https://doi.org/10.1038/s41586-019-1184-5

Molecular recording of mammalian embryogenesis

Michelle M. Chan[1,2,14], Zachary D. Smith[3,4,5,14], Stefanie Grosswendt[6], Helene Kretzmer[6], Thomas M. Norman[1,2], Britt Adamson[1,2,13], Marco Jost[1,2,7], Jeffrey J. Quinn[1,2], Dian Yang[1,2], Matthew G. Jones[1,2,8], Alex Khodaverdian[9,10], Nir Yosef[9,10,11,12], Alexander Meissner[3,4,6]* & Jonathan S. Weissman[1,2]*

Ontogeny describes the emergence of complex multicellular organisms from single totipotent cells. This field is particularly challenging in mammals, owing to the indeterminate relationship between self-renewal and differentiation, variation in progenitor field sizes, and internal gestation in these animals. Here we present a flexible, high-information, multi-channel molecular recorder with a single-cell readout and apply it as an evolving lineage tracer to assemble mouse cell-fate maps from fertilization through gastrulation. By combining lineage information with single-cell RNA sequencing profiles, we recapitulate canonical developmental relationships between different tissue types and reveal the nearly complete transcriptional convergence of endodermal cells of extra-embryonic and embryonic origins. Finally, we apply our cell-fate maps to estimate the number of embryonic progenitor cells and their degree of asymmetric partitioning during specification. Our approach enables massively parallel, high-resolution recording of lineage and other information in mammalian systems, which will facilitate the construction of a quantitative framework for understanding developmental processes.

The development of a multicellular organism from a single cell is an astonishing process. Classic lineage-tracing experiments using *Caenorhabditis elegans* revealed surprising outcomes, which include deviations between lineage and functional phenotype, but nonetheless benefitted from the highly deterministic nature of development in this organism[1]. More-complex species generate larger and more elaborate structures that progress through multiple transitions, which raise questions about the coordination between specification and commitment to ensure faithful recapitulation of an exact body plan[2,3]. Single-cell RNA sequencing (scRNA-seq) has permitted unprecedented explorations into cell-type heterogeneity, and has produced profiles of developing flatworms[4,5], frogs[6], zebrafish[7,8] and mice[9,10]. More recently, CRISPR–Cas9-based technologies have been applied to record cell lineage[11–14], and combined with scRNA-seq to generate fate maps in zebrafish[15–17]. However, these technologies include only one or two bursts of barcode-diversity generation, which may be limiting for other applications or organisms.

An ideal molecular recorder for addressing developmental questions in complex multicellular organisms would possess the following characteristics: (1) minimal effect on cellular phenotype; (2) high-information content to account for hundreds of thousands of cells; (3) a single-cell readout for simultaneous profiling of functional state[15–17]; (4) flexible recording rates that can be tuned to a broad temporal range; and (5) continuous generation of diversity throughout the experiment. The last point is especially relevant for mammalian development, in which spatial plans are gradually and continuously specified and may originate from small, transient progenitor fields. scRNA-seq has revealed populations of cells that have a continuous spectrum of phenotypes, which implies that differentiation does not occur instantaneously and further highlights the need for an evolving recorder[18].

Here we generate and validate a method for simultaneously reporting cellular state and lineage history in mice. Our CRISPR–Cas9-based recorder is capable of generating high-information content and can perform multi-channel recording with readily tunable mutation rates. We use the recorder as a continuously evolving lineage tracer to observe the fate map that underlies mouse embryogenesis through gastrulation, recapitulating canonical paradigms and illustrating how lineage information may facilitate the identification of novel cell types.

A transcribed and evolving recorder

To achieve our goal of a tunable molecular recorder that is capable of creating high-information content, we used Cas9 to generate double-stranded breaks that result in heritable insertions or deletions (indels) after repair[11–17]. We record within a 205-base-pair, synthetic DNA 'target site' that contains three 'cut sites' and a static 8-base-pair 'integration barcode', which is delivered in multiple copies using piggyBac transposition (Fig. 1a, b). We embedded this sequence into the 3′ untranslated region of a constitutively transcribed fluorescent protein to enable profiling from the transcriptome. A second cassette encodes three independently transcribed and complementary guide RNAs to permit recording of multiple distinct signals[19] (Fig. 1a, b).

Our system is capable of high-information storage owing to the diversity of heritable repair outcomes and the large number of target sites, which can be distinguished by the integration barcode (Fig. 1c). DNA repair generates hundreds of unique indels, and the distribution for each cut site is different and non-uniform; some regions lead to highly biased outcomes, whereas others create a diverse series[20–22] (Fig. 1c, Extended Data Fig. 1). To identify sequences that can tune the mutation rate of our recorder for time scales that are not pre-defined

[1]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA, USA. [2]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, USA. [5]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [6]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin, Germany. [7]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA, USA. [8]Integrative Program in Quantitative Biology, University of California, San Francisco, San Francisco, CA, USA. [9]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, CA, USA. [10]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [11]Chan Zuckerberg Biohub, San Francisco, CA, USA. [12]Ragon Institute of Massachusetts General Hospital, MIT and Harvard University, Cambridge, MA, USA. [13]Present address: Department of Molecular Biology, Lewis Sigler Institute, Princeton University, Princeton, NJ, USA. [14]These authors contributed equally: Michelle M. Chan, Zachary D. Smith. *e-mail: meissner@molgen.mpg.de; jonathan.weissman@ucsf.edu

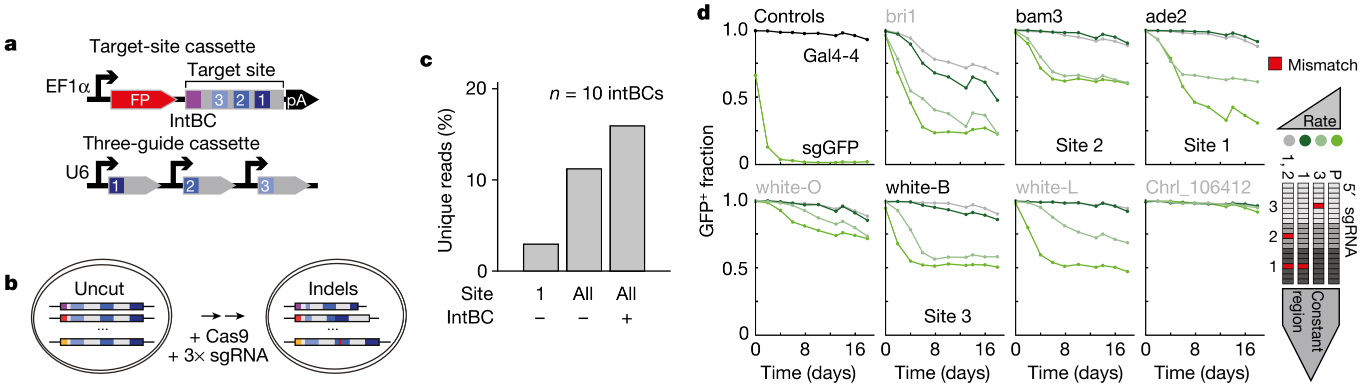


Fig. 1 | Optimization of a multi-purpose molecular recorder. a, Target-site (top) and three-guide (bottom) cassettes. The target site consists of an integration barcode (intBC) and three cut sites for Cas9-based recording. Three different single-guide RNAs (sgRNAs) are each controlled by independent promoters (in this study, mouse U6, human U6 and bovine U6). FP, fluorescent protein; sites 1–3 are indicated. **b**, Molecular recording principle. Each cell contains multiple genomic, intBC-distinguishable target-site integrations. sgRNAs direct Cas9 to cognate cut sites to generate insertion (red) or deletion mutations. Here Cas9 is either ectopically delivered or induced by doxycycline. **c**, Percentage of uniquely marked reads recovered after recording within a K562 line with ten intBCs for six days. Information content scales with number of sites and presence of the intBC. All, all three cut sites. **d**, sgRNA mismatches alter mutation rate. Seven protospacers (bri1, bam3, ade2, white-O, white-B, white-L and chrl_106412) were integrated into the coding sequence of a GFP reporter to infer mutation rate by the fraction of GFP-positive cells over a 20-day time course. Single or dual mismatches were made in guides according to proximity to the protospacer adjacent motif: region 1 (proximal), region 2 and region 3 (distal). Guides against Gal4-4 and the GFP coding sequence (sgGFP) act as negative and positive controls. Sequence names in black were incorporated into the target site. P, perfect complementarity between guide RNA and protospacer; 3, mismatch in region 3; 1, mismatch in region 1; 1,2, simultaneous mismatches in regions 1 and 2.

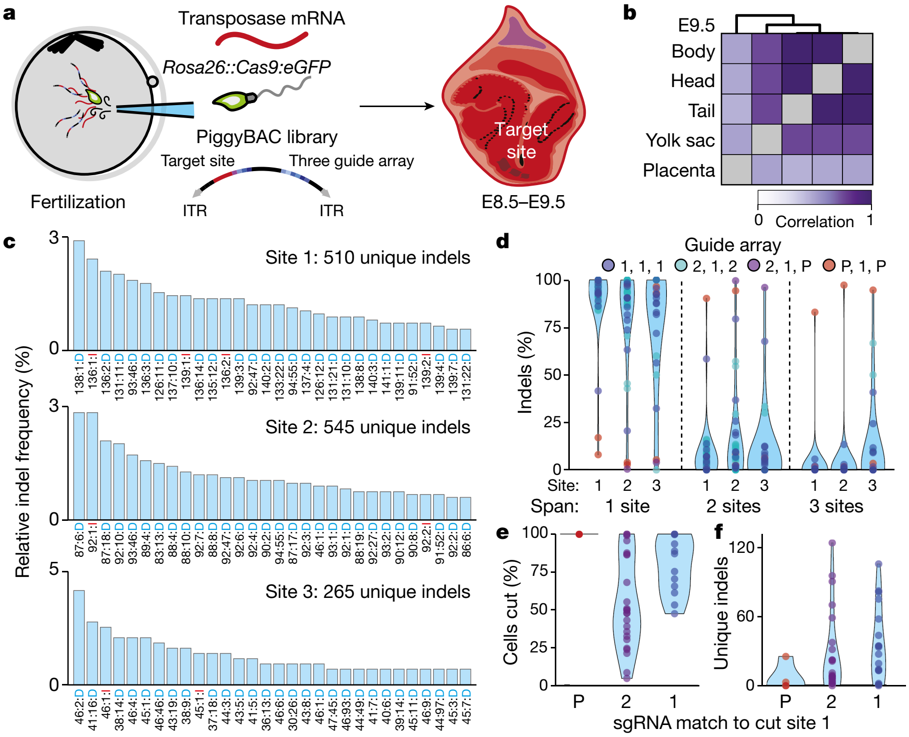


Fig. 2 | Lineage tracing in mouse from fertilization through gastrulation. a, Lineage tracing in mouse experiments. The target site (within the 3′ untranslated region of mCherry) and the three-guide array are encoded into a single piggyBac transposon vector. The vector, transposase mRNA and *Rosa26::Cas9:eGFP* sperm are injected into oocytes to ensure early integration and tracing in all subsequent cells after zygotic genome activation. Transferred embryos are then recovered after gastrulation. ITR, inverted terminal repeat. **b**, Pearson correlation coefficient heat map of indel proportions recovered from bulk tissue of an E9.5 embryo (see Extended Data Fig. 2). **c**, Indel frequency distribution estimated from 40 independent target sites from all embryos. Each site produces hundreds of outcomes for high-information encoding. See Extended Data Fig. 4 and Supplementary Methods for frequency calculation. The indel code along the *x* axis is as follows: alignment coordinate: indel size:indel type (red I, insertion; blue D, deletion). **d**, Proportion of indels that span one, two or three sites, shown per site. Each dot denotes 1 of 40 independent intBCs and sums to 100% across site-spanning indels. Colours indicate the guide array: P, no mismatches; 1, mismatch in region 1; 2, mismatch in region 2. **e**, Percentage of cells with mutations according to guide complementarity. Indel proportions within one mouse depend on timing: mutations that happen earlier in development are propagated to more cells. Dots represent site-1 measurements from independent intBCs; $n = 4$, 24, and 18 for no mismatches, region-2 and region-1 mismatches, respectively. **f**, Indel diversity is inversely related to cutting efficiency for site 1, as in **e**. Early mutations owing to fast cutting are propagated to more cells, which leads to smaller numbers of unique indels.

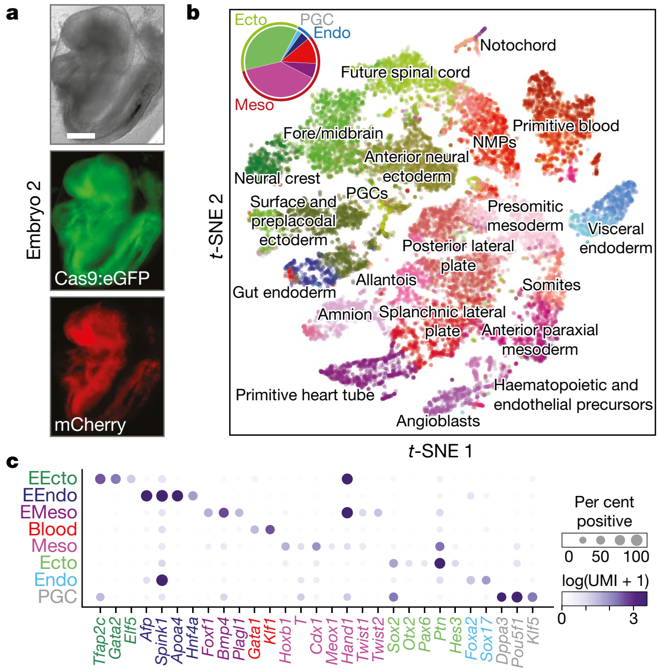


Fig. 3 | Assigning cellular phenotype by scRNA-seq. a, Images of a lineage-traced E8.5 embryo (embryo 2 of 7 for which single-cell data were collected, see Extended Data Fig. 3), including for Cas9:eGFP and the mCherry target site. Scale bar, 0.2 mm. **b**, *t*-SNE plot of scRNA-seq from embryo in **a**. Only large or spatially distinct clusters are labelled. Inset, pie chart of germ layers. Lighter and darker shades represent embryonic and extra-embryonic components, respectively. Mesoderm is further separated to include blood (red). See Extended Data Fig. 5b for additional embryos. $n = 22{,}264$ cells. **c**, Dot plot of canonical tissue-specific markers. Grouping clusters of diverse tissue types into germ layers reduces the fraction of marker-positive cells, but the specificity to their respective states remains high—especially when considered combinatorially. The size of the circle denotes the fraction of marker-positive cells, and colour intensity indicates normalized expression (cluster mean). Ecto, embryonic ectoderm; EEcto, extra-embryonic ectoderm; EEndo, extra-embryonic endoderm; endo, embryonic endoderm; meso, embryonic mesoderm; EMeso, extra-embryonic mesoderm; PGC, primordial germ cell; NMPs, neuromesodermal progenitors; UMI, unique molecular identifier.

(and may extend from days to months), we screened several guide RNA series that contained mismatches to their targets[23] by monitoring their activity on a GFP reporter over a 20-day time course. We selected those series that demonstrated a broad dynamic range (Fig. 1d). Slower cutting rates may also improve viability in vivo, as frequent Cas9-mediated double-stranded breaks can cause cellular toxicity[24,25]. To demonstrate information recovery from single-cell transcriptomes, we stably transduced K562 cells with our technology and generated a primary cell-barcoded cDNA pool via the 10x Genomics platform, which enabled us to assess global transcriptomes and specifically amplify mutated target sites (Extended Data Fig. 1c).

Tracing cell lineages during development

We next applied our technology to map cell fates during early development in mouse, from totipotency onwards. We integrated multiple target sites into the genome, delivered constitutive Cas9–GFP-encoding sperm into oocytes to initiate cutting, and isolated embryos for analysis at approximately embryonic day (E)8.5 or E9.5 (Fig. 2a, Supplementary Methods). To confirm our lineage-tracing capability, we amplified the target site from bulk placenta, yolk sac and three embryonic fractions from an E9.5 embryo, and recapitulated their expected relationships using the similarity of their indel proportions (Fig. 2b, Extended Data Fig. 2).

Following this in vivo proof of principle, we generated single-cell data from additional embryos (Extended Data Fig. 3). We collected scRNA-seq data for 7,364–22,264 cells from 7 embryos (between approximately 15.8% and 61.4% of the total estimated cell count) and recovered 167–2,461 unique lineage identities ($\geq$1 target site recovered for 15–75% of cells from 3 to 15 integration barcodes, Extended Data Fig. 4). Many target sites are captured at low levels or are heterogeneously represented, which we improved by changing the promoter from a truncated form of EF1α to a longer version that contains an intron[26] (see embryo 7 in Extended Data Fig. 4).

We estimated the likelihood distribution of indels by combining data from all seven embryos. Many indels are shared with K562 cells; however, their likelihoods differ, which suggests that cell type or developmental status may influence repair outcomes[20] (Fig. 2c, Extended Data Figs. 1, 4f). Our ability to independently measure and control the rate of cutting across the target site is preserved in vivo, and there is minimal interference between cut sites except when using combinations of the fastest guides, which may lead to end-joining between simultaneous double-stranded breaks (Fig. 2d). The fastest cutters result in higher proportions of cells with identical indels, indicating that mutations are arising earlier in development and correspondingly reducing indel diversity (Fig. 2e, f). Importantly, the lineage tracer retains recording capacity beyond the temporal interval that we study here, as most embryos still have cells with unmodified cut sites (Fig. 2e).

Simultaneous scRNA-seq to assign state

To ascertain cell function, we next used annotations from a compendium of wild-type mouse gastrulation (E6.5–E8.5). We assigned cells from lineage-traced embryos by their proximity to each cell-state expression signature and determined the age of each embryo by their tissue proportions compared to the wild-type reference[27] (Fig. 3a–c). We proceeded with six of our seven embryos, as these appeared to be morphologically normal and included every expected tissue type: two mapped most closely to E8.5 and the remaining four mapped to E8.0 (Extended Data Fig. 5). Placenta was not specifically isolated, but is present in four out of six embryos and serves as a valuable outgroup to establish our ability to track transitions to the earliest bifurcation.

We also developed breeder mice that would enable exploration of all stages of development by injecting target sites into Cas9$^-$ backgrounds. This approach substantially increased the number of stably integrated target sites (to about 20). The resulting mice can be crossed with Cas9-expressing strains to yield viable Cas9$^+$ F_1 litters that maintain continuous, stochastic indel generation into adulthood, which demonstrates that cutting does not noticeably interfere with normal mouse development (Extended Data Fig. 6).

Single-cell lineage reconstruction

We developed phylogenetic reconstruction strategies to specifically exploit the characteristics of our lineage tracer: the presence of categorical indels, the irreversibility of mutations and the presence of missing values (Extended Data Fig. 7, Supplementary Methods). We determined the best reconstruction by summing the log-likelihoods for all indels that appear in the tree, using likelihoods estimated from embryo data (Extended Data Figs. 4, 7). When cell-type identity from scRNA-seq is overlaid onto the tree, we observe functional restriction during development and fewer cell types being represented as we move from root to leaves (Fig. 4a, b, Extended Data Fig. 8).

Strategies for ordering cells based on scRNA-seq—such as trajectory inference—typically assume that functional similarity reflects close lineage[18]. To investigate this question directly, we used a modified Hamming distance to measure pairwise lineage distances and compared them to the RNA-seq correlations. In general, cells that are separated by a smaller lineage distance have transcriptional profiles that are more similar to each other, although this relationship is clearer for some embryos than others (Fig. 4c, Extended Data Fig. 9). This result is consistent with the notion of a continuous restriction of potency as cells differentiate into progressively specialized types.

We also developed a shared progenitor score that estimates the degree of common ancestry between different tissues by evaluating the number

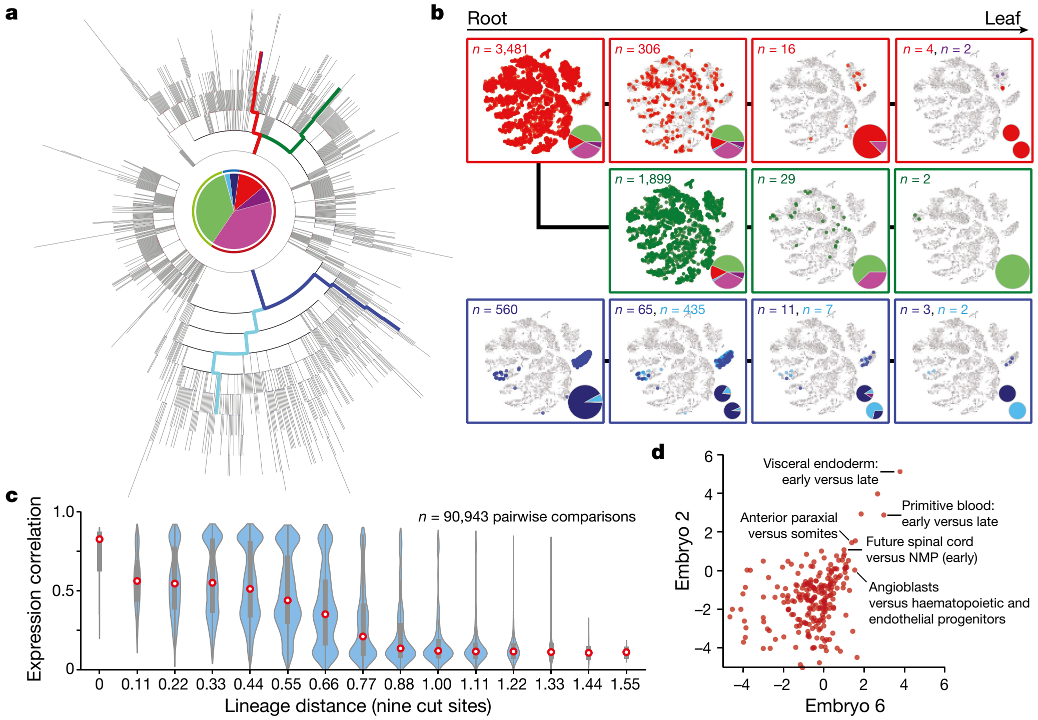


Fig. 4 | Single-cell lineage reconstruction of mouse embryogenesis. **a**, Reconstructed lineage tree comprising 1,732 nodes for embryo 2, with example lineages highlighted. Each branch represents an indel generation event. **b**, Example paths from tree in **a** highlighted by colour. Cells for each node in the path are overlaid onto the plot from Fig. 3b, and tissue proportions are shown as a pie chart. Tissue representation decreases with increased tree depth, which indicates functional restriction. Bifurcating sublineages are included for the top and bottom paths. In the top (red) path, this bifurcation occurs within the final branch after primitive blood specification. In the bottom (blue) path, bifurcation happens early within bipotent cells that become either gut or visceral endoderm. **c**, Violin plots of the pairwise relationship between lineage and expression for single cells. Lineage distance uses a modified Hamming distance normalized to the number of shared cut sites. Pearson correlation decreases with increasing lineage distance, which shows that closely related cells are more likely to share function. Red dot highlights the median, edges show the interquartile range and whiskers show the full range. **d**, Comparison of shared progenitor scores ($\log_2$-transformed) between our two most information-dense embryos (embryo 2, $n = 1,400$ alleles; embryo 6, $n = 2,461$ alleles). Cells from closely related transcriptional clusters (for example, between the early and late states of either the primitive blood or the visceral endoderm) derive from common progenitors and score as highly related in both embryos. We also observe a close link between mesoderm and ectoderm that may reflect shared heritage between neuromesodermal progenitors and more-posterior neural ectodermal tissues, such as the future spinal cord[43].

and specificity of shared nodes in the tree (Supplementary Methods). Despite the stochastic timing of indel formation, this approach can reproducibly recover emergent tissue relationships, such as the possible shared origins between anterior somites and paraxial mesoderm or between neuromesodermal progenitors and the future spinal cord (Fig. 4d). The full map of shared progenitor scores can be clustered to create a comprehensive picture of tissue relationships during development (Extended Data Fig. 8d).

State and lineage do not always conform

Although our reconstructed tissue relationships generally recapitulate canonical knowledge, extra-embryonic endoderm and embryonic endoderm display consistent and unexpectedly close ancestry despite their independent origins from the hypoblast and embryo-restricted epiblast, respectively (Fig. 5a, Extended Data Fig. 9). Manual inspection of the trees revealed a subpopulation of cells that appears transcriptionally as embryonic endoderm, but which lineage analysis places within extra-embryonic branches (shown in blue in Fig. 4b). Consistent with this finding, a previous targeted study using marker-directed lineage tracing identified a latent extra-embryonic contribution to the developing hindgut during gastrulation, although it was not possible to broadly evaluate transcriptional differences between cells with different origins[28].

In this study, the scRNA-seq profiles that we collected in tandem with the lineage readout enabled us to assess the degree of convergence towards a functional endoderm signature and to identify distinguishing genes. Endoderm-classified cells that are derived from extra-embryonic origin are most similar to the endoderm cell type, but also share a slightly higher similarity with yolk sac that is not apparent within the *t*-distributed stochastic neighbour embedding (*t*-SNE) projection of the full embryo (Fig. 5b, Extended Data Fig. 10). Given the independent origins of extra-embryonic and embryonic endoderm, we might expect a subtle but persistent transcriptional signature that reflects their distinct developmental history. When we separate endoderm cells according to their lineage, we identify two X-chromosome-linked genes—*Trap1a* and *Rhox5*, which are general markers for extra-embryonic tissue[29,30]—that are consistently upregulated across embryos in the endoderm of extra-embryonic origin (Kolmogorov–Smirnov test, Bonferroni-corrected $P < 0.05$, Fig. 5c, d). Notably, in other RNA-seq studies, these relationships are not captured by whole-embryo clustering and are only found by specific examination of the hindgut[9,31] (Extended Data Fig. 10). These observations confirm that our lineage tracer can successfully pinpoint instances of convergent transcriptional regulation.

Towards a quantitative fate map

Simultaneous single-cell lineage tracing with transcriptional phenotype provides the opportunity to infer the cellular potency and specification biases of ancestral cells as reconstructed by our fate map[32,33]. Each node within the tree represents a unique lineage identity that stems from a single reconstructed progenitor cell, which allows us to estimate the lower boundaries of the progenitor field size (Supplementary Methods). We investigated the founding number of progenitors during the earliest transitions in cellular potential. We defined totipotency as a node that gives rise to both embryonic and placental cell types, and then tiered pluripotency into 'early' and 'late' according to the presence of extra-

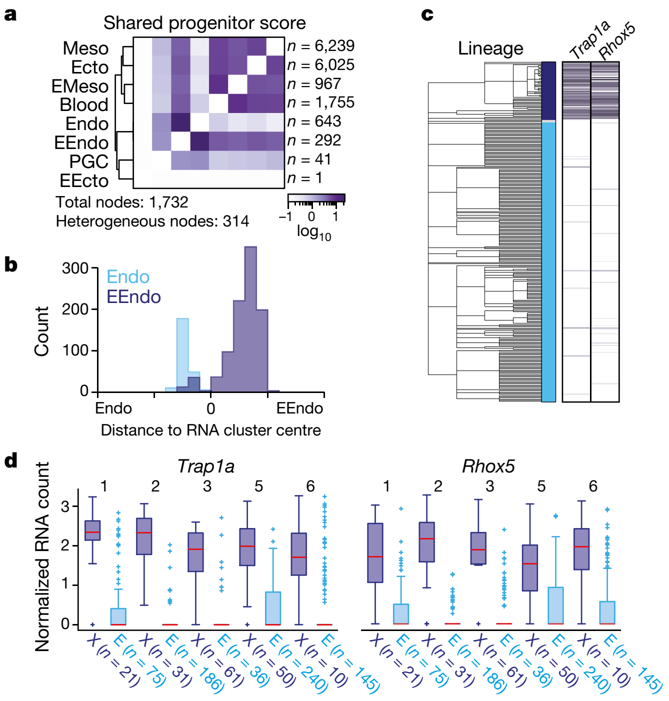


Fig. 5 | Disparities between transcriptional identity and lineage history within the endoderm. a, Shared progenitor-score heat map for embryo 2 reconstructs expected relationships. The number of nodes that include cells from different lineages is highlighted (heterogeneous nodes). See Extended Data Fig. 9 for additional embryos. *n*, number of cells assigned to each tissue. **b**, For cells from embryo 2, the relative distance from the mean expression profile of either the endoderm or the extra-embryonic endoderm cluster according to origin (endo or EEndo). **c**, Endoderm-cell lineage tree from embryo 2 with expression heat map for two extra-embryonic marker genes. Middle bar indicates lineage: dark blue, extra-embryonic; light blue, embryonic; grey, ambiguous. **d**, Expression box plots for *Trap1a* and *Rhox5* confirms consistent differential expression across lineage-traced embryos according to their embryonic or extra-embryonic ancestry. Red line highlights median, edges show the interquartile range, whiskers show the Tukey fence, and crosses denote outliers. *n* values give the number of recovered endoderm cells of either embryonic (E) or extra-embryonic (X) origin per embryo. Numbers above plots indicate the embryo replicate number.

embryonic endoderm[34] (Fig. 6a). The contributions of these founders to extant lineages are asymmetric, which suggests that progenitors may be biased towards specific fates but retain the ability to generate other cell types. Lower bound estimates from our data suggest a range of 1–6 totipotent cells, 10–20 early pluripotent progenitors and 18–51 late pluripotent progenitors (Fig. 6b). The variable number of multipotent cells at these stages may reflect an encoded robustness that ensures the successful assembly of the functioning organism—particularly given that a single pluripotent cell can generate all the somatic lineages in an embryo[35]. Future studies that use more replicates generated by breeding may enable statistical approaches for evaluating these organism-scale developmental considerations.

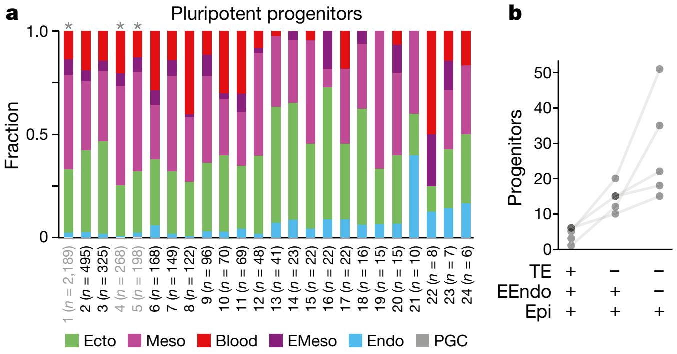


Fig. 6 | Lineage bias and estimated size of progenitor fields. a, Relative tissue distribution of cells descended from reconstructed or 'profiled' pluripotent progenitor cells for embryo 2. 'Profiled' is a unique lineage identity comprised of multiple cells that are directly observed in the data. Pluripotent cells form all germ layers but show asymmetric propensities towards different cell fates; this possibly reflects positional biases. Nodes highlighted in grey (with asterisk above the bar) give rise to primordial germ cells (lineages 1, 4 and 5 include 9, 1 and 1 PGCs each, respectively). Colour assignments as in Fig. 3. *n*, number of cells. **b**, Estimated progenitor field sizes for three types of early developmental potency for the six embryos analysed in this study. Totipotent cells give rise to all cells of the developing embryo, including trophectodermal (TE) lineages. Pluripotent progenitors are partitioned into early and late by the generation of extra-embryonic endoderm in addition to epiblast (epi). Dots represent single embryos; solid grey line connects estimates from the same embryo.

Discussion

In this study, we present cell-fate maps that underlie mammalian gastrulation, using a technology for high-information and continuous recording. Several ideas have emerged, including the transformative nature of CRISPR–Cas9-directed mutation combined with an scRNA-seq readout[15–17], how information about the history of a cell recorded by this technology can complement RNA-seq profiles to characterize cell type, and an early framework for quantitatively understanding stochastic transitions during mammalian development.

The modularity of our recorder enables substitutions that will increase its breadth of applications. Here we use three constitutively expressed guide RNAs to record continuously over time, but future modifications could use environmentally responsive promoters that sense stress, neuronal action potentials or cell-to-cell contacts[36], or combine these approaches for multifactorial recording. Similarly, Cas9-derived base editors[37]—including those that create diverse mutations[38]—could allow for content recording in cells that are particularly sensitive to nuclease-directed DNA double-stranded breaks[24,25].

Our cell-fate map identifies a phenotypic convergence of independent cell lineages, which showcases the power of unbiased organism-wide lineage tracing to separate populations that appear similar by scRNA-seq data alone. Specifically, we substantiate the extra-embryonic origin of a subset of cells that resemble embryonic endoderm. Although the initial specification of these lineages is known to rely on redundant regulatory programs, these lineages are temporally separated by several days, emerge from transcriptionally and epigenetically distinct progenitors, and form terminal cell types with highly divergent functions. The identification of highly predictive markers that segregate by origin, such as *Trap1a*, provides a clear outline for further exploration through spatial transcriptomics[39–41]. More generally, our approach can be used to investigate other convergent processes or to discriminate heterogeneous cell states that represent persistent signatures of independent developmental pathways, which will be critical for the assembly of a comprehensive cell atlas[42]. The scope of trans-differentiation within mammalian ontogenesis remains largely unexplored, but can be practically inventoried using our system.

Our technology is designed to quantitatively address previously opaque questions in ontogenesis. Higher-order issues of organismal regulation—such as the location, timing and stringency of developmental bottlenecks, as well as the corresponding likelihoods of state transitions to different cellular phenotypes—can be modelled from the assembly of historical relationships. Our hope is that characterization of these attributes will lead to insights that connect large-scale developmental phenomena to the molecular regulation of cell-fate decision-making.

Reporting summary

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

Data availability

The data are available in the Gene Expression Omnibus database under accession numbers GSE117542 (for lineage-traced embryos) and GSE122187 (for the gastrulation compendium). Any other relevant data are available from the corresponding authors upon reasonable request.

Code availability

The greedy reconstruction algorithm (named Cassiopeia) is available at https://github.com/YosefLab/Cassiopeia. Other code will be shared upon request.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1184-5.

Received: 2 July 2018; Accepted: 10 April 2019;
Published online 13 May 2019.

1. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol*. **100**, 64–119 (1983).
2. Pijuan-Sala, B., Guibentif, C. & Göttgens, B. Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nat. Rev. Mol. Cell Biol*. **19**, 399–412 (2018).
3. Zernicka-Goetz, M. Patterning of the embryo: the first spatial decisions in the life of a mouse. *Development* **129**, 815–829 (2002).
4. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaq1736 (2018).
5. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
6. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
7. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
8. Wagner, D. E. et al. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
9. Ibarra-Soria, X. et al. Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol*. **20**, 127–134 (2018).
10. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **172**, 1091–1107 (2018).
11. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR–Cas in human cells. *Science* **353**, aag0511 (2016).
12. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
13. Frieda, K. L. et al. Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).
14. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
15. Raj, B. et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol*. **36**, 442–450 (2018).
16. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
17. Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nat. Biotechnol*. **36**, 469–473 (2018).
18. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
19. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
20. van Overbeek, M. et al. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* **63**, 633–646 (2016).
21. Schimmel, J., Kool, H., van Schendel, R. & Tijsterman, M. Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells. *EMBO J*. **36**, 3634–3649 (2017).
22. Lemos, B. R. et al. CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl Acad. Sci. USA* **115**, E2040–E2047 (2018).
23. Gilbert, L. A. et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
24. Ihry, R. J. et al. p53 inhibits CRISPR–Cas9 engineering in human pluripotent stem cells. *Nat. Med*. **24**, 939–946 (2018).
25. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med*. **24**, 927–930 (2018).
26. Kim, S.-Y., Lee, J.-H., Shin, H.-S., Kang, H.-J. & Kim, Y.-S. The human elongation factor 1 alpha (EF-1α) first intron highly enhances expression of foreign genes from the murine cytomegalovirus promoter. *J. Biotechnol*. **93**, 183–187 (2002).
27. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol*. **36**, 411–420 (2018).
28. Kwon, G. S., Viotti, M. & Hadjantonakis, A.-K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev. Cell* **15**, 509–520 (2008).
29. Eakin, G. S. & Hadjantonakis, A.-K. Sex-specific gene expression in preimplantation mouse embryos. *Genome Biol*. **7**, 205 (2006).
30. Li, C.-S. et al. *Trap1a* is an X-linked and cell-intrinsic regulator of thymocyte development. *Cell. Mol. Immunol*. **14**, 685–692 (2017).
31. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
32. Soriano, P. & Jaenisch, R. Retroviruses as probes for mammalian development: allocation of cells to the somatic and germ cell lineages. *Cell* **46**, 19–29 (1986).
33. Jaenisch, R. Mammalian neural crest cells participate in normal embryonic development on microinjection into post-implantation mouse embryos. *Nature* **318**, 181–183 (1985).
34. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
35. Wang, Z. & Jaenisch, R. At most three ES cells contribute to the somatic lineages of chimeric mice and of mice produced by ES-tetraploid complementation. *Dev. Biol*. **275**, 192–201 (2004).
36. Baeumler, T. A., Ahmed, A. A. & Fulga, T. A. Engineering synthetic signaling pathways with programmable dCas9-based chimeric receptors. *Cell Reports* **20**, 2639–2653 (2017).
37. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
38. Hess, G. T. et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).
39. Hou, J. et al. A systematic screen for genes expressed in definitive endoderm by Serial Analysis of Gene Expression (SAGE). *BMC Dev. Biol*. **7**, 92 (2007).
40. Wang, G., Moffitt, J. R. & Zhuang, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Sci. Rep*. **8**, 4847 (2018).
41. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
42. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
43. Tzouanacou, E., Wegener, A., Wymeersch, F. J., Wilson, V. & Nicolas, J.-F. Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis. *Dev. Cell* **17**, 365–376 (2009).

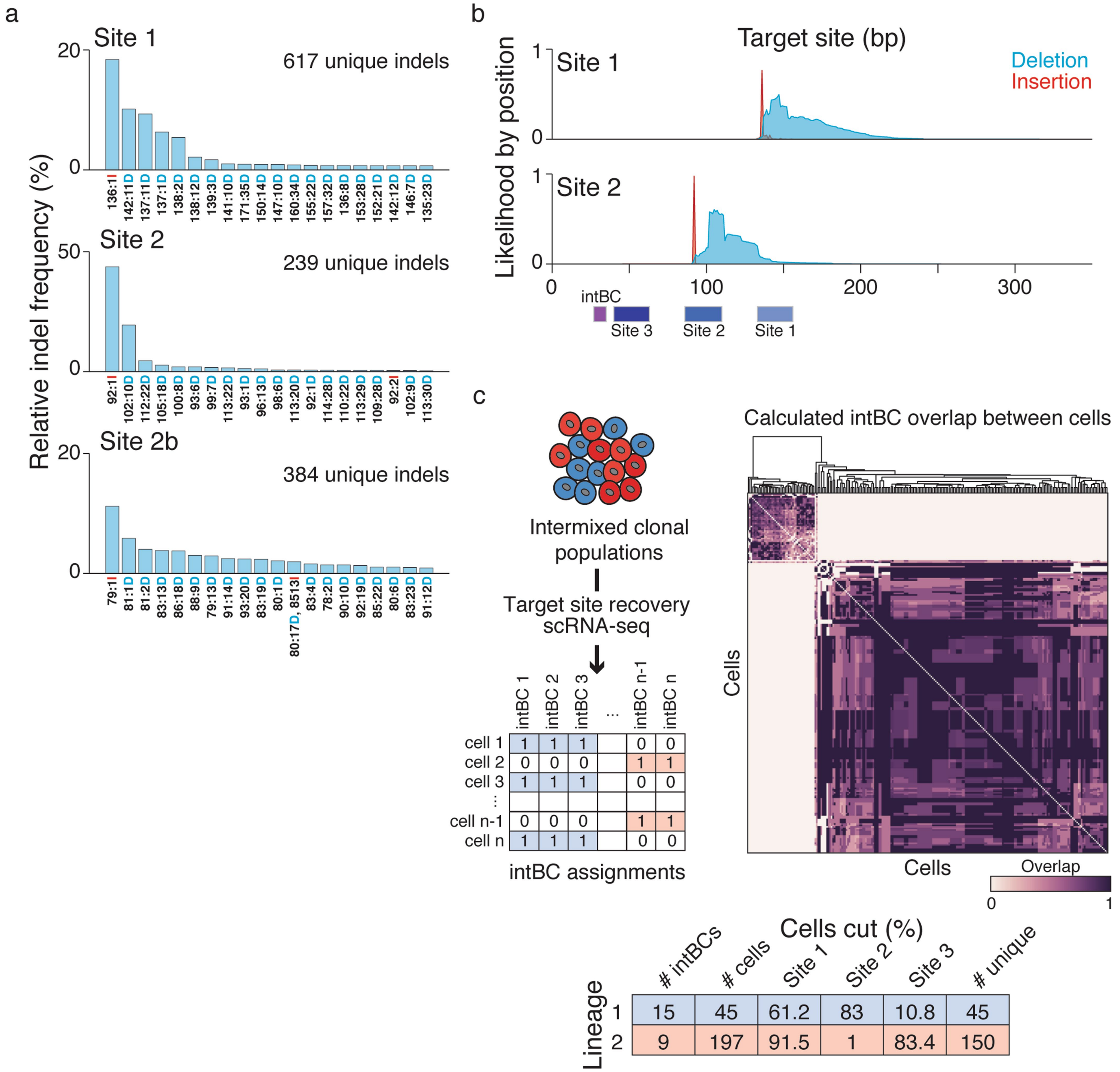
Acknowledgements We thank members of the Weissman, Meissner and Yosef laboratories, particularly J. Charlton and A. Arczewska for assistance with animal imaging, A. Kumar for explorations of extra-embryonic endoderm, and L. Gilbert and M. Horlbeck for guidance on technology design, as well as E. Chow and D. Bogdanoff from the UCSF Center for Advanced Technology for sequencing. This work was funded by National Institutes of Health Grants R01 DA036858 and 1RM1 HG009490-01 (J.S.W.), P50 HG006193 and R01 HD078679 (A.M.), F32 GM116331 (M.J.) and F32 GM125247 (J.J.Q.), and Chan-Zuckerberg Initiative 2018-184034. J.S.W. is a Howard Hughes Medical Institute Investigator. M.M.C. is a Gordon and Betty Moore fellow of the Life Sciences Research Foundation. T.M.N. is a fellow of the Damon Runyon Cancer Research Foundation. S.G., H.K. and A.M. are supported by the Max Planck Society.

Reviewer information *Nature* thanks Anna-Katerina Hadjantonakis, Patrick Tam, Valerie Wilson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.M.C., Z.D.S., A.M. and J.S.W. were responsible for the conception, design and interpretation of the experiments, and wrote the manuscript. M.M.C. and Z.D.S. conducted experiments, and M.M.C. developed the analysis with input from Z.D.S. S.G. and H.K. provided annotations for RNA-seq data, and assisted in experimental and analytical optimization. B.A., T.M.N. and M.J. provided vectors, experimental protocols and advice. J.J.Q. and D.Y. prepared several sequencing and transposon libraries and were engaged in discussion. M.G.J., A.K. and N.Y. provided phylogenetic reconstruction strategies.

Competing interests The authors declare no competing interests.

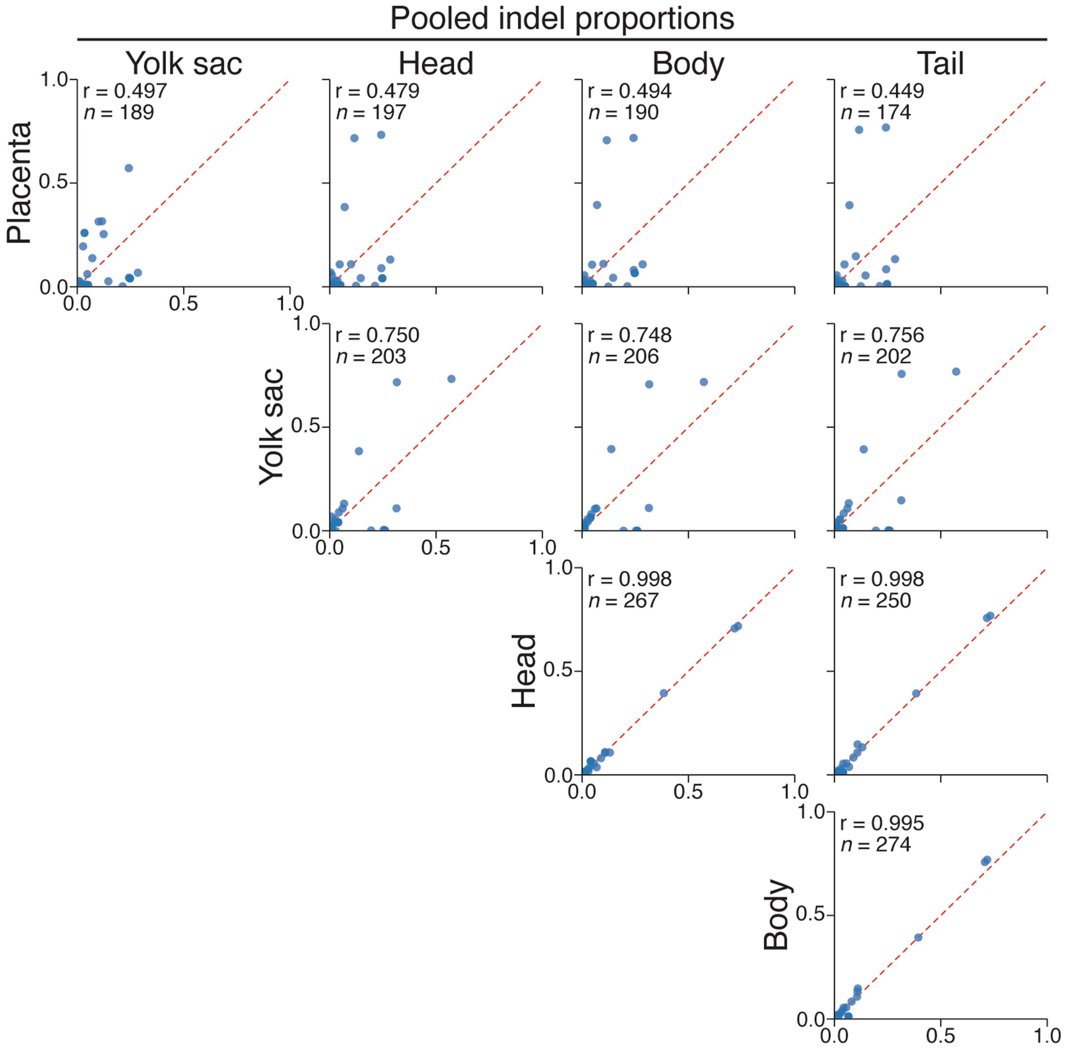
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1184-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1184-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.M. or J.S.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



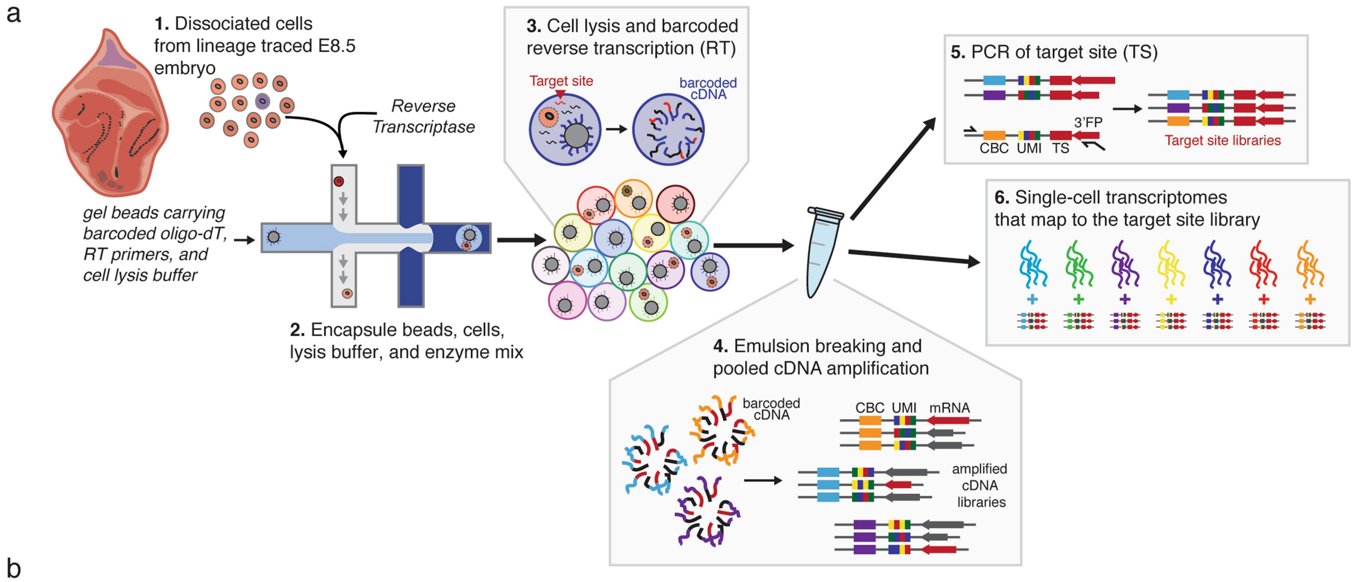
	intBC 1	intBC 2	intBC 3	...	intBC n-1	intBC n
cell 1	1	1	1		0	0
cell 2	0	0	0		1	1
cell 3	1	1	1		0	0
⋮						
cell n-1	0	0	0		1	1
cell n	1	1	1		0	0

Lineage	# intBCs	# cells	Site 1	Site 2	Site 3	# unique
1	15	45	61.2	83	10.8	45
2	9	197	91.5	1	83.4	150

Extended Data Fig. 1 | Target-site indel likelihoods from in vitro experiments. a, Histograms for the relative indel frequency for protospacer sites 1, 2 and 2b within the target site. In this experiment, sgRNA-expressing vectors respective to each position were delivered into K562 cells. Repair outcomes and frequencies are different for each site, but every site produces hundreds of discrete outcomes. The top 20 most-frequent indels for each site are shown. Site 3 was not profiled in this experiment. **b**, For sites 1 and 2, histograms representing the likelihood that any specific base in the target site is deleted (blue) or has an insertion (red) that begins at that position. The position of the intBC and protospacer sequences (sites) within the target site are represented as a schematic along the bottom, with the protospacer adjacent motif (PAM) for each site proximal to the intBC. Indels start at the double-stranded-break point, three bases from the PAM sequence. **c**, Simultaneous and continuous molecular recording of multiple clonal populations in K562 cells. We transduced K562 cells with a high-complexity library of unique intBCs, sorted them into wells of 10 cells each and propagated them for 18 days. At the end of the experiment, we detected two populations by their intBCs, which implies that only two clonal lineages expanded from the initial population of ten, and confirmed generation of target-site mutations. Top left, strategy for partitioning a multi-clonal population. Target sites are amplified from a single-cell barcoded cDNA library and the intBCs in each cell are identified as present or absent. Top right, heat map of the overlap of intBCs between all cells. The cells segregate into two populations that represent the descendants of two progenitor cells from the beginning of the experiment. Bottom, table summarizing results of the experiment, including the generation of indels over the experiment duration. These data additionally showcase our ability to combine dynamic recording with tracing based on traditional static barcodes.



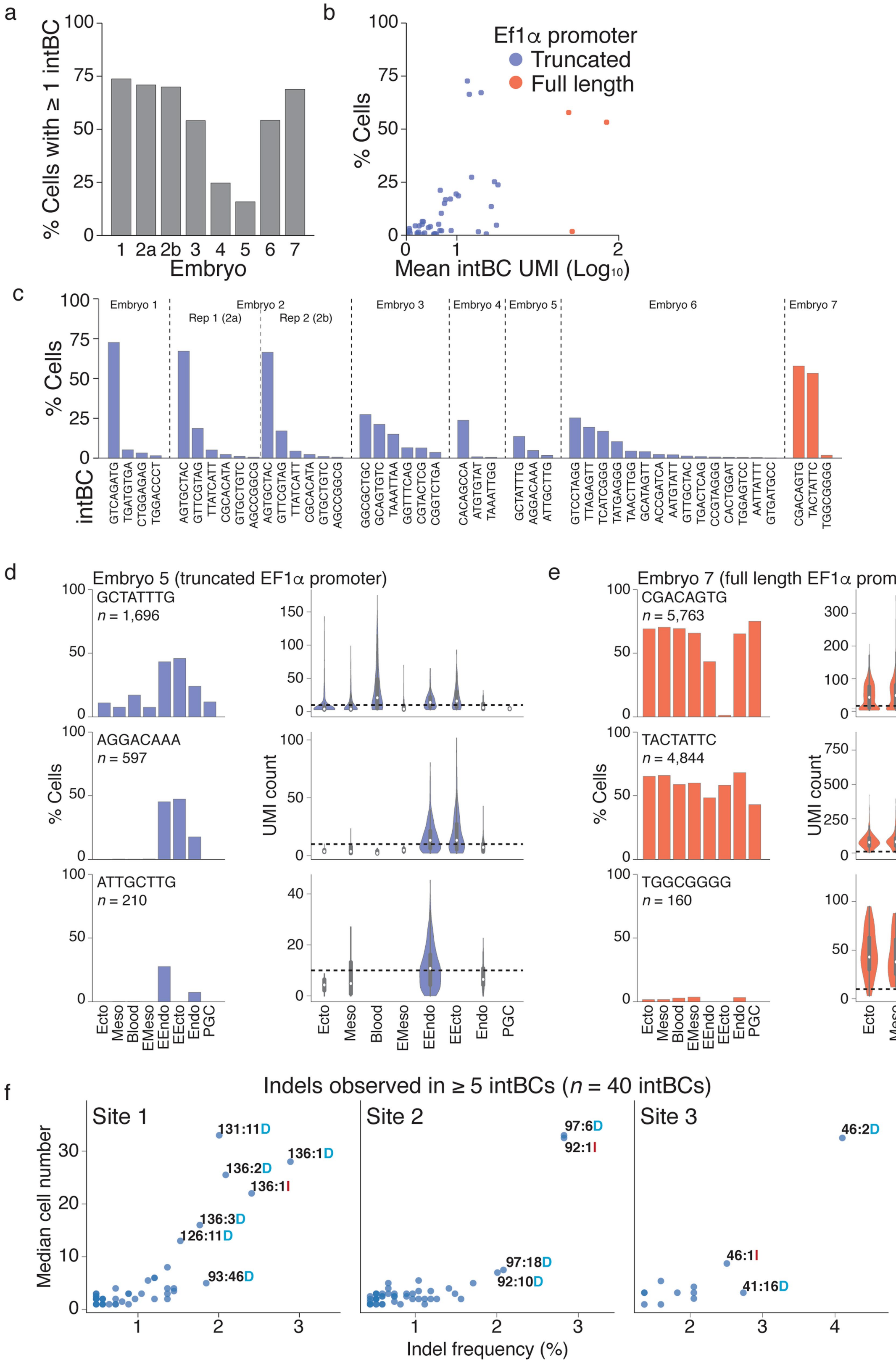
Extended Data Fig. 2 | Capturing early differentiation by pooled sequencing of indels generated within an E9.5 embryo. Scatter plots of indel proportions from dissected bulk tissue of an E9.5 embryo. Placenta is the most distantly related from embryonic tissues, followed by the yolk sac; the three embryonic compartments share the highest similarity. *n*, number of indels used in the comparison; *r*, Pearson correlation of the relative indel proportions. Each of the three sites is considered independently per intBC. A heat map representing the correlation coefficients appears in Fig. 2b.


a
1. Dissociated cells from lineage traced E8.5 embryo
Reverse Transcriptase
gel beads carrying barcoded oligo-dT, RT primers, and cell lysis buffer
2. Encapsule beads, cells, lysis buffer, and enzyme mix
3. Cell lysis and barcoded reverse transcription (RT)
Target site
barcoded cDNA
4. Emulsion breaking and pooled cDNA amplification
barcoded cDNA
CBC UMI mRNA
amplified cDNA libraries
5. PCR of target site (TS)
3'FP
CBC UMI TS
Target site libraries
6. Single-cell transcriptomes that map to the target site library

b

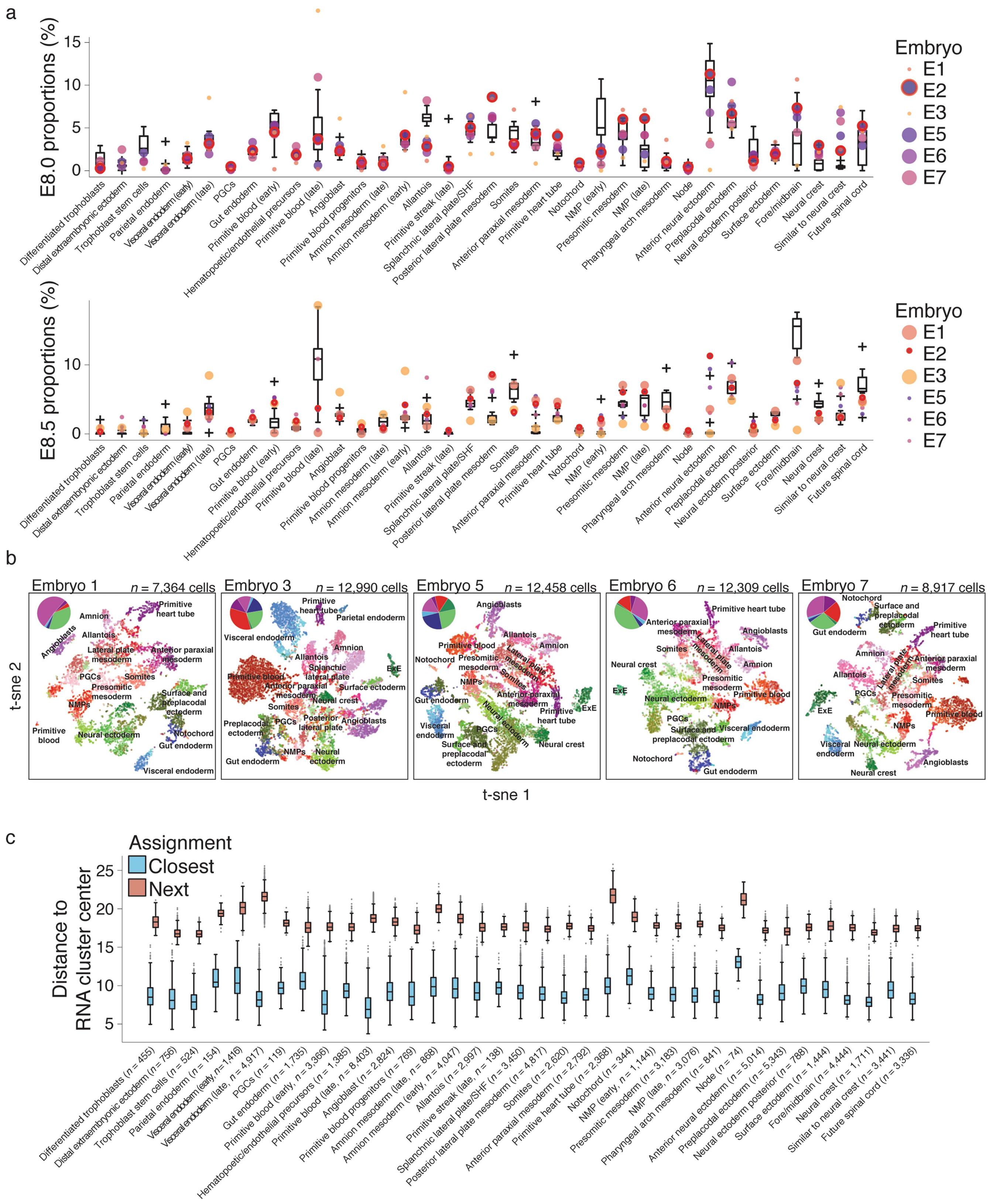
Embryo			Target site							RNA					
Rep	Cells	sgRNA Array	Reads	UMI	intBC	Cells	Alleles	per intBC	Indels per site	Reads	% Aligned	Cells	Reads (median)	UMI (median)	Genes (mean)
1	20,100	P,1,P	70,643,957	226,817	4	5,431	517	17.5	8.5	278,986,000	66.7	7,364	37,885	12,778	3,491
2a	52,200	1,1,1	20,138,720	219,331	6	8,120	895	41.5	15.5	236,945,147	70.2	11,458	20,679	9,355	2,645
2b	52,200	1,1,1	18,993,868	179,051	6	7,843	858	44	18	228,721,397	70.7	11,216	20,392	9,091	2,618
3	50,700	1,1,1	21,394,355	204,778	6	7,022	1,757	169.5	52	215,585,542	70.8	12,990	16,596	5,703	1,978
4	13,680	2,1,2	13,402,235	59,619	3	2,073	180	15	6	188,712,510	76.8	8,399	22,486	10,179	2,590
5	24,990	2,1,2	14,794,267	75,828	3	1,975	167	41	10	217,409,595	74.8	12,458	17,451	7,523	2,408
6	21,000	2,1,2	10,758,332	254,230	15	6,674	2,461	26	10	182,234,141	74.0	12,309	14,804	6,493	2,274
7	56,400	2,1,P	13,050,705	1,007,630	3	6,143	170	21	11	281,462,426	70.1	8,917	31,564	13,893	3,339

Extended Data Fig. 3 | Experimental overview. a, Schematic of platform used for generation of scRNA-seq libraries and corresponding target-site amplicon libraries, adapted from a previous study[19]. The barcoded and amplified cDNA library is split into two fractions: one fraction is used to generate a global transcription profile and the other is used to specifically amplify the target site. CBC, cell barcode; UMI, unique molecular identifier. **b**, Summary of lineage-traced embryos detailing the type of guides used, the sampling proportion and sequencing results. Cells from embryo 2 were run on two 10x lanes. Embryo 4 was omitted from further analysis owing to the absence of cells identified as primitive heart tube. The sgRNA array is listed in order from site 1 to site 3: P, perfect complementarity between guide RNA and protospacer; 2, mismatch in region 2; 1, mismatch in region 1.



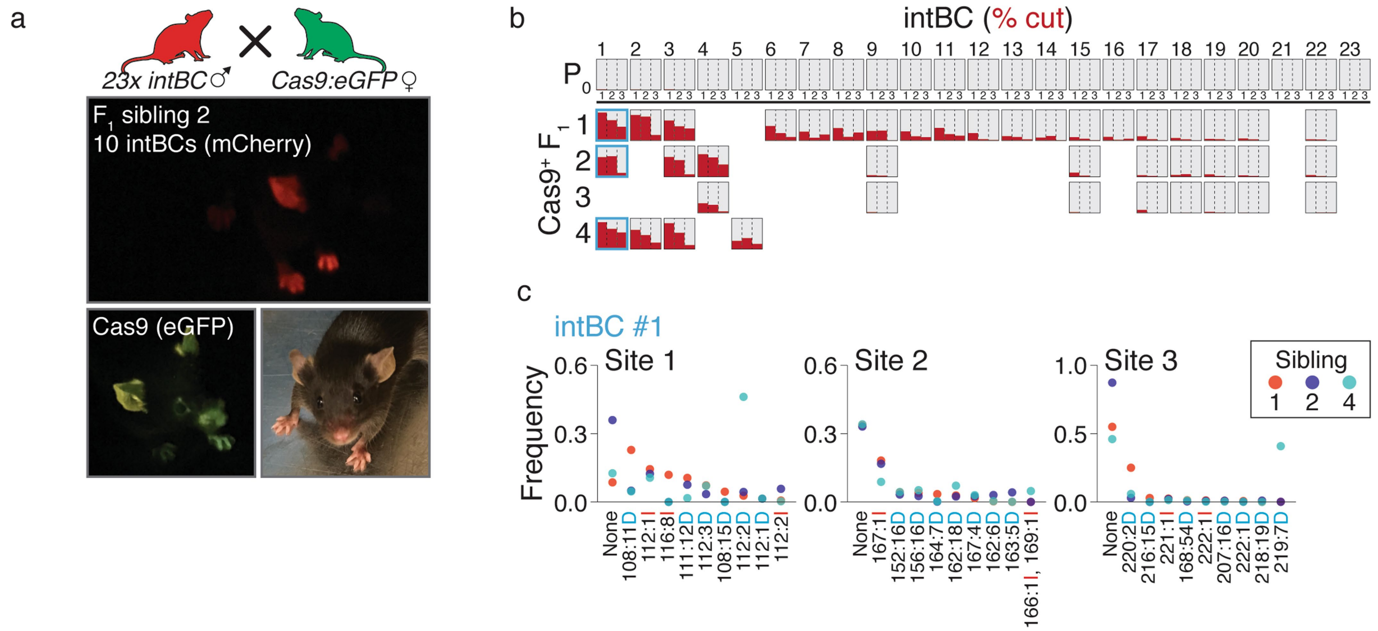
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Target-site capture in mouse embryos. **a**, Percentage of cells with at least one target site captured. Cells from embryo 2 were run on two 10x lanes. **b**, Scatter plot showing the relationship between the mean number of UMIs (a proxy for expression level) sequenced per target site and the percentage of cells in which the target site is detected, which we refer to as 'target-site capture'. In general, as the mean number of UMIs increases, the percentage of cells also increases. Using a full-length, intron-containing EF1α promoter in mouse embryos leads to a higher number of UMIs, which generally results in better target-site capture. **c**, Percentage of cells for which a given intBC is detected across all seven embryos profiled in this study. **d**, Target-site capture and expression level across tissues for embryo 5, which uses a truncated EF1α promoter to direct transcription of the target site. Each row corresponds to a different intBC, indicated in the top left of the histogram. Left, the percentage of cells in each tissue for which the target site is captured. Right, violin plots representing the distribution of UMIs for the target site in each tissue. Dashed line refers to a ten-UMI threshold. The target site may be expressed at different levels in a tissue-specific manner, which leads to higher likelihoods of capture in certain tissues. Biased capture of target sites that carry the intBCs AGGACAAA and ATTGCTTG may also be explained by mosaic integration after the first cell cycle as their capture is preferential to extra-embryonic lineages that are restricted early in development. White dot indicates the median UMI count for cells from a given tissue, edges indicate the interquartile range, and whiskers denote the full range of the data. **e**, Target-site capture and expression level across tissues for embryo 7, which drives target-site expression from an intron-containing EF1α promoter. Each row corresponds to a different intBC, indicated in the top left of the histogram. Left, the percentage of cells in each tissue for which the target site is captured. Right, violin plots representing the distribution of UMIs for the target site in each tissue as in **d**. Dashed line is a visual threshold for ten UMIs. Although tissue-specific expression may explain some discrepancy in target-site capture, high expression (as estimated from the number of UMIs) can still correspond to low capture rates, as observed for the intBC TGGCGGGG. One possibility is that particular indels may destabilize the transcript and lead to either poor expression or capture. **f**, Scatter plots that show the relationship between estimated relative indel frequency and the median number of cells that carry the indel. Because the indel frequency within a mouse is dependent on the timing of the mutation, we cannot calculate the underlying indel frequency distribution using the fraction of cells within embryos that carry a given indel. Instead, we estimate this frequency by the presence or absence of an indel using all of the target-site integrations across mice, which reduces biases from cellular expansion but assumes that any given indel occurs only once in the history of each intBC. Because the number of integrations is small, we might expect our estimates to be poor. Here we see that the number of cells marked with an indel increases with indel frequency, which suggests that our frequency estimates are underestimated for particularly frequent indels. This is probably due to the fact that we cannot distinguish between identical indels in the same target site that may have resulted from multiple repair events (convergent indels). The most frequent insertions are of a single base and tend to be highly biased towards a single nucleotide (for example, 92:1:I is uniformly an 'A' in 5 out of 7 embryos, and never below 88%).



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

Extended Data Fig. 5 | scRNA-seq tissue assignment and wild-type comparison. a, Box plots representing tissue proportions from E8.0 (top) and E8.5 (bottom) wild-type embryos ($n = 10$ each), with lineage-traced embryos mapping to each state overlaid as dots. Wild-type embryos display a large variance in the proportions of particular tissues, and the proportions of our lineage-traced embryos generally fall within the range of those recovered from wild type. Large circles indicate embryos that were scored as either E8.0 or E8.5, and the bold red overlay highlights embryo 2, which is used throughout the text. Note that many processes—such as somitogenesis and neural development—are continuous or ongoing between E8.0 to E8.5. For example, from E8.0 to E8.5, the embryonic proportions of anterior neural ectoderm and fore- and midbrain are inversely correlated, as one cell type presumably matures into the other. Many of our embryos scored as E8.0 exhibit intermediate proportions for both tissue types, which supports the possibility that these embryos are slightly less developed than E8.5 but more developed than E8.0. For box plots, the centre line indicates the median, edges indicate the interquartile range, whiskers indicate the Tukey fences, and crosses denote outliers. **b**, *t*-SNE plots of scRNA-seq data with corresponding tissue annotations for the six lineage-traced embryos used in this study. Insets, pie charts of the relative proportions for different germ layers. Mesoderm is further separated to include blood (red). Although 36 different states are observed during this developmental interval, only broad classifications of particular groups (for example, neural ectoderm or lateral plate mesoderm) are overlaid to provide a frame of reference. In general, the relative spacing and coherence of different cell states are consistent across different embryos. **c**, Box plots of the Euclidean distance between single-cell transcriptomes and the average transcriptional profile of their assigned cluster (cluster centre) in comparison to their distance from the average of the next-closest possible assignment. Comparison is to the same 712 informative marker genes that were used to assign cells to states, and includes all cells used in this study (Supplementary Methods). Middle bar highlights the median, edges indicate the interquartile range, whiskers indicate the Tukey fences, and grey dots denote outliers. n values refer to the cumulative number of cells assigned to each state across all seven embryos for which single-cell data were collected, including for embryo 4, which was ultimately withheld from further analysis owing to the lack of primitive heart tube development.



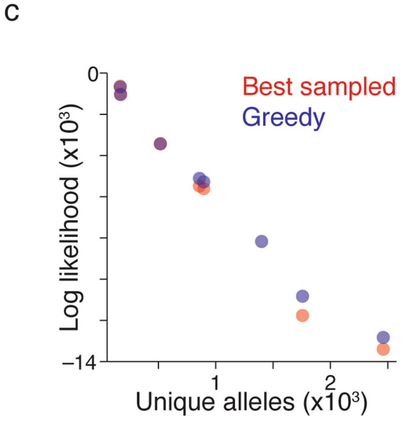
Extended Data Fig. 6 | Continuous indel generation by breeding. **a**, Strategy for generating lineage-traced mice through breeding. The target site and guide array cassette are integrated into mouse zygotes as in Fig. 2a using sperm from C57BL/6J mice to generate P_0 breeder mice, which are capable of transmitting high-copy genomic integrations of the technology. Then, P_0 mice are crossed with homozygous transgenic mice that constitutively express Cas9 to enable continuous cutting from fertilization onwards in F_1 progeny. Sibling 2 of a cross between a P_0 male and a Cas9:eGFP female is shown. **b**, Bar charts show the degree of mutation (per cent cut, red) for a P_0 male (top row) and four F_1 offspring generated by breeding with a Cas9:eGFP female before weaning (21 days post partum). Each row represents a mouse and each column represents a target site. Each sibling inherits its own subset of the 23 parental target-site integrations, and demonstrates different levels of mutation throughout gestation and maturation. **c**, Indel frequencies for the ten most-frequent indels from three siblings in a common target-site integration (column 1 in **b**). Each mouse shows a large diversity of indels, and the different frequencies observed in each mouse demonstrate an independent mutational path.

a

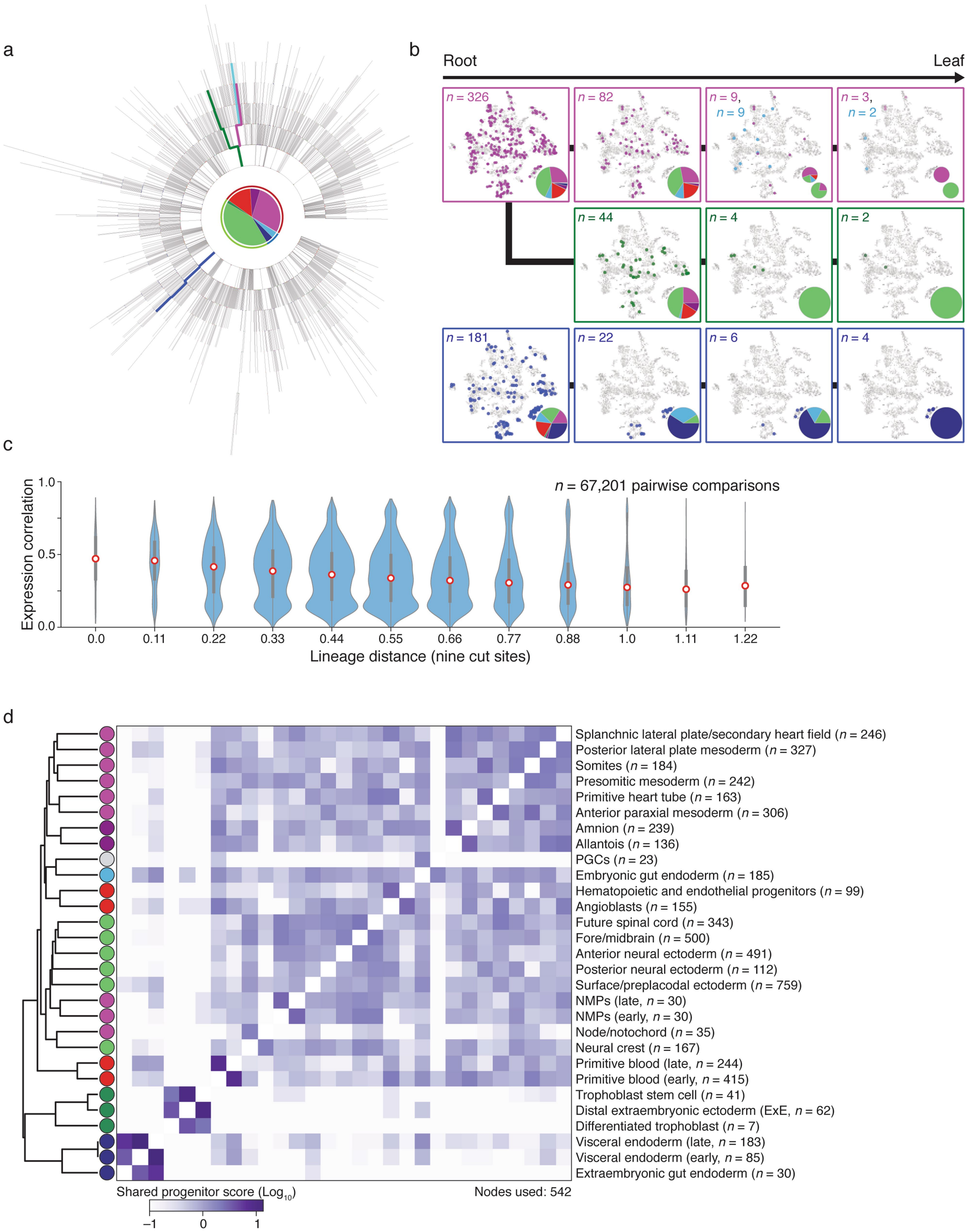
	scGESTALT[15] (Raj *et al.*)	LINNEAUS[17] (Spanjaard *et al.*)	ScarTrace[16] (Alemany *et al.*)	Kalhor *et al.*[12]	This manuscript (Chan *et al.*)
Organism	Zebrafish	Zebrafish	Zebrafish	Mouse	Mouse
Single cell	Yes	Yes	Yes	No	Yes
Continuous recording	No	No	No	Yes	Yes
Cut sites (number)	10	16-32	8	60	9-45
Recovery rate	6-28%	14.5-99.6%*	100%**	N/A	15.8-73.7%***
Integration barcode	No	No	No	Yes	Yes
Distributed barcode	N/A; single integration	Yes	Tandem integrations	Yes	Yes
Designed for recutting	No	No	No	Yes	No
Reconstruction strategy	Camin-Sokal	Custom (max. parsimony)	Clonal analysis****	Manhattan distance	Custom greedy

b

		Sampled			Greedy	
Rep	**Alleles**	**Simulations**	**Nodes**	**Log likelihood**	**Nodes**	**Log likelihood**
1	517	145,384	686	**–3,438**	655	–3,440
2a	895	112,247	1,246	–5,615	1,134	**–5,287**
2b	858	119,123	1,203	–5,490	1,089	**–5,119**
2	1,400	N/A	N/A	N/A	1,732	**–8,176**
3	1,757	62,920	2,601	–11,766	2,319	**–10,831**
5	167	150,000	150	**–651**	156	–686
6	2,461	42,820	2,935	–13,399	2,690	**–12,831**
7	170	150,000	203	**–1,037**	201	–1,054

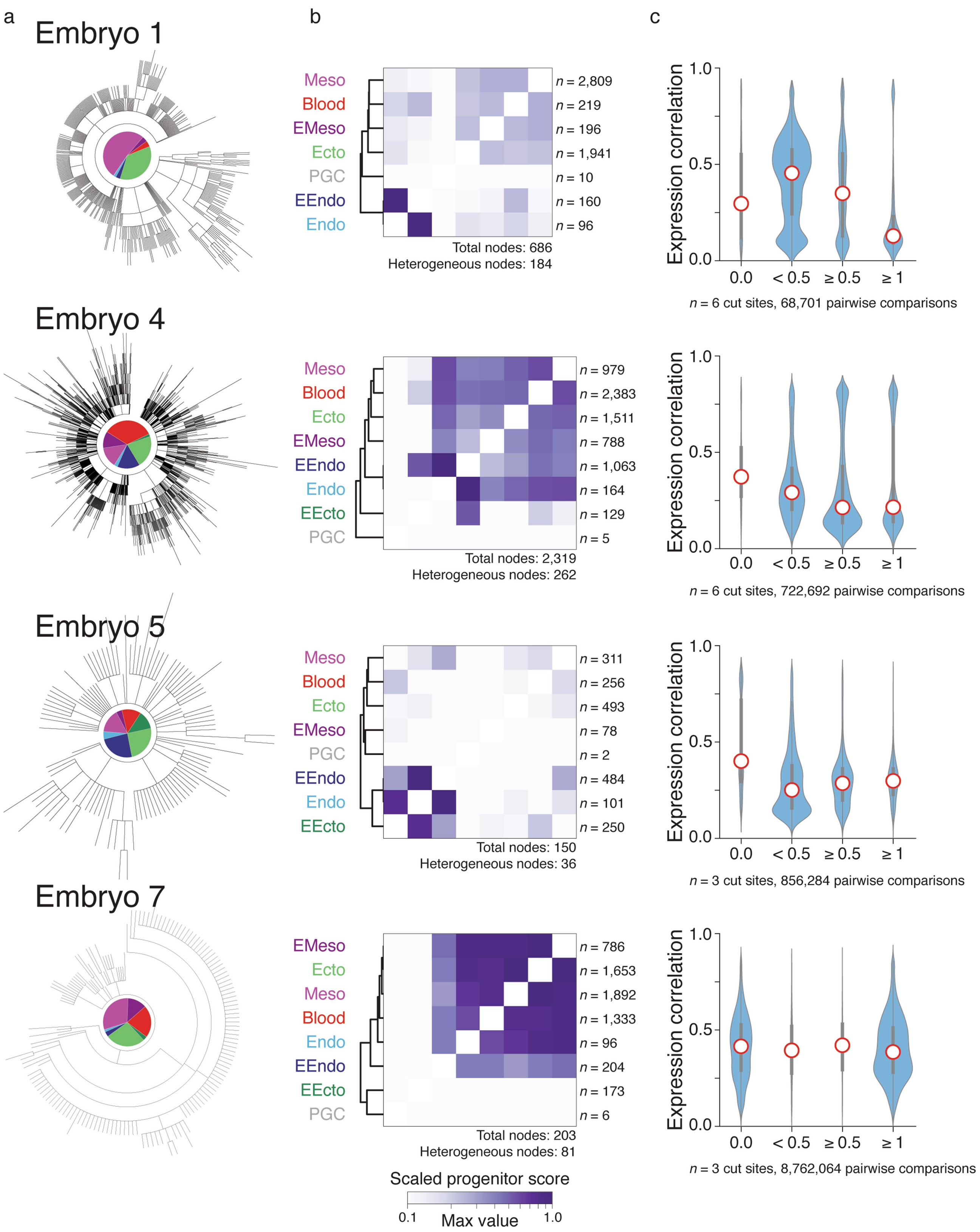


Extended Data Fig. 7 | Performance of tree-building algorithms used on embryonic data. a, Table summarizing contemporary Cas9-based lineage tracers that have been applied to vertebrate development, highlighting attributes that differ between the studies. See Supplementary Methods for a more detailed overview of key characteristics of our technology. Single asterisk denotes that the study reports the average fraction recovered by tissue for integrations that cannot be distinguished, such that percentages reported here are effectively equivalent to our '≥1 intBC' metric. Double asterisk indicates that the value refers to a plate-based DNA-sequencing approach that can be applied to all methods to improve target-site recovery. Triple asterisk denotes a range of cells in which at least one intBC is confidently detected and scored. Quadruple asterisk denotes that the study presents a tree reconstruction method, but includes results that predominantly rely on clonal analysis. **b**, Table of allele complexity, number of nodes and log-likelihood scores for embryos. Tree likelihoods are calculated using indel frequencies estimated from all embryo data (Fig. 2c, Extended Data Fig. 4, Supplementary Methods). Bold scores indicate the reconstruction algorithm selected for each embryo (trees shown in Fig. 4, Extended Data Figs. 8, 9). **c**, log-likelihood of trees generated using either the greedy or biased sampling approach as a function of complexity, which is measured as the number of unique alleles. There is near-equivalent performance of the two algorithms for low-complexity embryos, but the greedy algorithm produces higher-likelihood trees for embryos with larger numbers of unique alleles.



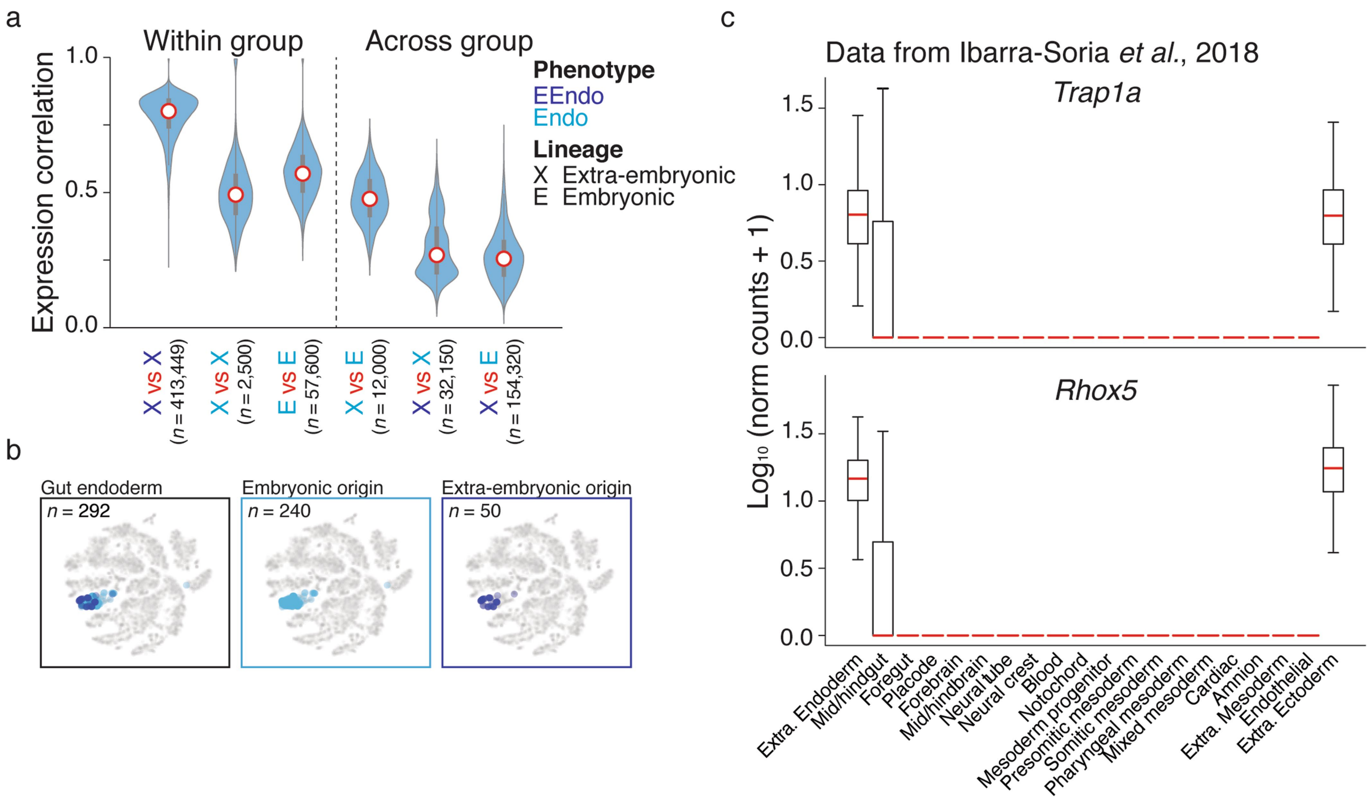
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Single-cell lineage reconstruction of early mouse development for embryo 6. a, Reconstructed lineage tree comprising 2,690 nodes generated from our most information-dense embryo (embryo 6) that we used to compare shared progenitor scores with embryo 2 in Fig. 4d. Each branch represents an independent indel generation event. **b**, Example paths from root to leaf from the selected tree (highlighted by colour). Cells for each node in the path are overlaid onto the *t*-SNE representation in Extended Data Fig. 5, with the tissue proportion for cells within each node included as a pie chart (colours are as in Fig. 3b). In the top path (pink), the lineage bifurcates into two independently fated progenitors that either generate mesoderm (secondary heart field and primitive heart tube) or neural ectoderm (anterior neural ectoderm and neural crest). Note that the middle path (green) also represents an earlier bifurcation from the same tree, and eventually produces neural ectoderm (neural crest and future spinal cord). These paths begin with a pluripotent node that can generate visceral endoderm, but subsequently lose this potential. The bottom path (dark blue) begins in an equivalently pluripotent state but becomes restricted towards the extra-embryonic visceral-endoderm fate. **c**, Violin plots that represent the relationship between lineage and expression for individual pairs of cells as calculated for embryo 2 in Fig. 4c. Expression Pearson correlation decreases with increasing lineage distance, which shows that closely related cells are more likely to share function. Red dot highlights the median, edges indicate the interquartile range, and whiskers indicate the full range. **d**, Comprehensive clustering of shared progenitor scores for embryo 6, which has the greatest number of unique alleles and samples multiple extra-embryonic tissue types. Shared progenitor score is calculated as the sum of shared nodes between cells from two tissues, normalized by the number of additional tissues that are also produced (a single shared progenitor score is calculated as $2^{-(n-1)}$, in which n is the number of clusters present within that node). In general, extra-embryonic tissues that are specified before implantation—such as extra-embryonic endoderm or ectoderm—co-cluster away from embryonic tissues and within their own groups, whereas the amnion and allantois of the extra-embryonic mesoderm cluster with other mesodermal products of the posterior primitive streak. The co-clustering of anterior paraxial mesoderm and somites may reflect the continuous nature of somitogenesis from presomitic mesoderm during this period, with production of only the anterior-most somites by E8.5. Note that the gut endoderm cluster has been further portioned according to embryonic or extra-embryonic lineage (Fig. 5).



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Summary of results from additional mouse embryos. a, **b**, Representative highest-likelihood tree analyses for additional embryos, including reconstructed trees as shown in Fig. 4a (**a**) and shared progenitor score heat maps as shown in Fig. 5a (**b**), normalized to the highest score for each embryo to account for differences in total node numbers. Here the shared progenitor score is calculated as the number of nodes that are shared between tissues, scaled by the number of tissues within each node (a single shared progenitor score is calculated as $2^{-(n-1)}$, in which n is the number of clusters present within that node). In general, the clustering of shared progenitors is recapitulated across embryos, with mesoderm and ectoderm sharing the highest relationship and either extra-embryonic ectoderm or extra-embryonic endoderm representing the most-deeply rooted and distinct outgroup, although these scores are sensitive to the number of target sites, the rate of cutting and the number of cells in the cluster. By shared progenitor, primordial germ cells (PGCs) are also frequently distant from other embryonic tissues; however, this often reflects the rarity of these cells, which restricts them to only a few branches of the tree in comparison to better-represented germ layers. The number of heterogeneous nodes from which scores are derived is included for each heat map. **c**, Violin plots that represent the pairwise relationship between lineage distance and transcriptional profile as shown for embryo 2 in Fig. 4c. Lineage distance is calculated using a modified Hamming distance, and transcriptional similarity by Pearson correlation. The exact dynamic range for lineage distance depends on the number of intBCs included and the cutting rate of the three-guide array. Here distances are binned into perfect (0), close ($0 > x > 0.5$), intermediate ($0.5 \leq x < 1$), and distant ($x \geq 1$) relationships for all cells that contain either three or six cut sites, depending on the embryo. As lineage distance increases, transcriptional similarity decreases, which is consistent with functional restriction over development. Red dot highlights the median, edges indicate the interquartile range, and whiskers denote the full range.



Extended Data Fig. 10 | Expression characteristics of extra-embryonic and embryonic endoderm. a, Violin plots that represent the pairwise scRNA-seq Pearson correlation coefficients for within- or across-group comparisons according to lineage (X, extra-embryonic; E, embryonic) and cluster assignment (light blue, gut endoderm; dark blue, visceral endoderm). Within-group comparisons for cells with the same lineage and transcriptional cluster identity are shown on the left, and across-group comparisons are presented on the right. Notably, extra-embryonic cells with gut-endoderm identities show higher pairwise correlations to embryonic cells with gut-endoderm identities (column 4) than they do to visceral-endoderm cells, with which they share a closer lineage relationship (column 5). Red dot highlights the median, edges indicate the interquartile range, and whiskers denote the full range. *n*, number of pairwise comparisons between cells in embryo 2. **b**, *t*-SNE plots of scRNA-seq data for embryo 2, with gut-endoderm cells highlighted. Endoderm cells segregate from the rest of the embryo, and cannot be distinguished by embryonic (light blue) or extra-embryonic (dark blue) origin. *n*, number of cells for embryo 2. Cells of ambiguous origin are not included in the two right-most plots. **c**, Expression box plots for the extra-embryonic markers *Trap1a* and *Rhox5* from an independent scRNA-seq survey of E8.25 embryos (data and annotations taken from a previous study[9]). Both genes are heterogeneously present in cells identified as mid- and hindgut but uniformly present in canonical extra-embryonic tissues, which is consistent with the presence of a subpopulation of cells of extra-embryonic origin that resides within this otherwise-embryonic cluster. Red lines highlight the median, edges indicate the interquartile range, and whiskers denote the Tukey fence. Outliers were removed for clarity.

nature research

Corresponding author(s): Jonathan Weissman, Alexander Meissner

Last updated by author(s): Mar 18, 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

scRNA-seq data was processed and aligned using 10x Cell Ranger v2. The filtered gene-barcode matrices were then processed in Seurat v2.0 (https://satijalab.org/seurat/) for data normalization (global scaling method “LogNormalize”), dimensionality reduction (PCA), and generation of t-sne plots, which use the first 16 principal components. Amplicons were additionally processed using cutadapt v1.14 to remove sequence beyond the polyA (http://cutadapt.readthedocs.io/en/stable/) and BioPython v1.7 to build a consensus sequence from multiple, trimmed UMIs using parameters described in the methods. We use emboss water (v6.6.0) to align sequences to the target site reference sequence with the following parameters, which were determined empirically: [–asequence targetSiteRef.fa –sformat1 fasta –bsequence consensusUMI.fa –sformat2 fasta –gapopen 15.0 –gapextend 0.05 –outfile sam –aformat sam]. Additional processing of the resulting alignment files was done in perl.

Data analysis

Following processing using a custom software pipeline described above, data was analyzed using Python.

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

Lineage tracing data is available in GEO with accession number GSE117542. Wild type embryo data is available under GSE122187. All figures use raw data generated in this 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

No sample size calculation was performed, the number of embryos reported is the number that we were able to generate in a reasonable cost and time frame. Each embryo generates a stochastic mutational path from the delivery of the target site to collection at E8.5, requiring the innovation of novel analytical tools and perspectives to confirm the reproducible generation of similar lineages, which was accomplished via strategies such as the shared progenitor score between different tissues. Each embryo collected demonstrated our reproducible ability to recover indels according to the rate of the guide series and to assign them to matched transcriptional profiles from the same cells. Moreover, the general composition of embryos was also consistent and fell between E8.0 and E8.5 when compared to a wild type compendium.

Data exclusions

We excluded one of the seven embryos for which we generated single cell data from detailed lineage analysis because it did not produce cells of the primitive heart tube, suggesting a developmental abnormality that may be due to the mutational nature of the randomly integrating, target site containing piggyBAC transposon.

Replication

We demonstrate the reproducible nature of our technology to be recovered from early embryos and to be assigned to a consistent make up of developmental cell types. The reproducibility of lineage relationships is more difficult to evaluate due to the stochastic nature of indel generation, though we confirm general trends between high complexity embryos by comparing shared progenitor scores as a proxy for the ancestral relationships between different tissues. Analysis related to the reproducibility of lineage relationship is presented for each of the six morphologically normal embryos in Figures 4, Extended Data Figures 8 and 9

Randomization

As our objective was to recover high complexity embryos with as large a number of integrated target sites as possible, embryos were selected for inclusion in this study based upon the uniform brightness and high intensity of the target-site linked reporter. Our study does not follow a hypothesis driven design, and as such, no groupings of embryos were made therefore randomization was not applicable.

Cells were assigned to states according to their Euclidean Distance to the 712 marker genes described in the methods and available as a supplementary file in GSE122187. The robust nature of these assignments were confirmed by comparing the distance for each assigned cell to its closest and next closest cluster center (see Extended Data Figure 5c).

To estimate shared progenitor scores, we downsampled the number of cells from each tissue before calculating: 150 cells were randomly sampled from each tissue and the tree was pruned to only include the sampled cells. For tissues with less than 150 cells, all cells were included. For embryo 2, we downsampled to 300 cells since it is a merger of two biological replicates and is therefore doubly sampled. The shared progenitor score was calculated from the pruned tree and the process was repeated 1000 times for each embryo. The median progenitor score is presented in the heatmap and was used instead of the mean to prevent potential outlier effects.

During tree reconstruction, we attempted one of two different approaches, "Biased Search Through Phylogenetic Space" and "Greedy." In Biased search, we generated trees by selecting indels either randomly or according to their frequency normalized weight (the fraction of alleles an indel is found divided by its independent frequency, see Figure 2c). In these cases, we generated >30,000 simulated trees and calculated the log likelihood of each by summing the likelihoods of all indels that appear in the tree and reported the one with the highest likelihood. We also employed a greedy algorithm that recursively splits cells into mutually exclusive groups based upon the presence or absence of a specific mutation, prioritizing mutations that appear frequently within the embryo but are improbably according to their independent likelihood (see Figure 2c). This approach yields only one tree, which was only selected if it performed better than the best tree recovered by our sampling approach. Finally, the cumulative tree for embryo 2 could only be generated with this approach is it includes too many cells to enable robust sampling over ~100,000 simulations.

Blinding

Tree building and cell state assignments operate with the same parameters independently of the embryo used. As such, there is no need to blind the investigator to the data being handled. Individual parameters were not altered according to the specific features of a given sample, with the following minor exceptions:

The number for assigning shared progenitor scores was set according to the overall complexity (number of cells within each tissue) to 300 for embryo 2 and 150 for all other embryos because embryo 2 was sampled ~2x more deeply.

The number of frequency normalized weighted tree simulations for each embryo depended on the number of alleles: higher allele numbers underwent fewer simulations due to increased processing times. In these cases, the greedy algorithm consistently yielded trees with appreciably higher probabilities.

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)	The K562 cell line originated from ATCC.
Authentication	Cytogenetic profiling by array comparative genomic hybridization closely matches previous characterizations of the K562 cell line (Naumann et al., 2001).
Mycoplasma contamination	Cell lines tested negatively for mycoplasma
Commonly misidentified lines (See ICLAC register)	None used

Animals and other organisms

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

Laboratory animals	Oocytes were isolated from B6D2F1 strain female mice (age 6 to 8 weeks, Jackson Labs) , sperm was isolated from 2 8 week old Gt(ROSA)26Sortm1.1(CAG=cas9*,EGFP)Fezh/J strain mouse (Jackson labs) or C57BL/6J strain mice. Blastocysts were transferred into CD-1 strain female mice (age 6-10 week old).
Wild animals	None used
Field-collected samples	None used
Ethics oversight	All procedures follow strict animal welfare guidelines as approved by Harvard University IACUC protocol (#28-21).

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

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	K562 cells were filtered to make a single cell suspension.
Instrument	LSR-II flow cytometer (BD Biosciences)
Software	FlowCytometryTools (http://eyurtsev.github.io/FlowCytometryTools/)
Cell population abundance	To isolate reporter cell lines, the population abundance was <10% of the unsorted population.
Gating strategy	Cells were sorted against a negative control, or gating thresholds were obvious from bimodality in the cell population for highly expressed fluorescent proteins.

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.