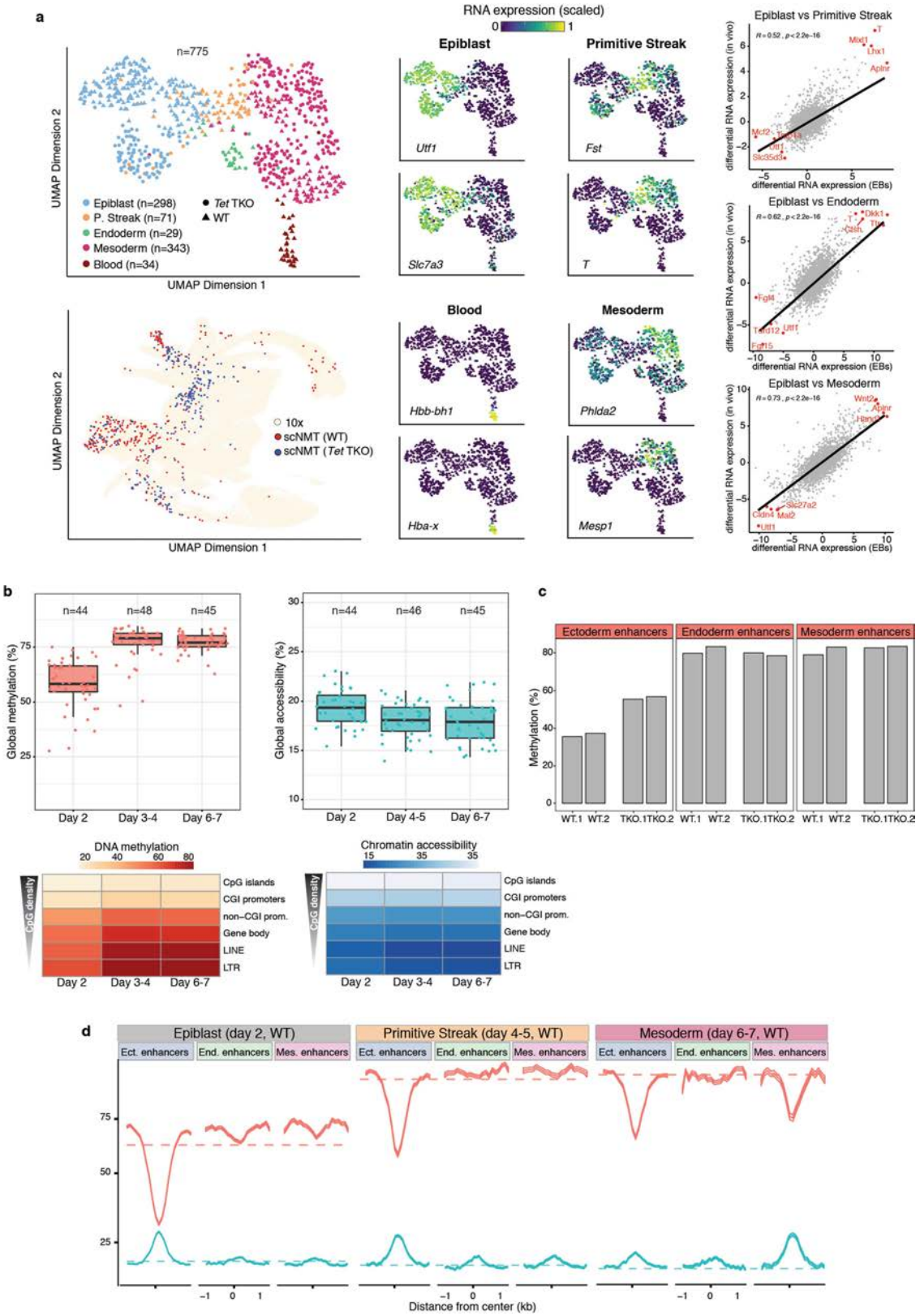


Extended Data Fig. 10 | E7.5 ectoderm enhancers contain a mixture of pluripotency and neural signatures with different epigenetic dynamics. a, Scatter plot showing H3K27ac levels for individual ectoderm enhancers ($n = 2,039$) quantified in serum-grown ES cells (pluripotency enhancers, x axis) versus 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.

OPEN ACCESS Freely available online

PLoS BIOLOGY

The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko[1☯], Sylvain Foret[2☯], Robert Kucharski[3], Stephan Wolf[4], Cassandra Falckenhayn[1], Ryszard Maleszka[3]*

1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, **2** ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, **3** Research School of Biology, the Australian National University, Canberra, Australia, **4** Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany

Abstract

In honey bees (*Apis mellifera*) the behaviorally and reproductively distinct queen and worker female castes derive from the same genome as a result of differential intake of royal jelly and are implemented in concert with DNA methylation. To determine if these very different diet-controlled phenotypes correlate with unique brain methylomes, we conducted a study to determine the methyl cytosine (mC) distribution in the brains of queens and workers at single-base-pair resolution using shotgun bisulfite sequencing technology. The whole-genome sequencing was validated by deep 454 sequencing of selected amplicons representing eight methylated genes. We found that nearly all mCs are located in CpG dinucleotides in the exons of 5,854 genes showing greater sequence conservation than non-methylated genes. Over 550 genes show significant methylation differences between queens and workers, revealing the intricate dynamics of methylation patterns. The distinctiveness of the differentially methylated genes is underscored by their intermediate CpG densities relative to drastically CpG-depleted methylated genes and to CpG-richer non-methylated genes. We find a strong correlation between methylation patterns and splicing sites including those that have the potential to generate alternative exons. We validate our genome-wide analyses by a detailed examination of two transcript variants encoded by one of the differentially methylated genes. The link between methylation and splicing is further supported by the differential methylation of genes belonging to the histone gene family. We propose that modulation of alternative splicing is one mechanism by which DNA methylation could be linked to gene regulation in the honey bee. Our study describes a level of molecular diversity previously unknown in honey bees that might be important for generating phenotypic flexibility not only during development but also in the adult post-mitotic brain.

Citation: Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, et al. (2010) The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers. PLoS Biol 8(11): e1000506. doi:10.1371/journal.pbio.1000506

Academic Editor: Laurent Keller, University of Lausanne, Switzerland

Received May 25, 2010; **Accepted** August 24, 2010; **Published** November 2, 2010

Funding: Work in FL's lab was supported by a grant from the Ministerium fur Wissenschaft, Forschung und Kunst Baden-Wurttemberg. Work in RM's lab was supported by the Australian Research Council grant DP1092706. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: DMG, differentially methylated gene; DNMT3, DNA methyltransferase 3; mCpG, methylated CpG; o/e, observed/expected

* E-mail: maleszka@rsbs.anu.edu.au

☯ These authors contributed equally to this work.

Introduction

Many animal species have evolved the capacity to generate organisms with contrasting morphological, reproductive, and behavioral phenotypes from the same genome. However, the question of how such strikingly different organismal outputs occur with no standard genetic changes remains one of the key unresolved issues in biology.

The nutritionally controlled queen/worker developmental divide in the social honey bee *Apis mellifera* is one of the best known examples of developmental flexibility in any phylum. Despite their identical nature at the DNA level, the queen bee and her workers are strongly differentiated by their anatomical and physiological characteristics and the longevity of the queen [1]. Furthermore, the behaviors of queens and workers are remarkably divergent, varying from the navigational proficiency of foragers to the colony-bound omnipresent chemical influences of the queen which control many aspects of the colony's existence. A diet of royal jelly during larval development clearly influences the epigenetic status of the queen's cells without altering any of the hardwired characteristics of her genome. As a result, two contrasting organismal outputs, fertile queens and non-reproductive workers, are generated from the same genome.

Recently, we have shown that diet is not the only modulator of developmental trajectories in honey bees. By silencing the activity of DNA methyltransferase 3 (DNMT3), a key component of epigenetic machinery controlling global gene reprogramming, we were able to generate adult bees with queen characteristics [2]. This relatively simple perturbation of the DNA methylation system not only mimicked the dietary effect of royal jelly on phenotype but also changed the cytosine methylation pattern of an illustrator gene. Furthermore, analysis of gene expression in both queens and workers suggested that their alternative developmental pathways are associated with subtle transcriptional changes in a particular

Author Summary

The queen honey bee and her worker sisters do not seem to have much in common. Workers are active and intelligent, skillfully navigating the outside world in search of food for the colony. They never reproduce; that task is left entirely to the much larger and longer-lived queen, who is permanently ensconced within the colony and uses a powerful chemical influence to exert control. Remarkably, these two female castes are generated from identical genomes. The key to each female's developmental destiny is her diet as a larva: future queens are raised on royal jelly. This specialized diet is thought to affect a particular chemical modification, methylation, of the bee's DNA, causing the same genome to be deployed differently. To document differences in this epigenomic setting and hypothesize about its effects on behavior, we performed high-resolution bisulphite sequencing of whole genomes from the brains of queen and worker honey bees. In contrast to the heavily methylated human genome, we found that only a small and specific fraction of the honey bee genome is methylated. Most methylation occurred within conserved genes that provide critical cellular functions. Over 550 genes showed significant methylation differences between the queen and the worker, which may contribute to the profound divergence in behavior. How DNA methylation works on these genes remains unclear, but it may change their accessibility to the cellular machinery that controls their expression. We found a tantalizing clue to a mechanism in the clustering of methylation within parts of genes where splicing occurs, suggesting that methylation could control which of several versions of a gene is expressed. Our study provides the first documentation of extensive molecular differences that may allow honey bees to generate different phenotypes from the same genome.

group of genes encoding conserved physio-metabolic proteins [2,3]. These findings prompted us to examine the hypothesis that significant behavioral differences between queens and workers are partly underpinned by differences between their brain epigenomes that have arisen from basically identical genomes during development. The choice of brain tissue is critical because it is a non-dividing, largely diploid tissue and is thus free of any complications that arise from differential genomic replication that may characterize polytene and endopolyploid tissues (nearly all adult tissues of insects are non-diploid). In the context of methylomes, the use of whole bodies, or abdomens, creates an unacceptable mixture of methylomic signatures that simply cannot be deconvoluted in regards to function in any biologically meaningful manner.

We used bisulfite converted brain DNA of both castes together with Solexa (Illumina GA) sequencing technology [4] to generate a DNA methylation map at single-nucleotide resolution across the *Apis* genome. This powerful approach has recently been used to compare DNA methylation profiles across a group of selected species, including DNA from a worker honey bee whole body [5]. The results confirm the antiquity of DNA methylation in eukaryotes [6,7] and provide more experimental evidence that this epigenomic modification is utilized in a lineage-specific manner [8–10].

Here we confirm that in contrast to heavily methylated mammalian genomes [11], only a small and specific fraction of the honey bee genome is methylated [5,10,12,13]. Furthermore, the methylated cytosines occur in a group of genes showing a higher level of conservation than non-methylated genes. Nearly 600 of those genes show significant methylation differences in the brains of queens and workers, suggesting that their transcription might be epigenetically modulated in a context-dependent manner. Additional deep sequencing of selected genes in all three castes—queens, workers, and drones (haploid males)—suggests that brain methylation patterns are unique to each behavioral system. We discuss our findings in the context of epigenetic influences on global regulatory networks and their ability to generate contrasting phenotypic and behavioral outcomes from the same genome.

Results

Characterization of Brain Methylomes in Queens and Workers

The sequencing of bisulfite converted *Apis* DNA yielded a dataset of 131 million reads after filtration and quality checks, 68.5% of which were mapped to unique genomic regions. The total sequence output was 18.8 giga bases (10.2 Gb for the queen and 8.6 Gb for the worker) yielding a combined 20× coverage of the 260 Mb genome. Our reads also contained multiple coverage of thousands of unmethylated repeated elements (ALUs and mariners) giving false-positive rates of only 0.1% for the queen DNA and 0.2% for the worker DNA. Figure S1A shows the distribution of the coverage depth for all cytosines on both strands, whereas distribution of the CpG nucleotides is shown in Figure S1B. More than 90% of the 10,030,209 CpGs in the *Apis* genome were covered by at least two sequencing reads, allowing for the methylation status of individual sites to be determined with confidence.

The characteristics of the brain methylomes of queens and workers are shown in Tables 1 and 2. Three firm conclusions can be drawn. First, of the over 60 million cytosines that exist in the *Apis* genome, only approximately 70,000 are methylated. Second, nearly all the methylated cytosines occur in CpG dinucleotides. Third, the overriding majority of these methylated sites are in exons. Finally, the number of methylated cytosines in *Apis* is nearly three orders of magnitude lower than in the human genome [11]. This relatively small number of mCs overcomes the large technical hurdles that exist in both mammalian and plant genomes where the number of methylated sites that need to be examined in terms of their importance to biological phenomena is in the hundreds of millions.

As shown in Table 1 the quantities of methylated CpGs (mCpGs) in queen and worker brain DNA are very similar, 69,064 and 68,222, respectively, with 54,312 mCpGs in common. Similarly, the methylation levels of mCpG are almost identical in both castes (Figure S2). Methylation in honey bees appears to be restricted to cytosines associated with CpG dinucleotides, with no significant non-CpG or asymmetric methylation detected in either genomic or mitochondrial DNA (Table 1). Therefore, we conclude that methylation at non-CpG sites is either extremely rare or non-existent in the honey bee genome. In accord with previous analyses [2,5,12,13], methylated sites in *Apis* appear to be exclusively located in exons with only infrequent mCpGs detected in intronic regions (Table 2). Most importantly, the methylated exons reside in genomic regions with low CpG observed/expected (o/e) ratios (Figure 1), whereas non-methylated exons fall into the category with high CpG o/e ratios. This bimodal profile is consistent with previous predictions based on bioinformatics analyses [10,12,13] and reflects the propensity of methylated Cs to be converted over time to thymines, resulting in a lower than expected density of the CpGs in methylated genes. However, the

Table 1. Cytosine DNA methylation in queens and workers in CG, CHG, and CHH genomic contexts (H = A, T, or C).

	Total	Methylated in Queens	Methylated in Workers	Methylated in Both Castes
CG	10,030,209	69,064	68,222	54,312
CHG	8,673,113	14	130	0
CHH	45,072,611	561	3,019[a]	0

The thresholds used for methylation calls are detailed in the Methylation Assessment section.
[a]Nearly all of the 3,019 CHH that were inferred to be methylated in worker brains on the basis of Solexa reads were found to be not methylated by an additional sequencing of selected amplicons using the 454 technology.
doi:10.1371/journal.pbio.1000506.t001

total number of methylated genes in *Apis* revealed by genome-wide bisulfite sequencing is 5,854 instead of the 4,000 predicted to be methylated on the basis of local CpG bias. One reason for this difference might be that some genes do not display significant CpG depletion as a result of evolutionary pressure to maintain a particular protein coding sequence.

The genome-wide profiling of mCpGs confirms that methylated genes in *Apis* encode proteins showing a higher degree of conservation than proteins encoded by non-methylated genes [10]. Figures S3, S4, S5 and Table S1 show the results of our cross-species comparisons for methylated and non-methylated genes (Figure S3), for high-CpG and low-CpG genes (Figure S4), and high-CpG methylated and non-methylated genes (Figure S5). Most of the highly conserved genes are expected to be utilized by most tissues. In contrast, less conserved genes expressed in specialized tissues, such as those encoding odorant-binding proteins or odorant receptors, are not methylated (not shown). The repeated elements, ALUs, and mariners that harbor most of the DNA methylation content in humans and plants are *not* methylated in the bee genome, certainly not in the brain (Figures S6). Similarly, the multi-gene families encoding rRNAs and tRNAs, mitochondrial DNA, and CpG islands show no evidence of methylation in the brain (Figure S6). Lastly, while methylation of sub-telomeric regions has been shown to be important for the control of telomere length and recombination [14], the honey bee telomeres are also not methylated (not shown). The lack of methylation in ALUs and transposons has also been reported in a recent study performed on DNA extracted from a worker's whole body [5]. Given the proposed role of cytosine methylation in defense against genomic parasites in plants and vertebrates [7], the lack of methylation in ALU repeats and mariner transposons suggests that these mobile elements do not significantly impact on genome stability in honey bees. Indeed the bee genome contains an unusually small percentage of common types of transposons and retrotransposons found in other insects, possibly as a result of a strong selective pressure against mobile elements in male bees (drones) that develop from unfertilized eggs and carry a haploid set of chromosomes [15].

As in the human and *Arabidopsis* genomes [4,11], methylation in *Apis* shows evidence of periodicity, although due to a much lower density of modified CpGs in this species the periodicity of 10 nucleotides (one helical DNA turn) is not obvious. However, a 3-base periodic pattern is clearly detectable, reflecting a preferential methylation of CpGs occupying the first and second position of the arginine codons (autocorrelation data in Figure S7).

Detailed Analysis of Methylation Patterns in Selected Amplicons by Deep Bisulfite Sequencing

To validate our Solexa-based methylation results, we designed primers for selected regions of eight nuclear and four mitochondrial genes and re-sequenced the PCR-generated amplicons using 454 technology. As illustrated in Figure 2, the 454 sequencing profiles are essentially identical with the Solexa-based results. All nuclear genes show differential methylation in the brains of queens and workers, including those cases where the methylation is almost absent, such as GB18602 in queen brains (Figure 2). No methylation was detected by this approach in the four selected mitochondrial amplicons (not shown).

To further expand our analysis, we increased the 454 bisulfite sequencing coverage of the eight nuclear genes selected for validation and also included DNA from drone brains. We obtained several thousand high-quality reads for 24 amplicons (eight genes in three castes), with the total coverage ranging from 48 to 2,427×. The results shown in Figure 3 reveal both the dynamics and uniqueness of the methylation patterns in each cast. Out of the eight genes with differential worker/queen methylation, three show similar methylation patterns in workers and drones, but a distinct methylation pattern in queens (Figure 3A). Three additional genes show similar methylation patterns in queens and drones, but a distinct pattern in workers (Figure 3B). Two out of eight analyzed genes (GB11061 - seryl-tRNA synthetase and GB15356 - syd, chromosome segregation; Figure 3C) show distinct

Table 2. Cytosine DNA methylation in CG dinucleotides (mCG) in the exonic, intronic, and "intergenic" regions of queens and workers.

Genomic Location	Queens	% mCGs	% of All mCGs	Workers	% mCGs	% of all mCGs
Exons	54,378	8.6	78.74	51,658	8.16	75.72
Introns	5,992	0.2	8.68	6,720	0.22	9.85
Introns + exons	60,370	1.64	87.41	58,378	1.57	85.57
Intergenic regions[a]	8,694	0.16	12.59	9,844	0.17	14.43

[a]The annotation of the *Apis* transcriptome is largely limited to the coding regions, and it is likely that some of the intergenic regions may correspond to untranslated segments of mRNAs.
doi:10.1371/journal.pbio.1000506.t002

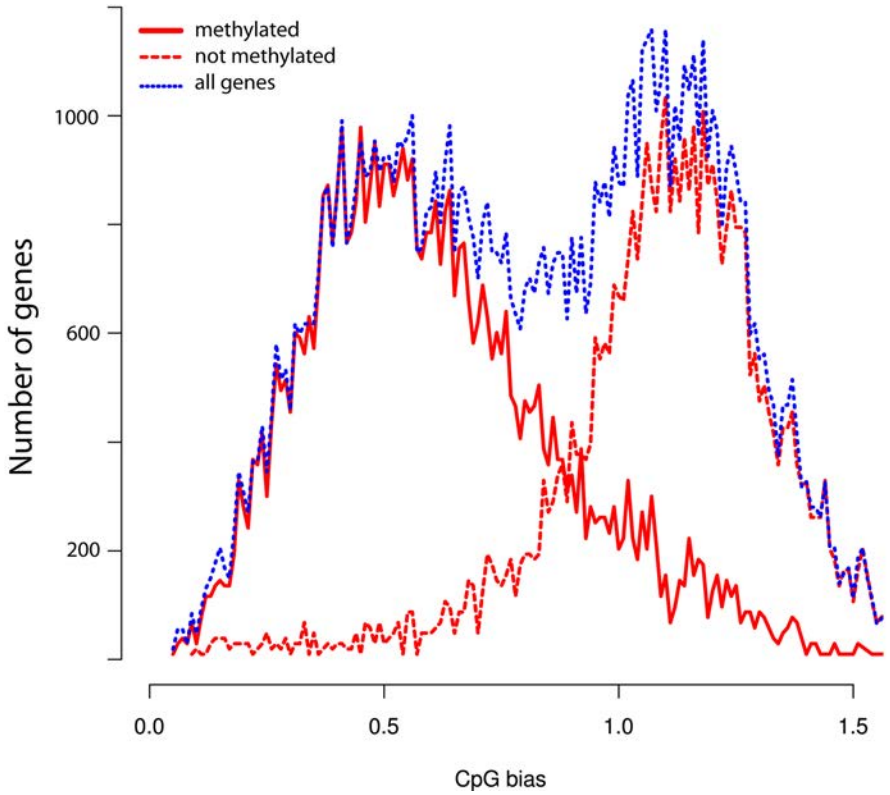


Figure 1. CpG (o/e) bias of protein-coding regions in the honey bee genome. Since the profiles for both queens and workers are virtually identical, only the queen profile is shown.
doi:10.1371/journal.pbio.1000506.g001

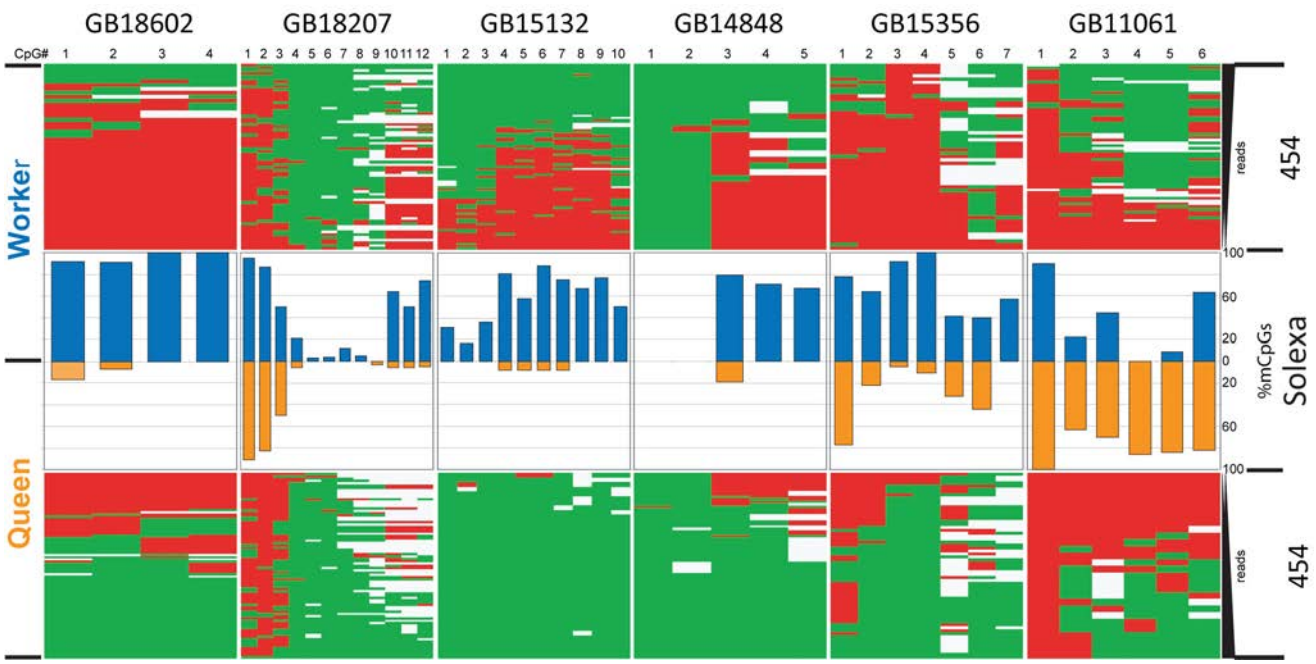


Figure 2. Comparison of CpG methylation profiles in differentially methylated genes generated by two technologies, Solexa genome-wide shotgun sequencing and 454 sequencing of PCR produced amplicons. The ''heat maps'' represent the 454 sequencing of PCR amplified segments, whereas the bars illustrate the Solexa reads. The eight nuclear genes for this experiment were chosen from the list of DMGs shown in Tables 3 and S2, taking into account the availability of convenient CpG-containing regions for primer design. Six genes are shown in this figure and the others in Figure 3. Gene annotations: GB18602 - membrane protein; GB18207 - cadherin; GB15132 - TAP42 (TOR signaling); GB14848 - clathrin assembly protein; GB15356 - syd, chromosome segregation; GB11061 - seryl-tRNA synthetase.
doi:10.1371/journal.pbio.1000506.g002

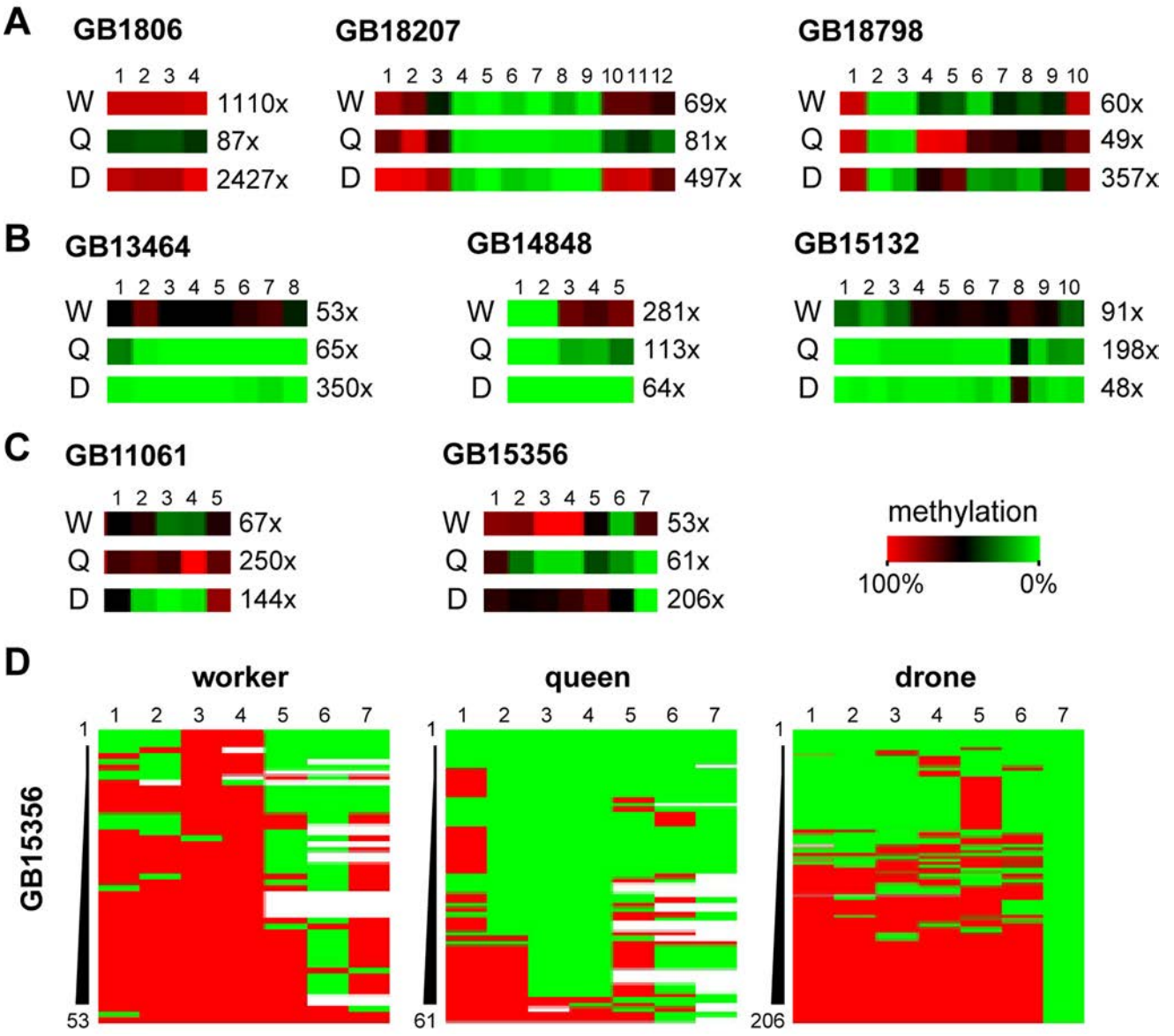


Figure 3. Detailed analysis of deep sequencing of selected genes. The bisulfite converted amplicons of selected genes were sequenced using 454 technology. The selection was based on differential methylation in brains of queens and workers, but DNA from male brains (drones) was also used in this experiment. The panels illustrate the uniqueness of brain methylation patterns in bees. 3A: Genes showing similar methylation patterns in workers and drones, but a distinct methylation pattern in queens. 3B: Genes with similar methylation patterns in queens and drones, but a distinct pattern in workers. 3C: Gene with distinct methylation patterns in all three castes. Panel 3D shows the full methylation heatmaps of GB15356. This result is discussed in the chapter "Detailed Analysis of Methylation Patterns in Selected Amplicons by Deep Bisulfite Sequencing." Gene annotations: GB18798 - ubiquitin conjugation factor; GB13464 - RhoGAP93B. For other genes, see Figure 2.
doi:10.1371/journal.pbio.1000506.g003

methylation patterns in all three castes. The latter finding was also confirmed by the analysis of full methylation heatmaps of GB15356 (Figure 3D). GB15356 is strongly methylated in workers, with many reads showing complete methylation in the 5′-half of the amplicon (Figure 3D). In queens, GB15356 methylation is strongly reduced and many reads show no methylation at all. Intriguingly, drones show a bimodal methylation pattern with approximately half of the reads methylated and the other half unmethylated (Figure 3D). These results further illustrate caste-specific differences in methylation patterns and suggest a complex role of DNA methylation in the regulation of caste-specific epigenomic differences in the brain.

Identification of Differentially Methylated Genes

To determine if there is a link between DNA methylation patterns and the striking morphological and behavioral polymorphisms of queen bees and workers, we examined the levels of CpG methylation in all annotated transcription units in both brains using high stringency criteria (Supporting Information). This approach generated a list of 561 differentially methylated genes (DMGs, Tables 3 and S2) showing significant methylation differences between the two castes. With the exception of highly expressed genes encoding ribosomal proteins, DMGs in *Apis* are expressed at low or moderate levels across all analyzed tissues (Tables 3 and S2). In several cases their transcriptional activities

Table 3. Differentially methylated genes in brains of queens and workers.

				Relative Expression[b]				
Gene ID[a]	**No. of CpGs**	**Antenna**	**Brain**	**HPG**	**Larva**	**Ovary**	**Thorax**	**Gene Annotation**
GB18602	30	1	1	1	1	1	1	Transmembrane protein YhhN
GB18303	13	1	1	1	1	1	1	Activator protein of Rab-like small GTPases
GB13368	9	2	1	2	10	1	3	3-hydroxyacyl-CoA dehydrogenase, NAD-binding
GB13215	34	1	1	1	1	1	1	Glycine cleavage system P-protein,
GB15588	9	1	1	1	1	1	1	Low-density lipoprotein receptor domain class A
GB15132	24	1	1	1	1	1	1	TAP42 (regulates the TOR signaling pathway)
GB12560	12	1	1	1	1	9	1	RNA-binding protein
GB11648	13	1	1	1	1	2	1	Catalase
GB19645	12	1	1	1	1	1	1	Phosphodiesterase 6
GB12929	39	1	1	1	1	1	1	Paralytic - Na channel
GB11421	31	1	1	1	1	1	1	Tight junction associated protein
GB19503	33	1	1	1	1	1	1	Heat shock protein 8
GB13740	24	1	1	1	1	1	1	Dysfusion, TF with PAS domain
GB10394	8	1	1	1	1	1	1	TNF-receptor-associated factor 1
GB16628	9	10	6	8	10	10	10	Ribosomal protein L6

Only the top 15 genes are shown; see Table S2 for list of 561 genes that fall into this category. Based on microarray data from Foret et al. [10]. The genome assembly v.02 was used throughout this study.

[a]GB numbers refer to the proteins at BeeBase: genomes.arc.georgetown.edu/drupal.

[b]Genes were ranked into 10 bins based on their expression levels from low (1) to high (10).

doi:10.1371/journal.pbio.1000506.t003

were found to be significantly up-regulated in some tissues relative to others. For example, the expression of 3-hydroxyl-CoA dehydrogenase (GB13368) is much higher in the larva than in the adult brain, and RNA-binding protein (GB12560) is significantly up-regulated in the ovaries relative to other tissues (Table 3). Almost all DMGs encode highly conserved, well-characterized proteins that have been implicated in core processes such as metabolism, RNA synthesis, nucleic acids binding, and signal transduction (Table S2). While a number of genes could not be clearly assigned to functional categories, their high level of conservation across phyla indicates that they are nevertheless likely to be involved in essential cellular processes (e.g. GB18943, GB13480, and GB18037). Several differently methylated genes encode proteins previously shown to be involved in either brain development or activity-dependent neural functions in both vertebrates and invertebrates. These include the Ephrin receptor GB1258516 [16], a nicotinic acetylcholine receptor GB19703, "no extended memory" GB16408 that is encoded by cytochrome B561 in *Drosophila*, two NMDA receptors GB19334 and GB15722, and a membrane channel GB12287 that mediates cell adhesion. When defective, GB12287 results in the "big brain" phenotype (Table S2). We note that Dynactin, used in our previous study [2] to illustrate the methylation differences between the two castes during larval growth in both royal jelly-fed and RNAi-treated individuals, does not show differential methylation in the brain. However, two genes, GB11197 and GB13866, encoding proteins associated with the large Dynein complex to which Dynactin also belongs are differentially methylated in the brain. Thus, the multi-protein Dynein complex appears to be epigenetically modulated during larval growth and in adult brains.

CpG Bias and Epigenetic Modulation

Recently, Elango et al. [13] on the basis of bioinformatic analyses of a dataset of differentially expressed genes in brains of queens and workers proposed that "high-CpG genes in *A. mellifera* generally are more prone to epigenetic modulation than low-CpG genes." We have tested this hypothesis using our new caste-specific brain methylome data. The results summarized in Table S3 suggest that (a) the methylation of a gene is a decreasing function of its CpG richness (Figure S8), (b) the "caste-specific genes" [13] that are methylated have a lower CpG content than the non-methylated genes (Table S3), and (c) DMGs are over-represented in the low CpG genes (Table S3). Therefore, our results do not support the hypothesis of Elango et al. [13]. However, it is noteworthy that although the DMGs are generally CpG-depleted, they tend to be less CpG-depleted than those genes that are not differentially methylated (Table S3). This intermediate CpG density observed in DMGs underscores the uniqueness of this class of genes and suggests that they might be methylated in a distinct manner from the rest of methylated genes. This class of genes showing differential patterns of methylation associated with phenotypic polymorphism is thus of special importance in the study of complex context-dependent phenotypes.

Unraveling the Link between CpG Methylation and Splicing

To explore the relationship between differential methylation and expression patterns in queens and workers, we examined in more detail the first gene on the DMG list (GB18602) encoding a putative transmembrane protein with the YhhN domain conserved from bacteria to mammals. Figure 4 shows the distribution

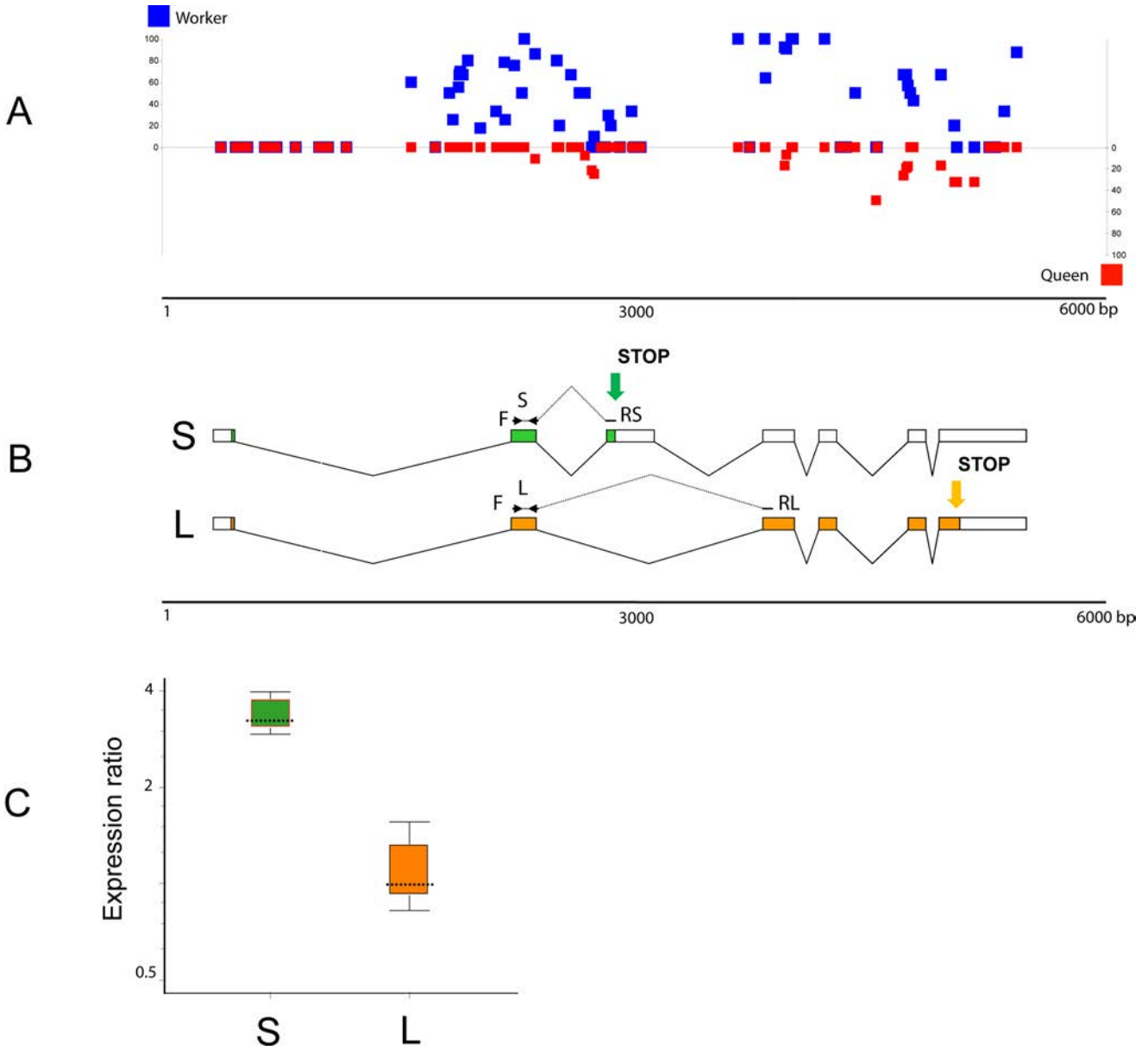


Figure 4. Expression profile of an alternatively spliced and differentially methylated gene GB18602 in queen and worker brains. (A) The CpG methylation pattern indicating the level of methylation for individual CpGs (blue squares, workers; red squares, queens). (B) Gene model of GB18602 showing the two spliced variants S (short protein) and L (long protein) and the positions of PCR primers used for variant-specific amplifications. The green and orange arrows indicate the positions of two alternative Stop codons. (C) Relative expression of the two spliced variants in brains of queens and workers examined by real-time PCR. The level of transcript S (green) encoding the truncated protein is significantly up-regulated in the queen brain, whereas the L variant (orange) is expressed at the same level in both castes. The queen expression represents a combined set of data from three independent queen samples: 4 mo old (1 brain), 12 mo old (2 brains), and swarm queens of unknown age (2 brains). Workers were 8 d old (6 brains in 3 replicates). The reference gene was calmodulin [2]. Whisker-box plot of expression ratio values: dotted line, median value; box, inter-quartile range of values; whiskers, outer 50% of observations. For more details, see Table S4.
doi:10.1371/journal.pbio.1000506.g004

of mCpGs against the GB18602 gene model (Figure 4A and 4B) and the relative expression of two spliced variants in both castes (Figure 4C). The L variant (L) encoding a long protein shows identical expression levels in both queens and workers, whereas the S variant (S) encoding a short protein is significantly up-regulated in queen relative to worker brain (Figure 4C). The majority of the differentially methylated sites in the GB18602 locus map to the region spanning the additional cassette-exon that contains a Stop codon for the short protein encoded by the S transcript, suggesting a correlation between methylation and the outcome of alternative splicing of this gene in *Apis*. The increased level of methylation spanning the conditional splicing event (insertion or skipping of the cassette-exon) in the worker brain may impede the inclusion frequency of this exon into the mature transcript. Since the L variant is expressed at the same levels in both castes, the increased methylation in workers appears to be specifically affecting splicing, but not transcription. The observed differential pattern of expression of both transcripts in the brains of queens and workers (Figure 4C) supports this idea. Although the function of this gene is not known, the expression profiles of the

Drosophila melanogaster ortholog CG7582 suggest that it encodes a protein involved in fat and sugar metabolism [17]. In the fly, which has no CpG methylation, this gene is not alternatively spliced and shows the highest levels of expression in the nervous system (FlyAtlas.org). In contrast, the human ortholog of GB18602, designated TMEM86A, produces alternatively spliced variants, including one encoding a truncated protein similar to the honey bee variant S. In addition to GB18602 we found numerous other examples of methylated genes in *Apis* in which most or even all clusters of mCpGs show a non-random, highly significant tendency to be near differentially spliced exons (Figure S9). Another salient finding relevant to methylation of intron-containing genes is the differential methylation of the multi-gene histone family in *Apis*. As illustrated in Table 4 and Figure S10, all intron-containing histone genes are methylated, whereas intronless histone genes show no evidence of methylation. It is noteworthy that the methylated histone genes in *Apis* belong to a distinct class of histone variants. Unlike the canonical histones these variants are expressed constitutively and independently of replication and act as multifunctional regulators in a range of processes including DNA repair, transcription initiation and termination, meiotic recombination, etc. [18]. It is believed that they represent lineage specific innovation that is important for each organism's evolutionary specialization [18].

Discussion

The discovery of a functional DNA methylation system in honey bees and other invertebrates [1,7–10,19] has brought a fresh perspective to the study of epigenetic regulation of development and behavior. It reinforced the view that this covalent modification of DNA is an ancient and widely utilized evolutionary mechanism that was present in the basal Metazoa and has been recruited to serve diverse functions in modern organisms, including regulation of gene expression, cell differentiation, and silencing of transposons [20–22]. However, the trajectories from methylation changes to complex phenotypes are indirect, multi-level, and virtually unknown. For example, the hundreds of millions of methylated cytosines in the human genome and their large variation in different cell types in vivo pose a major challenge to uncovering those changes causative to phenotype. By contrast, the honey bee *Apis mellifera* shares its basic methylation enzymology with humans, yet as shown in this and other studies [5,10,12,13] only a small and specific fraction of its genome is methylated. The present results show that honey bees utilize methyl tags to mark a core of mostly conserved and ubiquitously expressed critical genes whose activities cannot be switched off in most tissues. Recent data suggest that in spite of their permanent expression these genes might not be required at the same level

Table 4. Annotation of the histone gene family in *Apis mellifera*.

Class	Proposed Subclass	Type	Apis Histone Genes	NCBI RefSeq mRNA	Proposed Gene ID	Proposed Protein ID	Splicing Status	Methyl CpG
H1	H1	canonical	GB12700[a]	XM_001121111	H1.1	H1.A		
			GB12218	XM_001122184	H1.2	H1.B		
H2A	H2A	canonical	GB18806	XM_001120186	H2A.1	H2A.A		
			GB12818	XM_001120346	H2A.2	H2A.B		
			GB13800	XM_001119899	H2A.3	H2A.C		
			N/A[c]	XM_001120934	H2A.4	H2A.D		
	H2A.X	**variant**	GB18954	XM_624697	H2A.X	H2A.X	**Spliced**	**Yes**
	H2A.Z/H2AV	**variant**	GB12991	XM_624164	H2A.Z	H2A.Z	**Spliced**	**Yes**
	Pseudogene		Pseudogene		ψH2A			
H2B	H2B	canonical	GB12700[a]	XM_001120238	H2B.1	H2B.A		
			GB13012	XM_001120889	H2B.2	H2B.B		
			N/A[c]	XM_001122218	H2B.3	H2B.C		
			GB12922[b]	XM_001119846	H2B.4	H2B.D		
			GB11889	XM_001120014	H2B.5	H2B.D		
H3	H3	canonical	GB11223	XM_001120304	H3.1	H3		
			GB14620	XM_001121026	H3.2	H3		
			N/A[c]	XM_001120132	H3.3	H3		
	H3.3	**variant**	GB12948	XM_001120696	H3.3.1	H3.3	**Spliced**	**Yes**
			GB11228	XM_624496	H3.3.2	H3.3	**Spliced**	**Yes**
	CENPA	**variant**	GB18566	N/A	CENPA	CENPA	**Spliced**	**Yes**
H4	H4	canonical	GB20104	XM_001120066	H4.1	H4		
			GB12644[a]	XM_01119948	H4.2	H4		
			GB14107	XM_001120988	H4.3	H4		
			GB17789	NM_001011609	H4.4	H4		

See Figure S10 for additional details. These Bee Base protein entries are either incorrect or missing:
[a]chimeras,
[b]truncated,
[c]not available.
doi:10.1371/journal.pbio.1000506.t004

throughout development, or under changing environmental conditions [23–25].

In honey bees, feeding of newly hatched larvae destined to become queens with royal jelly leads to metabolic acceleration and increased growth driven by global but relatively subtle changes in the expressional levels of a large number of ubiquitous genes [2,3,10]. These initial stages of larval development are later followed by the activation of more specific pathways to lay down caste-specific structures [3,10]. Interestingly, adult queen bees continue to be fed royal jelly, suggesting that this highly specialized diet is important for maintaining their reproductive as well as behavioral status. One possibility is that adult queens adjust their brain methylomes according to external instructions from their diet. One of the ingredients of royal jelly, phenyl butyrate [26], is a known histone deacetylase inhibitor and growth regulator that has been implicated in improving cognitive deficits in mice [27] and in life extension of *Drosophila* [28]. Although the significance of phenyl butyrate in royal jelly is not yet understood, it is conceivable that this complex diet evolved to provide two important functions for honey bees. It primarily serves as the source of nutrients for queen development but also as the regulator of epigenetic networks controlling gene expression in the brain. In addition to having different morphologies, reproductive capacities, and distinct behaviors, the genetically identical queen and worker honey bees also have different synaptic densities in their brains. In a recent study, Groh and Rossler [29] proposed that such developmental, diet-induced heterochrony results in fewer synapses in olfactory centers in queens, which may result in poorer performance on olfactory learning tasks compared to workers.

Recent studies using rodent models provided strong support for an idea that the nervous system has co-opted epigenetic mechanisms utilized during development for activity-dependent brain functions, including the generation and maintenance of long-term behavioral memories in adulthood [30,31]. Not surprisingly, DNA methylation has also been found to be involved in memory processing in honey bees [32], highlighting the significance of this epigenomic setting in conserved brain functions. These findings also provided evidence that DNA methylation, once believed to be an inert process after cellular differentiation, is dynamically regulated in the adult brain. Although both DNA methylation and chromatin remodeling have been implicated in these processes, the specific biological mechanisms underlying such adaptations remain largely unknown.

Our study provides experimental evidence that at least 560 differentially methylated ubiquitously expressed genes are involved in generating molecular brain diversity in female honey bees. Although it is still unclear how methylation might be linked to the gene regulatory networks, it has been proposed that DNA methylation together with changes in the histone profiles has the capacity to adjust DNA accessibility to cellular machinery by changing chromatin density [33–35]. Our findings support this notion and suggest that this mechanism provides an additional level of transcriptional control to fine tune the levels of messenger RNAs, including differentially spliced variants, encoded by the conserved genes. The association of mCpG clusters with alternatively spliced exons and genes containing introns in *Apis* is reminiscent of the distribution of mCpGs around the exon/intron junctions in human genes [36]. Epigenetic control of both splicing and mRNA levels might be utilized in different lineages, suggesting that a direct relationship between gene methylation and transcription is a widely spread phenomenon in both the animal and plant kingdoms [8,37].

Cytosine methylation may interact with other epigenetic features, such as distinctive histone modification signatures that have been shown to correlate with the splicing outcome in a set of human genes [33–35]. The correlation between methylation and splicing is further highlighted by the differential methylation of two classes of histone genes in *Apis*. We find that only intron-containing histone variants are methylated, whereas intronless canonical histone genes are not methylated. Interestingly, histone variants have been implicated in multiple conserved roles in eukaryotes [18] and therefore are part of the cellular maintenance systems together with other ubiquitously expressed genes. In a broader context, methylated cytosines may specify information to set up, proliferate, and regulate splicing patterns during cellular processes such as development and differentiation.

Thus, rather than switching the genes on and off by promoter methylation, the intragenic methylation in *Apis* operates as a modulator of gene activities. As a result the entire topology of a complex brain network can be reprogrammed by subtle adjustments of many genes that act additively to produce a given phenotype [38]. Such adjustable DNA methylation levels generating variability in the transcriptional output of methylated genes could underlie genetically inherited propensity to phenotypic variability in accord with the recently proposed model of stochastic epigenetic variations as a heritable force of evolutionary change [39].

The technical advantages of the low number of methylated cytosines in the genome, together with diet-controlled phenotypes arising from the same genome, make the honey bee an extremely tractable, simplified in vivo system in which to examine fundamental principles underpinning transitions from methylomes to organismal plasticity. In particular, the absence of promoter methylation in honey bees brings into focus gene body methylation as an important mechanism controlling various aspects of transcription. The utility of honey bees for understanding the intricacies of this process in the behavioral context can now be experimentally tested.

Materials and Methods

Source of DNA

Total DNA was extracted from dissected gland-free brains of 50 age-matched egg-laying queens (2.5 wk old) and from fifty 8-d-old workers. These individuals represent early stages of the reproductive life of queen bees and mature young workers capable of performing foraging tasks [19].

Sequencing of Bisulfite Converted DNA Libraries Using the Solexa GAIIx Platform (Illumina)

5 µg of high molecular weight DNA were used for fragmentation using the Covaris S2 AFA System in a total volume of 100 µl. Fragmentation-run parameters: Duty cycle 10%; Intensity: 5; Cycles/burst: 200; Time: 3 min; number of cycles: 3, resulting in a total fragmentation-time of 180 s. Fragmentation was confirmed with a 2100 Bioanalyzer (Agilent Technologies) using a DNA1000 chip. Fragment sizes were 140 bp on average for queen and worker DNAs, respectively. The fragmented DNAs were concentrated to a final volume of 75 µl using a DNA Speed Vac. End repair of fragmented DNA was carried out in a total volume of 100 µl using the Paired End DNA Sample Prep Kit (Illumina) as recommended by the manufacturer. For the ligation of the adaptors, the Illumina Early Access Methylation Adaptor Oligo Kit and the Paired End DNA Sample Prep Kit (Illumina) were used, as recommended by the manufacturer. For the size selection of the adaptor-ligated fragments, we used the E-Gel Electrophoresis System (Invitrogen) and a Size Select 2% precast agarose gel (Invitrogen). Each fragmented DNA was loaded on two lanes of

the E-gel. Electrophoresis was carried out using the "Size Select" program for 16 min. According to the standard loaded (50 bp DNA Ladder, Invitrogen), 240 bp fragments were extracted from the gel, pooled, and directly transferred to bisulfite treatment without further purification. For the bisulfite treatment we used the EZ-DNA Methylation Kit (Zymo) as recommended by the manufacturer with the exception of a modified thermal profile for the bisulfite conversion reaction. The conversion was carried out in a thermal cycler using the following thermal profile: 95°C for 15 s, 50°C for 1 h, repeat from step 1, 15×, 4°C for at least 10 min. The libraries were subsequently amplified, using the Fast Start High Fidelity PCR System (Roche) with buffer 2, and Illuminas PE1.1 and PE2.1 amplification primers. PCR thermal profile: 95°C for 2 min, 95°C for 30 s, 65°C for 20 s, 72°C for 30 s, then repeat from step 2, 11×, 72°C for 7min, hold at 4°C. PCR reactions were purified on PCR purification columns (MinElute, Qiagen) and eluted in 20 µl elution buffer (Qiagen).

Validation of the Libraries

1 µl of the libraries were analyzed on a 2100 Bioanalyzer (Agilent Technologies) using a DNA1000 chip. The fragment sizes were 240 bp and 243 bp for the queen and worker libraries, respectively. The estimated concentrations of the libraries were 0.8 ng/µl for the queen library and 5.8 ng/µl for the worker library.

Sequencing and Analysis

We used 8 pM of single stranded DNA per lane for Solexa sequencing. In total we sequenced 6 lanes. Worker: 1. single end - 36 bp - 10,187,567 reads (×2); 2. paired end - 76 bp - 7,960,842 reads (×2); 3. paired end - 76 bp - 7,444,938 reads (×2); 4. paired end - 76 bp - 11,642,135 reads (×2). Queen: 1. paired end - 76 bp - 16,752,247 reads (×2); 2. paired end - 76 bp - 16,778,784 reads (×2). For sequencing we used a Solexa Genoma Analyzer GAIIx with a v2 Paired End Cluster Generation Kit - GA II (Illumina) and v3 36 bp Cycle Sequencing Kits (Illumina). Extraction of sequences was done using Illumina Pipeline v1.4 software. Image analysis and basecalling was done using Illumina SCS v2.5 software.

Mapping

Reads were mapped using BSMAP-1.0240 with minor modifications [40]. A number of trimming and mapping options were assessed, and the conditions yielding the highest genome coverage depth was used for further processing (-s 12 -v 5 -k 6, for word size, number of mismatches, and number of words). Only the reads mapping uniquely were used. Mapping was carried out on a Linux cluster running Debian 5.0 (lenny).

Methylation Assessment

To increase the accuracy of methylation calls, only those cytosines fulfilling neighborhood quality standards NQS41 were counted [41]; namely, we only took into account bases of quality 20 or more, flanked by at least three perfectly matching bases of quality 15 or more. Deamination efficiency was assessed using the observation that the genomic repeats are not methylated in the honeybee (Figure S3). The deep coverage of these repeated sequences allowed us to estimate that the deamination rate is 99.76% for the queens and 99.71% for workers. The methylation status of each cytosine was then assessed by comparing the number of methylated and non-methylated reads to a binomial distribution with a probability of success equal to the deamination rate and a number of trials equal to the number of reads mapping to that cytosine and adjusting the resulting p values for multiple testing with the method of Benjamini and Hochberg [42]. An adjusted p value of 0.05 was used as a threshold for methylation calls. All statistical computations were carried out using the R language (www.r-project.org).

Honeybee ESTs and predicted genes were loaded into a Mysql database and visualized with Gbrowse (www.gmod.org), where CpG methylation levels in queens and workers were added as separate tracks.

Differential Methylation

Base-wise differences between queen and workers were estimated using Fisher exact tests. Gene-wise differences were assessed by generalized linear models of the binomial family, where methylation levels were modeled as functions of two categorical variables: caste and CpG position. p values were adjusted for multiple testing with the method of Benjamini and Hochberg [42].

Amplicon Sequences Selection

Illumina sequencing and BSMAP mapping results were confirmed by 454 sequencing of a set of bisulfite amplicons. Amplicon sequences were selected using raw methylome data and the following criteria: minimum coverage - 5 mapped reads for each queen and worker sample; minimum 2 mCpGs within a maximum of ~600 bp of sequence showing at least 50% difference in methylation levels between the two samples. In addition, four regions of mtDNA were selected. All primers and other details are listed in Table S4.

Other Protocols

All molecular protocols are described elsewhere [2,9,10,43].

Supporting Information

Figure S1 Coverage of all cytosines. (A) Cumulative distribution of the coverage of all cytosines, on either strand of the genome, in workers and queens. On the *x*-axis, coverage refers to the coverage depth that is the number of reads uniquely mapped to a given cytosine. The *y*-axis is the cumulative distribution; for instance, approximately 50% of all cytosines are covered by less than 5 reads, and about 80% are covered by less than 10 reads. (B) Cumulative distribution of the coverage of all CpGs in the genome, in workers and queens. On the *x*-axis, coverage refers to the coverage depth that is the number of reads uniquely mapped to a given CpG dinucleotide. The *y*-axis is the cumulative distribution (for instance, approximately 50% of the CpGs are covered by less than 15 reads, and about 80% are covered by less than 25 reads).
Found at: doi:10.1371/journal.pbio.1000506.s001 (0.41 MB PDF)

Figure S2 Methylation levels of methylated CpGs. Distribution of the methylation level of methylated CpGs. The methylation level is the proportion of methylated reads mapping to a given CpG. Over 30% of the CpGs are fully methylated.
Found at: doi:10.1371/journal.pbio.1000506.s002 (0.16 MB PDF)

Figure S3 Number of methylated and non-methylated Apis genes with BLAST hits to different species at various E-value thresholds. The amino acid sequences of the genes were compared. Fisher exact tests were conducted to assess whether significantly more methylated genes have a BLAST hit than non-methylated genes. Statistically significant tests at the 5% level are denoted with a star, and non-significant tests are

shown with a dot. The details of this analysis can be found in Table S3.
Found at: doi:10.1371/journal.pbio.1000506.s003 (0.13 MB PDF)

Figure S4 Number of high and low CpG honey bee genes with BLAST hits to different model species at various E-value thresholds. The amino acid sequences of the genes were compared. Fisher exact tests were conducted to assess whether significantly more low CpG genes have a BLAST hit than high CpG genes. Statistically significant tests at the 5% level are denoted with a star, and nonsignificant tests are shown with a dot. The details of this analysis can be found in Table S3.
Found at: doi:10.1371/journal.pbio.1000506.s004 (0.13 MB PDF)

Figure S5 Number of high CpG methylated and non-methylated honey bee genes with BLAST hits to different model species at various E-value thresholds. The amino acid sequences of the genes were compared. Fisher exact tests were conducted to assess whether significantly more high CpG methylated genes have a BLAST hit than high CpG non-methylated genes. Statistically significant tests at the 5% level are denoted with a star, and non-significant tests are shown with a dot. The details of the analysis can be found in Table S3.
Found at: doi:10.1371/journal.pbio.1000506.s005 (0.13 MB PDF)

Figure S6 Coverage (red) and methylation ratio (green) along various kinds of repetitive elements. The methylation ratio is the proportion of the reads where a cytosine is either methylated or unconverted. The *y*-axes are logarithmic in base 10 (the *x*-axis is truncated to the nearest multiple of 50, just like the *y*-axis is truncated to the nearest integer).
Found at: doi:10.1371/journal.pbio.1000506.s006 (0.63 MB PDF)

Figure S7 Periodicity of methylation patterns. (A) Autocorrelation of CpG methylation status over 1 kb. (B) Autocorrelation over 100 bp. Figures A and B show that the correlation of methylation status of neighboring CpGs increases sharply between 1 bp and 20 bp, then drops rapidly between 40 bp and 100 bp, and then slowly fades away. CpGs within a neighborhood of 2 bp to 100 bp are thus more likely to share the same methylation status than more distant CpGs. (C) Fourrier transform of autocorrelation showing a clear periodicity peak at 33 cycles per 100 bp (every 3 bp). (D) Distribution of codon position of mCs, and distribution of methylation level depending on the position. These two panels indicate that the distance between methylated CpGs is often a multiple of three and that the methylated cytosine corresponds most frequently to the first nucleotide of an arginine codon.
Found at: doi:10.1371/journal.pbio.1000506.s007 (0.33 MB PDF)

Figure S8 Correlation between CpG o/e and proportion of methylated CpGs. Genes with a lower CpG content tend to have a higher proportion of methylated CpGs. The red line is a polynomial regression through the points. The Akaike Information Criterion for model selection and a (monotonously decreasing) polynome of degree three was identified as the best model.
Found at: doi:10.1371/journal.pbio.1000506.s008 (0.71 MB PDF)

Figure S9 Distribution of methylated CpGs relative to splicing sites. For 169 genes, each containing a single well-defined alternative splicing event, the distance of all mCpGs to the centre of the alternatively spliced intron was computed, and the median of all these distances was calculated. A null distribution of this median distance was constructed using a randomization procedure (Manly, 2007): the methylation status of mCpGs of these genes were randomly shuffled 1,000 times, and the corresponding median distances computed. The observed value (1,224) is smaller than the smallest of the null distribution (1,259); the probability of the methylated CpGs to be as close or closer to the alternatively spliced intron as in this dataset is thus less than 0.001.
Found at: doi:10.1371/journal.pbio.1000506.s009 (0.77 MB PDF)

Figure S10 Annotation of the histone gene family in Apis mellifera showing the methylation profiles. See Table 4 for details.
Found at: doi:10.1371/journal.pbio.1000506.s010 (0.34 MB PDF)

Table S1 Sequence conservation of methylated and non-methylated genes. (A) Number of high and low CpG Apis genes with blast hits to different species at various E-value thresholds. The amino acid sequences of the genes were compared. Fisher exact tests were conducted to assess whether significantly more low CpG genes have a blast hit than high CpG genes. (B) Number of methylated and non-methylated honey bee genes with blast hits to different model species at various E-value thresholds. The amino acid sequences of the genes were compared. Fisher exact tests were conducted to assess whether significantly more methylated genes have a blast hit than non-methylated genes. (C) Number of high CpG methylated and non-methylated honey bee genes with blast hits to different model species at various E-value thresholds. The amino acid sequences of the genes were compared. Fisher exact tests were conducted to assess whether significantly more high CpG methylated genes have a blast hit than high CpG non-methylated genes.
Found at: doi:10.1371/journal.pbio.1000506.s011 (0.19 MB DOC)

Table S2 Differentially methylated genes in queens and worker brains. A generalized linear model of the binomial family was used to identify genes that are differentially methylated between castes. The methylation level of each gene was modeled as a function of the caste and of each of its CpG dinucleotides. In the table, "Caste" indicates whether the caste is a statistically significant factor explaining differences in methylation levels, "CpG" represents the different dinucleotides of that gene, and "Caste * CpG," the interaction factor, indicates whether the CpG dinucleotides behave differently between castes. GB numbers refer to the proteins at BeeBase: genomes.arc.georgetown.edu/drupal. Genes were ranked into 10 bins based on their expression levels from low (1) to high (10). No value in the relative expression column indicates those genes that are not represented on the microarray. Based on microarray data from Foret et al. [10].
Found at: doi:10.1371/journal.pbio.1000506.s012 (1.13 MB DOC)

Table S3 Evaluation of the Elango et al. hypothesis. (A) CpG o/e in methylated genes. (B) Differential methylation and differential gene expression.
Found at: doi:10.1371/journal.pbio.1000506.s013 (0.04 MB DOC)

Table S4 Details on genes used for deep 454 sequencing.
Found at: doi:10.1371/journal.pbio.1000506.s014 (0.03 MB XLS)

Acknowledgments

We thank Berit Haldemann, Andre Leischwitz, and Matthias Schaefer for their help with various aspects of sequencing. We thank George Gabor L. Miklos for stimulating discussions and critical comments on the manuscript and Joanna Maleszka for providing the biological materials used in this study. We also thank Fiona Wilkes for editorial help, Ros Attenborough for drafting the non-technical summary and three anonymous reviewers for their constructive comments.

Author Contributions

The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: FL RM. Performed the experiments: SF RK SW CF. Analyzed the data: FL SF RK. Contributed reagents/materials/analysis tools: FL RM. Wrote the paper: RM.

References

1. Maleszka R (2008) Epigenetic integration of environmental and genomic signals in honey bees: the critical interplay of nutritional, brain and reproductive networks. Epigenetics 3: 188–192.
2. Kucharski R, Maleszka J, Foret S, Maleszka R (2008) Nutritional control of reproductive status in honey bees via DNA methylation. Science 319: 1827–1830.
3. Barchuk AR, dos Santos Cristino A, Kucharski R, Simões ZLP, Maleszka R (2007) Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. BMC Dev Biol 7: 70.
4. Cokus SJ, Feng S, Zhang X, Chen Z, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215–219.
5. Zemach A, McDaniel IE, Silva P, Zilberman D (2010) Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328: 916–919.
6. Colot V, Rossignol JL (1999) Eukaryotic DNA methylation as an evolutionary device. Bioessays 21: 402–411.
7. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74: 481–514.
8. Suzuki MM, Kerr AR, De Sousa D, Bird A (2007) CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res 17: 625–631.
9. Wang Y, Jorda M, Jones PL, Maleszka R, et al. (2006) Functional CpG methylation system in a social insect. Science 314: 645–647.
10. Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R (2009) Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. BMC Genomics 10: 472.
11. Lister R, Pelizzola M, Dowen RH, Hawkins RD, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315–322.
12. Wang Y, Leung FC (2009) In silico prediction of two classes of honeybee genes with CpG deficiency or CpG enrichment and sorting according to gene ontology classes. J Mol Evol 68: 700–705.
13. Elango N, Hunt BG, Goodisman MA, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc Natl Acad Sci U S A 106: 11206–11211.
14. Gonzalo S, Jaco I, Fraga MF, Chen T, et al. (2006) DNA methyltransferases control telomere length and telomere recombination in mammalian cells. Nature Cell Biol 8: 416–424.
15. Honey Bee Genome Consortium (2006) Insights into social insects from the genome of the honey bee Apis mellifera. Nature 443: 931–949.
16. Vidovic M, Nighorn A, Koblar S, Maleszka R (2007) Eph receptor and Ephrin signaling in developing and adult brain of the honeybee (*Apis mellifera*). Dev Neurobiol 67: 233–251.
17. Zinke I, Schütz CS, Katzenberger JD, Bauer M, Pankratz MJ (2002) Nutrient control of gene expression in Drosophila: microarray analysis of starvation and sugar-dependent response. EMBO J 21: 6162–6173.
18. Talbert PB, Henikoff S (2010) Histone variants–ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol 11: 264–275.
19. Gabor Miklos GL, Maleszka R (2010) Epigenomic communication systems in humans and honey bees: from molecules to behaviour Hormones & Behavior;(doi:10.1016/j.yhbeh.2010.05.016).
20. Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. Science 187: 226–232.
21. Riggs AD (1975) X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 14: 9–25.
22. Regev A, Lamb M, Jablonka E (1998) The role of DNA methylation in invertebrates: developmental regulation or genome defense? Mol Biol Evol 15: 880–891.
23. Butte AJ, Dzau VJ, Glueck SB (2001) Further defining housekeeping, or "maintenance," genes: focus on "a compendium of gene expression in normal human tissues." Physiol Genomics 7: 95–96.
24. Sugden K, Pariante C, McGuffin P, Aitchison K, D'Souza U (2008) Housekeeping gene expression is affected by antidepressant treatment in a mouse fibroblast cell line. J Psychopharmacol;doi:10.1177/0269881108099690.
25. Gibson G (2008) The environmental contribution to gene expression profiles. Nat Rev Genet 9: 575–581.
26. Burzynski SR, Patil S, Ilkowska-Musial E, Chitur S, et al. (2008) Pathway analysis of the effect of chromatin remodeling agent phenylbutyrate on the brains of honeybees. Society for Neuroscience, Washington 2008, Abstract 494.2/UU42.
27. Ricobaraza A, Cuadrado-Tejedor M, Pe'rez-Mediavilla A, Frechilla D, et al. (2009) Phenylbutyrate ameliorates cognitive deficit and reduces tau pathology in an Alzheimer's disease mouse model. Neuropsychopharmacology 34: 1721–1732.
28. Kang HL, Benzer S, Min KT (2002) Life extension in Drosophila by feeding a drug. Proc Natl Acad Sci U S A 99: 838–843.
29. Groh C, Rössler W (2008) Caste-specific postembryonic development of primary and secondary olfactory centers in the female honeybee brain. Arthropod Struct Dev 37: 459–468.
30. Miller CA, Sweatt JD (2007) Covalent modification of DNA regulates memory formation. Neuron 53: 857–869.
31. Miller CA, Campbell SL, Sweatt JD (2008) DNA methylation and histone acetylation work in concert to regulate memory formation and synaptic plasticity. Neurobiol Learn Mem 89: 599–603.
32. Lockett GA, Helliwell P, Maleszka R (2010) Involvement of DNA methylation in memory processing in the honey bee. Neuroreport 21: 812–816.
33. Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. Nature Rev Genet 10: 295–304.
34. Ball MP, Li JB, Gao Y, Lee JH, et al. (2009) Cytosine methylation may interact with other epigenetic features, such as histone modifications. Nature Biotech 27: 361–368.
35. Luco RF, Pan Q, Tominaga K, Blencowe BJ, et al. (2010) Regulation of alternative splicing by histone modifications. Science 327: 996–1000.
36. Laurent L, Wong E, Li G, Huynh T, et al. (2010) Dynamic changes in the human methylome during differentiation. Genome Res 20: 1–12.
37. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39: 61–69.
38. Wittkopp PJ (2007) Variable gene expression in eukaryotes: a network perspective. J Exp Biol 210: 1567–1575.
39. Feinberg AP, Irizarry RA (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc Nat Acad Sci U S A 107: 1757–1764.
40. Xi Y, Li W (2009) BSMAP: whole genome Bisulfite Sequence MAPping program. BMC Bioinformatics 10: 232.
41. Altshuler D, Pollara V, Cowles C, Van Etten, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513–516.
42. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B 57: 289–300.
43. Kucharski R, Maleszka R (2005) Microarray and rtPCR analyses of gene expression in the honey bee brain following caffeine treatment. J Mol Neurosci 27: 269–276.