DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

# Molecular Recording of Mammalian Embryogenesis

## Citation

Chan, Michelle M., Zachary D. Smith, Stefanie Grosswendt, Helene Kretzmer, Thomas Norman, Britt Adamson, Marco Jost, Jeffrey J. Quinn, Dian Yang, Alexander Meissner, Jonathan S. Weissman. 2019. Molecular Recording of Mammalian Embryogenesis. Nature 570, no. 7759: 77-82.

## Published Version

doi: 10.1038/s41586-019-1184-5

## Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:41999978

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

## Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

# Molecular recording of mammalian embryogenesis

Michelle M. Chan[1,2]*, Zachary D. Smith[4,5,6]*, Stefanie Grosswendt[7], Helene Kretzmer[7], Thomas Norman[1,2], Britt Adamson[1,2], Marco Jost[1,2,3], 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[4,5,7], Jonathan S. Weissman[1,2]

[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]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA, USA

[4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[5]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA

[6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA

[7]Department of Genome Regulation, Max Planck Institute for Molecular Genetics, Berlin 14195, Germany

[8]Integrative Program in Quantitative Biology, University of California, San Francisco, San Francisco, California, USA

[9]Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, California, USA

[10]Center for Computational Biology, Berkeley, California, USA

[11]Chan Zuckerberg Biohub, San Francisco, California, USA

[12]Ragon Institute of Massachusetts General Hospital, MIT and Harvard, Cambridge, Massachusetts, USA

*These authors contributed equally to this work

Correspondence: jonathan.weissman@ucsf.edu (J.S.W.), meissner@molgen.mpg.de (A.M.)

**Ontogeny describes the emergence of complex multicellular organisms from single totipotent cells. In mammals, this field is particularly challenging due to the indeterminate relationship between self-renewal and differentiation, variation of progenitor field sizes, and internal gestation. Here, we present a flexible, high information, multi-channel molecular recorder with a single cell (sc) readout and apply it as an evolving lineage tracer to define a mouse cell fate map from fertilization through gastrulation. By combining lineage information with scRNA-seq profiles, we recapitulate canonical developmental relationships between different tissue types and reveal the nearly complete transcriptional convergence of endodermal cells from extra-embryonic and embryonic origins. Finally, we apply our cell fate map 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 to facilitate a quantitative framework for understanding developmental processes.**

Development of a multicellular organism from a single cell is an astonishing process. Classic lineage tracing experiments using *C. elegans* revealed surprising outcomes, including deviations between lineage and functional phenotype, but nonetheless benefited from the highly deterministic nature of this organism's development[1]. Alternatively, more complex species generate larger, more elaborate structures that progress through multiple transitions, raising questions regarding 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, producing 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[11-13], and combined with scRNA-seq to generate fate maps in zebrafish[14-16]. 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 these questions would possess the following characteristics: 1) minimal impact 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[14-16]; 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, where spatial plans are gradually and continuously specified and may originate from small, transient progenitor fields. Moreover, scRNA-seq has revealed populations of cells with a continuous spectrum of phenotypes, implying that differentiation does not occur instantaneously, further motivating the need for an evolving recorder[17].

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

## Results

### A transcribed, multi-channel, and continuously evolving molecular recorder

To achieve our goal of a tunable, high information content molecular recorder, we utilized Cas9 to generate insertions or deletions (indels) upon repair of double-stranded breaks, which are inherited in the next generation of cells[11-16]. We record within a 205 base pair, synthetic DNA “target site” containing three “cut sites” and a static 8 base pair “integration barcode” (intBC), which are delivered in multiple copies via piggyBac transposition (**Fig. 1a, b**). We embedded this sequence into the 3’UTR 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 (**Fig. 1a, b**)[18].

Our system is capable of high information storage due to the diversity of heritable repair outcomes, and the large number of targeted sites, which can be distinguished by the intBC (**Fig. 1c**). DNA repair generates hundreds of unique indels, and the distribution for each cut site is different and nonuniform: some produce highly biased outcomes while others create a diverse series (**Fig. 1c**, **Extended Data Fig. 1**)[19-21]. To identify sequences that can tune the mutation rate of our recorder for timescales that are not pre-defined, and may extend from days to months, we screened several guide RNA series containing mismatches to their targets[22] by monitoring their activity on a GFP reporter over a 20-day timecourse and selected those that demonstrated a broad dynamic range (**Fig. 1d**). Slower cutting rates may improve viability *in vivo*, as frequent Cas9-mediated double-strand breaks can cause cellular toxicity[23,24]. 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, allowing us to assess global transcriptomes and specifically amplify mutated target sites (**Extended Data Fig. 1c**).

## Tracing cell lineages in mouse development

We next applied our technology to map cell fates during mouse early development 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 ~embryonic day (E)8.5 or E9.5 (**Fig. 2a**, **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 Figure 2**).

Following this *in vivo* proof of principle, we generated single cell data from additional embryos (**Extended Data Figure 3**). We collected scRNA-seq data for 7,364 – 12,990 cells from 7 embryos (~15.8% – 61.4% of the total cell count) and recovered 167 – 2,461 unique lineage identities (≥1 target site recovered for 15% – 75% of cells from 3 to 15 intBCs, **Fig. 2c, Extended Data Figure 4**). Many target sites are either lowly or heterogeneously represented, which we improved by changing the promoter from a truncated form of Ef1α to an intron-containing version (see embryo 7, **Extended Data Figure 4**)[25].

We estimated the indel likelihood distribution by combining data from all seven embryos. Many indels are shared with K562 cells, though their likelihoods differ, suggesting that cell type or developmental status may influence repair outcomes (**Fig. 2d**, **Extended Data Figure 1, 4f**)[19]. Our ability to independently measure and control the rate of cutting across the target site is preserved *in vivo*, with minimal interference between cut sites except when using combinations of the fastest guides that may lead to end-joining between simultaneous double strand breaks (**Fig. 2e**). The fastest cutters result in higher proportions of cells with identical indels, indicating earlier mutations in development, which correspondingly reduce indel diversity (**Fig. 2f, g**). Importantly, the lineage tracer retains additional recording capacity beyond the temporal interval studied here, as most embryos still have unmodified cut sites (**Fig. 2f**).

## Assigning cellular states by simultaneous scRNA-seq

Next, to ascertain cell function, we utilized 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 aged each embryo by their tissue proportions compared to each stage (**Fig. 3a-c**)[26]. We proceeded with six of our seven embryos, as they 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 of six embryos, serving as a valuable outgroup to establish our ability to track transitions to the earliest bifurcation.

We also developed breeder mice that would enable facile exploration of all stages of development by injecting target sites into Cas9 negative backgrounds. This approach substantially increases the number of stably integrated target sites (~20). Resulting mice can be crossed with Cas9 expressing strains to yield viable Cas9$^+$ F1 litters that maintain continuous, stochastic indel generation into adulthood, demonstrating that cutting does not noticeably interfere with normal animal development (**Extended Data Fig 6**).

## Single cell lineage reconstruction of mouse embryogenesis

We developed phylogenetic reconstruction strategies to specifically exploit the characteristics of our lineage tracer, namely categorical indels, irreversibility of mutations, and presence of missing values (**Extended Data Figure 7**, **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 Figures 4 and 7**). When cell type identity from scRNA-seq is overlaid onto the tree, we observe functional restriction during development, with fewer cell types represented as we move from root to leaves (**Fig. 4a, b, Extended Data Figure 8**).

scRNA-seq-based strategies for ordering cells, such as trajectory inference, typically assume that functional similarity reflects close lineage[17]. To investigate this question directly, we used a modified Hamming distance to measure pairwise lineage distance and compared them to RNA-seq correlation. Generally, cells separated by a smaller lineage distance have more similar transcriptional profiles, though this relationship is clearer for some embryos than others (**Fig. 4c**, **Extended Data Figure 9**). This result is consistent with the notion of continuous restriction of potency as cells differentiate into progressively differentiated types.

We also developed a shared progenitor score that estimates the degree of common ancestry between different tissues by evaluating the number and specificity of shared nodes in the tree (**Methods**). Despite the stochastic timing of indel formation, this approach can reproducibly recover emergent tissue relationships, such as possible shared origins between anterior somites and paraxial mesoderm or 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**).

**Transcriptional state and developmental origin do not always correspond**

While our reconstructed tissue relationships generally recapitulate canonical knowledge, extra-embryonic and embryonic endoderm display consistent and unexpectedly close ancestry despite their independent origins from the hypoblast and embryo-restricted epiblast (**Fig. 5a, Extended Data Figure 9**). Manual inspection of the trees revealed a subpopulation of cells that appear transcriptionally as embryonic endoderm but that lineage analysis places within extra-embryonic branches (**Fig 4c, blue**). Consistent with this finding, an earlier, targeted study using marker-directed lineage tracing identified latent extra-embryonic contribution to the developing hindgut during gastrulation, although it was not possible to broadly evaluate their transcriptomes[27].

Here, scRNA-seq profiles collected in tandem with the lineage readout allow us to assess the degree of convergence towards a functional endoderm signature and identify distinguishing genes. Endoderm-classified cells derived from extra-embryonic origin are most similar to the endoderm cell type, but do share slightly higher similarity with yolk sac that is not apparent within the t-sne projection of the full embryo (**Fig. 5b, Extended Data Figure 10**). Given these independent origins, we might expect a subtle, but persistent, transcriptional signature reflecting their developmental history. Strikingly, when we separate endoderm cells according to their lineage, we identify two X-linked genes, Trap1a and Rhox5, general markers for extra-embryonic tissue[28,29] that are consistently upregulated in the extra-embryonic origin endoderm across embryos (K–S test, Bonferroni corrected *P*-value <0.05, **Fig. 5d, e**). 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 (**Extended Data Figure 10**) [9,30]. 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 phenotype provides the unique opportunity to infer the cellular potency and specification biases of ancestral cells as reconstructed by our fate map[31,32]. Each node within the tree represents a unique lineage identity stemming from a single reconstructed progenitor cell, allowing us to estimate lower boundaries of their field size (**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 extra-embryonic ectodermal/placental cell types and tiered pluripotency into "early" and "late" according to the presence of extra-embryonic endoderm (**Fig. 6a**)[33]. The contributions of these founders to extant lineages are asymmetric, suggesting that even though a progenitor may be biased towards a specific fate, it retains the ability to generate other cell types. Lower bound estimates from our data suggest a range of 1–6 totipotent cells, 10–20 early, and 18–51 late pluripotent progenitors (**Fig. 6b**). The variable number of multipotent cells at these stages may reflect an encoded robustness that ensures successful assembly of the functioning organism, particularly given that a single pluripotent cell can generate all somatic lineages in an embryo[34]. Future studies using more replicates generated by breeding may enable statistical approaches to evaluate these organism-scale developmental considerations.

## Discussion

In this study, we present cell fate maps underlying mammalian gastrulation using a technology for high information and continuous recording. Several key ideas have emerged, including the transformative nature of CRISPR-Cas9-directed mutation with a single cell RNA-seq readout[14-16], how information about a cell's history 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 allows for 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 employ environmentally-responsive promoters that sense stress, neuronal action potentials, or cell-to-cell contacts[35], or combine these approaches for multifactorial recording. Similarly, Cas9-derived base editors[36], including those that create diverse mutations[37] could allow for content-recording in cells that are particularly sensitive to nuclease-directed DNA double strand breaks[23,24].

Our cell fate map identifies phenotypic convergence of independent cell lineages, showcasing the power of unbiased organism-wide lineage tracing to separate populations that appear similar in scRNA-seq alone. Specifically, we substantiate the extra-embryonic origin of a subset of cells that resemble embryonic endoderm. While the initial specification of these lineages are known to rely on redundant regulatory programs, they 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[38,39,40]. More generally, our approach can be used to investigate other convergent processes or to discriminate heterogeneous cell states that represent persistent signatures of a cell's past, which will be critical for the assembly of a comprehensive cell atlas[41]. The scope of transdifferentiation within mammalian ontogenesis remains largely unexplored, but can be practically inventoried using our system.

Ultimately, 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 modeled from the assembly of historical relationships. Our hope is that characterization of these attributes will lead to new insights that connect large-scale developmental phenomena to the molecular regulation of cell fate decision-making.

## References

1. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology* **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.e17

(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–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. Raj, B. *et al*. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36,** 442–450 (2018).

15. 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).

16. Spanjaard, B. *et al*. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36,** 469–473 (2018).

17. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541,** 331–338 (2017).

18. Adamson, B. *et al*. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167,** 1867–1882.e21 (2016).

19. van Overbeek, M. *et al*. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol. Cell* **63,** 633–646 (2016).

20. 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).

21. Lemos, B. R. *et al*. CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl. Acad. Sci. U.S.A.* **115,** E2040–E2047 (2018).

22. Gilbert, L. A. *et al*. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 647–661 (2014).

23. Ihry, R. J. *et al*. p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat. Med.* **337,** 816 (2018).

24. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **19,** 1 (2018).

25. Kim, S.-Y., Lee, J.-H., Shin, H.-S., Kang, H.-J. & Kim, Y.-S. The human elongation factor 1 alpha (EF-1 alpha) first intron highly enhances expression of foreign genes from the murine cytomegalovirus promoter. *J. Biotechnol.* **93,** 183–187 (2002).

26. 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).

27. 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).

28. Eakin, G. S. & Hadjantonakis, A.-K. Sex-specific gene expression in preimplantation mouse embryos. **7,** 205 (2006).

29. Li, C.-S. *et al*. Trap1a is an X-linked and cell-intrinsic regulator of thymocyte development. *Cell. Mol. Immunol.* **14,** 685–692 (2017).

30. Pijuan-Sala, B. *et al*. A single-cell molecular map of mouse gastrulation and early

organogenesis. *Nature* **566,** 490–495 (2019).

31. 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).

32. Jaenisch, R. Mammalian neural crest cells participate in normal embryonic development on microinjection into post-implantation mouse embryos. *Nature* **318,** 181–183 (1985).

33. Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4,** 487–492 (2009).

34. 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. *Developmental Biology* **275,** 192–201 (2004).

35. Baeumler, T. A., Ahmed, A. A. & Fulga, T. A. Engineering Synthetic Signaling Pathways with Programmable dCas9-Based Chimeric Receptors. *Cell Rep* **20,** 2639–2653 (2017).

36. 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).

37. Hess, G. T. *et al*. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat*. *Methods* **13,** 1036–1042 (2016).

38. 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).

39. 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).

40. 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).

41. Regev, A. *et al*. The Human Cell Atlas. *Elife* **6,** 503 (2017).

42. 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).

## Supplementary References

43. Schiml, S., Fauser, F. & Puchta, H. The CRISPR/Cas system can be used as nuclease for in plantagene targeting and as paired nickases for directed mutagenesis in Arabidopsis resulting in heritable progeny. *Plant J* **80,** 1139–1150 (2014).

44. Ren, X. *et al*. Performance of the Cas9 Nickase System in Drosophila melanogaster. *G3* **4,** 1955–1962 (2014).

45. Kimura, Y., Hisano, Y., Kawahara, A. & Higashijima, S.-I. Efficient generation of knock-in transgenic zebrafish carrying reporter/driver genes by CRISPR/Cas9-mediated genome engineering. *Sci Rep* **4,** 206–7 (2014).

46. Dong, Z., Dong, X., Jia, W., Cao, S. & Zhao, Q. Improving the efficiency for generation of genome-edited zebrafish by labeling primordial germ cells. *International Journal of Biochemistry and Cell Biology* **55,** 329–334 (2014).

47. Bao, Z. *et al*. Homology-Integrated CRISPR–Cas (HI-CRISPR) System for One-Step Multigene Disruption in Saccharomyces cerevisiae. *ACS Synth. Biol.* **4,** 585–594 (2014).

48. Zhou, H., Liu, B., Weeks, D. P., Spalding, M. H. & Yang, B. Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Research* **42,** 10903–10914 (2014).

49. Jiang, W. *et al*. Demonstration of CRISPR/Cas9/sgRNA-mediated targeted gene modification in Arabidopsis, tobacco, sorghum and rice. *Nucleic Acids Research* **41,** e188–e188 (2013).

50. Feng, Z. *et al*. Efficient genome editing in plants using a CRISPR/Cas system. *Nature Publishing Group* **23,** 1229–1232 (2013).

51. DiCarlo, J. E. *et al*. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Research* **41,** 4336–4343 (2013).

52. Jacobs, J. Z., Ciccaglione, K. M., Tournier, V. & Zaratiegui, M. Implementation of the CRISPR-Cas9 system in fission yeast. *Nature Communications* **5,** 1–5 (1AD).

53. Wang, H. *et al*. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153,** 910–918 (2013).

54. Platt, R. J. *et al*. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159,** 440–455 (2014).

55. Yoshida, N. & Perry, A. C. F. Piezo-actuated mouse intracytoplasmic sperm injection (ICSI). *Nat Protoc* **2,** 296–304 (2007).

56. Gusfield, D. Efficient algorithms for inferring evolutionary trees. *Networks* **21,** 19–28 (1991).

57. Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome Biol*. **17,** 86 (2016).

## Acknowledgments

We would like to thank members of the Weissman, Meissner, and Yosef labs, particularly J. Charlton and A. Arczewska for assistance with animal imaging, A. Kumar for explorations of extraembryonic endoderm, and L. Gilbert and M. Horlbeck for guidance on technology design, as well as Eric Chow and Derek 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.

## 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.Q. and D.Y. prepared several sequencing libraries and were engaged in discussion. M.G.J, A.K, and N.Y provided phylogenetic reconstruction strategies.

## Author Information

The authors declare no competing interests. Correspondence and request for materials should be addressed to Jonathan Weissman (jonathan.weissman@ucsf.edu) and Alexander Meissner (meissner@molgen.mpg.de).

## Figure Legends

**Figure 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, mU6, hU6, and bU6).

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 10 intBCs for 6 days using the following information: site 1 only with intBCs masked, sites 1-3 (All) with intBCs masked, and sites 1-3 (All) with intBCs considered. Information content scales with number of sites and presence of the intBC.

d. sgRNA mismatches alter mutation rate. Seven protospacers were integrated into the coding sequence of a GFP reporter to infer mutation rate by the fraction of positive cells over a 20 day time course. Single or dual mismatches were made in guides according to proximity to the PAM: region 1 (proximal), region 2, and region 3 (distal). Guides against Gal4-4 and the GFP coding sequence act as negative and positive controls. Bold sequences were incorporated into the target site.

**Figure 2: Lineage tracing in mouse from fertilization through gastrulation**

a. Lineage tracing in mouse experiments. The target site (within mCherry’s 3’UTR) and the three guide cassettes are encoded into a single piggyBac transposon vector (ITRs, inverted terminal repeats). 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.

b. Pearson correlation coefficient heatmap of indel proportions recovered from bulk tissue of an E9.5 embryo (see also **Extended Data Figure 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 Figure 4** and **Methods** for frequency calculation. The indel code along the x-axis is as follows: “Alignment Coordinate: Indel Size Indel type (**I**nsertion or **D**eletion).”

d. Proportion of indels that span one, two, or three sites, shown per site. Each dot denotes one of 40 independent intBCs and sums to one across site-spanning indels. Colors 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 P, 2, and 1 region mismatches.

f. Indel diversity is inversely related to cutting efficiency for site 1 as in **e**. Early mutations due to fast cutting are propagated to more cells, leading to smaller numbers of unique indels.

**Figure 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 was collected, see **Extended Data Figure 3**), including for Cas9:EGFP and the mCherry:target site.
b. t-sne plot of scRNA-seq from embryo in **a**. Only large or spatially distinct clusters are labeled. (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 Figure 5b** for additional embryos.
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. Size: fraction of marker-positive cells, color intensity: normalized expression (cluster mean). XEcto, extra-embryonic ectoderm/placenta; XEndo, extra-embryonic endoderm/yolk sac; PGC, primordial germ cell; Endo, embryonic endoderm; Ecto, embryonic ectoderm; Meso, embryonic mesoderm; XMeso, extra-embryonic mesoderm.

**Figure 4: Single cell lineage reconstruction of mouse embryogenesis**

a. Reconstructed lineage tree comprised of 1,732 nodes for embryo 2 with three lineages highlighted. Each branch represents an indel generation event.
b. Example paths from tree in **a** highlighted by color. Cells for each node in the path are overlaid onto the plot from **Figure 3b**, with tissue proportions as a pie chart. Tissue representation decreases with increased tree depth, indicating 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, showing that closely related cells are more likely to share function. Red dot highlights the median, edges the interquartile range, and whiskers 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 (ex. primitive blood or visceral endoderm, which have early and late states) 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 (NMPs) and more posterior neural ectodermal tissues, such as the future spinal cord[42].

**Figure 5: Disparities between transcriptional identity and lineage history within the extra-embryonic endoderm**

a. Shared progenitor score heatmap for embryo 2 reconstructs expected relationships. The number of nodes that include cells from different lineages is highlighted (Heterogeneous nodes). See **Extended Data Figure 9** for additional embryos.

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 XEndo).

c. Endoderm cell lineage tree from embryo 2 with expression heatmap for two extra-embryonic marker genes. Middle bar indicates lineage: dark blue, extra-embryonic; light blue, embryonic; grey, ambiguous.

d. Expression boxplots 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 the interquartile range, whiskers the Tukey Fence, and crosses outliers. N's, the number of recovered XEndo origin cells of either embryonic (E) or Extraembryonic (X) function per embryo.

**Figure 6: Lineage bias and estimated size of progenitor pools**

a. Relative tissue distribution of cells descended from reconstructed or profiled pluripotent progenitor cells for embryo 2. "Profiled" is a unique lineage identity of multiple cells directly observed in the data. Pluripotent cells form all germ layers, but show asymmetric propensities towards different cell fates, possibly reflecting positional biases. Nodes highlighted in grey with asterisk overlasy give rise to primordial germ cells (lineages 1, 4, and 5 include 9, 1, and 1 PGCs each). Color assignments as in **Figures 3.**

b. Estimated progenitor field sizes for three types of early developmental potency. Totipotent cells give rise to all cells of the developing embryo, including trophectodermal (TE) lineages. Pluripotent progenitors are partitioned into early and late by generation of extra-embryonic endoderm (XEndo) in addition to epiblast (Epi). Dots represent single embryos; solid grey line connects estimates from the same embryo.

## Extended Figure Legends

### Extended Data Figure 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 region. In this experiment, single guide RNA 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. The indel code along the x-axis is as follows: "Alignment Coordinate: Indel Size Indel type (**I**nsertion or **D**eletion)." Site 3 was not profiled in this experiment.

b. Histograms representing the likelihood that any specific base in the target site is deleted (blue) or has an insertion (red) which begins at that position, for sites 1 and 2, respectively. The position of the integration barcode (intBC) and protospacer sequences (sites) within the target site is represented as a schematic along the bottom, with the PAM for each site proximal to the intBC. Indels, specifically insertions, start at the double strand break point 3-bases upstream of the PAM sequence.

c. Simultaneous and continuous molecular recording of multiple clonal populations in K562 cells. We transduced K563 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, implying that only two clonal lineages expanded from the initial population of 10, and confirmed generation of target site mutations. (Left) Strategy for partitioning a multi-clonal population into their clonal populations. Target sites are amplified from a single cell barcoded cDNA library and the intBCs in each cell is identified as present or absent. (Middle) Heatmap of the percent overlap of intBCs between all cells. The cells segregate into two populations representing the descendants of two progenitor cells from the beginning of the experiment. (Right) 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 Figure 2: Capturing early differentiation by pooled sequencing of indels generated within an E9.5 embryo**

Scatterplots 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, with the three embryonic compartments sharing the highest similarity.

**Extended Data Figure 3: Experimental overview**

a. Schematic of platform used for generation of single cell RNA-seq libraries and corresponding target site amplicon libraries, adapted from Adamson et al., 2016 (Ref 18). The barcoded and amplified cDNA library is split into two fractions prior to shearing: one fraction is used to generate a global transcription profile and the other is used to specifically amplify the target site.
b. Summary table of lineage traced embryos detailing the type of guides used, the sampling proportion, and sequencing results. Embryo 4 was omitted from further analysis due to the absence of cells identified as primitive heart tube.

**Extended Data Figure 4: Target site capture in mouse embryos**

a. Percentage of cells with at least one target site captured.
b. Scatterplot showing the relationship between the mean number of unique molecular identifiers (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." Generally, as the mean number of UMIs increases, the percentage of cells also increases. Using a full length, intron-containing Ef1a promoter in mouse embryos leads to a higher number of UMIs, which generally results in better target site capture.
c. Percent of cells for which a given integration barcode (intBC) is detected across all seven embryos profiled in this study.
d. Target site capture and expression level across tissues for Embryo 5, which utilizes a truncated Ef1a 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 represented the distribution of UMIs for the target site in each tissue. Dashed line refers to a 10 UMI threshold. The target site may be expressed at different levels in a tissue-specific manner, leading to higher likelihoods of capture in certain tissues. Examples such as the target sequences carrying the intBCs AGGACAAA and ATTGCTTG may also be explained by mosaic integration after the first cell cycle, as these follow a developmental logic and are preferentially expressed in extraembryonic tissues. White dot indicates the median UMI count for cells from a given germ layer, edges the interquartile range, and whiskers the full range of the data.

e. Target site capture and expression level across tissues for embryo 7, which drives the target site expression from a full length Ef1a 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 represented the distribution of UMIs for the target site in each tissue as in **d**. Dashed line is a visual threshold for 10 UMIs. While tissue specific expression may explain some discrepancy in target site capture, high expression (as estimated from number of UMIs) may still correspond to low capture rates, as observed for the intBC TGGCGGGG. One possibility is that certain indels may destabilize the transcript and lead to either poor expression or capture.

f. Scatterplots showing the relationship between estimated relative indel frequency and the median number of cells that carry the indel. Since 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 still assumes that any given indel occurs only once in the history of each intBC. Since the number of integrations is small (<50), we might expect our estimates to be poor. Here we see that the number of cells marked with an indel increases with indel frequency, suggesting that our frequency estimates are under-estimated for particularly frequent indels. This is likely due to the fact that we cannot distinguish between identical indels in the same target site that may have resulted from multiple repair outcomes (convergent indels). The most frequent insertions are of a single base and tend to be highly biased towards a single nucleotide (eg. 92:1I is uniformly an “A” in 5 out of 7 embryos and never < 88%).

**Extended Data Figure 5: single cell RNA-seq tissue assignment and wild type comparison**

a. Boxplots 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 large variance in the proportions of certain tissues and 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, respectively, and the bold red overlay highlights embryo 2, which is used throughout the text. Note that many processes are continuous or ongoing between E8.0 to E8.5, such as somitogenesis and neural development. For example, from E8.0 to E8.5, the embryonic proportions of anterior neural ectoderm and fore/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, supporting the possibility that these embryos are somewhat less developed than E8.5 but more developed than E8.0. For boxplots, center line indicates the median, edges the interquartile range, whiskers the Tukey Fences, and crosses the outliers.

b. Plots (t-sne) of single cell RNA-seq with corresponding tissue annotations for the six lineage traced embryos used in this study. (Inset) Pie chart of the relative proportions for different germ layers. Mesoderm is further separated to include blood (red). While 36 different states are observed during this developmental interval, only broad classifications of certain groups (eg. “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. Boxplots of the Euclidean distance between single cell transcriptomes and the average transcriptional profile of their assigned cluster (cluster center) in comparison to their distance from the average of the next closest possible assignment. Comparison is to the same 712 informative marker genes used to assign cells to states and includes all cells used in this study. Middle bar highlights the median, edges the interquartile range, whiskers the Tukey Fences, and grey dots the outliers. N's refer to the cumulative number of cells assigned to each state across all 7 embryos for which single cell data was collected, including for embryo 4.

**Extended Data Figure 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 **Figure 2a** using C57Bl/6J sperm to generate $P_0$ breeder mice, which are capable of transmitting high copy genomic integrations of the technology. Then, $P_0$ animals are crossed with homozygous, constitutively expressing Cas9 transgenic animals to enable continuous cutting from fertilization onwards in $F_1$ progeny. Shown is Sibling 2 of a cross between a $P_0$ male and a Cas9:EGFP female.

b. Bar charts showing the degree of mutation (% cut, red) for a $P_0$ male (top row) and 4 $F_1$ offspring generated by breeding with a Cas9:EGFP female prior to 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 10 most frequent indels from 3 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 animal demonstrates an independent mutational path.

**Extended Data Figure 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. Refer to **Methods** for a more detailed overview of key characteristics of our technology. * 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. ** Reports a plate-based DNA-sequencing approach that can be applied to all methods to improve target site recovery. *** Range of cells where at least one intBC is confidently detected and scored. **** Presents a tree reconstruction method, but results predominantly 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 (see **Extended Data Figure 5** and **Methods**). Bold scores indicate the reconstruction algorithm selected for each embryo (see **Figure 4**, and **Extended Data Figures 8** and **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 Figure 8: Single cell lineage reconstruction of early mouse development for embryo 6**

a. Reconstructed lineage tree comprised of 2,690 nodes generated from our most information-dense embryo (embryo 6), which we used to compare shared progenitor scores with embryo 2 in **Figure 4d**. Each branch represents an independent indel generation event, and each node contains a pie chart of the germ layer proportions for the cells contained within it (colors are as in **Figure 3b**).
b. Example paths from root to leaf from the selected tree (highlighted by color). Cells for each node in the path are overlaid onto the t-sne representation in **Extended Data Figure 5**, with the tissue proportion at each node in the tree included as a pie chart. In the top most path (pink), the lineage bifurcates into two independently fated progenitors that either generate mesoderm (secondary heart field/splanchnic plate mesoderm 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. Alternatively, the bottom path (dark blue) begins in an equivalently pluripotent state but becomes restricted towards the extraembryonic visceral endoderm fate.

c. Violin plots representing the relationship between lineage and expression for individual pairs of cells as calculated for embryo 2 in **Figure 4c**. Expression Pearson correlation decreases with increasing lineage distance, showing that closely related cells are more likely to share function. Red dot highlights the median, edges the interquartile range, and whiskers the full range.

d. Comprehensive clustering of shared progenitor scores for Embryo 6, which has the greatest number of unique alleles and samples multiple extraembryonic tissues. 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 shared progenitor score is calculated as $2^{-(n-1)}$ where n is the number of clusters present within that node). In general, extraembryonic tissues that are specified before implantation, such as extraembryonic endoderm or ectoderm, co-cluster away from embryonic tissues and within their own groups, while the amnion and allantois of the extraembryonic 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 most anterior somites by E8.5. Note that the gut endoderm cluster has been further portioned according to embryonic or extra embryonic lineage (see **Figure 5**).

**Extended Data Figure 9: Summary of results from additional mouse embryos**

Representative highest likelihood tree analyses for additional embryos, including:

a. Reconstructed trees as shown in **Figure 4a**.

b. Shared progenitor score heatmaps as shown in **Figure 5a**, 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 number of tissues within each node (a shared score is calculated as $2^{-(n-1)}$ where n is the number of clusters present within that node). In general, 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, though these scores are sensitive to the number of target sites and the rate of cutting. By shared progenitor, PGCs are also frequently distant from other embryonic tissues, but this often reflects the rarity of these cells, which restricts them to only a few branches of the tree in comparison to more represented germ layers. The number of heterogeneous nodes from which scores are derived is included for each heatmap.

c. Violin plots representing the pairwise relationship between lineage distance and transcriptional profile as shown for embryo 2 in **Figure 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 containing either 3 or 6 cut sites, depending on the embryo. As lineage distance increases, transcriptional similarity decreases, consistent with functional restriction over development. Red dot highlights the median, edges the interquartile range, and whiskers the full range.

**Extended Data Figure 10: Expression characteristics of extra-embryonic and embryonic endoderm**

a. Violin plots representing 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, while across group comparisons are presented on the right. Notably, extraembryonic 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 the interquartile range, and whiskers the full range.

b. Plots (t-sne) 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 extraembryonic (dark blue) origin.

c. Expression boxplots for the extra-embryonic markers Trap1a and Rhox5 from an independent single cell RNA-seq survey of E8.25 embryos (Ibarra-Soria et al., 2018, Ref [9]). Both genes are heterogeneously present in cells identified as mid/hindgut but uniformly present in canonical extra-embryonic tissues, consistent with a subpopulation of cells of extra-embryonic origin residing within this otherwise embryonic cluster. Red lines highlights the median, edges the interquartile range, and whiskers the Tukey Fence. Outliers were removed for clarity.

## Methods

### Plasmid design and construction

Because the principles governing Cas9 efficiency and subsequent indel generation are not absolute, we screened fourteen protospacers for potential inclusion in our target site, including nine protospacers known to function with moderate efficiency and five additional protospacers hypothesized to function[43-52]. Each protospacer was checked against the human and mouse genomes using bowtie to limit off target effects. A gene block library of the fourteen protospacers (no additional bases between sequences) with an 8 base pair randomer was ordered from IDT representing target site version 0.0.

The target site (tS) v0.0 vector backbone was derived from a previously described Perturb-seq lentiviral vector (pBA439, Addgene, Cat#85967)[18] with the following changes: the cassette for mU6-sgRNA-EF1a-PURO-BFP was removed and replaced with EF1a-tSv0.0-sfGFP using Gibson assembly with the target site in the coding sequence of sfGFP for use in the fluorescent reporter assay (PCT10, sequence available upon request).

A gene block library of five protospacers (ade2-whiteL-bam3-bri1-whiteB; no additional bases between sequences) with an 8 base pair randomer was ordered from IDT representing target site version 0.1. Protospacers in positions 1 (ade2), 3 (bam3), and 5 (whiteB), are used for cutting in subsequent experiments and are referred to as sites 1, 2, and 3.

Target site (tS) v0.1 was also cloned into pBA439 with the following changes: the cassette for mU6-sgRNA-EF1a-PURO-BFP was removed and replaced with EF1a-sfGFP-tSv0.1, followed by BGH pA on the original backbone (PCT12). Here the target site sits in the 3' UTR of GFP. To improve the delivery of multiple targets into the same cell, we swapped the v0.1 target site cassette into a commercially available piggyBac transposon vector (Systems Biosciences, #PB533A-2) with the following changes: IRES-Neo was swapped for either GFP (PCT16) or mCherry (PCT29). The backbone was digested with restriction enzymes and target site v0.1 gene block was PCR-amplified to add Gibson arms. Following Gibson assembly, the plasmids were transformed into at least 100uL of Stbl2 competent cells (Thermo Fisher, Cat#10268019), and plated onto 1-2 large plates (Fisher, #NC9372402) with LB/Carbenicillin to generate high complexity target site libraries (PCT17, and PCT30, respectively).

The three-guide expression vector design and cloning protocol were adapted from [18] to utilize guides against the three sites in the target site. The guide for site 1 (ade2) is under the control of the mU6 promoter, site 2 (bam3) under the control of hU6 promoter, and site 3 (whiteB) under the control of bU6-2 promoter. All guides are constitutively expressed in this system. Additionally, the triple-guide cassette was moved onto the piggyBac backbone described above.

Two further modifications of the plasmids described above were used in this study. First, in an attempt to decrease the cutting percentage variation between embryos, we cloned the triple-guide expression cassette without BFP into PCT29, and then cloned in the target site with intBCs to generate the resulting vectors (PCT41-43, for guide combinations (P,1,P), (1,1,1), and (2,1,2), respectively). In the second modification, we changed the truncated form of Ef1a in PCT29 to a promoter sequence comprised of the ubiquitous chromatic opening element (UCOE) and a full-length, intron-containing Ef1α and cloned in a triple-guide expression cassette for the guide combination (2,1,P), followed by cloning in of the target site to make PCT60. In these modifications, target site plasmid libraries (PCT41-43, PCT60) were transformed and expanded in 1-2L of liquid LB/Carb culture rather than on large plates.

A new target site design, v1.1, was utilized for further experiments to generate $P_0$ breeders (see below). A gene block library of three protospacers (ade2-bri1-whiteB; 30-60 bases between sequences) with a 14 base pair randomer was ordered from IDT representing target site version v1.1. For this target site, site1 is ade2, site2 is bri1, and site3 is whiteB. We cloned v1.1 into the same backbone as PCT60 with guide combinations (2,3,3) or (2,1,2) to make PCT61 and PCT62, respectively.

## Cell culture, DNA transfections, and viral production

The production of lentiviral particles or transfection of plasmids as is as described in[18].

## K562 GFP reporter assay

To construct the target site GFP reporter cell line, a doxycycline(Dox)-inducible Cas9 K562 cell line was stably transduced with PCT10 (8% infected, <0.1 MOI), and GFP positive cells were sorted using fluorescence activated cell sorting on a BD FACSAria2. For each protospacer in the target site, 1-4 guides was designed to achieve a series of mutation efficiencies and cloned into single guide expression vectors[22]. On Day -4, the reporter cell line was plated into wells and stably transduced with a different guide against target site v0.0, GFP-targeting protospacer EGFP-NT2 (positive control), or Gal4-targeting protospacer (negative control) in each well. On Day -2, cells were selected for guide cassette integration using 3 ug/mL puromycin. On Day 0, 50ng/mL Dox was added to induce Cas9 expression, and maintained through the course of the experiment. GFP fluorescence was recorded on a LSR-II flow cytometer (BD Biosciences) on every 2nd day starting at day 0, except day 13 was recorded in place of day 12. Data was analysed in Python using FlowCytometryTools (http://eyurtsev.github.io/FlowCytometryTools/). For guide virus produced in this experiment, labels were systematically shifted during production resulting in incorrect ordering of guide effect on GFP fluorescence, which was corrected for presentation in the manuscript. We confirmed the activity order of the guide series for three guides (ade2, bam3, and bri1) in sequencing experiments where new virus was prepared.

**K562 single cutting pooled assay**

To construct the cell line used here, a Dox-inducible Cas9 K562 cell line was stably transduced with PCT12 (6% infected, <0.1 MOI), and GFP positive cells were sorted on a BD FACSAria2. On Day -5, the cell line was plated and stably transduced with a different guide against target site v0.1, or GFP-targeting protospacers in each well. On Day -2, cells were selected for guide cassette integration using 3 ug/mL puromycin. On Day 0, 50ng/mL Dox was added to induce Cas9 expression, and maintained through the course of the experiment. Wells were sampled every 3-6 days for 20 days with cell pellets frozen down. Genomic DNA was isolated from frozen cell pellets, and the target site was PCR-amplified to make sequencing libraries (refer to **Pooled embryo library preparation** below for library prep protocol), which were sequenced on the Illumina Miseq. Timepoint samples were pooled and reads with no indels were removed to calculate relative indel frequencies.

**K562 multiple target site integration cell line**

To construct a cell line with multiple integrations, we nucleofected 200,000 Dox-inducible Cas9 K562 cells with 1500ng PCT17 and 200ng piggyBAC transposase using set program T-016 (Lonza #V4SC-2096; Systems Biosciences, #PB210PA-1).

### K562 triple guide cutting assay, and multi-clonal lineage tracing experiment

Multiple-integration cells described above were stably transduced with a triple guide expression vector (Perfect-Perfect-Perfect; fastest cutting) and recovered for 2 days.  GFP (target site) and BFP (triple guide) double positive cells were sorted using fluorescence activated cell sorting on a BD FACSAria2.  For the multi-clonal lineage tracing experiment, 10 cells were sorted into wells containing 200uL of pre-conditioned media on a 96 well plate (12 wells total).  At day 18, wells were inspected under the microscope and the 3 wells with the largest populations were selected for single cell analysis on the 10x Chromium.  Two of the lanes suffered wetting failures, and the remaining sample was taken through library preparation described below (refer to **Target site amplicon library preparation**).  The library was sequenced on the Illumina Miseq and would benefit from additional sequencing.

For the pooled experiment, ~112,000 cells were sorted into a tube, spun down, resuspended in fresh media, split into two wells with 50ng/mL Dox added to one of the wells.  Cells were collected 6 days post-sort, genomic DNA was isolated, and the target site was PCR-amplified to make sequencing libraries (refer to **Pooled embryo library preparation**), which were sequenced on the Illumina Miseq.  The 10 intBCs with the most reads were used for analysis.

### Embryo and $P_0$ breeder generation

Protocols are adapted from those described in ref [53] To enable *in vivo* lineage tracing, B6D2F1 strain female mice (age 6 to 8 weeks, Jackson Labs) were superovulated by sequential intraperitoneal injection of Pregnant Mare Serum Gonadotropin (5IU per mouse, Prospec Protein Specialists) and Human Chorionic Gonadotropin (5IU, Millipore) 46 hours apart. Twelve hours after delivery of the second hormone, MII stage oocytes were isolated and injected with *in vitro* transcribed piggyBAC transposase mRNA (100 ng/ul) prepared in an injection buffer (5 mM Tris buffer, 0.1 mM EDTA, pH = 7.4). Decapitated sperm isolated from an 8 week old *Gt(ROSA)26Sortm1.1(CAG=cas9*,EGFP)Fezh/J* strain mouse (Jackson labs, ref [54]) was resuspended with the purified piggyBAC library in the same injection buffer at concentrations ranging from 0.5 to 1.4 ug/uL.

Transposase-injected oocytes were then fertilized by piezo-actuated intracytoplasmic sperm injection (ICSI) as previously described ref [55]. Injected embryos were cultured in 25 uL EmbryoMax® KSOM drops (Millipore) covered in mineral oil (Irvine Scientific) in batches of 25-50 embryos. After 84 or 96 hours, successfully cavitated blastocysts were screened for uniform fluorescence of the target sequence cassette and transferred into one uterine horn of 6-10 week old pseudopregnant CD-1 strain female mice (Charles River). Uterine transfer results in an ~24 hour lag, so the day of transfer was scored as E2.5 and embryos were dissected from euthanized animals 6 or 7 days later at ~E8.5 or E9.5, depending on the experiment. All techniques utilized standard micromanipulation equipment, including a Hamilton Thorne XY Infrared laser, Eppendorf Transferman NK2 and Patchman NP2 micromanipulators, and a Nikon Ti-U inverted microscope.

The generation of breeders was conducted identically by coinjecting target design v1.1 piggyBAC plasmids with sperm from C57BL6/J strain males (Jackson labs), transferring uniformly bright mCherry blastocysts into CD-1 strain mice, and allowing live pups to be brought to term. Genotyping was conducted using tail tip genomic DNA purified using the Quick DNA Miniprep Plus kit (Zymogen) isolated prior to weaning. Animals with large intBC counts (n=23 for the male used in **Extended Data Figure 6**) were then bred into either male or female *Gt(ROSA)26Sortm1.1(CAG=cas9*,EGFP)Fezh/J* strain animals to generate live pups with continuous cutting. Fluorescence of live animals was confirmed and documented using a dual fluorescent protein flashlight (Nightsea).

**Pooled embryo library preparation**

RNeasy Mini Kit (Qiagen, #74104) was used to isolate RNA from whole embryos or dissected tissue for embryonic tissue. Alternatively, genomic tail tip DNA was used for $P_0$ breeders or Cas9+ $F_1$ animals. Following purification and/or first strand synthesis of cDNA from 1 ug of RNA (Promega), the target site was amplified using a 2-stage PCR protocol. In the $1^{st}$ stage, <100ng of diluted DNA template was amplified using 0.6 uM forward and reverse primers and Kapa HiFi HotStart ReadyMix according to the following PCR protocol: (1) 98C for 3 min, (2) 98C for 30 s, 69C for 30 s, 72C for 15 s (16 cycles for cDNA, 24 cycles for genomic DNA), (3) 72C for 5 min. Following 0.7X SPRI selection, the elute served as template for $2^{nd}$ stage PCR, using 0.6uM barcoded P5 and P7 secondary primers and Kapa HiFi HotStart ReadyMix according to the following PCR protocol: (1) 98C for 3 min, (2) 98C for 30 s, 60C for 30 s, 72C for 30 s (4-6 cycles), (3) 72C for 5 min. PCR products underwent 0.6X SPRI-selection and were eluted in 20-40uL of elution buffer to produce the final library. Libraries were sequenced on the Illumina HiSeq 2500 (Rapid Run) or Miseq, with the following run parameters: Read 1: 175 cycles, i7 index: 8 cycles, i5 index: 8 cycles, Read 2: 175 cycles.

For v1.0 target sites, the following primary primers were used:

MC38:

CGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCAGGAGCGGATTGCTTCGAACC

MC39:

TCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAACCACTACCTGAGCACCCAGTC

For v1.1 target sites, the following primary primers were used:

P5_PCT48-49_F:

TCGTCGGCAGCGTC**AGATGTGTATAAGAGACAGAATCCAGCTAGCTGTGCAGC**

ODY120_PCT48_R_PB:

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCGATGGACGATTGCGGAAGACAG

Secondary amplification was conducted using the following primers:

P5 primer:

AATGATACGGCGACCACCGAGATCTACAC[ILLUMINA INDEX]*TCGTCGGCAGCGTC*AGATGTGTA)

P7 primer

CAAGCAGAAGACGGCATACGAGAT[ILLUMINA INDEX]*GTCTCGTGGGCTCGG*AGATGTGTATAAG

**Single cell embryo dissociation**

Embryos are washed through several drops of PBS after isolation to reduce debris and put into ~100 uL PBS droplets on a microscope slide and screened for uniform fluorescence of the target site cassette on an Olympus IX71 inverted microscope running Metamorph. Selected embryos were dissociated to single cell suspensions by adding 100 uL of TrypLE (Invitrogen, #12605010) and pipetting the embryo or embryo pieces every 5 minutes for ~30 minutes until complete dissociation was visually confirmed. Trypsin was deactivated by adding 100 uL PBS+BSA is added to the droplet and moving cells into a 1.5 mL eppendorf tube, followed by several rounds of additional collection with 100 to 200 uL of PBS+BSA to a final volume of 1 mL. The dissociated cells are filtered through a Flowmi filter tip (Bel-Art Products, #H13680-0040) into a new tube, and spun down for 5 minutes at 1200 rpm on a tabletop centrifuge. Following the spin, 900uL of PBS+BSA is removed and the remaining volume is resuspended with an additional 900uL of PBS+BSA. The suspension is spun for 5 minutes at 1,200 rpm, 800 uL of PBS+BSA is removed, the remaining volume is spun for 5 minutes at 1,200 rpm, and PBS+BSA is removed until only ~30 uL of volume remains. 2 uL of the final resuspended cells were used for counting using a hemocytometer. We load ~17,000 cells into the 10x machine (Chromium Single Cell 3' Library & Gel Bead Kit v2) for a targeted recovery of 10,000 cells.

**scRNA-seq library preparation and sequencing**

Single cell RNA-seq libraries were prepared according to the 10x user guide, except for the following modification. After cDNA amplification, the cDNA pool is split into two fractions. 15uL of EB buffer is added to one of the fractions of 20uL of the cDNA pool, and scRNA-seq library construction proceeds as directed in the 10x user guide. RNA-seq libraries were sequenced on the Illumina HiSeq 4000 system.

## Target site amplicon library preparation

The target site-specific amplification protocol was adapted from [11]. 50-100 ng of template from the cDNA pool, 0.3 uM P5-truseq-long (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT), 0.6 uM MC63 (TCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGCAGGAGCGGATTGCTTCGAACC) was split across four parallel PCR reactions, and was amplified using Kapa HiFi HotStart ReadyMix according to the following PCR protocol: (1) 95C for 3 min, (2) 98C for 15 s, then 69C for 15 s (8-12 cycles). Reactions were re-pooled during 0.9X SPRI selection, and eluted into 60 uL. A second PCR with the elute as the template, 0.3 uM P5 (AATGATACGGCGACCACCGA), 0.6 uM barcoded P7 (CAAGCAGAAGACGGCATACGAGAT[ILLUMINA INDEX]GTCTCGTGGGCTCGGAGATGTGTATAAG) was split across four parallel PCR reactions, and amplified using Kapa HiFi HotStart ReadyMix according to the following PCR protocol: (1) 95C for 3 min, (2) 98C for 15 s, then 69C for 15 s (6 cycles). Reactions were re-pooled during 0.9X SPRI selection and then fragments of length 200-600bp were selected using the BluePippin. Target site libraries were sequenced on the Illumina HiSeq 2500 (Rapid Run), with the following run parameters: Read 1: 26 cycles, i7 index: 8 cycles, i5 index: 0 cycles, Read 2: 350 cycles.

## scRNA-seq library data processing

scRNA-seq data was processed and aligned using 10x Cell Ranger v2. The filtered gene-barcode matrices were then processed in Seurat (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.

### scRNA-seq tissue assignment

An independent project conducting scRNA-seq profiling of gastrulation identified 42 distinct tissues in wild type mice. We utilized the mean expression profile for each tissue and the list of 712 marker genes used for assignment of cells to tissues (see instructions for assignment in GSE122187). For each cell in lineage traced mouse embryos, we calculated the Euclidean distance between the cell's expression profile and the mean expression profile for each tissue using the 712 marker gene set, and assigned the cell to the tissue identity with the minimum distance. Expression values were transformed to log space using log(normalized UMI count + 1) before calculating the Euclidean distance. Comparisons between the best matched tissue to the next best match are presented for all data collected here in **Extended Data Figure 5** to highlight the precision of this approach.

### Embryo gastrulation stage assignment

The wild type mouse gastrulation compendium consists of five time points, profiling every 0.5 days from E6.5 to E8.5 with at least 10 embryos collected for each time point. Tissue proportion is calculated as the number of cells assigned to the tissue divided by the total number of cells in the embryo. The median tissue proportion was calculated for each time point treating each tissue independently. For each lineage-traced embryo, the Euclidean distance between its tissue proportions and the median tissue proportion for each time point was calculated and the embryo was assigned to the time point with the minimum cumulative distance.  All lineage-traced embryos were assigned to either E8.0 or E8.5 stages.

**Target site data processing**

A custom software pipeline was built to align and call indels in the target site.  The logic is as follows: (1) Assign cell barcode and UMIs to each read, (2) find the consensus sequence for each UMI, (3) align the consensus sequence to the target site reference sequence, (4) identify most likely integration barcodes (intBC) and create custom reference sequences, (5) repeat alignment against all reference sequences and select highest scoring alignment for each UMI, (6) call intBC and indels in the target site, (7) correct the intBC and allele using UMIs which appear in the same cell, (8) remove doublets.  Details appear below:

(1) Assigning cell barcode and UMIs to each read.  Specific amplification libraries of the target site amplicon were processed using 10x Cell Ranger software to assign cell barcodes and UMIs to each read.  The target site is designed to be orthogonal to the human and mouse genome, and does not align in Cell Ranger processing.  Unaligned reads from the Cell Ranger output bam file are parsed into fastq format with the cell barcode and UMI identifiers appended to the read name.

(2) Finding the consensus sequence for each UMI. To potentially increase the speed of consensus sequence finding, we attempt to trim reads to the same length for each UMI.  The read is trimmed to remove sequence beyond the polyA tail using cutadapt software

(http://cutadapt.readthedocs.io/en/stable/) with the following parameters: [–a AAAAAAAAA –e 0.1 –o trimmedFile.fq –untrimmed-output=untrimmedFile.fq –m 20 –max-n=0.3 –trim-n]. Reads that do not contain polyA sequence appear in the untrimmed file and are subjected to a second round of read trimming using a sequence which appears in the 3' end of the target site assuming the sequence has not been deleted from DNA repair, with cutadapt run using the following parameters: [-a GCTTCGTACGCGAAACTAGCGT -e 0.1 -o trimmedFile2.fq --untrimmed-output=untrimmedFile2.fq -m 20 --max-n=0.3 --trim-n --no-indels]. The adapter sequence used in the last round of trimming is then concatenated back on to the trimmed sequence to improve target site alignment in the next step. If >=60% of trimmed sequences for a given UMI are the same sequence, then the sequence is taken as the consensus sequence. Otherwise, a multiple sequence alignment is performed using BioPython and the consensus sequence is extracted from the alignment. Ambiguous bases are reported if there is <50% agreement for any position in the alignment.

(3) Aligning to the target site reference. We use the emboss implementation of the smith-waterman algorithm to align sequences to the target site reference sequence with the following parameters, which were determined empirically: [emboss water –asequence targetSiteRef.fa –sformat1 fasta –bsequence consensusUMI.fa –sformat2 fasta –gapopen 15.0 –gapextend 0.05 –outfile sam –aformat sam]. In this first alignment, the ambiguous sequence NNNNNNNN is used to represent the intBC. A minor bug had to be corrected in the emboss implementation to successfully output sam format. For target site v1.1, the gapopen penalty was increased to 20 and the gapextend penalty to 1.

(4) Identifying the most likely intBC. A perl script is used to parse the intBC from the alignment. The intBCs with the highest number of UMIs are substituted into the target site reference sequence to make custom reference sequences.  This step was included because upon manual inspection, there were obvious misalignments due to the ambiguous intBC sequence, which were corrected upon substitution of a real sequence.

(5) Selecting the highest scoring alignment for each UMI. Repeat smith waterman alignment against all custom reference sequences and select alignment with the highest score for each UMI.

(6) Calling indels and intBCs. A perl script is used to parse the intBC and indels from the alignment using the CIGAR string.  The boundaries for each site is defined and indels overlapping site boundaries are called as an indel in that site.  Sequence of the indel is not considered.

(7) Correcting indels using multiple reads with the same UMI from the same cell.  UMIs are filtered for alignment score and only cells that are in the matched scRNA-seq data set are kept. An intBC is corrected to an intBC with a higher UMI count in the cell if the intBCs are within an edit distance of 2 and the alleles are the same. An allele is the combination of indels in sites 1, 2, and 3.  An allele is corrected to an allele with a higher UMI count in the cell if the intBC is the same and the allele is within a 1-indel difference.  Only UMIs with greater than or equal to 3 UMIs are kept.

(8) Eliminating doublets. Cells that report two alleles for the same intBC are removed if the dominant allele is <80% of the total UMI count for the intBC. This removes 4.1-18.3% of cells in our embryos.

### Tree reconstruction strategies

#### 1. Biased search through phylogenetic space

We simulate the evolutionary process leading from a collection of uncut target sites to the final data set. The set of mutations (including "no mutation") across all target sites in a cell is referred to as an allele. In the final tree, each branch represents a mutation, and each node represents the allele of a cell, which may be a reconstructed ancestral allele, i.e. it is not present in the data set.

Input: table of unique alleles

- each allele may represent multiple cells
- we cannot distinguish between identical indels in the same position that may result from independent mutation events (convergent indels) if they appear with an identical set of co-segregating indels

Algorithm:

```
- Create root node in tree representing an allele with 0 mutations (c_allele)
- remove alleles in the table that match c_allele

- While alleles remain in table:
    - choose indel from table that can be added to current allele
        - can only add indels in positions that have no mutation
    - create new node by adding indel into c_allele (c_allele2)
    - draw directed edge labeled with indel between nodes from c_allele to c_allele2
    - remove alleles in table that match c_allele2
        - includes alleles that match c_allele2 with missing values for positions that have no mutations
    - if indels in table can be added to c_allele2, then c_allele = c_allele2; else, c_allele does not change
    - when indels cannot be added to c_allele, traverse up edges to ancestral nodes until an allele to which an indel can be added is found
```

We presented two methods that are used to choose indels. The first method, “Random,” selects a position where an indel can be added, and then selects an indel from the data set for that position; both selections occur in an unbiased manner. The second method, “Frequency Normalized Weighted” (FNW), identifies all of the indels that can be added to the current allele and scores them according to the fraction of alleles they are found in divided by the expected independent frequency of the indel (see **Fig. 2c**). These scores are used as weights to bias selection of the indel. The reasoning behind FNW is that indels that are found in many cells (or alleles) are more likely to have occurred early, but this has to be balanced against their expected likelihood of occurring. FNW biases the search towards more likely trees. To further increase the search for good trees, we first remove all indels that are unique to a single allele since we can assume that these indels occur at the leaves of the tree. The indels are added as branches leading to leaves in the final tree before the final tree likelihood is calculated.

The log likelihood of the tree is calculated as the sum of the likelihoods of all the indels that appear in the tree. The likelihood of each mutation is estimated from the embryo data set (**Fig. 2c**).

It is worth noting that the number of trees that are possible grossly exceeds 30,000; however, the search is biased towards finding good trees and performs markedly better than those that are randomly generated. Using high scoring trees to direct the search towards better ones, such as by grafting high scoring branches, could further improve our algorithm's ability to identify high scoring trees.

**2. Greedy search to reconstruct larger trees**

Our greedy algorithm consists of building the tree top-down, recursively splitting cells into mutually exclusive groups based on the presence or absence of a specific mutation. In particular, these splits are prioritized by selecting mutations that appear frequently in the dataset, but are known to be an improbable outcome from a Cas9 mutagenesis event. This transform is equal to the product of the observed frequency of the mutation and the log prior-probability. The mutations prioritized this way, we reason, are very likely to have occurred only once and relatively early in the experiment. Under this assumption, these mutations are useful to a top-down approach as they efficiently create maximally informative tree-splits. In practice, we can calculate the prior-probabilities of mutations several ways but while describing this algorithm we assume the priors are provided (**Fig. 2c**).

To deal with missing values, we first split cells based on the presence or absence of a mutation. Then, for each cell that reports a missing value for this cut site, we assign the cell to the group with which it shares the greatest similarity. Here, we define similarity as the average number of mutations it shares with the cells in each group.  We follow this procedure until only one cell remains. Note that for application to the dataset described in this manuscript, we filled missing values with unique indels to force the algorithm to choose splits based on the presence of mutations rather than absence.

Theoretically, building the tree in this fashion is possible due to the special case of multistate compatibility afforded by our model of evolution, namely that mutations can only arise once at a particular site (i.e. Cas9 cannot re-cut a site). This context allows one to consider a traditional Gusfield algorithm[56] in which one infers phylogenies by selecting character-splits based on the most frequently occurring mutations. In a special regime of "perfect-phylogeny" (where every mutation arose exactly once), this algorithm is provably optimal and extremely efficient as compared to other algorithms (linear in the number of cells and mutations, or O(|number cells| * |number of mutations|)). In the case of multi-state characters, the notion of compatibility often breaks down as this typically implies that a character can mutate many times to different states. Yet, as described previously, in our system this cannot happen – namely, once a mutation is obtained at a site, it cannot be changed again along that evolutionary path. In this way, we can apply an approximated Gusfield algorithm to reconstruct trees, where perfect phylogeny is possible although still confounded in cases where the same mutation arises twice independently.

Trees are visualized using the Python ete library (http://etetoolkit.org/).

**Pairwise single cell lineage distance measure used for violin plots**

A cut site can take 2 forms, uncut or indel. The distance is a modification of hamming distance where uncut is considered a special state.

Distance = (2*(sites with different indels) + 1*(sites with indel vs uncut))/(number of sites recovered in both cells)

Pairwise expression correlation was estimated using the same 712 marker genes used to assign cell states and was only included if two single cell transcriptomes shared ≥10 gene measurements.

**Estimating ancestral tissue relationships**

Each node, including leaves, that includes more than one tissue type is considered a "progenitor." Progenitors found at the leaves are not reconstructed or inferred but result from the lack of new indels that distinguish between tissues (ie. the lineage tracer does not produce new indels past the progenitor stage).

The shared progenitor score is calculated between two tissues as the number of shared progenitors scaled by the number of tissues each progenitor contributes to, and is calculated using the following algorithm:

```
For each progenitor,
    tList = list of tissues progenitor contributes to
```

```
        pScore = 1/(2^len(tList)-1)
        for each pair of tissues in tList:
                progenitorScoreForPairOfTissues += pScore
Example for a single progenitor:
        tList = [Endo, Meso, XMeso]
        pScore = 1/(2^(3-1)) = ¼
        ProgenitorScoreEndoMeso += ¼
        ProgenitorScoreEndoXMeso += ¼
        ProgenitorScoreMesoXMeso += ¼
```

The resulting matrix is a shared progenitor score matrix. To transform the similarity matrix to a distance matrix, we use 1-(matrix/maxScoreInMatrix). The distance matrix is then hierarchically clustered using either ward or average as the cluster joining criteria..

To account for the potential effect of cluster sizes (for example, if we assume that differentiation occurs for all tissues instantaneously, then the larger cluster sizes for mesoderm and ectoderm would increase the likelihood of detecting a progenitor between the two tissues), we downsampled each tissue before calculating the shared progenitor score: 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. For higher resolution clusters (**Fig. 4d**, **Extended Data Fig. 8**), we downsampled 500 times instead of 1000 times.

Note that the number of nodes reported below the heatmaps in **Extended Data Figure 8** represents the number of progenitors that are found in the complete tree. The number of nodes used to calculate the shared progenitor score depends on the sampled set of cells chosen.

For high resolution shared progenitor scores calculated for embryos 2 and 6 (**Fig. 4d** and **Extended Data Fig. 8**), we bolstered some populations prior to calculating shared progenitor scores by merging some cluster identities if they represent the linear maturation of one tissue type to another, are primarily one cluster versus the other at the assigned developmental time point, and have very close transcriptional profiles. Specifically, we merged node with notochord, amnion mesoderm (early) with amnion mesoderm (late), primitive blood progenitor with primitive blood (early), and anterior paraxial with pharyngeal (arch) mesoderm. We also merged surface and preplacodal ectoderm due to the similarity of their transcriptional profiles and omitted "similar to neural crest 2" as this transcriptional cluster is ambiguously determined (the cluster is globally most similar to neural crest but not obviously comprised of specific marker genes).

**Endoderm lineage assignment and differentially regulated gene identification**

Endoderm cells can have one of three origins based upon our tree: extra-embryonic, embryonic, or ambiguous. Cells are considered extra-embryonic if there is a progenitor in its lineage whose descendants include ≥ 40% extra-embryonic cells. Cells have ambiguous origin if they descend directly from the root node. Otherwise, cells are considered to be from embryonic origin. We identified endoderm cells of extra-embryonic origin in all embryos but embryo 7.

We use the Kolmogorov-Smirnov test (Python scipy.stats.ks_2samp) to identify differentially regulated genes between embryonic and extra-embryonic origin endoderm cells. Only highly variable genes in the embryo are considered for testing, and genes are significant if they have a Bonferroni corrected p-value under 0.05.

### Multipotent field size estimation and asymmetry

Progenitors are considered pluripotent if their descendants include at least one mesoderm (Meso or XMeso or Blood) cell, one ectoderm (Ecto) cell, and one endoderm (Endo) cell. A pluripotent progenitor are considered early pluripotent if it also has at least one extra-embryonic endoderm descendant, and further considered totipotent if it has at least one extra-embryonic ectoderm descendant. To estimate the lower bound for the number of multipotent cells, we start at the bottom level of the tree and count the number of multipotent cells at that level. If multipotent cells exist, then the number of multipotent cells is propagated to its ancestor in the above level, otherwise we count 0 for that level. We add one progenitor for every level that includes a multipotent cell and other cells to represent the progenitor that lead those non-multipotent cells at that level. The number of multipotent cells is then the number of cells propagated to the root of the tree. Progenitor asymmetry is simply the proportion of descendants from each of the tissues for that node.

### Comparison to other technologies

Several CRISPR-Cas9 based lineage tracers have been developed, each with distinct strengths and weaknesses. In **Extended Data Figure 7**, we present a table summarizing the different technologies, and elaborate on the attributes that, in combination, distinguish our strategy here:

1. Target sites are marked with a unique integration barcode (intBC). The intBC allows us to phase our target sites and perform a direct comparison for each target site across cells. This greatly improves the information content of our system as it allows us to distinguish between the same indel if it appears in different target sites (**Fig. 1c**).
2. Guide RNAs are integrated into genomic DNA and constitutively expressed from totipotency, which enables our lineage tracer to be truly evolving over multiple cell generations. In technologies applied to zebrafish development, guideRNAs are expressed as a pulse, which leads to the generation of a large diversity of barcodes at one or two timepoints.
3. Multiple integrations of multi-cutsite target sites are distributed throughout the genome. Technologies that integrate a single target site with many cut sites or have tandem integrations are subject to collapse of information when one indel may affect neighboring cut sites or alternatively, simultaneous cutting of several cut sites remove large portions of their lineage tracer. While our technology is also vulnerable to these effects, we are better buffered against them by distributing the target sites throughout the genome. We also highlight that indel generation is largely independent within target sites when slower cutters are used (**Fig. 2d-f**).
4. Simultaneous, multi-population lineage tracing (**Extended Data Figure 1c**). Since target sites are labeled with integration barcodes, we can use the identity of these barcodes to deconvolute pools of cells upon sequencing. Alternatively, independent samples, such as

embryos that have unique sets of integration barcodes, can be pooled onto a single 10x lane to decrease the cost of reagents.

5. Multi-channel recording using our triple guide vector. In our current manuscript, we use the three channels for lineage tracing but different types of sensors can be developed to record multiple independent inputs.
6. Ability to trace over different time scales by tuning the mutation rate of the system through mismatches in the guide RNA.

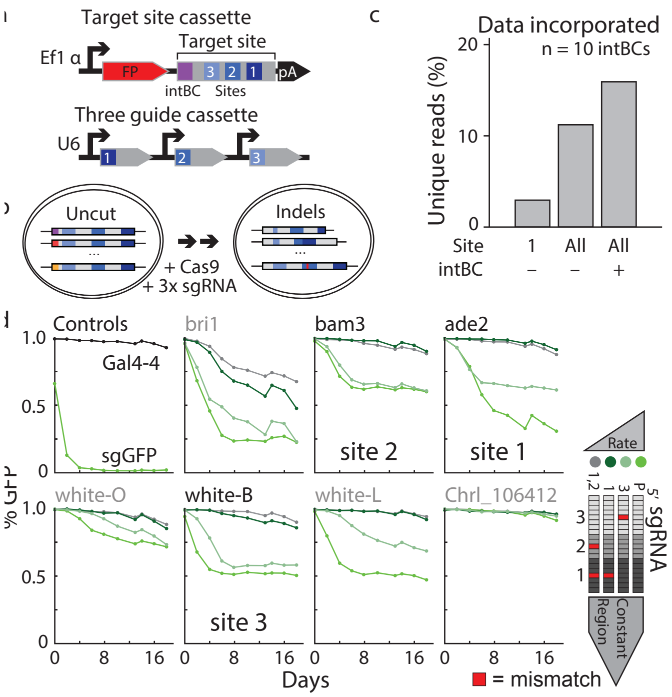
To fully utilize the information captured in our data set, we developed custom reconstruction strategies to identify the maximum likelihood tree (see **Tree reconstruction strategies** above). We estimate indel likelihoods using all of our embryo data (**Fig. 2c**), which allows us to estimate tree likelihoods rather than utilize maximum parsimony criteria. Phylogenetic algorithms developed for tumor evolution, such as SCITE[57], offer conceptual frameworks that are compelling to adapt for our technology.

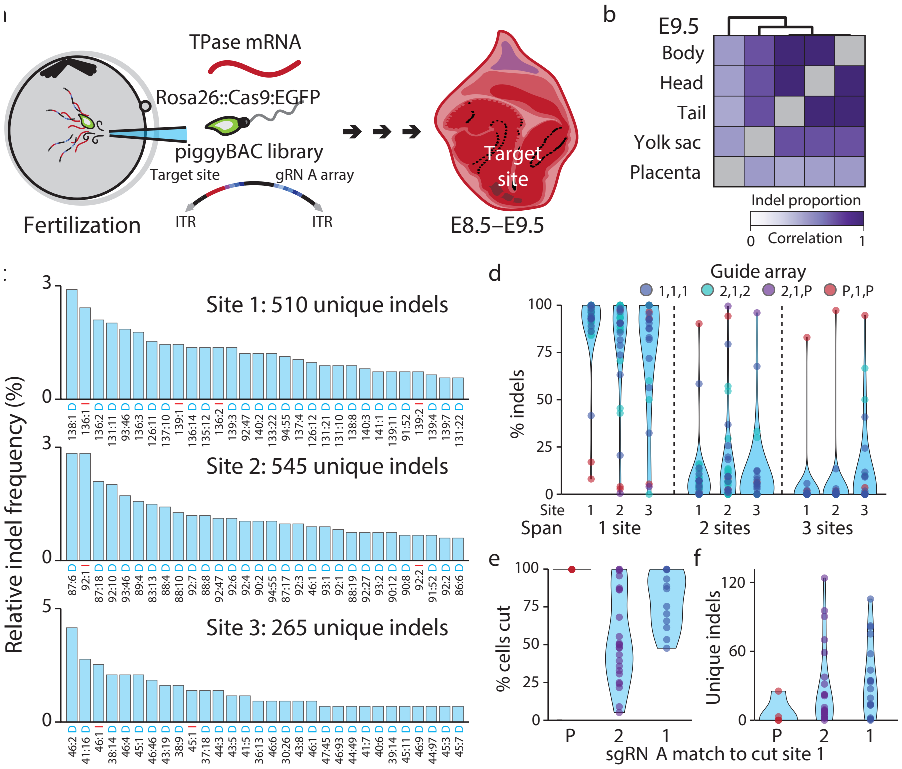
## Code availability

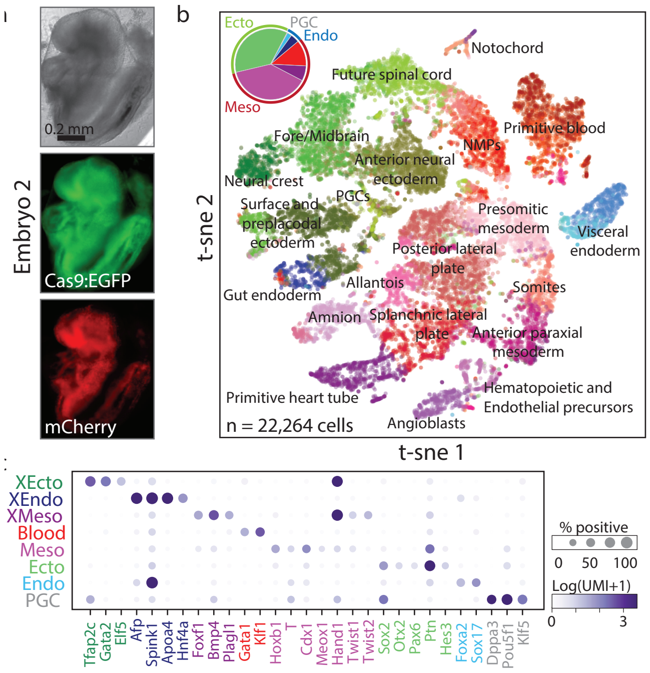
Code will be shared upon request.

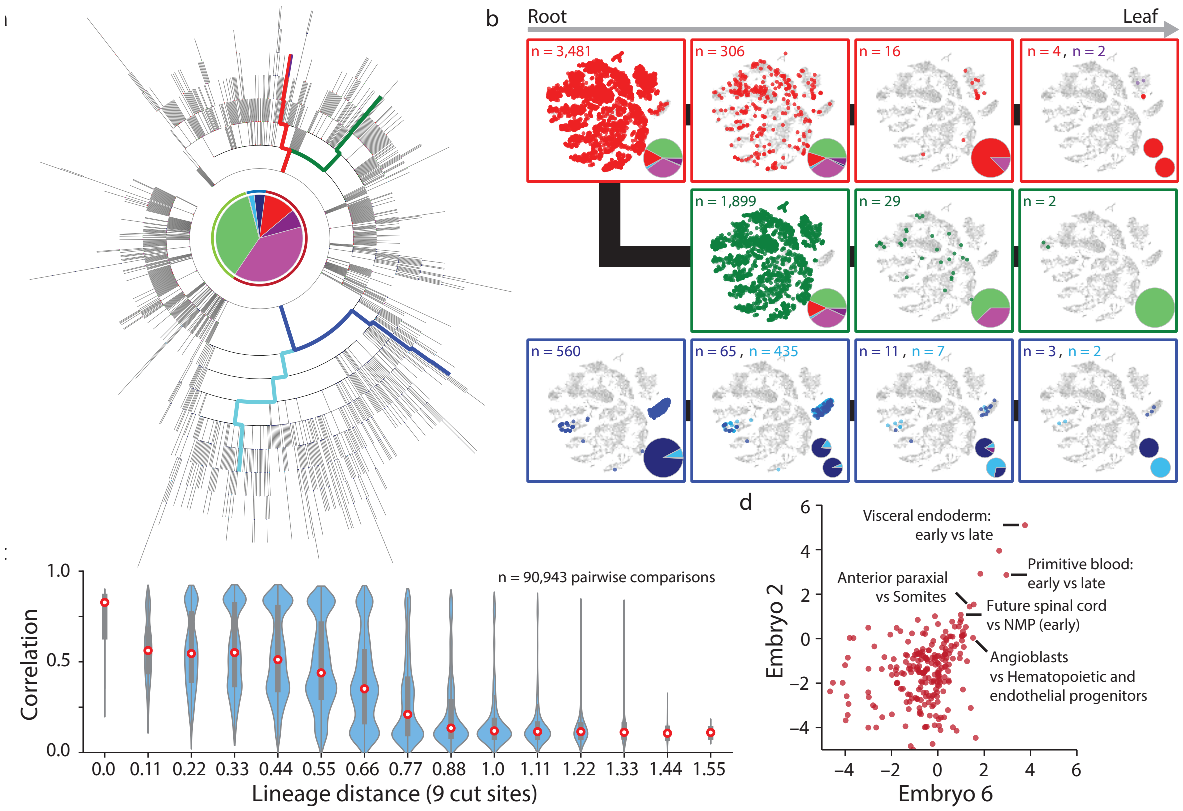
## Data Availability

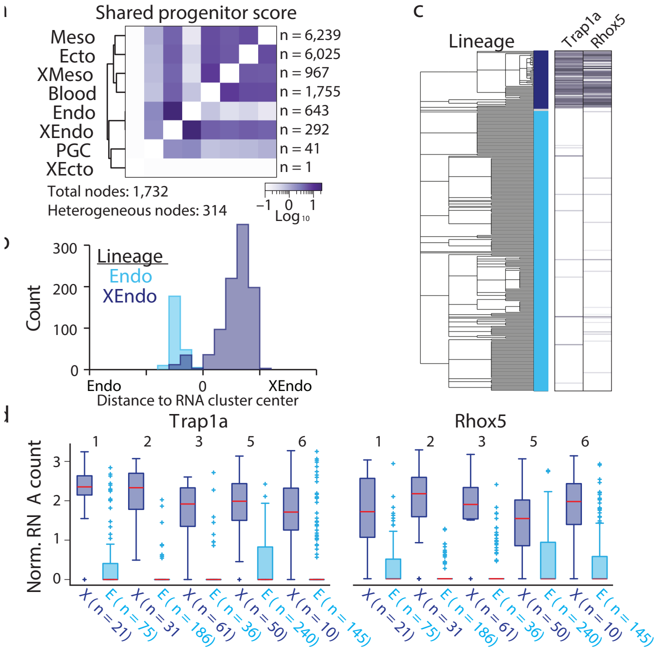
The data is available in the GEO database under accession numbers GSE117542 for lineage traced embryos and GSE122187 for the gastrulation compendium.

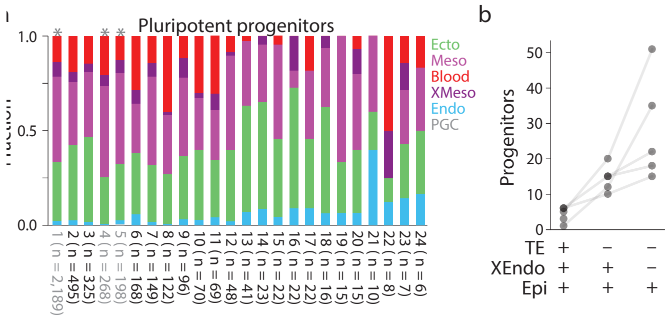

Target site cassette
Ef1 α
FP
Target site
3
2
1
pA
intBC
Sites
Three guide cassette
U6
1
2
3
Uncut
Indels
+ Cas9
+ 3x sgRNA
c
Data incorporated
n = 10 intBCs
Unique reads (%)
20
10
0
Site
intBC
1
–
All
–
All
+
Controls
Gal4-4
sgGFP
bri1
bam3
ade2
site 2
site 1
% GFP
1.0
0.5
0
white-O
white-B
white-L
Chrl_106412
site 3
0
8
16
Days
Rate
1,2
1
3
P
5' sgRNA
3
2
1
Constant Region
= mismatch


a
TPase mRNA
Rosa26::Cas9:EGFP
piggyBAC library
Target site
gRN A array
ITR
ITR
Fertilization
Target site
E8.5–E9.5
b
E9.5
Body
Head
Tail
Yolk sac
Placenta
Indel proportion
0
Correlation
1
c
Relative indel frequency (%)
3
0
Site 1: 510 unique indels
138:1 D
136:1 I
136:2 D
131:11 D
93:46 D
136:3 D
126:11 D
137:10 D
139:1 I
136:14 D
135:12 D
136:2 I
139:3 D
92:47 D
140:2 D
133:22 D
94:55 D
137:4 D
126:12 D
131:21 D
131:10 D
138:8 D
140:3 D
141:1 D
139:11 D
91:52 D
139:2 I
139:4 D
139:7 D
131:22 D
3
0
Site 2: 545 unique indels
87:6 D
92:1 I
87:18 D
92:10 D
93:46 D
89:4 D
83:13 D
88:4 D
88:10 D
92:7 D
88:8 D
92:47 D
92:6 D
92:4 D
90:2 D
94:55 D
87:17 D
92:3 D
46:1 D
93:1 D
92:1 D
88:19 D
92:27 D
93:2 D
90:12 D
90:8 D
92:2 I
91:52 D
92:2 D
86:6 D
5
0
Site 3: 265 unique indels
46:2 D
41:16 D
46:1 I
38:14 D
46:4 D
45:1 D
46:46 D
43:19 D
38:9 D
45:1 I
37:18 D
44:3 D
43:5 D
41:5 D
36:13 D
46:6 D
30:26 D
43:8 D
46:1 D
47:45 D
46:93 D
44:49 D
41:7 D
40:6 D
39:14 D
45:11 D
46:9 D
44:97 D
45:3 D
45:7 D
d
Guide array
● 1,1,1
● 2,1,2
● 2,1,P
● P,1,P
% indels
100
75
50
25
0
Site
1
2
3
1
2
3
1
2
3
Span
1 site
2 sites
3 sites
e
% cells cut
100
75
50
25
0
P
2
1
f
Unique indels
120
60
0
P
2
1
sgRN A match to cut site 1

b
Embryo 2
0.2 mm
Cas9:EGFP
mCherry
Ecto
PGC
Endo
Meso
Notochord
Future spinal cord
Primitive blood
Fore/Midbrain
NMPs
Anterior neural ectoderm
Neural crest
PGCs
Surface and preplacodal ectoderm
Presomitic mesoderm
Visceral endoderm
Posterior lateral plate
Gut endoderm
Allantois
Somites
Amnion
Splanchnic lateral plate
Anterior paraxial mesoderm
Hematopoietic and Endothelial precursors
Primitive heart tube
Angioblasts
n = 22,264 cells
t-sne 2
t-sne 1
XEcto
XEndo
XMeso
Blood
Meso
Ecto
Endo
PGC
Tfap2c
Gata2
Elf5
Afp
Spink1
Apoa4
Hnf4a
Foxf1
Bmp4
Plagl1
Gata1
Klf1
Hoxb1
T
Cdx1
Meox1
Hand1
Twist1
Twist2
Sox2
Otx2
Pax6
Ptn
Hes3
Foxa2
Sox17
Dppa3
Pou5f1
Klf5
% positive
0
50
100
Log(UMI+1)
0
3


b
Root
Leaf
n = 3,481
n = 306
n = 16
n = 4 , n = 2
n = 1,899
n = 29
n = 2
n = 560
n = 65 , n = 435
n = 11 , n = 7
n = 3 , n = 2
d
Visceral endoderm:
early vs late
Primitive blood:
early vs late
Anterior paraxial
vs Somites
Future spinal cord
vs NMP (early)
Angioblasts
vs Hematopoietic and
endothelial progenitors
Embryo 2
Embryo 6
6
4
2
0
−2
−4
−4
−2
0
2
4
6
n = 90,943 pairwise comparisons
Correlation
1.0
0.5
0.0
0.0
0.11
0.22
0.33
0.44
0.55
0.66
0.77
0.88
1.0
1.11
1.22
1.33
1.44
1.55
Lineage distance (9 cut sites)


Shared progenitor score
Meso
Ecto
XMeso
Blood
Endo
XEndo
PGC
XEcto
n = 6,239
n = 6,025
n = 967
n = 1,755
n = 643
n = 292
n = 41
n = 1
Total nodes: 1,732
Heterogeneous nodes: 314
−1 0 1
Log 10
c
Lineage
Trap1a
Rhox5
Lineage
Endo
XEndo
Count
300
200
100
0
Endo
0
XEndo
Distance to RNA cluster center
Trap1a
Rhox5
1
2
3
5
6
Norm. RN A count
3
2
1
0
X (n = 21)
E ( n = 75)
X (n = 31
E ( n = 186)
X (n = 61)
E ( n = 36)
X (n = 50)
E ( n = 240)
X (n = 10)
E ( n = 145)


b
Pluripotent progenitors
*
*
*
1.0
0.5
0.0
Ecto
Meso
Blood
XMeso
Endo
PGC
1 (n = 2,189)
2 (n = 495)
3 (n = 325)
4 (n = 268)
5 (n = 198)
6 (n = 168)
7 (n = 149)
8 (n = 122)
9 (n = 96)
10 (n = 70)
11 (n = 69)
12 (n = 48)
13 (n = 41)
14 (n = 23)
15 (n = 22)
16 (n = 22)
17 (n = 22)
18 (n = 16)
19 (n = 15)
20 (n = 15)
21 (n = 10)
22 (n = 8)
23 (n = 7)
24 (n = 6)
Progenitors
50
40
30
20
10
0
TE + – –
XEndo + + –
Epi + + +