LoxCode *in vivo* clonal barcoding resolves mammalian epiblast contribution to fetal organs

Tom S. Weber[1,2]*^, Christine Biben[2,3]*, Denise C. Miles[1,2] , Stephen Zhang[1], Patrick Tam[4,5], Samir Taoudi[2,3]#, Shalin H. Naik[1,2]#^

[1]Immunology Division, WEHI, Parkville, Melbourne. Australia. 3052
[2]Department of Medical Biology, University of Melbourne, Melbourne, Australia. 3052
[3]Epigenetics and Development Division, WEHI, Parkville, Melbourne. Australia. 3052
[4]Embryology Research Unit, Children's Medical Research Institute, University of Sydney. Westmead, NSW. Australia. 2145
[5]School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Sydney NSW Australia 2006.
*Contributed equally
#Contributed equally
^Correspondence: naik.s@wehi.edu.au, weber.ts@wehi.edu.au

Abstract:

How the mammalian epiblast gives rise to fetal organs during embryogenesis has been investigated using reporters of marker genes, or through single cell or spatial RNA sequencing to infer differentiation trajectories, but less is known about the clonal fate of mammalian epiblast cells *in vivo*. Here we develop a high diversity, high throughput, Cre recombinase-driven DNA LoxCode barcoding technology for *in vivo* clonal lineage tracing. Using this LoxCode mouse model cells in pre-gastrulation E5.5 embryos were barcoded *in utero* and assessed un bulk via PCR or via single-cell RNA sequencing for their contribution to a comprehensive collection of tissues and cell types in the E12.5 organogenesis-stage embryo. While few, typically large clones contributed to a diverse range of cell types in multiple germ layer derivatives, many clones exhibited reproducible patterns of lineage restriction. Most prominent were clonal fate biases towards either blood, ectoderm lineages, mesenchymal tissues or limbs, likely reflecting branch points during development. At the single-cell level, clones exhibited heterogeneity in terms of tissue contributions, gene expression profiles, and in some instances left-right and/or anterior-posterior asymmetries. Our study demonstrates the power and versatility of LoxCode barcoding in investigating native clonal fate and provides a deep clonal interrogation of the contribution of the mammalian epiblast to fetal organs.

Introduction:

Mammalian embryogenesis is an orchestrated process that generates and assembles the diverse cell types derived from the epiblast into a body plan that constitutes the blueprint of embryonic development. An important and yet outstanding question is whether all epiblast cells are pluripotent or whether there are cells that are already restricted to certain lineages prior to gastrulation. Answering this question in the mammalian embryo has been hampered by the inability to track cells of the epiblast at single-cell resolution over a substantial timeline. Seminal studies have shed light on the early stages of mouse embryogenesis, including the production of a presumptive fate map for the epiblast embryo[1], the discovery of how spatial position of cells in the epiblast influences the acquisition of developmental fate[2], single-cell RNA sequencing and fluorescent in situ hybridization of embryonic cells and tissues at different stages during development[3-8], and the lineage tracing of single cell clones of the epiblast across the window of gastrulation[9,10]. Despite the advances made possible through these studies of prospective cell fate, including a more recent *in vivo* pedigree analysis[4], technical challenges remain for implementing clonal lineage analysis through both prospective and/or retrospective tracing of lineage trajectories *in vivo* across gastrulation to organogenesis[5]. As an imperative, parallel assessment of the development of multiple clones is required to answer such questions.

Technological advances in cellular barcoding have empowered the tracing of clonal lineages[11,12]. Hereby, a progenitor population of cells is tagged with unique and heritable barcodes prior to further proliferation and differentiation. Subsequent comparison of the barcode repertoire of multiple clones contributing to cellular progeny allows determination of clonal relationships, e.g., derivation from common versus separate ancestors, estimation of clone sizes, timing of lineage allocation and diversification, and spatial location of the lineage precursor. Unlike viral barcoding methods that transduce cells *ex vivo* and which are then engrafted into a recipient, recent *in vivo* technologies enable barcode generation in cells in their native environment. This feature offers the ability to investigate clonal biology in a *native* setting. These barcoding techniques encompass prospective (multispectral barcoding[13], *in vivo* viral delivery[14], transposons[15,16], CRISPR/Cas9[4,17-21], Cre/Lox recombination[22,23]) and retrospective (natural barcodes[24], cell history recorder[25,26]) approaches. Many also allow the simultaneous assessment of the transcriptome and clonal origin of single cells for leveraging the ability to link cell state to cell fate, or to determine shared molecular features among clones[27].

We previously derived the design principles of a Cre/Lox system of barcode generation and determined mathematically that different DNA elements flanked by LoxP sites in alternating directions would create maximal barcode diversity[28]. Importantly, we predicted a combinatorial increase in the number of theoretical barcodes with each additional barcode element added to the cassette such that e.g. 14 LoxP sites and 13 code elements would yield a theoretical barcode diversity of over 30 billion when partially recombined[28] (*Methods*). Another innovative feature of our design included the use of DNA elements of very short size resulting in a compact construct that allows efficient recombination at low concentrations of Cre, generates over 2000 size-stable barcodes resistant to additional Cre excisions, and is fully readable by high-throughput short read sequencing.

Here we describe how LoxCode clonal lineage tracing offers several critical attributes that permit resolution of clonal pattern of cell/tissue lineages. Upon LoxCode generation in the epiblast at E5.5, and LoxCode extraction from tissue at E12.5 either through PCR or by single cell RNA-seq, and applying a suite of novel computational methods, we have established i) the patterns of contribution of epiblast cellular descendants to a full range of

embryonic tissues, ii) the inter-relationships of fetal tissues and organs, iii) an appraisal of which cell types are derived from individual epiblast cells, which has informed iv) cell type relationships at a clonal level.

RESULTS:

LoxCode design and experimental setup

Based on optimal design principles for Cre/LoxP cellular barcoding[28], the LoxCode was constructed through serial insertions of seven LoxP/Barcode paired inserts such that the final cassette was composed of 14 LoxP sites of alternating orientation and were separated by 13 short barcode elements of 8 bp and 14 bp (Fig. 1a, *Methods*). Upon Cre exposure, random recombination between the LoxP sites led to inversions and excisions, resulting in DNA sequence reshuffling with a theoretical DNA diversity of over 30 billion barcodes (Methods and Table 1). The LoxCode construct was introduced into the *Rosa26* safe harbor locus in C57BL/6 mice to enable *in vivo* lineage tracing though both PCR-based detection in genomic DNA as well as transcriptional read-out of LoxCodes by virtue of transcription driven by the endogenous *Rosa26* promoter.

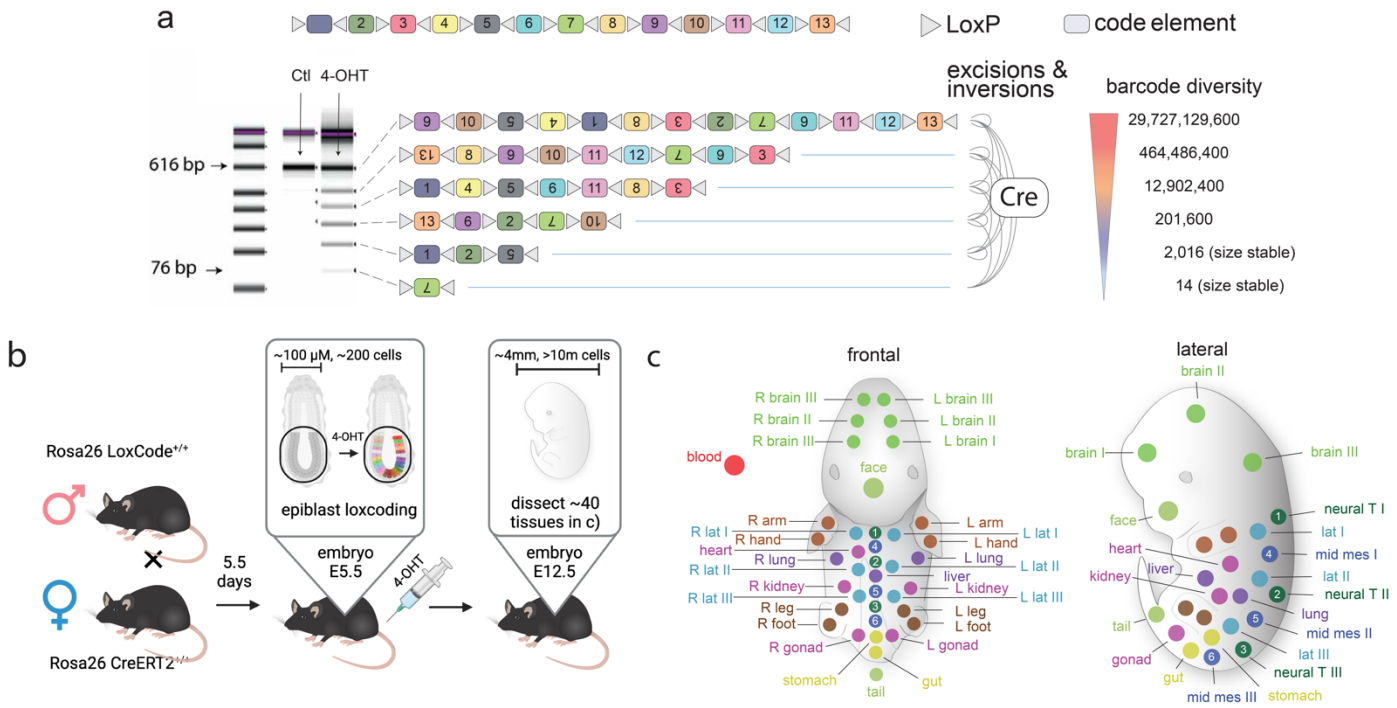


Figure 1. LoxCode design and experimental setup. a) The LoxCode cassette is composed of 14 LoxP sites in alternating orientation flanking 13 small (8-14 nucleotides) barcode elements. Cre exposure leads to random excisions and inversions, and different sized bands can be analysed via gel electrophoresis. Hypothetical LoxCodes of different sizes with inverted barcode elements depicted upside down. Computed theoretical barcode diversities for each barcode size are provided (right column). **b)** Experimental design: b) Barcoding of all cells (epiblast cells are highlighted in colour) in E5.5 pre-gastrulation embryos was activated by administration of 4-hydroxytamoxifen (4-OHT) to the pregnant mice, followed by dissection of E12.5 embryos into **c)** 40 body parts and organs for barcode retrieval and single-cell RNA sequencing analysis.

To study the clonal fate of E5.5 embryonic cells *in vivo*, LoxCode homozygous females were crossed to *Rosa26*-Cre-ER^{T2} homozygous males to produce heterozygous LoxCode/*Rosa26*-Cre-ER^{T2} embryos (Fig. 1b). LoxCode generation was induced at E5.5 via intravenous injection (IV) of low doses (50-100 µg) of 4-Hydroxytamoxifen (4-OHT) to both minimise developmental toxicity (no overt toxicity was observed) and to allow partial LoxCode recombination – the latter attribute enhances barcode diversity and complexity. To assess the clonal fate of E5.5 embryonic cells, E12.5 embryos were dissected into ~40 tissues and organs (Fig. 1c) which were either: i) lysed, prior to barcode PCR and sequencing (Fig. 2, 3); or ii) dissociated into single cell suspensions for concurrent scRNA-seq and barcode detection (Fig. 4).

While barcoding of the E5.5 embryo using *Rosa26*-Cre-ER^{T2} would not be limited to the epiblast cells, prior knowledge of lineage development during embryogenesis indicates that the extraembryonic ectoderm has no or little contribution to the germ layer tissues while the visceral endoderm may contribute to the endoderm of the gastrulating embryo[29,30]. Thus, the clonal descendants present in the organogenesis-stage embryo are therefore likely predominantly derived from the epiblast cells and we thus focused the analysis of cell clones in the embryo proper at the late-organogenesis stage.

Epiblast clonal fate heterogeneity

From three embryos, we recovered a total of 703 barcodes present in both PCR replicates of each tissue (~200-300 barcodes per embryo) of all expected sizes and calculated a minimum number of recombination events (excisions and/or inversions) for each barcode ranging from one up to eight (Supplementary Fig. 1 a-b). After removal of frequent barcodes that may arise from more likely or only few recombination events, so-called 'repeat use' or 'barcode collisions', using a new empirical method of filtering (see Methods) ~30% of barcodes remained, with an overlap of 10 barcodes between every two embryos, and only 1% (5 out of 265) present in all three embryos (Fig. 2a). Furthermore, clones contributing to at least two tissues represented one third of those barcodes, with very little overlap between embryos. This indicated efficient enrichment of single-use barcodes (Fig. 2a).

In a pluripotent model where every epiblast clone contributes broadly to every cell/tissue type, albeit with varying clone size, one would expect every barcode to be detected in all 40 different tissues or body parts. While we detected a few, typically large, multi-lineage clones in each embryo (~5% of all clones), this was not the rule: allocation of E5.5 epiblast cells ranged from only ~3.5% clones towards the gut up to a maximum of ~20% of all clones towards the brain (Fig. 2b). A minority of clones were restricted to a single tissue type which ranged from 1-3% of all clones for most tissues. Strikingly however, a significant fraction (12%) of LoxCodes were present exclusively in the circulating blood cells, which at E12.5 is predominantly extra-embryonic yolk sac-derived erythroid cells[31], indicating particularly early fate restriction (Fig. 2c). This is in line with previous studies[32,33].

Overall, the majority of LoxCodes of varying clone sizes spanning three orders of magnitude were unequally detected in a restricted number of tissues (Fig. 2d) deviating from a scenario where all epiblast cells display multi-lineage potency. This observation was further corroborated by simulations where cell numbers equivalent to the empirically derived clone sizes were drawn randomly from the distribution of cells across all ~40 tissues (Fig. 2d, p-vale<0.001).

Next, we explored whether, from the large number of possible clonal fates, a limited set of reproducible fate patterns were realised *in vivo*. For this, epiblast clones detected in at least two tissues were classified by Leiden

clustering into 5 main clusters of fate according to their contribution to 11 tissue types (Fig. 2b) and projected onto 2 dimensions by repurposing the Uniform Manifold Approximation and Projection (UMAP[34,35]) algorithm (Fig. 2e) and summarized in a bubble plot (Fig. 2f). We observed neural tissue-biased (cluster 1), multi-potent (cluster 2), limb-biased (cluster 3), and blood-biased (cluster 5) clones. Cluster 4 clones contributed to many tissues, but less to splanchnopleure and intermediate mesoderm derivatives (organs). Examples of Cluster 1-5 clonal contributions are shown (Fig. 2g).

Taken together, this analysis indicates that there were a small set of reproducible clonal fate patterns of E5.5 epiblast cells and that, within multi-lineage clones, strong biases were found among certain tissues, most notably between neural versus limb-biased multi-outcome clones (Fig. 2e-g).

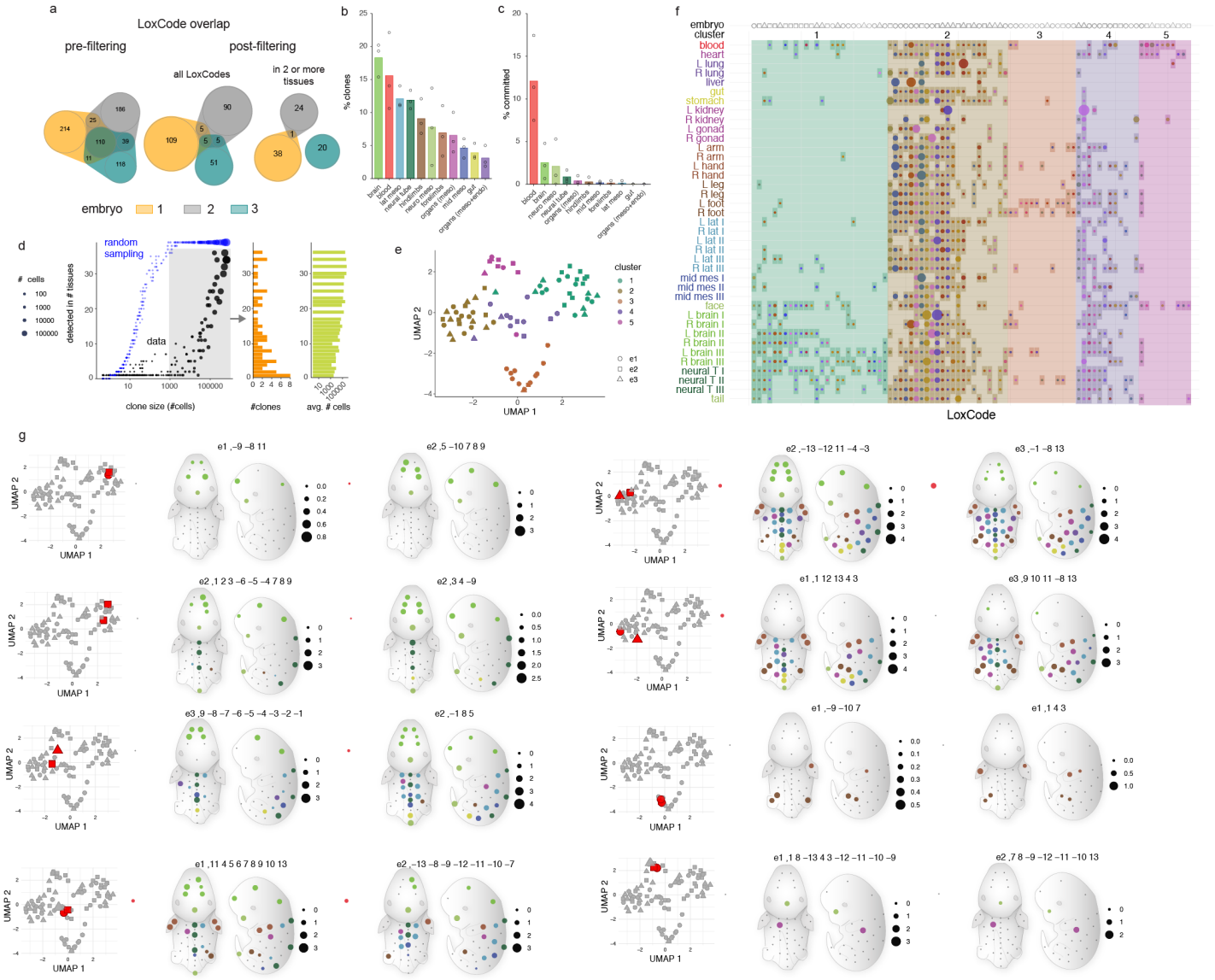


Figure 2. Patterns of epiblast clonal fate to organs and tissues. a) LoxCodes were excluded based on empirical frequency (see Methods) resulting in few overlapping barcodes between biological replicates. **b)** Proportion of clones contributing to indicated tissue

types. **c)** Proportion of clones uniquely detected in the indicated tissues. d**)** Comparison of the clone size and number of tissues each LoxCode was detected in, versus a simulation of random sampling of cells from all tissues of the corresponding clone sizes. Histograms on the right show the number of clones and average size for clones with more than 1000 cells. **e)** UMAP projection of clonal outcomes with 5 clusters of clonal output. **f)** Bubble plot/binary heatmap of clonal contribution to embryonic tissues. LoxCodes (columns) are ordered by 1D UMAP according to patterns of their clonal output to the indicated tissues (rows) from 3 embryos. Darkened rectangles indicate all detected clones (binary visualisation), and circles inside are proportional to clonal output to the indicated tissues (quantitative visualisation). Coloured backgrounds (six) correspond to the cluster annotations from e**). g)** Examples of clonal fate in a pseudo-anatomical map, where circles are scaled proportional to the logarithm of frequency of the indicated LoxCode (inverted elements prefixed by '-' and barcodes preceded by embryo of origin e1, e2, e3, respectively).

Fetal organ lineage branch point hierarchies informed by clonal relatedness of tissues

Clonal fates can vary qualitatively by the tissue types (lineages), and quantitatively, by the number of cells of the clone within each tissue (Fig. 2). While the genealogy of clones may be complex, we hypothesized that by comparing and summarizing these clone-associated metrics between pairs of tissues would allow inference of their ancestral relationships and reveal the order of the underpinning developmental processes. For example, tissues with highly concordant clonal composition likely reflect a shared developmental trajectory with 'late' branching into the tissues or cell types, whereas no or low concordance between two tissues may reflect 'early' lineage branching, and/or late-stage clonal expansion that may lead to disparity in clonal composition and frequencies.

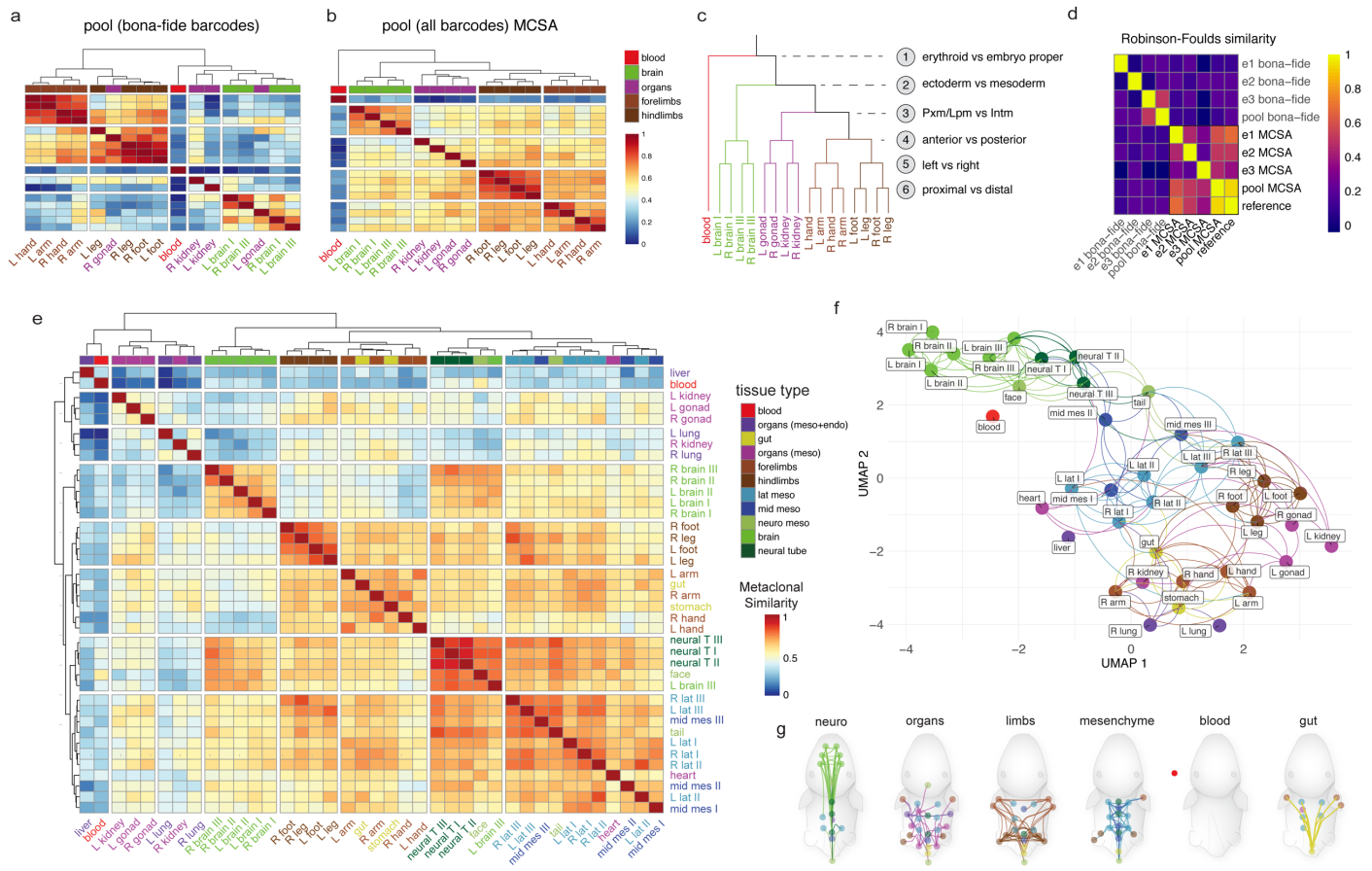
As a proof-of-principle, we first assessed similarities between clonal frequencies (using the cosine similarity as a measure) in a selection of samples that are well studied in terms of segregation during development, namely blood, brain (fore- and hindbrain, left and right), limbs (forelimbs and hindlimbs, left and right), and gonad and kidneys. Encouragingly, we found that samples of the same tissue type were more similar in their clonal composition compared to samples originating from a different tissue (e.g., brain tissues vs limbs), likely reflecting developmental branching between the tissue types (Fig. 3a). Unfortunately, clustering was relatively noisy both on individual embryos and for all three embryos pooled.

We noted however that the clustering of tissue types became more robust when all barcodes (including those otherwise excluded due to high frequency[36] that are termed 'repeat use' barcodes or 'barcode collisions', see Methods) were incorporated for computing the cosine similarity (Fig. 3b). Because frequent barcodes may correspond to several individual clones harbouring the same barcode by chance, i.e., what we refer to as *meta-clones*, we termed this approach **m**eta-**c**lonal **s**imilarity **a**nalysis (MCSA). To quantify the power of MCSA more systematically we compared branching hierarchies derived from filtered and unfiltered barcodes and from individual embryos and pooled against each other using the Robinson-Foulds similarity, a commonly used measure for the similarity between trees. In this comparison we also included a hierarchy based on results from Fig 3b and expected from literature constituted of 6 sequential bifurcation points: initiated with early branching of erythropoietic lineages, followed by ectoderm and mesoderm specification, paraxial/lateral plate mesoderm and intermediate mesoderm, then anterior posterior, then left right, and finally proximal distal branching (Fig. 3 c-d). This analysis highlighted that hierarchies derived from unfiltered barcodes were, despite the risk of barcode collisions, more consistent between individual embryos and the pool and were more similar to the model hierarchy in Fig. 3c.

The fact that stringent filtering, which is the norm in *in vivo* barcoding data analyses[19,22,23], lead to less robust results was counterintuitive initially due technical artifacts introduced due to barcode collisions. However, we speculated that a trade-off exists between the number of clones that are available to compute the similarity and the chance that some barcodes may represent more than a single clone. To test this idea, we simulated correlated gaussian random data points, assigned to each point a barcode using probabilities based on the empirical barcode list, and compared dendrograms generated using Robinson-Foulds distances between ground truth (i.e., in the absence of barcode collisions), filtered and unfiltered data (Supplementary Fig. 2a). We found that, especially under noisy conditions, hierarchies computed using unfiltered data indeed represented ground truth more accurately than filtered data (Supplementary Fig. 2b), with consideration that the values that would otherwise be direct measures of similarities were now akin to a robust 'rank order' of similarities.

Next, we applied MCSA to the full data set, and utilized 3 complementary approaches to visualize the results: through a hierarchically clustered heatmap (Fig. 3e), through UMAP to project tissues (rather than individual clones as in Fig 2e) in 2D and connect MCSA-based nearest neighbours (Fig. 3f), and a pseudo-anatomical representation of tissue interconnected using edge bundling of nearest neighbours and split by a reduced set of tissue classes (Methods) (Fig. 3g). For completeness we also provide the heatmap using *bona-fide* clones in Supplementary Fig. 3c with the same ordering as in Fig 3e.

All three visualizations revealed similar groupings, consisting of neural clusters, limb clusters, mesenchymal and organ cluster, and blood clearly isolated from the other tissues. However, there were also differences. While the heatmap separated splanchnopleure/extraembryonic mesoderm from neurectoderm/mesoderm derivatives, the UMAP suggested a more global neurectoderm/mesoderm separation, with splanchnopleural tissues closely associated with paraxial mesoderm derived tissues. Due to its robustness to noise and more flexible 2D layout rather than a linear ordering, the UMAP representation is more likely to visualize complex tissue relationships more accurately.



.

Figure 3. Major developmental branch points established through E12.5 metaclonal similarity analysis. Clustered heatmap of cosine similarity matrix computed for a selection of tissues pooled from three embryos using **a)** bona-fide and **b)** all clones (meta-clonal similarity analysis, MCSA, see text and Methods). **c)** Branching hierarchy containing 6 sequential branching levels. **d)** Robinson-Foulds similarity between hierarchies derived from bona-fide and all barcodes. **e)** MCSA of all tissues visualised as a heatmap, **f)** UMAP with lines connecting to 4 nearest neighbours with an MCS >0.75, or **g)** a pseudo-anatomical map where tissues of a given type (legend in a), were connected to their 4 nearest neighbours with an MCS >0.75. Edge colour indicates the tissue type of the outgoing node per legend in e). Data represent the aggregation of 3 biological replicates that were also assessed individually (Supplementary Fig. 3). Abbreviations: Pxm, paraxial mesoderm; Lpm, lateral plate mesoderm; Intm, intermediate mesoderm

The resulting data, amenable through *in vivo* LoxCode clonal lineage tracing, determined a comprehensive series of highly correlated and uncorrelated tissue groupings (Fig. 3e-f). These included:

(a) a brain tissues/neural tube/tail/face group that represent tissues of a neuroectoderm origin;

(b) limbs/neural tube/lateral mesoderm/medial mesoderm that represent tissues derived from paraxial mesoderm;

(c) gonad/kidney/gut containing derivatives of intermediate/lateral mesoderm.

These global groupings putatively demonstrate which tissue/organ groups have shared/similar clonal ancestry. Tissue groups that were less correlated such as liver/blood, brain/kidneys and gonad/lung were suggestive of distinct ancestry. Examples of the least correlated samples were:

(d) endoderm containing tissues (liver/lungs) that were distinct from mesodermal and neuroectodermal tissues, but are associated with gut/stomach;

(e) anterior neuroectodermal derivatives (forebrain/midbrain) were clustered separately from the neural tube. The separate clustering of these two groups may be due to the distinct origin of anterior neurectoderm derived from the epiblast and the neural tube derived from the neuromesodermal progenitor[10];

(f) neural tissues and limbs displayed distinct clonal composition;

(g) fore- and hindlimb displayed separate clonal origin and sidedness of clones in the limb tissues, possibly related to the separation of progenitors to contralateral lateral plate mesoderm.

Overall, this analysis by MCSA suggested that a putative hierarchical model of developmental branch points can be inferred from a single snapshot of barcode induction. In the future, incorporation of clonal data at additional time points would empower the discovery of so far uncharacterized branch points and the analysis for elucidating the order and precise timing of each bifurcation.

Determination of cell type bias and clonal relationships at a single cell level

The advantage of performing bulk barcode analyses of millions of cells (compared with single cell analysis of thousands of cells due to cost and throughput) is that the full repertoire of clones can be determined[37]. That is, rare clones are captured through bulk analysis that would otherwise be lost when only sampling a small fraction of all cells if using single-cell omics analysis. However, bulk barcode analyses do not reveal which of the different cell types in each tissue contains which barcode. As the LoxCode construct is also transcribed, this technology is compatible with single cell RNA-sequencing which enables concurrent identification of cell type and clonal origin. In this study, we dissected E12.5 embryos into organs/tissues derived from an E5.5 barcoded embryo as outlined in Fig 1b. From this, single cell suspensions were generated, multiplexed using MultiSeq[38] (Methods, Supplementary Fig. 5a), and processed on a droplet-based scRNAseq platform to capture ~30,000 cells. Following quality control (Methods), 33 cell types (Fig. 4a) across all tissues (Fig. 4b, Supplementary Fig. 5b) were detected.

Full length barcodes were then selectively amplified from the single cell full length cDNA libraries prior to fragmentation of the cDNA library and both were sequenced separately on a NextSeq (for cDNA) and MiSeq (for LoxCodes) and mapped *in silico* by virtue of their shared 10X cell barcode. Full length LoxCode capture efficiency was 27% (>8500 cells), but we noted recovery as low as 6% in neuronal tissues. After filtering out frequent LoxCodes as in Fig 2, we were left with ~200 cells, spanning 45 clones, where 5 of them contained >5 cells.

Cellular contribution of these 5 clones are highlighted in Fig. 4c-d indicating clear biased distributions across tissues (p-value < 0.0001) and cell types (p-value < 0.0001). At the tissue level, clone “1 8 -13” was enriched in lungs, gut and lateral mesenchymal tissues, clone “1 -6 9” was prominent in fore- and hindlimbs, with low detection frequency across all organs. Clone “1 -6 9”, “1 4 5 2 3”, and “1 10 -5 -4 -13” were detected in face, but

only for the latter two the presence in brain could be confirmed. At the transcriptional level, all 5 clones had representatives in the pulmonary/gastrointestinal mesenchyme and chondrocyte pole. Clone “1 4 5 2 3” contributed to a diverse set of neuronal cells, while clone “1 -6 9” differentiated mostly to musculo-skeletal mesenchyme, chondrocyte pole and fibroblast pole cells consistent with its limb-biased nature.

Overall, despite the low number of clones that could be assessed, heterogeneity in transcriptional fates was striking. To explore if meta-clones could provide additional insights, we repeated the analysis for barcodes for which we were less confident that they originated from a single cell. First, we confirmed that tissue type contribution deviated from the population baseline (Fig. 4e). Bias towards forelimbs (light-brown color) or hindlimbs (dark-brown color) was apparent for meta-clone “1 2 3 4 5” and “-7 -2 13” respectively. Relative to baseline, meta-clone “1 6 7 8 13” showed increased presence in mesoderm-derived organs, coinciding with a low detection in lungs, liver and gut. Meta-clone “1 4 13” was not detected in limbs but in blood (1 cell), although this was not statistically significant. Interestingly most meta-clones were detected in gut. Among meta-clones we identified 5 additional candidates with apparent transcriptional bias (Supplementary Fig. 5a-c). While we cannot directly comment on the fate of these clones (i.e., they could represent the combined fate of two or more epiblast cells), their heterogeneity adds support to the previous observations (Fig 3) that most E5.5 epiblast clones are biased in their fate.

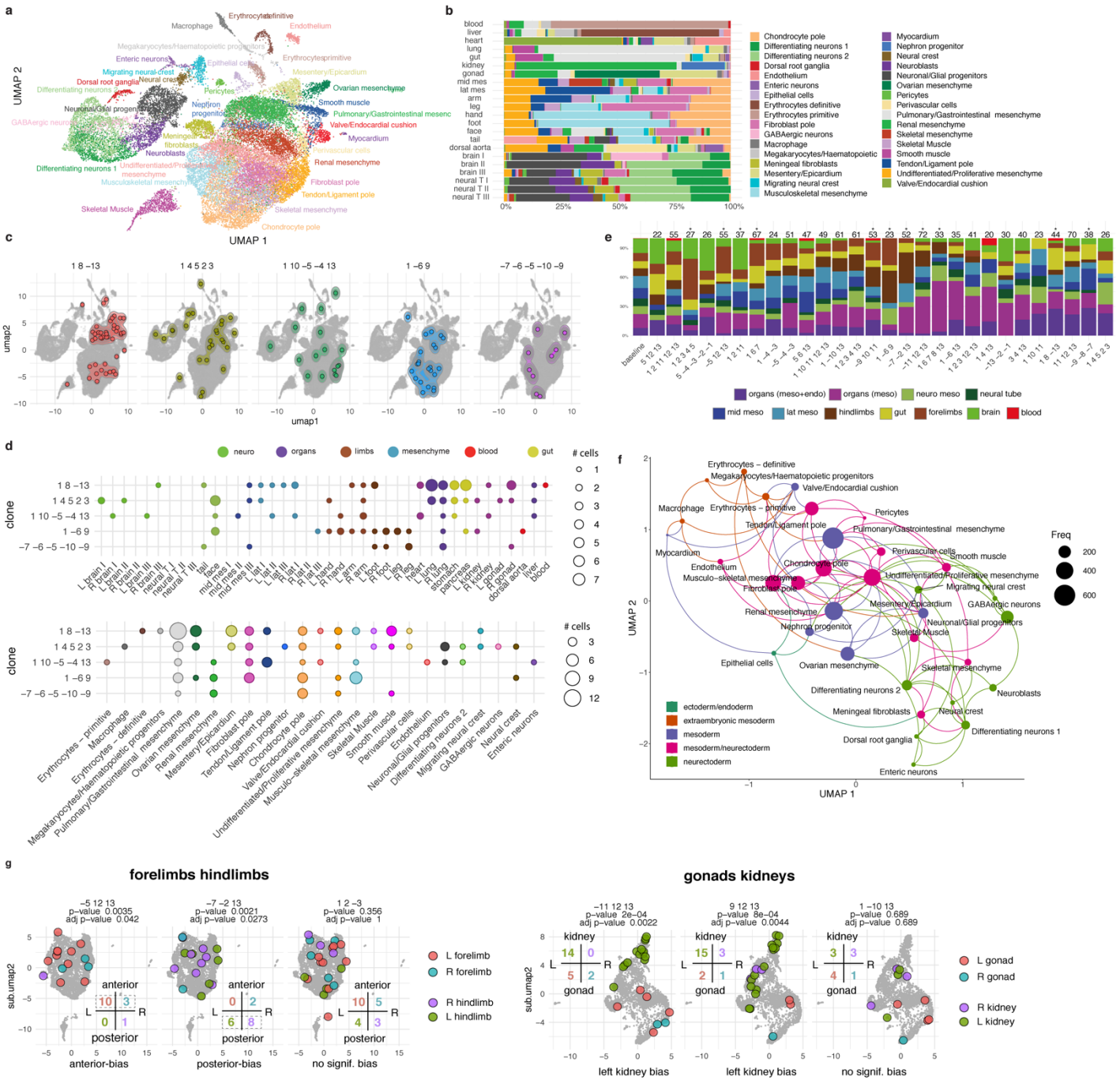


Figure 4. Single cell readout of epiblast fate bias and clonal relationships. a) UMAP of scRNA-seq derived from ~30,000 cells and annotated cell types, with **b)** their distribution across tissues (left and right and mesenchymal tissues pooled). **c)** Examples of single cell bona-fide clonal output corresponding to the indicated LoxCodes. Distribution of single cells per detected LoxCode in c) according to **d)** tissue (top) or cell type (bottom). **e)** Tissue type distribution of meta-clones with at least 20 cells and ordered using 1D UMAP. * indicates barcodes where distribution significantly deviates from a base-line distribution (adj. p-value < 0.05). **f)** scMCSA UMAP representation of cell types from b), with lines connecting each cell type's 3 nearest neighbours. Dot size is proportional to the cell

number in each cluster, and colouring is according to broad cell type categories as indicated in the legend. **g)** Left-right, anterior/posterior and tissue asymmetry of clonal contribution.

Global relationships between cell types were analysed using a single cell version of MCSA (scMCSA). In this approach meta-clonal similarities between transcriptional cluster (instead of tissues as in Fig. 3f) were computed and projected onto 2D using UMAP (Fig. 4f). Cell types broadly clustered into 1) haematopoietic and cardiac, 2) neuronal, and 3) mesenchymal types. Strong relatedness between haematopoietic and cardiac tissues was consistent with the notion that mesodermal cells that contribute to both tissues, are the first to be allocated during gastrulation, and might share common epiblast ancestors.

Comparing single-cell clonal data from left and right and anterior and posterior parts of the embryo allowed identification of asymmetries along the left-right and anterior-posterior body axis. Some clones displayed asymmetrical distribution to fore- and hindlimbs, and gonad and kidney (Fig. 4g), while most showed no significant spatial bias (Supplementary Fig. 6). Importantly, instances of asymmetrical contribution to limbs, gonads, and kidney were also detected in the bulk tissue analysis (Fig. 2f).

In sum, through LoxCode generation in the epiblast followed by detection and novel methods of analyses for their contribution to tissues and cell types of the late organogenesis stage embryo, we have uncovered the patterns of contribution of clones arising from the mouse epiblast cells and the inter-relationship of fetal tissues and organs in terms of their clonal features of development.

Discussion

We have developed the LoxCode mouse, an *in vivo* barcoding technology, to examine the breadth of clonal fates of cells in the pre-gastrulation E5.5 mouse embryo in a native setting at the tissue and transcriptome level. Our results highlight that pre-gastrulation epiblast cells are endowed with biased multi-lineage potency, leading to clonal contributions unevenly distributed across tissues and cell types in a multitude of germ layer derivatives. Most notable were clonal biases towards either blood, ectoderm lineages, mesenchymal tissues or limbs, likely reflecting developmental branch points during development. This was confirmed both via bulk analysis of tissues, which provides a large dynamic range of clonal composition, as well as single cell analyses, which allows resolution of the specific cell types that arise from each clone within that tissue.

When and how do these developmental branch points occur? Presently, our results do not allow discrimination between several non-mutually exclusive genealogies: epiblast molecular fate imprinting[39], persistence of a multi-lineage epiblast cell population that, by virtue of its physical location, becomes restricted in cell fate[40], lineage specific timing of allocation to specific cell/tissue type during development[41], and late-stage clonal expansion of the specified progenitor or the descendant tissue/cell types[42].

The concept of fate bias is well known to the field of development, and has been applied especially to the early phases[43,44]. In essence, for a cell at a given developmental stage, it is a summary of its pedigree at a later point in time, enumerating the leaves of the tree but ignoring intermediate progenitor cell fates and their relationships. In contrast to a binary view in which cells are classified according to their potential to differentiate either into one or another type, fate bias encodes more nuanced and richer information about the relative contributions. With the LoxCode mouse and the analysis tools developed here, we now provide a high-throughput approach to quantify fate bias at a pan-embryo scale that can be applied at any stage of development in the future.

The key features of the LoxCode mouse that empower this approach include barcode generation that is expeditious and efficient (within 6-12 hrs of Cre-ER^{T2} activity most barcodes recombine at least once) with a stable identity and high diversity (>30 billion LoxCodes, >2000 size-stable barcodes resistant to further Cre recombination), with barcodes detectable through PCR or single cell RNA-seq, and allowing for significantly higher throughput compared to existing Cre LoxP based barcoding[22,23] by using short read sequencing for barcode read-out. We anticipate the LoxCode mouse model to become a valuable experimental tool for studying clonal cell biology in mouse development, immunology and cancer.

As fate bias evolves over time, an important next challenge will be to connect fate bias across different stages of development. Pioneering technologies including evolvable hgRNA barcodes[18] already address this question, but are currently hampered by i) the non-randomness of NHEJ outcomes[45] which limits the diversity of founder cells, and ii) the lack of rapidly inducible Cas9 mouse models. Crossing the LoxCode mice with the MARC1 system or CARLIN[19] or other CRISPR based barcoding technologies could represent a powerful avenue to precisely map fate bias kinetics over time, by combining the best of Cre LoxP (i.e., efficiency, temporal control, high base-line diversity) with Cas9 based approaches (i.e., accumulation of mutations, exponential increase in diversity as a clone expands). In addition, constructs using alternative site-specific recombinases or integrase (e.g., Flp, Bxb1), or emerging technologies like transcriptional cell history recorders[25], could provide additional levels of resolution and the recording of molecular correlates of branching.

While we are far away from replicating the famous work by John Sulston in C. elegans [46] for mouse development, the current study provides a comprehensive view into E5.5 clonal fate heterogeneity in murine embryos and introduces the LoxCode technology as a powerful tool to characterize clonal heterogeneity at unprecedented depth and throughput.

Methods:

LoxCode cassette restriction cloning:

After several attempts to synthesise the LoxCode cassette via commercial service providers, which unfortunately failed presumably due to the palindromic nature of LoxP repeats, we sought to assemble the LoxCode cassette by sequential restriction cloning. For this we designed and ordered two partially reverse complementary 5' phosphorylated 90 bp oligo libraries (IDT, duplex oligos) containing two LoxP sites in opposite orientation (underligned), 8 mixed bases (Ns) between the LoxP sites, and flanked by 4 mixed bases. Furthermore each oligo contained 5 additional bases on its 5' end, such that when annealed, the duplex oligo would form sticky 5'-AATT-3' overhangs compatible with ligation into EcoRI digested plasmid. The base adjacent to the sticky end was asymmetrically chosen to be C and A (in bold), such that when ligated into EcoRI digested plasmid, the EcoRI (GAATTC) site would be recreated only on one end allowing for sequential integrations, i.e.

```
5'-AATTCNNNNATAACTTCGTATAATGTATGCTATACGAAGTTATNNNNNNNNATAACTTCGTATAGCATACATTATACGAAGTTATNNNNA-3'
    3'-GNNNNTATTGAAGCATATTACATACGATATGCTTCAATANNNNNNNNTATTGAAGCATATCGTATGTAATATGCTTCAATANNNNTTTAA-5'
```

Next the PUC18 plasmid was digested with EcoRI-HF restriction enzyme (NEB), and a standard ligation and cloning reaction was performed. Single colonies were screened for integrations and orientation via restriction digest. Plasmid from successful colonies with the recreated EcoRI site at the 3' end of the cassette were digested again with EcoRI, for a further round of cloning. This process was

repeated 7 times, to yield a sequence containing in total 14 LoxP sites, and 13 random barcode elements of either 8 (odd position, NNNNNNNN) or 14 bp (even position, NNNNGAATTTNNNN):

ATAACTTCGTATAATGTATGCTATACGAAGTTAT**ACTCCGCA**ATAACTTCGTATAGCATACATTATACGAAGTTAT**TCCAGAATTTGTAT**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**ACATCCAC**ATAACTTCGTATAGCATACATTATACGAAGTTAT**AAAGGAATTTCTCC**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**ATTTCCTC**ATAACTTCGTATAGCATACATTATACGAAGTTAT**GCCCGAATTTTTTC**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**GCTACTGG**ATAACTTCGTATAGCATACATTATACGAAGTTAT**ATGAGAATTTATGG**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**AACTAGAA**ATAACTTCGTATAGCATACATTATACGAAGTTAT**TGCAGAATTTCCTC**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**CGACACTT**ATAACTTCGTATAGCATACATTATACGAAGTTAT**AACGGAATTTTCAA**ATAACTTCGTATAATGTATGCTATACGAAGTTAT**CGTGTTTG**ATAACTTCGTATAGCATACATTATACGAAGTTAT

LoxCode mouse generation:

The LoxCode mouse line was generated following a previously established protocol[47]. In brief, double strand breaks were introduced at the *Rosa26* locus in C57BL/6J oocytes via microinjection of Cas9-gRNA ribonucleoproteins (gRNA: 5'-CTCCAGTCTTTCTAGAAGAT -3'), co-injected with a circular HDR donor vector (pBlueScriptIISK backbone) containing a splice acceptor, the LoxCode sequence, and a polyA signal flanked by 1000-1500bp Rosa26 specific homology arms. Oocytes were transplanted into pseudopregnant females, and 2 out of 26 viable pups displayed the expected integration via PCR and gel electrophoresis. Sanger sequencing was performed to confirm the integrity of the LoxCode sequence. The LoxCode line was then bred to homozygosity on a C57BL/6 background.

LoxCode theoretical and practical diversity:

The LoxCode construct is composed of 13 barcode elements, 7 of which are located at odd positions (i.e. 1, 3, 5, 7, etc), and 6 at even positions (i.e. 2, 4, …, 12). Recombination leads to inversion and excision that may change the locations and orientation of individual barcode elements. By design however, barcode elements at odd and even positions will always land up at odd and even positions respectively. We have previously shown through mathematical reasoning that all barcodes can be made by a series of inversions and excisions[28]. Thus it is straight-forward to compute the theoretical diversity of a cassette with n odd barcode elements and $m = n - 1$ even barcode elements by enumerating all permutations and considering forward and reverse orientations:

$$\frac{7!\,6!}{(7-n)!\,(6-m)!}\,2^{n+m}$$

Summing this formula for barcodes of size 1, 3, 5, 7, 9, and 13 yields a total of 30204722030, or over 30 billion theoretical barcodes. Note that due to DNA size constraint of Cre recombination, barcodes with 11 elements are not expected to be a stable product, although we detected them at low frequency.

The barcode diversity that is generated in an experiment depends on the number of recombination events (see Table 1), which in turn depend on Cre activity. Not enough Cre yields few barcodes of low diversity, too much Cre activity leads to an accumulation of barcodes with fewer elements that are less diverse. All barcodes can in theory be reached with 15 recombination events, however barcode with fewer elements typically reach their full diversity earlier. For instance, making all possible 201600 five-element barcodes requires 9 recombinations, while 5 recombinations are sufficient to create all 3-element barcodes (see Table 1).

The barcode diversity that is required for accurate lineage tracing depends on the number of clones that are tracked simultaneously. As a rule of thumb we recommend having 100 times excess barcodes relative to founding cells, which for the E5.5 computes to approximately 20,000 barcodes, or at least 3 recombination events per cell. We typically achieve a median of about 3 recombinations with a maximum of 8 or 9. Across all our experiments to date (including experiments outside of this work), we have recovered over 125,000 distinct barcodes, with a maximum number of recombinations of 11 (e.g. -3 12 -9 -10 7 4 -11 8 5).

In addition to theoretical barcode diversity, another important factor to consider are uneven probabilities. Some barcodes are much more likely to be made than others, reducing the effective diversity, as well as the proportion of cells that can confidently be tracked in a given population. We will discuss this aspect further in the barcode filtering section.

# recombinations	1	3	5	7	9	13
0	0	0	0	0	0	1
1	2	4	6	8	10	37
2	7	52	106	162	253	912
3	14	360	1212	2378	4418	19812
4	14	1506	11333	31928	70538	317578
5	14	2006	56242	252490	728021	3988251
6	14	2016	154973	1403382	5643564	36058157
7	14	2016	199695	4789612	29519488	238462747
8	14	2016	201587	9951050	106581710	1149595506
9	14	2016	201600	12504942	256471281	4056050952
10	14	2016	201600	12889288	409271561	10401368516
11	14	2016	201600	12902328	461386076	19356268864
12	14	2016	201600	12902400	464458544	26701414952
13	14	2016	201600	12902400	464486400	29495220993
14	14	2016	201600	12902400	464486400	29724820834
15	14	2016	201600	12902400	464486400	29727129600

Table 1. Theoretical barcode diversity for barcodes with 1 to 13 elements (columns) per number of recombination events (rows).

4-OH-Tamoxifen preparation and administration:

4-Hydroxytamoxifen Z-isomer (4-OHT, Sigma) was dissolved in 100% EtOH, diluted 1:1 with Kolliphor and stored in aliquots at –20 °C. On the day of the injection the dissolved 4-OHT was further diluted in PBS (50 µg/100 µl) and intravenously injected according to a procedure developed in Chevalier *et al.* (REF). Pregnant females received 50-100 µg of 4-OHT via tail vein injection 5 days after a plug was detected. Injected mice were separated from untreated females and monitored regularly.

Embryo dissections:

Rosa26CreERT2/Rosa26CreERT2 and *LoxCode/LoxCode* or *LoxCode/+* mice were set up and plug checked daily. Noon of the day of plug was referenced as 0.5 embryonic day (E0.5). Embryos were dissected in their yolk sac at E12.5 in PBS with 7% Foetal Calf Serum. The yolk sac was used as a genotyping sample when necessary (*LoxCode/+* mother). Embryos were staged using morphological criteria and assessed for normal development. The head was set aside and spilling blood was collected (blood sample). Heart, right and left lung lobes, liver, stomach (with spleen and pancreatic buds attached), intestine, left and right kidneys, left and right gonads, the region containing the dorsal aorta at the level of the aorta-gonad-mesonephros (AGM) area were dissected next. The rest of the body was cut transversely below the forelimb buds. From the anterior segment (forelimb level or level 1), left and right forelimb buds (termed L hand, R hand) were removed, followed by the left and right arms (L arm, R arm). The neural tube was removed (neural T I segment). Mesenchymal tissues were dissected as medial (left and right medial mesenchyme: L mid mes I and R mid mes I, encompassing somites) and lateral (left and right lateral mesenchyme: L lat I and R lat I, encompassing everything lateral/medial beyond the somites). The rest of the body was then cut just above the hindlimb buds level (between limb buds level or level 2), and corresponding neurectodermal or mesenchymal segments (neural T II, L mid mes II, R mid mes II, L lat II, R lat II) were dissected as above. The tail was removed and the third axial level (hindbud level or level 3) was dissected as per - level 1 (left and right hindlimbs (termed L foot, R foot), left and right hips (termed L leg, R leg), neural T III, L mid mes III, R mid mes III, L lat III, R lat III). The facial mesenchyme (face) was removed from the head, leaving the neurectoderm behind. Forebrain (brain I), midbrain (brain II) and hindbrain (brain III) segments were dissected according to morphological landmarks and halved along the midline (L brain I, R brain I, L brain II, R brain II, L brain III, R brain III). This resulted in a total of 43 embryonic samples.

Tissue lysis, PCR LoxCode amplification and sequencing library preparation:

Tissue samples were lysed in DirectPCR lysis reagent (cell) (Viagen Biotech), and the LoxCode sequence amplified using a 3-round PCR. The first round PCR (20 cycles) uses LoxCode specific primers that efficiently amplify the LoxCode sequence from genomic DNA. As an internal quality control, for each sample, two PCR reactions were performed.

FWD primer: TCTAGAGGATCCCCGGGTACCGA
REV Primers: TGATCCGCGCCTGGATGAAT

In the second round (5 cycles), overhang primers add a partial Illumina Read 1 and Read 2 sequencing primer site, as well as a stagger between the sequencing primers and the LoxCode sequence in order to increase diversity.

FWD Primers

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG TCTAGAGGATCCCCGGGTACCGA
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG N TCTAGAGGATCCCCGGGTACCGA
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG NN TCTAGAGGATCCCCGGGTACCGA
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG NNN TCTAGAGGATCCCCGGGTACCGA

REV Primers

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG GCCTGGATGAATTCGTGTATAACTTCG
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG N GCCTGGATGAATTCGTGTATAACTTCG
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NN GCCTGGATGAATTCGTGTATAACTTCG
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG NNN GCCTGGATGAATTCGTGTATAACTTCG

The third round (5 cycles) completes the sequencing primers and adds the P5 and P7 sequencing adaptors as well as introduces a unique sample index.

FWD primer:

AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC

REV Primers

CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG

To remove unused primers, primer dimers, and to counteract preferential amplification of shorter LoxCodes, amplicons were purified using SPRI beads at a 0.9:1 ratio between each round. All PCR reactions are carried out in a 20 µl volume, using Kapa Hifi Hot-Start Mastermix (Merck), using the following program 95 °C 2', N times (98 °C 20'', 65 °C 20'', 72 °C 30''), 72 °C 5', 4 °C hold, where N varied depending on the PCR round.
Sequencing libraries were pooled at equal volumes and sequenced on the Illumina MiSeq 600 cycle kit, using for Read 1 and Read 2 x any cycles respectively.

Dissociation of tissues:

For 10X analysis, dissected tissues were dissociated for 30 mins at 37C in 200ul of Liberase™ (Roche, 100ug/ml in PBS). 2ml of cold PBS (without ca/Mg) + 7% FCS + 2.5mM EDTA (FE buffer) was added to the tubes and carefully removed to rinse the enzyme off (the blood sample was centrifuged at this step to avoid cell loss). 1ml of fresh FE buffer was added. After 15-20 mins on ice, each tissue was mechanically dissociated and filtered on a 40µm sieve. Cells were pelleted (5 mins, 1500 rpm, 4C), counted with an automated cell counter (do you remember the brand, Tom?). 100K cells from each tissue was processed for MultiSeq labelling following the protocol described in McGinnis *et* al.. For each barcode oligo (5'-CCTTGGCACCCGAGAATTCCA**NNNNNNNN**A_{30}-3'), the oligo and anchor LMO (5'-TGGAATTCTCGGGTGCCAAGGgtaacgatccagctgtcact-Lipid-3') were diluted to 2 µM in 1:1 molar ratio in PBS. The co-

anchor LMO (5'-Lipid-AGTGACAGCTGGATCGTTAC-3') was diluted to 2 µM in PBS. Cells were resuspended in 200 µL PBS, 23 µl of anchor-barcode mixture added, and incubated on ice for 5 minutes. Then 23 µl of co-anchor solution was added and kept on ice. After an additional 5 minutes, labelling was quenched by addition of 1 mL of 1% BSA in PBS (ice cold). In all subsequent steps, cells were kept in 1% BSA on ice. After MultiSeq labelling, cells were resuspended in FACS buffer (PBS (without Ca/Mg) + 7% FCS) and labelled with 7AAD. Live cells were sorted on an Aria II cell sorter. 10K cells of each tissue were used to create 4 pools for 4 10X lanes: Pool 1 (blood, heart, right and left lung, liver, stomach, intestine, right and left kidney, right and left gonad, face), Pool 2 (dorsal aorta, L hand, R hand, L arm, R arm, L foot, R foot, L leg, R leg, L lat I, L lat II, face), Pool 3 (L lat III, R lat I, R lat II, R lat III, L mid mes I, L mid mes II, L mid mes III, R mid mes I, R mid mes II, R mid mes III, L brain I, face), Pool 4 (R brain I, L brain II, R brain II, L brain III, R brain III, neural T I, neural T II, neural T III, tail, face). The "Face" sample was seeded into each pool to allow for batch effect correction. Each pool was centrifuged (5 mins, 1500 rpm, 4C), resuspended in 32ul of PBS + 1% Bovine Serum Albumin and counted before loading on Chromium controller (10X Genomics) for scRNAseq. Sequencing libraries were prepared according to manufacturer's recommendations and sequenced on the Illumina platform (NextSeq 75 cycle kit).

Target amplification (scRNAseq):

The LoxCode sequence was amplified from full length cDNA using a combination of PCR, biotin capture, restriction digest and adaptor ligation. In a first round, transcripts contained the LoxCode sequence were enriched from full-length cDNA using a partial Illumina Read 1sequencing primer (present on every fragment in the full length cDNA library), and a reverse primer specific to exon 1 of the *Rosa26* locus. In a second round, overhanging forward primers completes the Read 1 sequencing primer binding site on the 5' end, and biotinylated LoxCode specific primers (located inside the LoxCode construct) further enriches LoxCode cDNA. In a next round, biotinylated amplicon is purified using MyONE C1 streptavidin beads (Thermofisher), removing unwanted DNA. Then LoxCode specific DNA was cleaved off the beads using an Eco53KI restriction site 5' of the LoxCode sequence. Finally stubby adaptors (IDT) were ligated using Blunt/TA ligation mix (NEB), and amplified using generic P5 and P7 primers. Libraries were sequenced on an Illumina MiSeq 600 cycle kit using custom sequencing primers.

Analysis: Alignment, Filtering, Correlation, UMAP, NN-clone,

Except for LoxCode alignment, all analyses and graphs were generated using R and RStudio. A complete list of scripts to recreate the plots included in the manuscript will be provided as Supplementary Materials in a future version of this preprint at the time of submission.

LoxCode Alignment:

Raw Illumina MiSeq paired-end sequencing reads were aligned to the known LoxCode elements using custom C++ code. Depending on the sequencing configuration, for full length LoxCodes with 13 barcode elements, 7 elements were determined from Read 1 and the remaining 6 elements from Read 2. For LoxCodes of size less than 13, a consensus barcode was determined, by comparing overlapping elements from Read 1 and Read 2.

Barcode Filtering:

Barcodes detected in only one of two PCR replicates were discarded, PCR duplicates normalized and barcode counts merged by taking the mean across the two samples. To filter out frequent LoxCodes for bona-fide clonal fata analysis, an empirical list of close to 30K common barcodes was generated from a total of 37 independent DNA samples from LoxCode-Rosa26CreErt2 cells (>50K cells per sample) induced at a similar recombination rate in the embryos but in adult mice. Barcodes detected in more than 10 independent samples were removed from the analysis.

Metaclone similarity analysis (MCSA):

Heatmap

Barcode counts per sample were normalized to counts per million (CPM), small barcodes with frequencies less than 0.1 per million discarded, log(1+x) transformed, and the cosine similarity between samples calculated and normalized. These meta-clonal similarity (MCS) values were then ordered using hierarchical clustering (method="ward.D2") and plotted as a heatmap.

UMAP (tissues)

MCS values were computed as for the heatmap and projected onto 2 dimensions through Uniform Manifold Approximation and Projection using the R package 'umap'.

Anatomical tissue connectivity map

An anatomical 2D tissue map was created using the locations of the analysed embryonic tissues and regions. A 4-NN-neighbour network based on MCS values was then projected onto this map connecting nodes between most similar tissues and regions. Edges with an MCS <0.75 were removed. For the purpose of visualization, edges between nodes were drawn using the edge bundling algorithm using the R package 'ggraph'.

UMAP (barcodes)

Counts across tissues per barcode were normalized to counts per million (CPM), log(1+x) transformed and projected onto 2 dimensions through Uniform Manifold Approximation and Projection using the R package 'umap'. Leiden clustering was performed from umap's computed NN-neightbour network using the R package 'leiden'.

Single Cell RNA sequencing analysis and cluster annotation

Sequencing data was aligned to mouse genome (mm10) with the cellranger software (version 5.0) and analysed using the R package 'Seurat'. Samples were demultiplexed using the R package deMULTIplex and custom code. Transcriptional clusters were annotated using established marker genes from the literature.

References:

1 Lawson, K. A., Meneses, J. J. & Pedersen, R. A. Clonal analysis of epiblast fate during germ layer formation in the mouse embryo. *Development* **113**, 891-911, doi:10.1242/dev.113.3.891 (1991).
2 Parameswaran, M. & Tam, P. P. Regionalisation of cell fate and morphogenetic movement of the mesoderm during mouse gastrulation. *Dev Genet* **17**, 16-28, doi:10.1002/dvg.1020170104 (1995).
3 Peng, G. *et al.* Molecular architecture of lineage allocation and tissue organization in early mouse embryo. *Nature* **572**, 528-532, doi:10.1038/s41586-019-1469-8 (2019).
4 Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* **570**, 77-82, doi:10.1038/s41586-019-1184-5 (2019).
5 Tam, P. P. L. & Ho, J. W. K. Cellular diversity and lineage trajectory: insights from mouse single cell transcriptomes. *Development* **147**, doi:10.1242/dev.179788 (2020).
6 Grosswendt, S. *et al.* Epigenetic regulator function through mouse gastrulation. *Nature* **584**, 102-108, doi:10.1038/s41586-020-2552-x (2020).
7 Mittnenzweig, M. *et al.* A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell* **184**, 2825-2842 e2822, doi:10.1016/j.cell.2021.04.004 (2021).
8 Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487-491, doi:10.1038/s41586-019-1825-8 (2019).
9 Saykali, B. *et al.* Distinct mesoderm migration phenotypes in extra-embryonic and embryonic regions of the early mouse embryo. *Elife* **8**, doi:10.7554/eLife.42434 (2019).
10 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, doi:10.1016/j.devcel.2009.08.002 (2009).
11 VanHorn, S. & Morris, S. A. Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. *Dev Cell* **56**, 7-21, doi:10.1016/j.devcel.2020.10.021 (2021).
12 Chen, C., Liao, Y. & Peng, G. Connecting past and present: single-cell lineage tracing. *Protein Cell* **13**, 790-807, doi:10.1007/s13238-022-00913-7 (2022).
13 Livet, J. *et al.* Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56-62, doi:10.1038/nature06293 (2007).
14 Ratz, M. *et al.* Clonal relations in the mouse brain revealed by single-cell and spatial transcriptomics. *Nat Neurosci* **25**, 285-294, doi:10.1038/s41593-022-01011-x (2022).
15 Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322-327, doi:10.1038/nature13824 (2014).
16 Rodriguez-Fraticelli, A. E. *et al.* Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212-216, doi:10.1038/nature25168 (2018).
17 Yang, D. *et al.* Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905-1923 e1925, doi:10.1016/j.cell.2022.04.015 (2022).
18 Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, doi:10.1126/science.aat9804 (2018).
19 Bowling, S. *et al.* An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. *Cell* **181**, 1693-1694, doi:10.1016/j.cell.2020.06.018 (2020).
20 Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, doi:10.1126/science.abc1944 (2021).
21 Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111, doi:10.1038/nature20777 (2017).
22 Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456-460, doi:10.1038/nature23653 (2017).

23 Pei, W. *et al.* Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by PolyloxExpress Barcoding. *Cell Stem Cell* **27**, 383-395 e388, doi:10.1016/j.stem.2020.07.018 (2020).
24 Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393-397, doi:10.1038/s41586-021-03786-8 (2021).
25 Choi, J. *et al.* A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98-107, doi:10.1038/s41586-022-04922-8 (2022).
26 Loveless, T. B. *et al.* Molecular recording of sequential cellular events into DNA. *bioRxiv* (2021).
27 Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, doi:10.1126/science.aaw3381 (2020).
28 Weber, T. S. *et al.* Site-specific recombinatorics: in situ cellular barcoding with the Cre Lox system. *BMC Syst Biol* **10**, 43, doi:10.1186/s12918-016-0290-3 (2016).
29 Gardner, R. L. & Rossant, J. Investigation of the fate of 4-5 day post-coitum mouse inner cell mass cells by blastocyst injection. *J Embryol Exp Morphol* **52**, 141-152 (1979).
30 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, doi:10.1016/j.devcel.2008.07.017 (2008).
31 McGrath, K. E. *et al.* A transient definitive erythroid lineage with unique regulation of the beta-globin locus in the mammalian embryo. *Blood* **117**, 4600-4608, doi:10.1182/blood-2010-12-325357 (2011).
32 Kinder, S. J. *et al.* The orderly allocation of mesodermal cells to the extraembryonic structures and the anteroposterior axis during gastrulation of the mouse embryo. *Development* **126**, 4691-4701, doi:10.1242/dev.126.21.4691 (1999).
33 Padron-Barthe, L. *et al.* Clonal analysis identifies hemogenic endothelium as the source of the blood-endothelial common lineage in the mouse embryo. *Blood* **124**, 2523-2532, doi:10.1182/blood-2013-12-545939 (2014).
34 Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*, doi:10.1038/nbt.4314 (2018).
35 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
36 Naik, S. H., Schumacher, T. N. & Perie, L. Cellular barcoding: a technical appraisal. *Exp Hematol* **42**, 598-608, doi:10.1016/j.exphem.2014.05.003 (2014).
37 Merino, D. *et al.* Barcoding reveals complex clonal behavior in patient-derived xenografts of metastatic triple negative breast cancer. *Nat Commun* **10**, 766, doi:10.1038/s41467-019-08595-2 (2019).
38 McGinnis, C. S. *et al.* MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods* **16**, 619-626, doi:10.1038/s41592-019-0433-8 (2019).
39 Chen, L., Wang, D., Wu, Z., Ma, L. & Daley, G. Q. Molecular basis of the first cell fate determination in mouse embryogenesis. *Cell Res* **20**, 982-993, doi:10.1038/cr.2010.106 (2010).
40 Bardot, E. S. & Hadjantonakis, A. K. Mouse gastrulation: Coordination of tissue patterning, specification and diversification of cell fate. *Mech Dev* **163**, 103617, doi:10.1016/j.mod.2020.103617 (2020).
41 Zhu, M. *et al.* Developmental clock and mechanism of de novo polarization of the mouse embryo. *Science* **370**, doi:10.1126/science.abd2703 (2020).
42 Dumas, L., Clavreul, S., Michon, F. & Loulier, K. Multicolor strategies for investigating clonal expansion and tissue plasticity. *Cell Mol Life Sci* **79**, 141, doi:10.1007/s00018-021-04077-1 (2022).
43 Tabansky, I. *et al.* Developmental bias in cleavage-stage mouse blastomeres. *Curr Biol* **23**, 21-31, doi:10.1016/j.cub.2012.10.054 (2013).
44 Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165**, 61-74, doi:10.1016/j.cell.2016.01.047 (2016).

45 Leeper, K. *et al.* Lineage barcoding in mice with homing CRISPR. *Nat Protoc* **16**, 2088-2108, doi:10.1038/s41596-020-00485-y (2021).
46 Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev Biol* **100**, 64-119, doi:10.1016/0012-1606(83)90201-4 (1983).
47 Kueh, A. J. *et al.* An update on using CRISPR/Cas9 in the one-cell stage mouse embryo for generating complex mutant alleles. *Cell Death Differ* **24**, 1821-1822, doi:10.1038/cdd.2017.122 (2017).

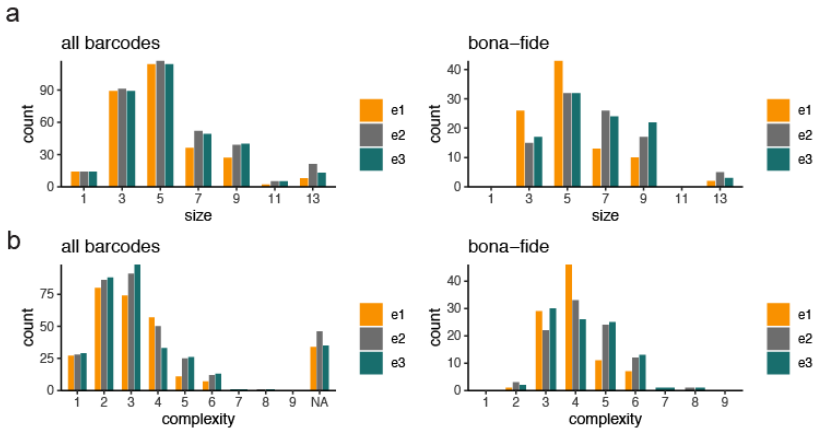
Data and code availability

Barcoding and scRNA data generated in this study, as well as the scripts to create the figures will be made available here in an updated version upon submission for peer review.

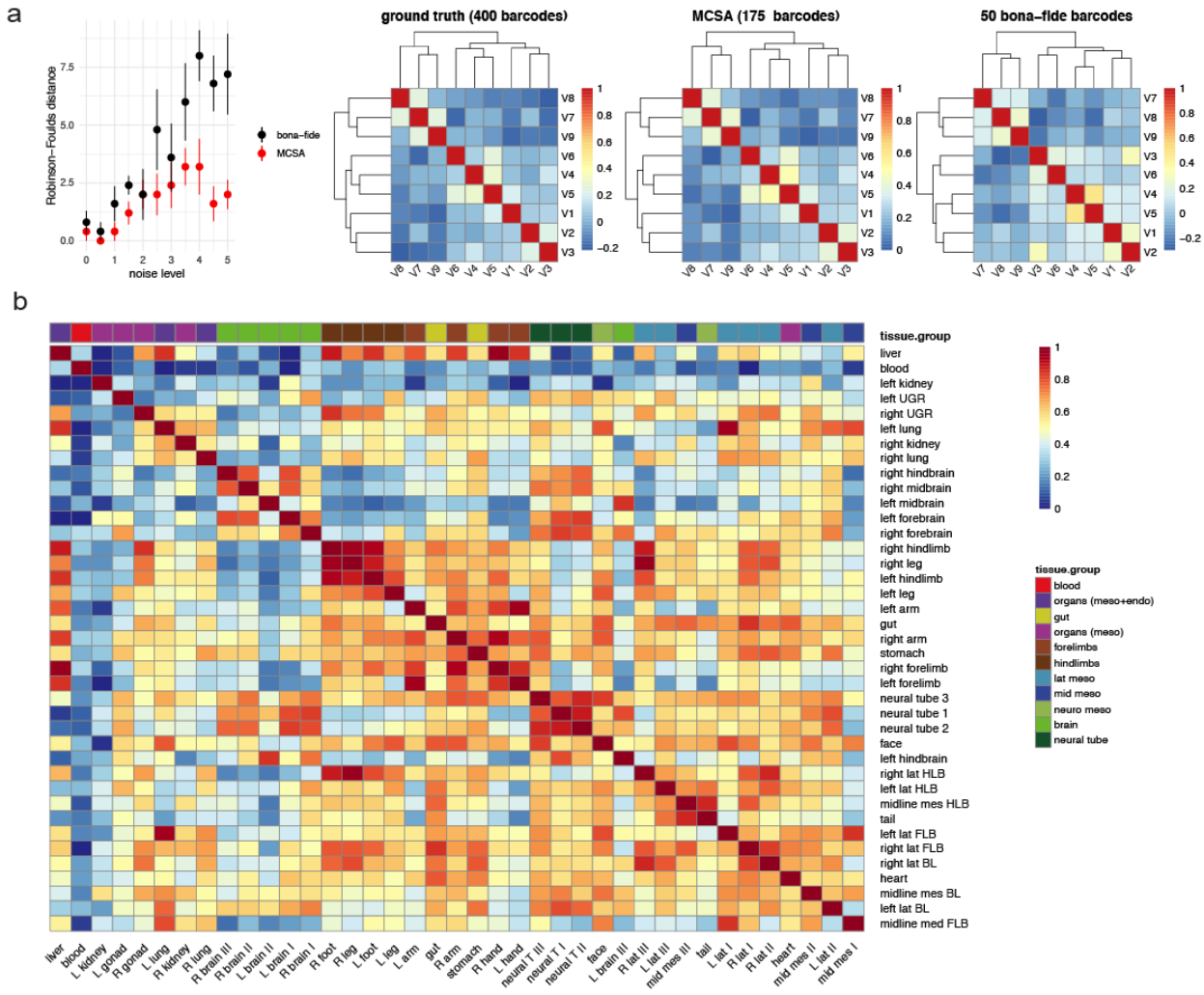
Acknowledgements
We thank the WEHI genomics and single cell core (SCORE), CRISPR mouse facility MAGEC, especially Andrew Kueh for mouse generation, and mouse facilities for animal husbandry. We thank Illumina, Inc for sequencing support. We thank Chris McGinnis and the Gartner lab (UCSF) for sharing MultiSeq reagents. We would also like to thank Ken Duffy for discussions and input during the early conceptual phase of the LoxCode technology. This work was funded by the CASS Foundation (SM/18/7776 for T.S.W.), National Health and Medical Research Council, Australia (2012196 (T.S.W), 1184736 (T.S.W. & S.H.N), 2009675 and 1145184 (S.H.N.), 1052195 (D.C.M.), 1129012 and 2011770 (S.T), the DHB Foundation (S.T) Human Frontiers Special Program (RGP0060/2012).

Contributions
T.S.W. with S.H.N. conceived the LoxCode approach. D.C.M., T.S.W., and S.H.N. built the LoxCode construct with S.G.'s advice. T.S.W., C.B, D.C.M, P.T, S.T., S.H.N. conceived the study design. C.B. performed all embryonic dissections, with T.S.W. for tissue and single cell processing. T.S.W optimised library preparation and sequencing. T.S.W. performed all analysis and generated figures, with S.Z. for the LoxCode recombination reference. C.B. provided cluster annotations, and T.S.W., C.B., S.T., P.T. and S.H.N. interpreted the data. T.S.W, D.C.M., S.T., and S.H.N. sourced funding. S.T. and S.H.N. supervised the study. All authors contributed to writing the manuscript.

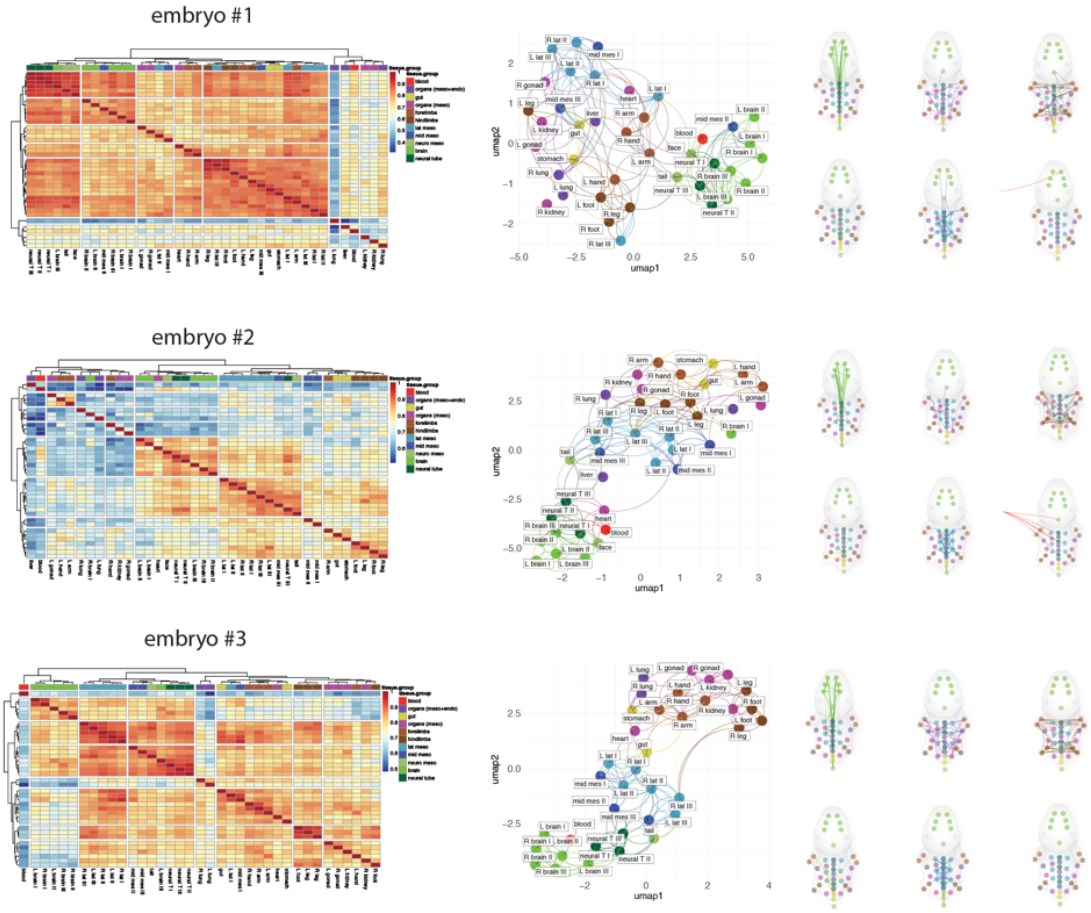
Supplementary Figures



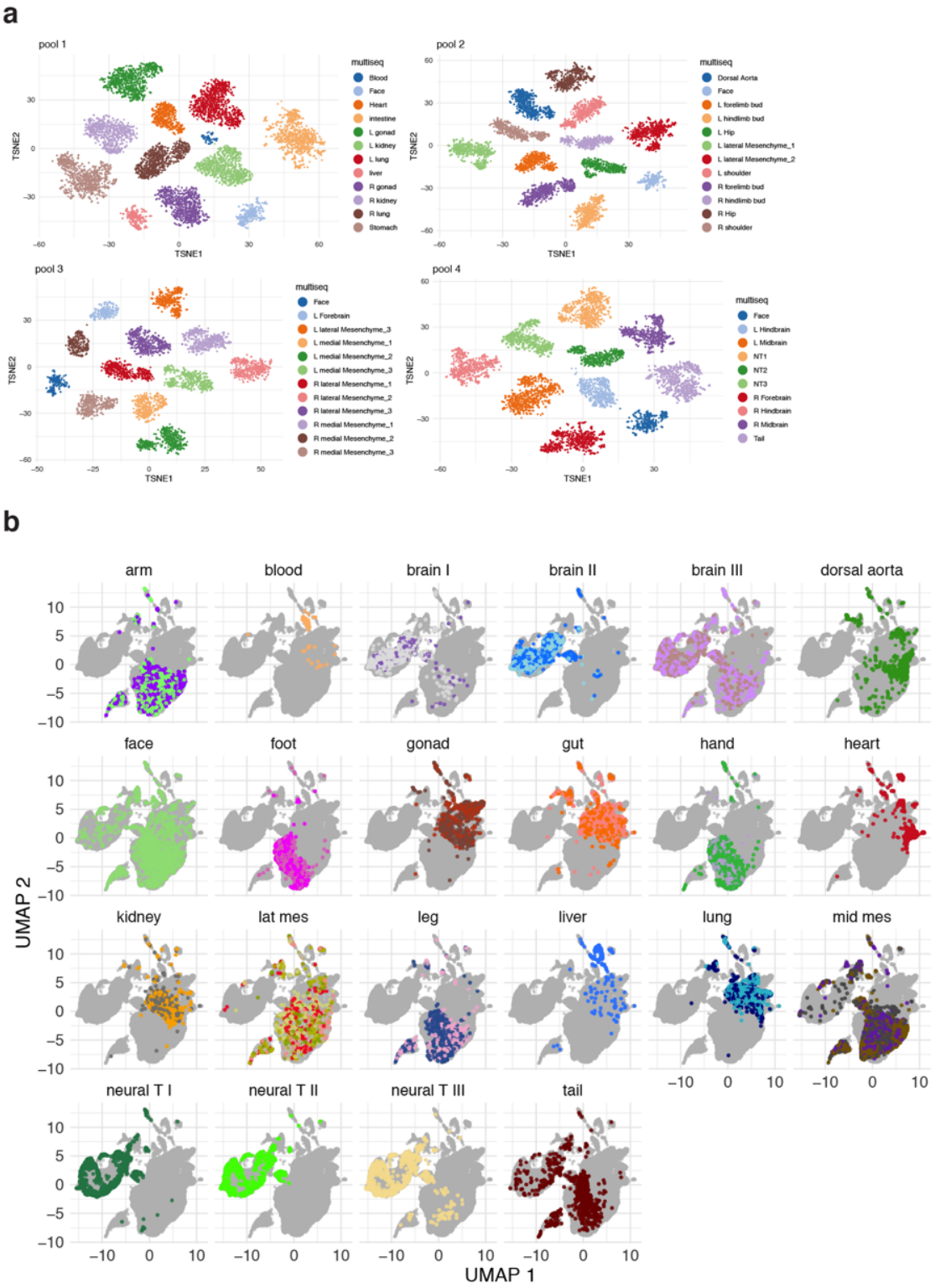
Supplementary Fig 1. Barcode filtering. a) Distribution of barcode sizes for all (PCR-filtered) and bona-fide (empirical-frequency-filtered) clones. **b)** Distribution of the minimum number of recombinations (complexity) for all (PCR-filtered) and bona-fide clones.



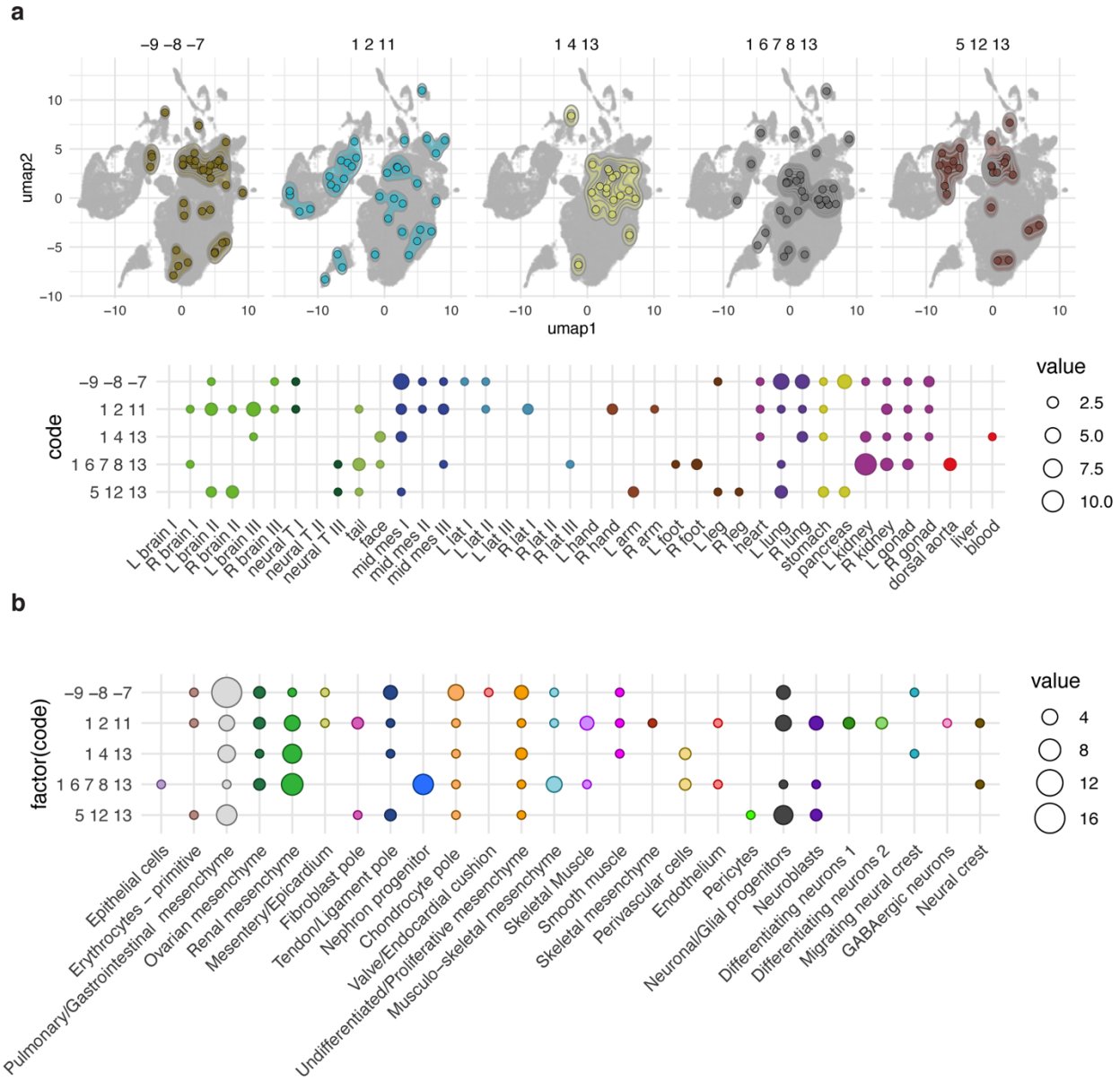
Supplementary Fig 2. Robinson-Foulds distance (mean ± SEM, 50 simulations) between branching hierarchies from simulated barcoding data and ground truth using 50 bona-fide or all barcodes (MCSA) (left). Examples of cosine similarity matrix computed from ground truth by MCSA and using 50 bona-fide barcodes at noise level 2. Note how MCSA represents the ordering/extent of relationships between tissues more faithfully compared to bona fide barcodes **b)** Cosine similarity matrix for all tissues using bona-fide barcodes ordered according to the corresponding MCSA heatmap in Fig. 2e.



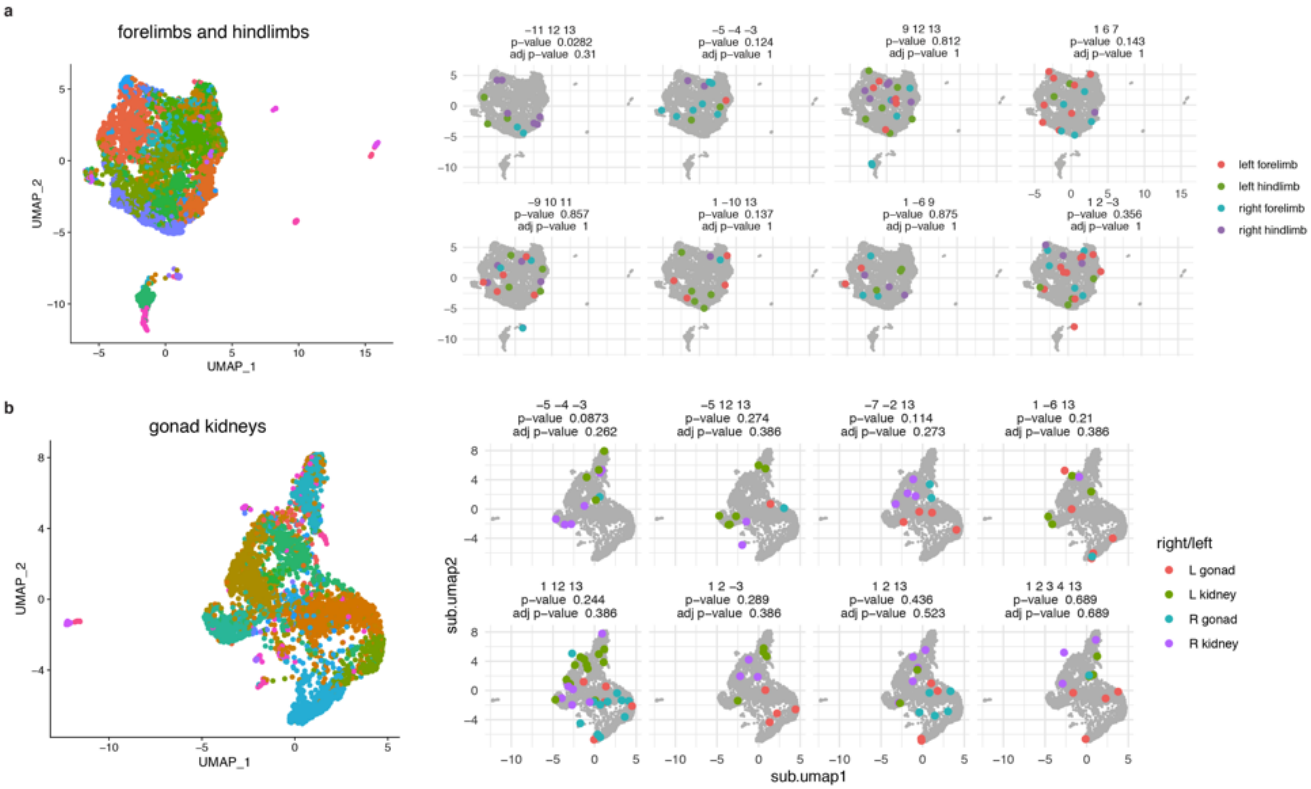
Supplementary Fig 3. MCSA heatmap, UMAP and pseudo-anatomical map of tissue relationships. These plots are as per Figure 2 and represent the corresponding 3 biological replicates that were pooled to generate Figure 2.


a
pool 1
multiseq
Blood
Face
Heart
intestine
L gonad
L kidney
L lung
liver
R gonad
R kidney
R lung
Stomach
TSNE1
TSNE2
pool 2
multiseq
Dorsal Aorta
Face
L forelimb bud
L hindlimb bud
L Hip
L lateral Mesenchyme_1
L lateral Mesenchyme_2
L shoulder
R forelimb bud
R hindlimb bud
R Hip
R shoulder
TSNE1
TSNE2
pool 3
multiseq
Face
L Forebrain
L lateral Mesenchyme_3
L medial Mesenchyme_1
L medial Mesenchyme_2
L medial Mesenchyme_3
R lateral Mesenchyme_1
R lateral Mesenchyme_2
R lateral Mesenchyme_3
R medial Mesenchyme_1
R medial Mesenchyme_2
R medial Mesenchyme_3
TSNE1
TSNE2
pool 4
multiseq
Face
L Hindbrain
L Midbrain
NT1
NT2
NT3
R Forebrain
R Hindbrain
R Midbrain
Tail
TSNE1
TSNE2
b
arm
blood
brain I
brain II
brain III
dorsal aorta
face
foot
gonad
gut
hand
heart
kidney
lat mes
leg
liver
lung
mid mes
neural T I
neural T II
neural T III
tail
UMAP 1
UMAP 2

Supplementary Fig 4. The 40 tissues that were dissected and made into a single cell suspension as per Figure 1, were then multiplexed using MULTI-seq, followed by pooling, running on droplet-based scRNAseq platform with a 3' capture scRNA-seq kit. a) Demultiplexing using MULTI-seq was effective in separating the tissues of origin, and b) were differentially distributed across transcriptional space (UMAP from Figure 4). Note, different colours are indicative of left and right (or I, II, III for mesenchymal tissues) samples of the indicated tissue e.g. left and right kidney.



Supplementary Fig 5. Meta-clonal fate heterogeneity. a) Examples of single cell meta-clonal output corresponding to the indicated LoxCodes. **b)** Distribution of single cells per detected LoxCode according to **t**issue (top) or cell type (bottom).



Supplementary Fig 6. Asymmetry in clonal contribution. Contributions of meta-clones to a) left and right fore- and hindlimbs, b) left and right gonad and kidneys. Meta-clones with significant bias are shown in Fig 4. g.