

## Lineage tracing of human embryonic development and foetal haematopoiesis through somatic mutations

Michael Spencer Chapman<sup>1,2,3,\*</sup>, Anna Maria Ranzoni<sup>1,4,5,\*</sup>, Brynelle Myers<sup>1,4,5</sup>, Nick Williams<sup>1</sup>, Tim Coorens<sup>1</sup>, Emily Mitchell<sup>1,3,4</sup>, Timothy Butler<sup>1</sup>, Kevin J Dawson<sup>1</sup>, Yvette Hooks<sup>1</sup>, Luiza Moore<sup>1,6</sup>, Jyoti Nangalia<sup>1,4,5</sup>, Philip S Robinson<sup>1,7</sup>, Elizabeth Hook<sup>6</sup>, Peter J Campbell<sup>1,4,\*†</sup>, Ana Cvejic<sup>1,4,5,\*†</sup>

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

2 Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, W12 0HS, UK

3 Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK

4 Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute, Cambridge, CB2 0AW, UK

5 University of Cambridge, Department of Haematology, Cambridge, CB2 0AW, UK

6 Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, CB2 0QQ, UK

7 University of Cambridge, Department of Paediatrics, Cambridge, CB2 0QQ, UK

\* Equal contribution

† Corresponding authors

Ana Cvejic, [as889@cam.ac.uk](mailto:as889@cam.ac.uk)

Peter Campbell, [pc8@sanger.ac.uk](mailto:pc8@sanger.ac.uk)

### Abstract

To date, ontogeny of the human haematopoietic system during foetal development has been characterized mainly through careful microscopic observations. Here we used whole-genome sequencing (WGS) of 511 single-cell derived haematopoietic colonies from healthy human foetuses of 8 and 18 post-conception weeks (pcw) coupled with deep targeted sequencing of tissues of known embryonic origin to reconstruct a phylogenetic tree of blood development. We found that in healthy foetuses, individual haematopoietic progenitors acquire tens of somatic mutations by 18 pcw. Using these mutations as barcodes, we timed the divergence of embryonic and extra-embryonic tissues during development and estimated the number of blood antecedents at different stages of embryonic development. Our analysis has shown that ectoderm originates from a smaller set of blood antecedents compared to endoderm and mesoderm. Finally, our data support a hypoblast origin of the extra-embryonic mesoderm and primitive blood in humans.

## Introduction

Human embryonic development starts with the division of the zygote to two daughter cells called blastomeres. After several further divisions, at approximately the 100-cell-stage, the cells physically separate into the inner cell mass (ICM) which is fated to become the embryo, and the trophoblast which develops into the chorionic sac and foetal component of the placenta. The ICM then undergoes further differentiation into the epiblast and hypoblast, which primarily contribute to embryonic and extra-embryonic tissues respectively. At 3 weeks of development, the epiblast differentiates into the three germ layers (ectoderm, mesoderm and endoderm). Ectoderm gives rise to the epidermis and nervous system; mesoderm generates the heart, kidney, skeleton, muscles and connective tissues; while endoderm forms the liver, lungs and inner layer of the digestive tract<sup>1</sup>.

The first haematopoietic cells (mainly nucleated erythrocytes and, to a lesser extent, macrophages and megakaryocytes<sup>2-4</sup>) start to form within blood islands in the yolk sac from 16 post-conception days (pcd). However, the developmental origin of these so-called primitive haematopoietic cells in humans is still debated. Yolk sac blood islands arise from the extra-embryonic mesoderm<sup>5</sup>. The precise origin of this extra-embryonic tissue remains unclear with some evidence in mice pointing to the primitive streak<sup>6,7</sup>, while histological studies in humans suggest it derives from the hypoblast<sup>7,8</sup>. Interestingly, human foetal histological studies have shown that at 8 pcw more than 90% of erythroid cells in circulating blood are still of primitive origin, while more than 80% of those found in liver derive from the aorta-gonad-mesonephros (the so-called ‘definitive’ wave of haematopoiesis). Thus, it has been speculated that definitive erythrocytes are temporarily retained in the liver<sup>9</sup>.

Establishment of the bone marrow, the main site of adult blood production, marks the end of the embryonic period in humans<sup>4</sup>. Haematopoietic colonisation of the bone marrow starts at ~10.5 weeks with the appearance of monocytes/macrophages, followed by haematopoietic stem and progenitor cells (HSPCs), although the exact onset of haematopoietic stem cell (HSC) activity in the bone marrow and the kinetics of migration between different organs remain unclear<sup>4,9,10</sup>. Here, we used the accumulation of somatic mutations in foetal HSPCs to gain further insights into the dynamics of human prenatal development and the origins of primitive and definitive haematopoiesis.

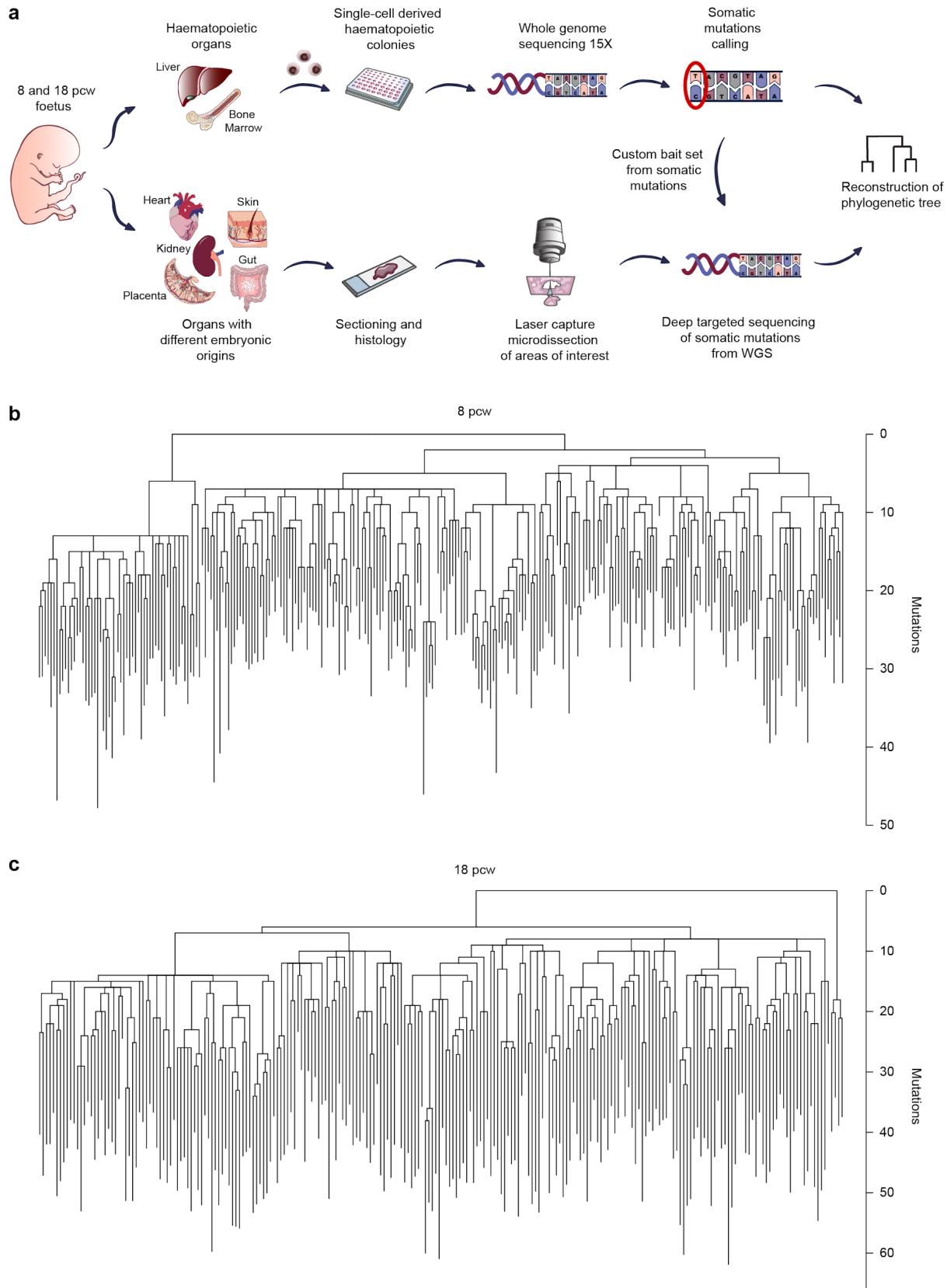
## Reconstructing the phylogeny of foetal HSPCs from whole-genome sequencing data

Somatic mutations accumulate in a cell’s genome over successive cell divisions<sup>11</sup>. These mutations are inherited by all future progeny of the cell as they are copied during DNA replication. Therefore, somatic mutations can effectively be used as barcodes in lineage

tracing studies<sup>12</sup>. By analysing patterns of mutations that have accumulated in somatic cells over time, it is possible to identify clonal relationships between cells and thereby reconstruct the phylogeny of a population<sup>13</sup>. In this study, we performed whole-genome sequencing (WGS) of single-cell-derived blood colonies from human foetal HSPCs, coupled with deep targeted sequencing of matched tissues of known embryonic origin, to build a phylogenetic tree of foetal development (Fig. 1a).

We first collected haematopoietic organs from two karyotypically normal male human foetuses of 8 and 18 pcw and used fluorescence-activated cell sorting (FACS) to isolate single HSPCs in liquid culture. Individual HSPCs were then expanded *in vitro* into colonies and sequenced by WGS to around 15× depth. From the liver of the 8 pcw foetus, we sequenced 277 colonies derived from single HSPCs (Lin-CD34+CD38+). At 18 pcw, both liver and bone marrow support blood production. Thus, we sequenced 234 colonies from phenotypically defined HSPCs (116 megakaryocyte-erythroid progenitors (MEPs), 105 common myeloid progenitors (CMPs) and 13 HSCs) from the liver (78 cells) and bone marrow of both femurs (79 and 77 cells) (Extended Data Fig. 1, Extended Data Table 1). The obtained sequencing depth (mean 22.6× for the 8 pcw colonies and mean 12.2× for the 18 pcw colonies) was sufficient to reliably call single-nucleotide variants (SNVs) (Extended Data Fig. 2). We applied stringent filtering criteria to remove germline mutations, recurrent artefacts from library preparation and sequencing, and mutations that were likely to have been acquired *in vitro*.

To validate the filtering strategy, we performed targeted resequencing of 501 of the 511 colonies to estimate the proportion of mutations that passed our filtering but were in fact subclonal or artefactual. Over 96% of mutations shared by more than one colony in the 8 pcw foetus and 100% of those in the 18 pcw foetus had variant allele fractions (VAFs) consistent with somatic mutations acquired *in vivo* (Extended Data Fig. 3a, b). Mutations passing filtering criteria were assigned to each cell. Mutations that were shared by multiple colonies were used to reconstruct phylogenetic trees of blood development (Fig. 1b, c). We constructed a tree in which the length of each branch is proportional to the number of somatic mutations accumulated over time within the inferred ancestral lineage. For clarity, here we refer to *lineage* as the direct descendants of a given ancestral cell.



**Fig.1 | Experimental workflow and phylogenetic trees.** **a**, Single HSPCs were isolated from an 8 pcw human foetal liver and from 18 pcw matched foetal liver and two femurs. Single-cell derived colonies were expanded in vitro and sequenced by WGS. Somatic mutations identified

were used to reconstruct clonal relationships between cells. From the same two foetuses, non-haematopoietic organs with different embryonic origins were isolated, sectioned and laser-capture micro-dissected in order to produce well-defined sections with known developmental origin. These tissues underwent deep targeted sequencing using a DNA bait set covering the somatic mutations identified by WGS of blood colonies. Obtained data were integrated into the phylogenetic trees to highlight the time of separation of embryonic tissues during human development. **b**, Phylogeny of 277 single HSPCs from the liver of the 8 pcw foetus. **c**, Phylogeny of 234 single HSPCs from the liver and bone marrow of the 18 pcw foetus. Each tip of the tree is a single-HSPC derived colony. Cells sharing a common ancestor are grouped in clades deriving from a single node. Branch lengths are proportional to the number of acquired somatic mutations.

In addition to HSPCs, we collected non-haematopoietic tissues of mesodermal, ectodermal, endodermal and extra-embryonic origin from the same two foetuses. We dissected the placenta, gut, kidney, heart and skin from the 8 pcw foetus; and kidney, heart and skin from the 18 pcw foetus. All tissues were fixed, paraffin-embedded, sectioned and cut using laser-capture microdissection (LCM) to obtain well-defined tissue structures (which we refer to as *microbiopsies* from now on) of specific extra-embryonic and embryonic origin. Microdissected tissues included the mesenchymal core from the placenta (deriving from the extra-embryonic mesoderm), syncytiotrophoblast (from the trophoblast), circulating blood (from the primitive wave of haematopoiesis), gut epithelium (from the endoderm), heart muscle (from the lateral plate mesoderm), kidney tubules (from the intermediate mesoderm), limb deep undifferentiated tissue (from mesoderm) intervertebral disc (from the paraxial mesoderm), epidermis (from the surface ectoderm) and skin peripheral nerve (from the neural ectoderm) (Extended Data Fig. 4). We then performed deep targeted sequencing (Extended Data Table 2) using custom DNA bait sets designed to include all shared and a majority of private somatically-acquired SNVs (please see Methods) detected in HSPC colonies by WGS.

### Acquisition of somatic mutations in early embryonic and foetal development

We estimated mutation burdens following adjustments for sensitivity and removal of *in vitro* mutations. In the 8 pcw foetus, the mean corrected SNV burden was 25 mutations per colony (standard deviation (sd) 6.1 mutations per colony) and in the 18 pcw foetus, the mean SNV burden was 42.6 mutations per colony (sd 7.7 mutations per colony) (Extended Data Fig. 5a). Interestingly, the observed standard deviations are broadly equivalent to that expected for a Poisson distribution, suggesting that all the lines-of-descent sequenced within each foetus have experienced the same mutation rates during development.

We detected a mean of 1.82 indels (range: 0 - 6) in the 8 pcw colonies and 1.23 (range: 0 - 5) in the 18 pcw colonies (Extended Data Fig. 5b). Due to the low numbers of indels, an equivalent sample-specific correction approach was not feasible. However, we used a Bayesian Poisson regression model to calculate a corrected mean indel burden of 2.09 indels per colony (95% confidence interval 1.90 - 2.29) for the 8 pcw foetus and of 2.13 indels per colony for the 18 pcw foetus (95% confidence interval 1.91 - 2.39). On average, the proportion of SNVs shared between multiple HSPC colonies was 0.62 for the 8 pcw foetus and 0.49 for the 18 pcw foetus (Extended Data Fig. 5c).

To further examine the genomic distribution of mutations identified by WGS we assessed the proportion of SNVs mapping to different genomic features. We identified a higher than expected proportion of SNVs in intergenic regions of the genome (Extended Data Fig. 6a). In contrast, we detected a lower than expected proportion of mutations in the introns, exons, and regions 5 kb upstream/downstream of the gene body (Extended Data Fig. 6b-c). This is consistent with previous reports of lower mutation rates in regions of transcription and accessible chromatin<sup>14-16</sup>.

The 96-profile mutation spectrum, which incorporates the trinucleotide context of the substitution, revealed a predominance of C>T transitions, primarily at CpG sites, which are associated with spontaneous deamination of methylated cytosine<sup>17</sup>. In both foetuses, the mutation spectrum of clonal mutations seen only in single colonies closely matched those of mutations shared by two or more colonies, confirming that the mutation spectrum does not noticeably change across the first half of gestation (Extended Data Fig. 7).

### Defining clonal relationships between foetal HSPCs

To determine clonal relationships between HSPCs, we constructed a phylogenetic tree for both 8 and 18 pcw foetuses (Fig. 1b, c). The root of a somatic lineage tree is, by definition, the zygote – it is likely that the initial bifurcation of the phylogeny therefore marks the first zygotic division to two blastomeres. Interestingly, the reconstructed phylogenetic trees revealed unequal contribution of each blastomere to the blood compartment (5:1 for 8 pcw and 59:1 for 18 pcw) indicating even higher asymmetry of embryonic contribution is possible for the first zygotic cell division than previously observed<sup>13,18</sup>.

The first 3 cell divisions (up to the 8-cell embryo) had a relatively high mutation rate, with a mean of 2.4 mutations acquired in each of the new daughter cells per cell division. This is comparable to estimates derived from VAFs of post-zygotic mutations in bulk sequencing<sup>19</sup>. From this point onwards in our two foetuses, there was a profusion of high-degree polytomies,

defined as branch-points in the phylogenetic tree generating more than two descendants. In the context of a somatic lineage tree, a polytomy can only occur when there are cell divisions that generate no detected somatic mutations. Of the 5 cells from the 8-cell stage that contributed significantly to the HSPC population there are four polytomies (12-, 10-, 5-, 5-degree) and one dichotomy. Modelling of these polytomies suggests that the mutation rate decreases to less than 0.9 mutations/cell division at this stage of embryonic development (approximate Bayesian computation, see Methods). This coincides with zygotic genome activation in humans<sup>20</sup>, which may enable expression of DNA repair machinery.

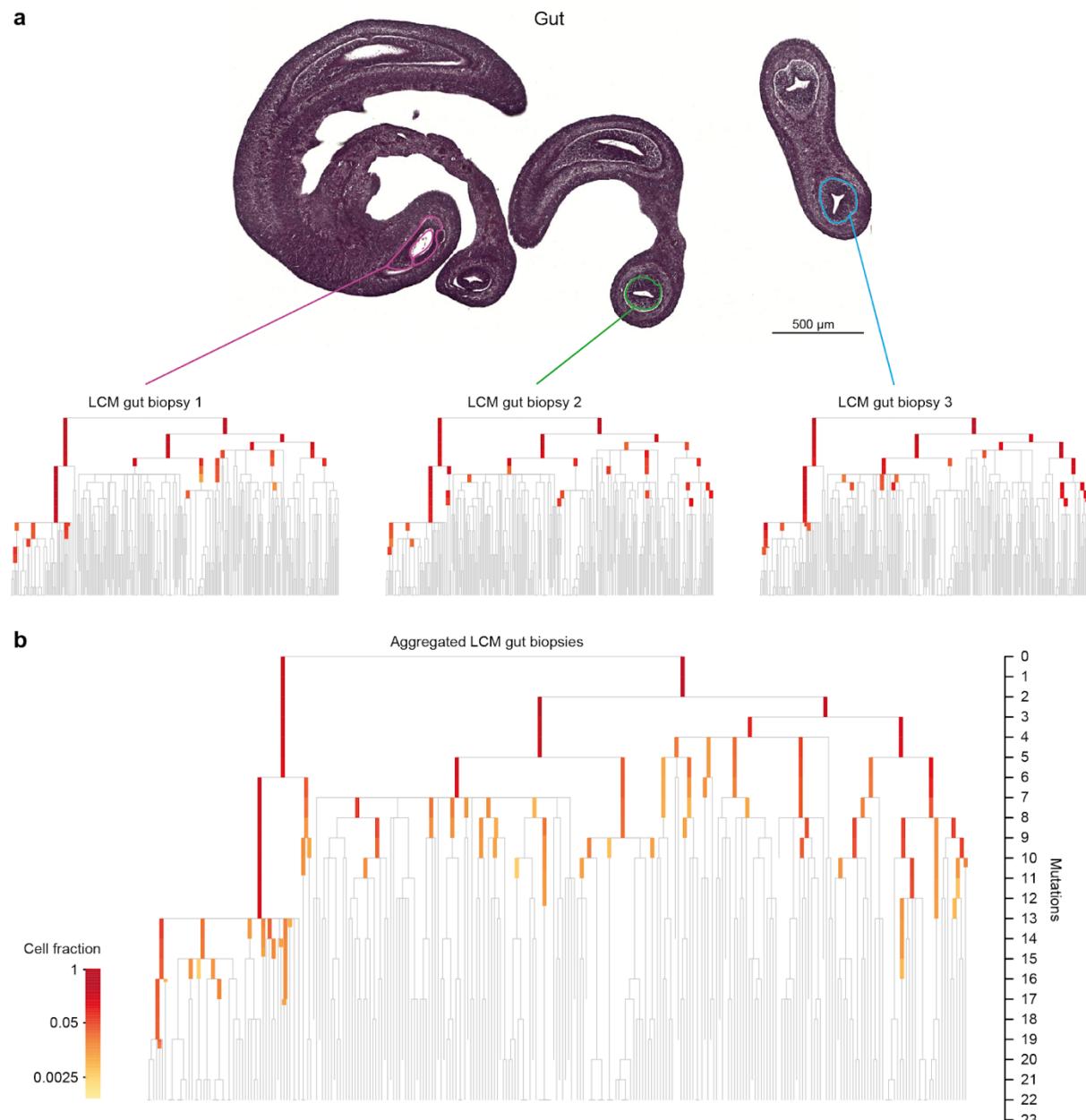
### **Timing the divergence of gut and blood during foetal development**

In order to precisely determine the timing of individual mutation acquisition relative to key developmental events (i.e. trophectoderm formation, gastrulation and organogenesis), we deep-sequenced non-haematopoietic tissues for the somatic mutations that had been identified in HSPCs (mean tissue sequencing depth 450×, Extended Data Table 2). As we have seen, the mutation rate during embryogenesis is sufficiently informative that the phylogenetic tree essentially represents a high-resolution lineage tree defining the initial cell divisions of the embryo. A given mutation occurs in a specific embryonic cell, which can be placed accurately on the lineage tree – this mutation then acts as a permanent clonal mark inherited by all descendants of that cell. Using targeted sequencing of different tissues, with different developmental origins, we can identify mutations that are present in multiple tissue layers, organs or regions – such mutations must have occurred in cells that preceded the developmental split between those tissues. Therefore, the tissues that shared the fewest mutations with HSPCs were likely to be the first to separate from ancestral lines-of-descent that gave rise to the haematopoietic cells (*blood antecedents* from now on) and greater sharing of mutations in a given tissue suggests it diverged from blood antecedents later in development.

Using laser-capture microdissection, we generated multiple microbiopsies of non-haematopoietic tissues, such as the epithelium of the gut, which derives from the endoderm (Fig. 2a). DNA from the microbiopsies underwent targeted sequencing to characterise the presence or absence of each mutation, as well as the fraction of cells carrying each mutation in that microbiopsy.

For gut epithelium, we observed that mutations close to the root of the HSPC phylogenetic tree were universally present in the different microdissections, and contributed to high fractions of cells in the individual biopsies (Fig. 2a). From this, we can infer that the earliest cells in the embryo, such as those in the 4-cell and 8-cell cleavage stages, contributed large numbers of

descendants to each of the gut microbiopsies (as well as blood HSPCs). As we traverse the phylogenetic tree from the root towards the tips, we found that the mutations became progressively less well represented in individual microbiopsies, before eventually becoming undetectable. Since each microbiopsy only samples a few hundred to thousand cells, and the average sequencing coverage was 70x per microdissection, we aggregated data from all 12 microbiopsies of gut epithelium to improve our statistical power for detecting mutations occurring later in embryogenesis (Fig. 2b).



**Fig.2 | Reconstructing lineage divergence through targeted sequencing.** **a**, Laser-capture microdissection was used to generate microbiopsies of 8 pcw gut epithelium. Somatic mutations shared with HSPCs were then detected by deep targeted sequencing. Mutations

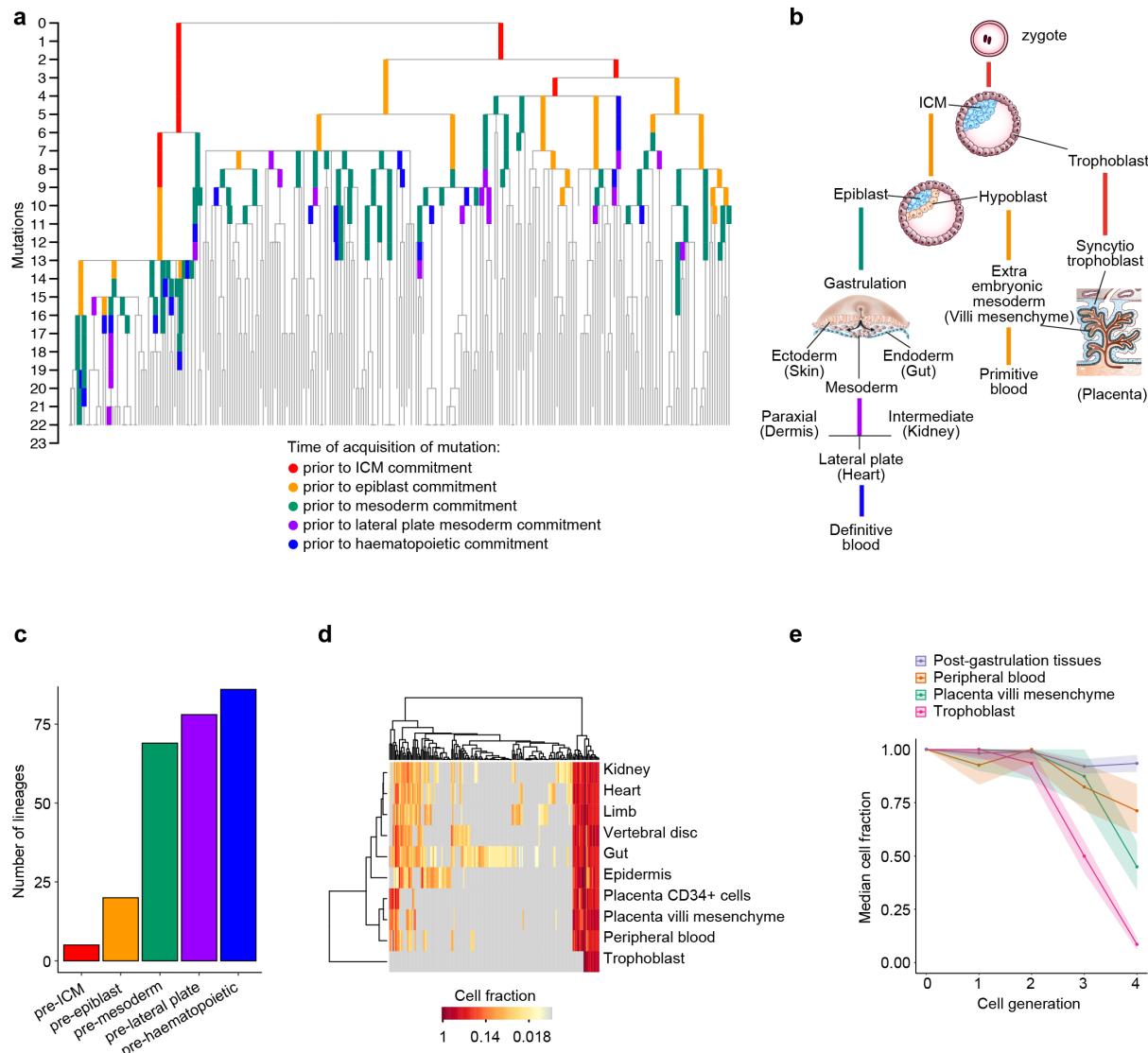
*detected in HSPCs that were shared with gut epithelium are highlighted on the phylogenetic trees. The trees only show the first 22 mutations to highlight early branching. Three representative microbiopsy cuts are shown. Individual mutations in the phylogenetic tree are coloured on a grey-yellow-red log scale according to the fraction of cells in the biopsy carrying that mutation (with grey being absent), as shown in the scale in panel b. Mean sequencing depth: 89× for microbiopsy 1 (pink), 56× for microbiopsy 2 (green) and 91× for microbiopsy 3 (blue). b, Mutations shared after aggregating reads across 12 microbiopsies. Mean sequencing depth of aggregated microbiopsies: 841×. LCM, laser-capture microdissection.*

The aggregated data for gut revealed that many lineages ( $\geq 60$ ) made substantial contributions to both gut epithelium and blood HSPCs. These lineages must have predated gastrulation, when endoderm and mesoderm split from one another. Following individual lines-of-descent from root towards tip, different patterns emerged – for some, the cellular contribution of successive generations in that line-of-descent decreased steadily, suggesting that both daughter cells of each successive cell division contributed to gut epithelium. For others, the cellular contribution to gut epithelium stayed relatively constant over several generations, suggesting that only the daughter cells in that line-of-descent contributed meaningfully to gut epithelium.

### **Timing specification of extra-embryonic and embryonic tissues**

We proceeded in the same way, using LCM and deep targeted sequencing, to characterise all the embryonic and extra-embryonic tissues available to us (Extended Data Table 2). This enabled us to time the progressive divergence of different tissues from blood antecedents during embryonic development (Fig. 3a, b).

In the mouse, the molecular distinction between trophoblast and ICM occurs at the 8-16 cell stage, but the exact timing in humans remains unknown<sup>21</sup>. In the 8 pcw foetus, we found that the trophoblast-derived tissue was the first to diverge from blood antecedents at the 4- to 16-cell stage embryo (corresponding to 2 to 4 generations from the zygote; *generation* is defined here as a round of cell division), thus following a timeline similar to mouse development. We can directly infer the presence of five distinct blood antecedents present at the time of trophoblast divergence in the embryo (Fig. 3c).



**Fig. 3 | Timing of divergence of embryonic and extra-embryonic lineages during development.** **a**, Phylogenetic tree showing the clonal relationships of 8 pcw HSPCs. Mutations identified in HSPCs by WGS that were also detected in non-haematopoietic tissues are coloured on the tree according to the earliest diverging tissue in which the mutation was reliably detected. Branch lengths are proportional to the number of mutations accumulated. The tree is cut at 22 mutations to highlight the early branching of the tree. **b**, Diagram of main human developmental events. Lineage separation is coloured as per the phylogenetic tree. In brackets are examples of tissues collected from each embryonic lineage. ICM, inner cell mass. **c**, Bar plot showing the minimum number of HSPC antecedents present at the time of divergence of different developmental lineages. **d**, Heatmap representation of the targeted sequencing data. Each column represents an individual mutation, and each row a single tissue. The colour shows the VAF of the mutation in that tissue with grey indicating that it was not detected. The tissues are clustered using soft cosine similarity (see methods). **e**, Line plot showing lineage loss of microdissected tissues at different times (represented as cell

*generations from the zygote). Shaded areas represent confidence intervals. ICM, inner cell mass.*

The next major event in embryogenesis is epiblast specification. We can infer that a minimum of 20 blood antecedent lineages already existed prior to epiblast specification (Fig. 3c). A recent study of *in vitro* human development up to the pre-gastrulation stage reported that specification of trophoblast, hypoblast and epiblast was evident by immunostaining of specific markers at 8 days post fertilisation. At this stage, the embryo comprised ~500 cells, with ~400 cells in the trophoblast, ~60 in the epiblast and ~40 in the hypoblast<sup>22</sup>. Therefore, there is a strong concordance between our estimates and the estimates based on immunostaining of epiblast-committed cells in embryos at the specific developmental stage. Our analysis, however, has the property that we can ascertain their ultimate contribution to organogenesis, not just their expression pattern at the time of specification. The concordance in numbers by the two methods does indeed suggest that the majority of cells expressing epiblast markers do in fact go on to contribute to detectable fractions of cells during organogenesis.

Proceeding further into embryonic development, we infer that at least 70 blood antecedents existed prior to mesoderm specification. Interestingly, the number of mutations shared between HSPCs and lateral, intermediate and paraxial mesoderm-derived tissues compared to endoderm- and ectoderm-derived tissues was similar, indicating that specification of each mesodermal component occurred within a few generations of mesoderm formation (Fig. 3a).

The origin of extra-embryonic mesoderm in humans has been debated for the last five decades, with mouse studies showing it derives from the primitive streak, while histological observations in humans point towards a hypoblast origin<sup>6,7,8</sup>. To address this issue, we compared the scope of shared mutations in the mesenchymal core of the placental villi and peripheral blood to that of tissues of known trophoblast and post-gastrulation origin. We observed that tissues derived from extra-embryonic mesoderm clustered separately from trophoblast-derived and post-gastrulation tissues (Fig. 3d and Extended Data Fig. 8a). In addition, 90% of trophoblast lineages were already lost in the tree by the fourth generation, in contrast to 25% and 50% of peripheral blood and placenta mesenchyme respectively and 10% of post-gastrulation lineages (Fig. 3e). Our data therefore support the hypothesis that the extra-embryonic mesoderm in humans emerges from the hypoblast<sup>8</sup>.

### **Relative contributions of cells to germ layers at gastrulation**

It is not known whether equal numbers of cells at gastrulation contribute to the tissues of all three germ layers. Interestingly, we found that the pattern of shared mutations in ectodermal

tissues was distinct from that observed in mesodermal and endodermal tissues (Fig. 4). In non-haematopoietic mesodermal and endodermal tissues, we detected mutations from the majority of the blood antecedents up until 5-6 generations from the zygote, though each antecedent gave rise to a small proportion of cells from sequenced tissues at 8 pcw (Figure 4a, 4e and 4g). This implies that many epiblast cells contributed to mesoderm and endoderm development, and made broadly similar, but individually small, fractional contributions to these germ layers.

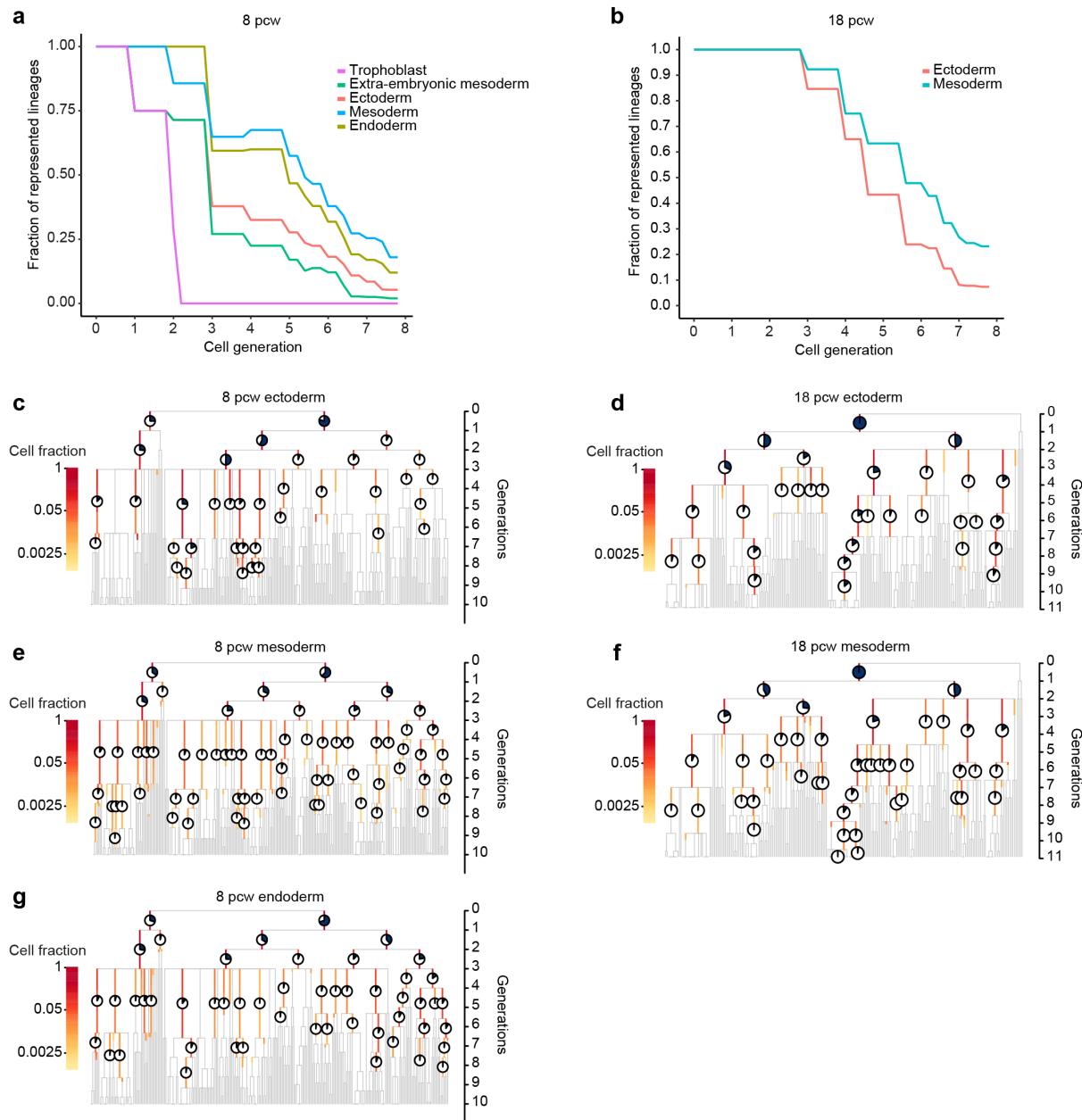
In ectoderm, however, mutations from fewer blood antecedents were detected, though with each giving rise to a relatively high cellular fraction of sequenced ectodermal microbiopsis (Figure 4b). Importantly, the individual lines-of-descent for ectoderm were detectable down to the same number of generations as for endoderm and non-haematopoietic mesoderm (typically 6-8 generations): there were just fewer of them. Moreover, following individual lines-of-descent from root to tip in the ectodermal deep sequencing showed that cellular contributions frequently stayed constant over several generations. This implies that the different pattern evident for ectoderm was not explicable by varying statistical power of mutation detection or the problem of lineages missing from our phylogeny. Instead, it suggests that relatively fewer cells from the epiblast contributed to the ectodermal tissues we sequenced than to endoderm or non-haematopoietic mesoderm.

A comparable pattern was seen in the 18 pcw foetus with two blood antecedents from around the ninth generation giving rise to a quarter of cells from sequenced ectodermal tissues. A similar limited set of blood antecedents predominated in all three ectodermal tissues that we sequenced (epidermis, hair follicle and peripheral nerve). While we cannot exclude the possibility that ectoderm antecedents migrate to a given anatomic region of the body, prior to their specification to distinct ectodermal derivatives, the more plausible conclusion would be that ectoderm originates from a smaller set of antecedents compared to endoderm and mesoderm.

### Trafficking of HSPCs across different haematopoietic sites

The colonisation of the long bones by haematopoietic cells from the liver starts around the 11th week of development<sup>9</sup>. From then on and up until birth, both the liver and the bone marrow are active sites of haematopoiesis. To investigate the kinetics of HSPC migration across early haematopoietic organs in 18 pcw foetus, we collected HSPCs from three separate sites of blood production. Out of the 234 HSPCs analysed, 78 were isolated from the liver, 77 from the bone marrow of femur 1 and 79 from femur 2. We did not observe statistically significant clustering ( $P=0.24$ ) by organ of origin in the phylogeny, suggesting extensive traffic of

haematopoietic stem and progenitor cells between liver and bone marrow (Extended Data Fig. 9a,b).



**Fig. 4 | Representation of haematopoietic lineages in developmentally-defined non-haematopoietic tissues.** **a**, Line plot showing the fraction of captured haematopoietic lineages in the 8 pcw phylogeny that are also detected in non-haematopoietic post-gastrulation tissues, grouped by germ layer. This is plotted over successive generations, during which the total number of lineages increases, but the proportion of lineages detected decreases. **b-d**, 8 pcw phylogenetic trees highlighting mutations detected in HSPCs that were shared with **b**, ectoderm-derived tissues, **c**, mesoderm-derived tissues and **d**, endoderm-derived tissues. The fraction of cells deriving from each lineage is plotted on a donut plot, with the black line representing the median cell fraction, and the dark and light blue segments representing the

*lower and upper 95% confidence intervals respectively. e, Line plot showing the fraction of captured haematopoietic lineages in the 18 pcw phylogeny that are also detected in non-haematopoietic tissues. For clarity the minor branch has not been considered as it does not contribute to any of the sequenced tissues. f-g, 18 pcw phylogenetic trees highlighting mutations detected in HSPCs that were shared with f, ectoderm-derived tissues and g, mesoderm-derived tissues. Trees in this figure were made ultrametric.*

## Discussion

Our data reveal an initial high rate of somatic mutation acquisition, which drops markedly after 3 generations, coinciding with maternal-to-zygotic transition. Although at this stage of development, zygotic transcription is still inactive, the extensive chromatin accessibility<sup>23,24</sup> and dilution of maternally-produced DNA-repair factors in dividing blastomeres could temporarily leave the genome more prone to accumulating mutations in the first generations. The reconstructed phylogenetic trees revealed an unequal contribution of each blastomere to the blood compartment which is in line with the stochastic nature of early cell fate decisions. This was particularly striking in the 18 pcw foetus, where just four out of 234 examined blood cells originated from one of the blastomeres. These four cells did not share any mutations with non-haematopoietic tissues of mesodermal and ectodermal origin (Extended Data Fig. 8) suggesting that the progeny of that blastomere had modest contribution to the embryonic tissues by either committing preferentially to trophoblast and hypoblast or dying out.

Based on the VAFs obtained from targeted sequencing in the 8 and 18 pcw foetus we further estimated that the separation between trophoblast and ICM occurs at the 8- to 16-cell stage with five distinct blood antecedents being present at the time. The number of blood antecedents in the embryo increased to 20 at the epiblast stage and a minimum of 70 at the time of mesoderm specification. Interestingly, our data suggest that ectoderm originates from a smaller set of blood antecedents compared to endoderm and mesoderm. Once the mesoderm was formed the specification of mesodermal components (i.e. lateral, intermediate and paraxial mesoderm) occurred within a few generations. Finally, we show that the primitive blood, captured in the circulation of the 8 pcw foetus, originates from hypoblast, as demonstrated by its higher level of mutation sharing with the extra-embryonic mesodermal tissues compared to tissues of known post-gastrulation origin. In this study we provide the conceptual framework for the future exploration of the outstanding questions in human development through the use of naturally occurring somatic mutations as barcodes.

## Acknowledgements

The authors would like to thank the Wellcome Sanger Institute (WSI) Cytometry Core Facility for their help with single-cell index sorting, the WSI DNA pipelines for their contribution in sequencing the data and the WSI Cancer, Ageing and Somatic Mutation programme IT and sample teams for their bioinformatic and logistical support. We would also like to thank Jana Eliasova for her precious support and help with the illustrations and the Human Developmental Biology Resource (HDBR) for providing samples.

## Funding

The study was supported by European Research Council project 677501 – ZF\_Blood (to A.C. and A.M.R), EMBO small grant (to A.C.) and core support grants from the Wellcome Trust to the Wellcome Sanger Institute and both Wellcome and MRC to the Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute. M.S.C. was supported by a Wellcome Clinical PhD Fellowship.

## Authors contributions

A.C., A.M.R. and P.C. conceived the study; A.M.R. performed all the experiments with help from B.M; M.S.C. carried out the computational analysis of the WGS data, under the supervision of P.C.; Y.H. prepared all histology sections for microdissection; T.B. assisted with the targeted sequencing strategy; T.B and P.R. assisted with the laser capture microdissections; E.H. and L.M. assisted with annotating histology slides; N.W., T.C. and E.M. assisted with somatic mutation calling from WGS data, the construction of the phylogeny and the assignment of mutations to the tree; N.W., J.N. and K.D. helped design and implement population simulation models to help interpret the phylogenies; A.C., A.M.R, and M.S.C. designed the figures and wrote the manuscript with inputs from the other authors. All authors approved the final version of the manuscript.

## Methods

### Ethics and tissue acquisition

Tissues used in this study were isolated from two karyotypically normal male human foetuses (8 and 18 pcw) following termination of pregnancy and informed written consent. All material was provided by the Joint MRC/Wellcome Trust (Grant MR/R006237/1) Human Developmental Biology Resource (<http://www.hdb.org>), in accordance with ethical approval by the NHS Research Health Authority, REC Ref: 18/LO/0822. All tissues were processed on the day of collection.

### **Generation of single-HSPC colonies**

Foetal liver tissue was passed through a 70  $\mu\text{m}$  strain with cold PBS (Gibco) to generate a single-cell suspension. Bone marrow was isolated by flushing cold PBS into the cavity of the two femurs (for the 18 pcw sample) and passed through a 70  $\mu\text{m}$  strain to generate a single cell suspension. Cells were centrifuged at 300 g for 5 minutes at 4°C and resuspended in Red Blood Cell lysis buffer (eBioscience) for 2 minutes. The lysis reaction was stopped by adding 20 ml of cold PBS. MACS columns (Miltenyi Biotec - 130-090-101) were used for live cell enrichment, following the manufacturer's instructions. For the 18 pcw tissues, a CD34+ cell enrichment step was also performed, using MACS columns (Miltenyi Biotec - 130-046-702) following the manufacturer's instructions. Cells were stained with antibodies in a total volume of 100  $\mu\text{m}$  PBS with 5% FBS (Gibco) for 30 minutes covered from light at 4°C and filtered into polypropylene FACS tubes (ThermoFisher) to a final volume of 500  $\mu\text{l}$  of PBS with 5% FBS. Single HSPCs were index-sorted with a BD Influx Sorter in liquid culture (as previously described<sup>25</sup>) into 96-well plates seeded with MS5 feeder layer (DSMZ) and cultured for 6-8 days at 37° and 5% CO<sub>2</sub>. At the end of the culture, cells were collected and filtered into polypropylene FACS tubes. Cells were sorted with a BD Influx Sorter into 96-well plates and lysed using the Arcturus PicoPure Kit (Applied Biosystems) according to the manufacturer's instructions.

### **Tissue processing for LCM**

Tissues from the 8 pcw (placenta, heart, kidney, gut, skin and vertebrae) and 18 pcw (skin, kidney and heart) samples were fixed in PAXgene Tissue FIX (Qiagen) for 4 and 24 hours respectively and embedded in paraffin using a Tissue-Tek VIP machine (Sakura). Sections of 5-10  $\mu\text{m}$  were cut using a microtome (Leica-RM2265), placed on PEN MembranSlide glass slides with a 2.0 $\mu\text{m}$  membrane (Leica) and dried in a laboratory oven (Genlab) at 37°C. Sections were dehydrated with a series of ethanol (VWR) washes and stained with Mayer's haematoxylin (Sigma) / Accustain Eosin Y (Sigma) before a final xylene (Sigma) cleaning step. Slides were dried overnight. High resolution imaging was performed using a slide scanner (Hamamatsu). Areas of interest were annotated on tissue scans and isolated with a microscope equipped with a laser-capture microdissection system (Leica LMD7). Microbiopsies were collected in 96-well plates and lysed using the Arcturus PicoPure Kit (Applied Biosystems) according to the manufacturer's instructions.

### **Library preparation and sequencing**

Library preparation for both the HSPC colonies and LCM microbiopsies was performed as previously described<sup>26</sup>. Paired end sequencing reads (150bp) were generated using either the

Illumina NovaSeq® 6000 platform (8 pcw foetus) or Illumina HiSeq® 4000 (18 pcw foetus) resulting in ~22.6× coverage and ~12.2× coverage per colony respectively. BWA-MEM was used to align sequences to the human reference genome (NCBI build37).

### **Mutation discovery and tree-building from whole-genome sequencing data**

Whole-genome sequencing data from single-cell colonies demonstrated variable levels of contamination by mouse DNA from the mouse feeder cell layer. Therefore, sequencing reads were filtered using the 'Xenome' algorithm which tests each read for mapping against the human and mouse reference genomes, and removes reads of clear mouse-origin, or that are ambiguous<sup>27</sup>. Remaining reads were aligned to the human reference genome and single-nucleotide variants (SNVs) and indels were called against an unmatched reference genome using the in-house pipelines CaVEMan and Pindel<sup>28,29</sup>. For all mutations passing quality filters in at least one sample, in-house software (cgpVAF, <https://github.com/cancerit/vafCorrect>) was used to produce matrices of variant and normal reads at each site for all HSPC colonies from that foetus.

Multiple post-hoc filtering steps were then applied to remove germline mutations, recurrent library prep and sequencing artefacts, and probable in vitro mutations, as detailed below:

- 1) A custom filter designed to remove artefacts specifically associated with the library prep process used for low input DNA concentrations (<https://github.com/MathijsSanders/SangerLCMFiltering>). Specifically, this includes those introduced by cruciform DNA structures.
- 2) A binomial filter was applied to aggregated counts of normal and variant reads across all samples. Sites with aggregated count distributions consistent with germline single nucleotide polymorphisms were filtered.
- 3) A beta-binomial filter was applied to retain only mutations whose count distributions across samples came from an over-dispersed beta-binomial distribution consistent with an acquired somatic mutation.
- 4) Mutations called at sites with abnormally high or low mean coverage were considered unreliable/ possible mapping artefacts and were filtered.
- 5) For each mutation call, normal and variant reads were aggregated from positive samples ( $\geq 2$  variant reads). Sites with counts inconsistent with a true somatic mutation were filtered.
- 6) Remaining mutations were only retained if there was at least one sample that met all minimum thresholds for variant read count ( $\geq 3$  for autosomes,  $\geq 2$  for XY

chromosomes), total depth ( $\geq 6$  for autosomes,  $\geq 4$  for XY chromosomes), and a binomial probability of  $>0.1$  that variant:normal counts reflect that expected for a genuine somatic mutation (0.5 for autosomes and 0.95 for XY chromosomes [not set to 1 to allow for artefacts]).

Filter threshold levels for filter steps 1-4 were set by using a combination of past experience, review of threshold parameter histograms, and experimentation. Mutational signatures and consistency with the phylogenetic tree were used for validation. In general, more stringent thresholds were used than in analyses on adult samples, due to the low number of anticipated true positives.

### **Phylogenetic tree construction**

For purposes of tree building, samples were assigned a genotype for each remaining mutation. All samples with  $\geq 2$  variant reads and a probability  $> 0.05$  that counts came from a distribution expected for a somatic mutation were genotyped as present. Samples with 0 variant reads and a depth of  $\geq 6$  at a given site were genotyped as absent. Samples not meeting either of these criteria were genotyped as 'unknown'.

A genotype matrix of shared mutations (a positive genotype  $\geq 1$  sample [i.e. that would be informative for tree-building]) was fed into the MPBoot program<sup>30</sup>. This can incorporate 'unknown' genotypes, and constructs a maximum parsimony phylogenetic tree with bootstrap approximation.

### **Mutation assignment back to the phylogenetic tree**

All mutations retained post-filtering were then assigned back to individual branches on the phylogenetic tree using the in-house developed R package 'treemut' (<https://github.com/NickWilliamsSanger/treemut>). This goes back to the original count data, and uses a maximum-likelihood approach to assign each mutation to a specific branch. Edge lengths are then made proportional to the number of mutations on the branch.

### **Inference of the mutation rate after the 8-cell-stage embryo**

After the third round of cell division (corresponding to the 8-cell-stage embryo), the 8 pcw HSPC phylogeny shows a profusion of high degree polytomies. Specifically, of the five blood antecedents from the 8-cell stage for which we have significant numbers of descendants, there are 12-, 10-, two 5-degree polytomies and a single dichotomy. The true degree of these polytomies may be underestimated if descendant lineages are not captured in the HSPC phylogeny. We ran 1000 simulations of 5 independent cells dividing and acquiring mutations

randomly according to a Poisson distribution with lambda corresponding to a specific average mutation rate, and recorded the proportion of simulations that resulted in polytomies at their roots of *at least* the degrees observed. We repeated these simulations for average mutation rates ranging from 0.5 to 1.5 mutations per cell division. These simulations showed that once the mutation rate rose to 0.9 mutations per cell division,  $\leq 5\%$  of simulations gave results consistent with the data, and once the mutation rate was 1.0, this proportion dropped to  $\leq 1\%$ .

### **Targeted sequencing**

Individual custom DNA capture panels were designed for each foetus according to the manufacturer's guidelines (SureSelect<sup>XT</sup> Custom DNA Target Enrichment Probes, Agilent). For the 8 pcw foetus a 0.85 Mb panel was designed to capture regions of all SNVs and indels passing the previously described filtering steps. For the 18 pcw foetus a 0.5 Mb bait set was designed to capture SNVs only, and included all shared SNVs and around half of private SNVs. No repeat-masking was used in the design process, and hence on-target read percentage was low (<10% for both panels).

### **Validation of somatic mutations called in haematopoietic colonies**

DNA libraries from the WGS-sequenced 8 pcw and 18 pcw HSPC colonies were combined into pools of 24-32 samples, hybridized to the appropriate capture panel, multiplexed on flow cells, and subjected to paired-end sequencing (150-bp reads) on the HiSeq4000 machine (Illumina). BWA-MEM was used to align sequences to the human reference genome (NCBI build37). cgpVAF was used to compile variant read counts and total read depth from all included mutant loci. For each mutation the total variant read and total depth counts were aggregated for all colonies expected to carry the mutation, based on the phylogeny structure.

For the 18 pcw foetus, 2137 private SNVs were included in the bait set. Of these, 1770 achieved a depth of  $\geq 8\times$  in the colony in which they were originally called (the "minimum depth" set); 273 achieved a depth of  $\geq 40\times$  in the colony in which they were originally called (the "high depth" set). The read counts from each set were used as input to a binomial mixture model to ascertain the proportion of mutations that were clonal (and therefore likely to be true somatic mutations), sub-clonal (and therefore *in vitro* mutations), or absent. The binomial mixture model uses an expectation maximisation algorithm to calculate the best combination of between 1 and 4 binomial distributions to fit the observed data. For each individual mutation it assigned the most likely binomial distribution for it to be derived from i.e. was it most likely clonal or sub-clonal. Similar results were obtained with both the "minimum depth" and "high depth" mutation sets (Extended Data Fig. 3).

For the 8 pcw foetus, 3274 private SNVs were included in the bait set. Of these, 1431 achieved a depth of  $\geq 8\times$  (minimum depth set) in the colony in which they were originally called, and 34 a depth  $\geq 40\times$  (high depth set). Otherwise analysis proceeded similarly as per the 18 pcw HSPC colonies.

Indels were included in the 8 pcw capture panel only. There were 246 indels in total: 196 private and 50 shared. Of these, 73 private and 33 shared indels on autosomes achieved a total depth of  $\geq 8\times$  in colonies in which they were called. Applying the binomial mixture model to these read counts showed that counts for both private and shared mutations were consistent with a single binomial distribution with probabilities of 0.49 and 0.475 respectively.

### **Detection of somatic mutations discovered in HSPC colonies in non-haematopoietic tissues**

DNA libraries from LCM cuts of non-haematopoietic tissues were combined into pools of 16 samples, hybridized to the appropriate capture panel, multiplexed on flow cells, and subjected to paired-end sequencing (150-bp reads) on both the HiSeq4000 and NovaSeq 6000 machines (Illumina). For the 68 DNA libraries from the 8 pcw foetus LCM cuts, one HiSeq4000 lane and two NovaSeq 6000 lanes were used. For the 48 DNA libraries from the 18 pcw foetus, two HiSeq4000 lanes were used. BWA-MEM was used to align sequences to the human reference genome (NCBI build37). cgpVAF was used to compile variant read counts and total read depth from all included mutant loci.

Mutation calling was performed using a depth-aware naive Bayesian classifier comparing the two competing models:

$H_0$ : Observed VAF is consistent with sequencing error; versus

$H_A$ : Observed VAF is consistent with the mutation being truly present in the sample.

The data comprised counts of total depth,  $n_{i,j}$ , and number of reads,  $x_{i,j}$ , for the  $i$ th mutation in the  $j$ th sample. We estimated the sequencing error rate for the  $i$ th mutation as coming from a  $BetaBinomial(\alpha_E, \beta_E, n_{i,j})$  distribution, using deep sequencing data from colonies that were known to be negative for the mutation by their position on the phylogenetic tree. Estimates were derived using the method of moments, empirically weighted for differences in sequence depth across colonies as described by Kleinman<sup>31</sup>. The data for a true mutation were assumed to be drawn from a  $BetaBinomial(\alpha_M, \beta_M, n_{i,j})$  distribution with  $\alpha_M = 1.5$ ;  $\beta_M = 10$ . The prior probability of a mutation being present in a tissue,  $\pi_{i,j}$ , was set as the proportion of HSPC colonies from the WGS phylogeny harbouring that mutation. Using this and the usual beta-binomial probability density function, then, the posterior probability for  $H_A$  is:

$$P(H_A | x_{i,j}; n_{i,j}) = \pi_{i,j} P(x_{i,j} | \alpha_M, \beta_M, n_{i,j}) / (\pi_{i,j} P(x_{i,j} | \alpha_M, \beta_M, n_{i,j}) + (1 - \pi_{i,j}) P(x_{i,j} | \alpha_E, \beta_E, n_{i,j}))$$

For aggregated tissues, the posterior probability of the mutation being present in the tissue as a whole was calculated as one minus the product of the probabilities of the mutation being absent in each individual microbiopsy. The knowledge of the phylogeny structure was then used to clean up calls that did not concord with mutation calls from neighbouring branches. Mutations that were called as present from distal branches when their ancestral branches were absent were removed; mutations which had been called as absent, but had confident mutation calls from both distal and proximal branches (or from the same branch) had their posterior probabilities elevated. This latter scenario usually occurred for mutations where the depth was unusually low and no positive reads had been found in a given tissue. Trees for figures 3 and 4 were then plotted with the VAF of individual mutations called as present overlayed on the underlying HSPC phylogeny.

### **Inference of numbers of lineages at different developmental stages**

The presence of mutations from the HSPC phylogeny in non-haematopoietic tissues demonstrates that the mutation was acquired in a blood antecedent cell that still had the capacity to differentiate into both blood and the non-haematopoietic tissue in question. By considering a phylogeny including only those branches shared with a specific non-haematopoietic tissue you can visualize the phylogeny of these multipotent cells. The number of descendant non-shared branches stemming directly from this shared phylogeny is a direct read of the minimum number of more committed HSPC antecedents arising from these multipotent antecedents.

We performed this analysis considering the shared phylogenies of non-haematopoietic tissues that developmentally diverge from blood at different points:

- (1) the trophoblast-shared phylogeny: the inner cell mass (ICM) /trophoblast divergence,
- (2) the mesenchymal core-shared phylogeny: the epiblast/ hypoblast divergence,
- (3) the endoderm- or ectoderm-shared phylogeny: the mesoderm /non-mesoderm divergence,
- (4) the kidney- or limb-shared phylogeny: the lateral-plate/ non-lateral-plate mesoderm divergence,
- (5) the heart-shared phylogeny: the haematopoietic lateral-plate mesoderm/ non-haematopoietic lateral-plate mesoderm divergence.

For the early lineage commitment events (ICM commitment and epiblast commitment) our HSPC phylogeny captures the vast majority of antecedents from these time points, and therefore the estimates are likely close approximations of the true values. For the later

commitment events, the phylogeny only captures a small fraction of antecedents, and therefore the numbers represent lower bounds on the true value.

### Clustering of targeted sequencing results from non-haematopoietic tissues

To evaluate the similarity of the pattern of shared mutations and their VAFs in the different non-haematopoietic tissues we used the soft cosine similarity statistic, a measure used extensively in natural language processing. In this measure, the similarity between two sets of observations is defined as:

$$\text{soft cosine}(a, b) = \frac{\sum_{i,j}^N s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^N s_{ij} a_i a_j} \sqrt{\sum_{i,j}^N s_{ij} b_i b_j}}$$

Here the  $i$  and  $j$  represent an iteration over all mutations; the  $a_i$  and  $b_j$  represent the natural log of the VAF of mutation  $i$  in tissue  $a$  versus mutation  $j$  in tissue  $b$  respectively; the  $s_{ij}$  represents a similarity score for mutation  $i$  and  $j$  which for our analysis we considered to be the inverse of the  $1 +$  the minimum total branch distance between the two mutations.

The euclidean distance between tissues was then calculated, and the tissues clustered by hierarchical clustering using the *hclust* function from the “stats” package within R. The calculated dendrogram is displayed alongside the heatmap in Figure 3d.

### Measure of lineages lost up to the 4th generation from the zygote

In the 8pcw foetus, we attempted to assign individual mutations to the round of cell division during which they originated. For the first two generations from the zygote (up to the 4-cell stage embryo) it was straight-forward to directly infer this with some confidence because (1) divisions are invariably associated with mutations due to the relatively high mutation rate i.e. they are dichotomies, and (2) the phylogeny captured many descendants from each of the 4-cell stage antecedents allowing each branch to be resolved. The fact that the VAFs across these branches sum to 0.5 across both embryonic and extra-embryonic tissues adds confidence that no lineages have been missed. For the 3rd generation (i.e. the 8-cell stage embryo) this was no longer true, as some of the 8-cell-stage antecedents contribute only to trophoblast. However, targeted sequencing data from matched trophoblast allowed an apparent single branch to be resolved into two separate branches resulting from 2nd and 3rd generation divisions. From the 4th generation, there are two major difficulties in assigning mutations to generations: (1) due to the drop in mutation rate, many divisions occur without the acquisition of new mutations making it impossible to know which of the polytomous branches is a result of which generation (2) we have fewer descendants from each lineage, meaning that each daughter branch may have mutations resulting from many successive cell

divisions. For the purposes of figure 3e, we considered all daughter branches of 3rd generation antecedents as potential 4th generation divisions, although in reality they are a mixture of 4th and higher divisions. To resolve branches that had mutations from multiple successive cell divisions we again used the targeted sequencing data: a binomial mixture model was applied to aggregated read counts from all non-haematopoietic tissues, analysing mutations from each branch in turn. Branches with read counts best explained by more than one binomial distribution (i.e. containing mutations present in different cell fractions) were annotated such that only mutations from the highest probability binomial distribution were considered as 4th generation mutations.

Once mutations were assigned to each generation, the total cell fraction captured by our HSPC phylogeny for mutations acquired in the 1st - 4th cell divisions was calculated by summing the cell fractions for each branch in a given generation. The uncertainty in these cell fractions was calculated by bootstrapping the read counts from each branch 1000 times, repeating the sum of cell fractions across branches for each bootstrap, and obtaining the 0.025 and 0.975 quantiles from the resulting distribution for a 95% percentile interval.

### **Construction of the phylogenies with branch lengths scaled by generations**

As described above, assignment of branches to specific generations is relatively straightforward for the first three generations. However, this becomes more challenging once there are frequent polytomies and fewer descendant cells from each branch.

To create an approximation of a “generation” tree where branches are placed along the y axis according to the generation in which the mutations arose we did the following:

- (1) Branches were manually assigned to generations as previously for the first 3 generations
- (2) Branches from polytomies were extended according to  $\log_2$  of the number of descendants e.g. for an 8-degree polytomy, each branch would be extended to length 3, the average number of generations. In reality, in this example, individual branches may result from 1 to 7 generations from the antecedent, depending on the true phylogeny structure underlying the polytomy.
- (3) Terminal branches were extended to make the tree ultrametric.

For the analysis of the lineages represented through different generations (Figure 4a-b), single branches from polytomies were therefore considered as belonging to multiple generations. This approach means that the absolute proportions of represented lineages by generation are inaccurate as different assumptions will falsely increase/ decrease these proportions. However, for comparison between tissues it remains a useful measure.

### **Analysis of mutation distribution in the genome**

To evaluate whether the mutation distribution across different genomic features was comparable to that expected by chance we compared our data to a simulation of random mutagenesis within the human genome.

In this simulation the *bedtools random* command was used to generate 160,000 random sites within the human genome. From these, 5000 were subsampled such that the 32 possible trinucleotide contexts were represented in the same proportions as in the data. The central pyrimidine within each trinucleotide was then virtually “mutated” to an alternative base, again according to probabilities matching the mutational signatures of the data to generate a vcf file of 5000 mutational-signature matched mutations. The resulting genomic locations were annotated using the VAGrENT algorithm (<https://github.com/cancerit/VAGrENT>) whereby each mutation is assigned a Sequence Ontology term to classify its consequence.

This simulation was repeated 500 times and the distribution of the resulting proportions of mutations in different Sequence Ontology categories was compared to those observed in the true mutation sets.

### **Analysis of molecular variance to look for phylogenetic clustering by anatomic location**

Analysis of molecular variance (AMOVA) was used to test for phylogenetic clustering by anatomical location in the 18 pcw foetus. We compared HSPCs that had been isolated from the liver, femur 1, and femur 2.

The phylogeny was first made ultrametric using a bespoke method that gives more weight to the lengths of early branches. The mutational distance between any two HSPC colonies was then calculated (this is the minimum mutational “walk” to get from one sample to the other) for all possible colony-colony pairs. The sum of squared distances for “within location” and “between location” pairs was calculated, divided by their degrees of freedom, and this was used to obtain an estimate of the proportion of variation explained by differences in cell locations (the F statistic).

P values were obtained by randomly re-labelling the cell locations 30,000 times and calculating the same statistic. The given p value is therefore a proportion of the random re-assignments that had a more extreme F statistic than that observed in the data.

### **Code availability**

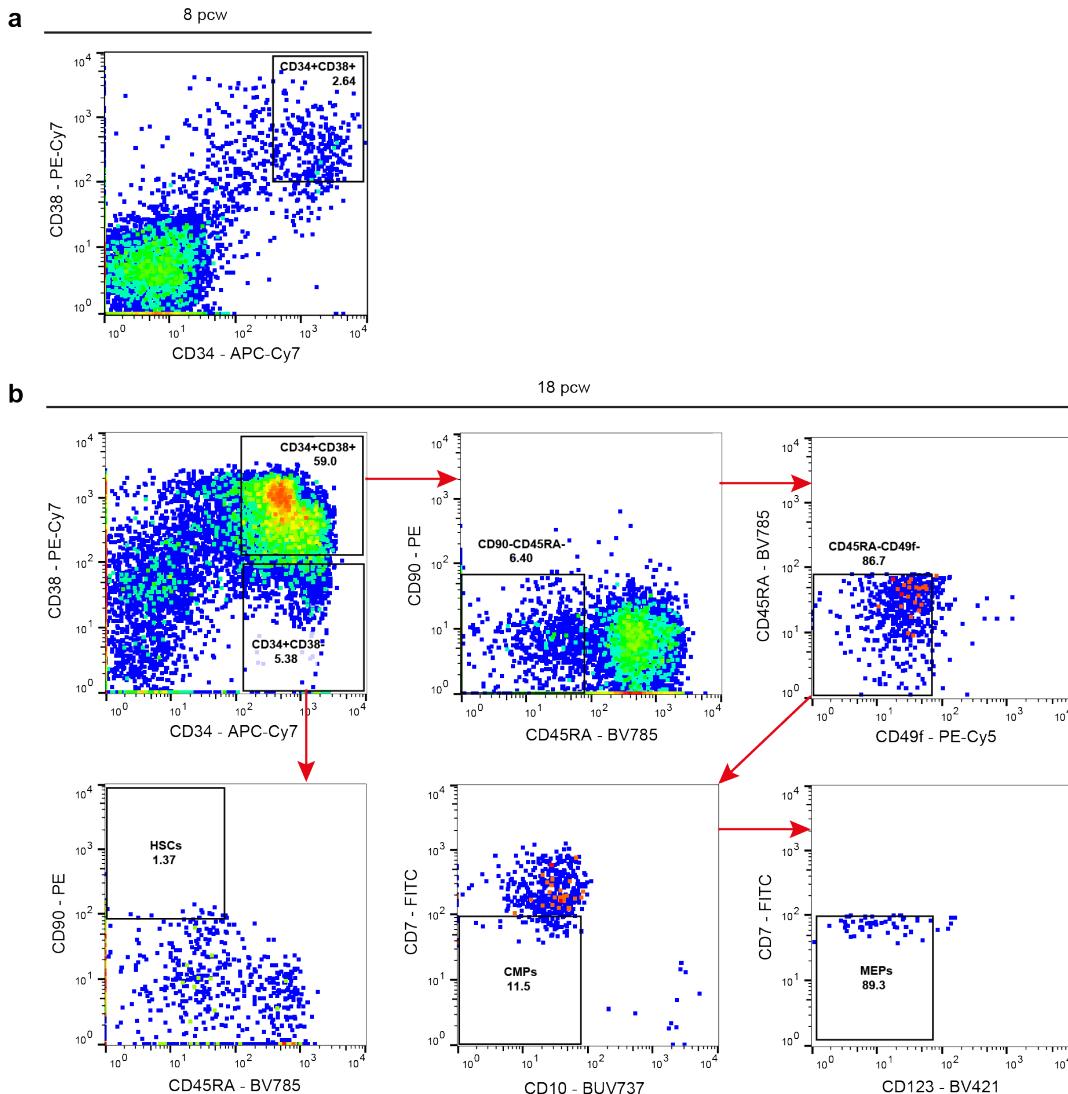
All scripts and some derived data sets are available at [https://github.com/mspencerchapman/Phylogeny\\_of\\_foetal\\_haematopoiesis](https://github.com/mspencerchapman/Phylogeny_of_foetal_haematopoiesis).

## **Data availability**

Whole genomes and targeted sequencing data will be available via European Genome-Phenotype Archive (EGA, <https://www.ebi.ac.uk/ega/>). Accession numbers not yet assigned.

1 **Extended Data**

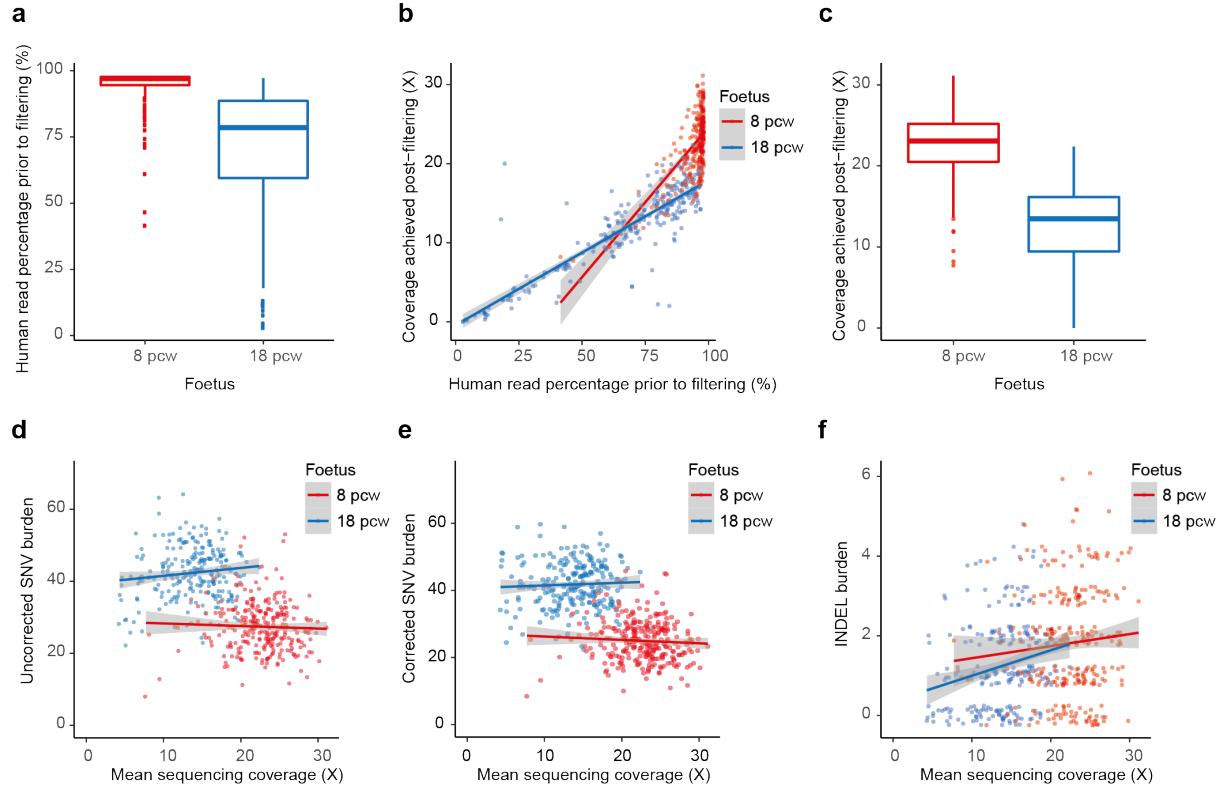
2



3

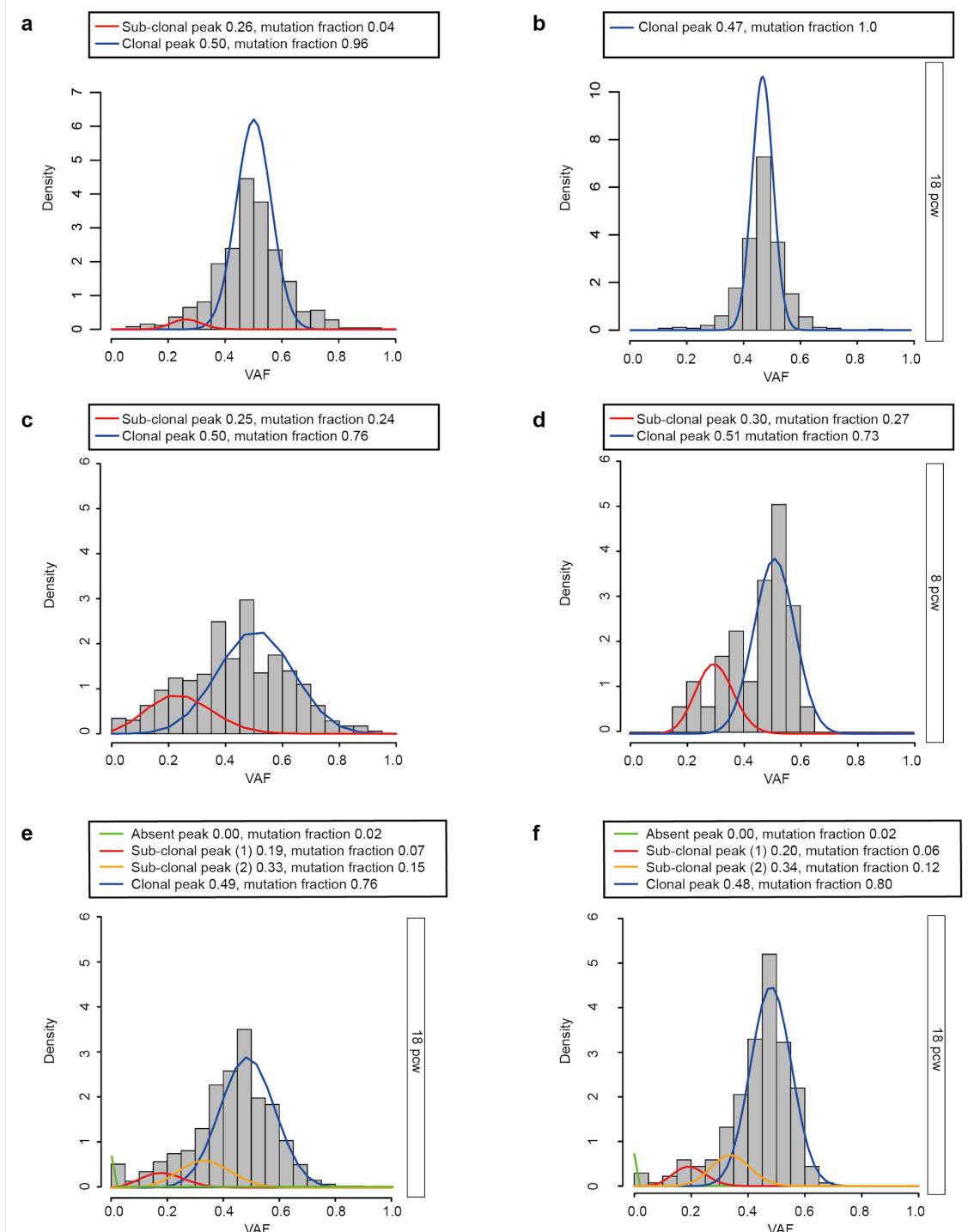
4

5 **Extended Data Fig. 1 | FACS sorting strategy for haematopoietic progenitor cells from**  
6 **foetal liver and bone marrow. A, Following the exclusion of debris and cell doublets by**  
7 **gating, anti-CD34 and anti-CD38 staining was used to sort single haematopoietic progenitor**  
8 **cells from the liver of a 8 pcw foetus. B, Sorting strategy for different stem and haematopoietic**  
9 **progenitor populations from matched liver and two femurs of a 18 pcw foetus. Following**  
10 **exclusion of debris and cell doublets, anti-CD34, anti-CD38, anti-CD90, anti-CD45RA, anti-**  
11 **CD49f, anti-CD7, anti-CD10 and anti-CD123 staining was used to sort haematopoietic**  
12 **progenitor populations. HSCs = haematopoietic stem cells, CMPs = common myeloid**  
13 **progenitors and MEPs = megakaryocyte-erythroid progenitors. n = 20,000 cells.**



14  
15 **Extended Data Fig. 2 | Sample contamination and sequencing coverage.** **a**, Box plot  
16 showing the percentage of human sequencing reads, prior to the exclusion of contaminating  
17 mouse reads from the feeder layer. **b**, Dot plot showing, for each colony of the two foetuses,  
18 the final sequencing coverage, post-exclusion of mouse reads, against the percentage of  
19 human reads. **c**, Box plot showing final sample coverage for the two foetuses. **d**, Dot plot  
20 showing the uncorrected SNV burden per colony against sample coverage. **e**, Dot plot  
21 showing the corrected SNV burden per colony against sequencing coverage. **f**, Dot plot  
22 showing the uncorrected indel burden per colony against sequencing coverage. SNVs = single  
23 nucleotide variations.

24

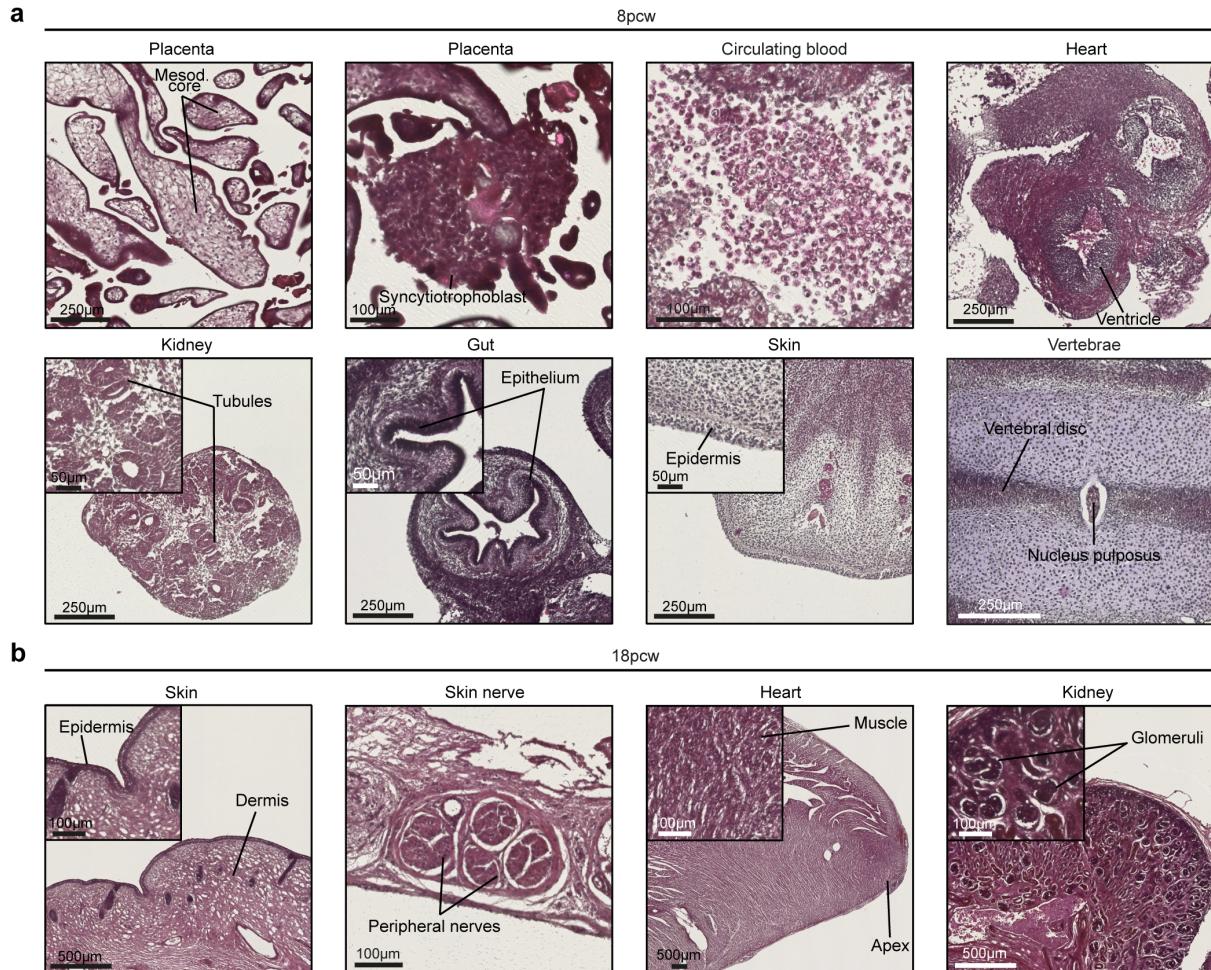


25

26

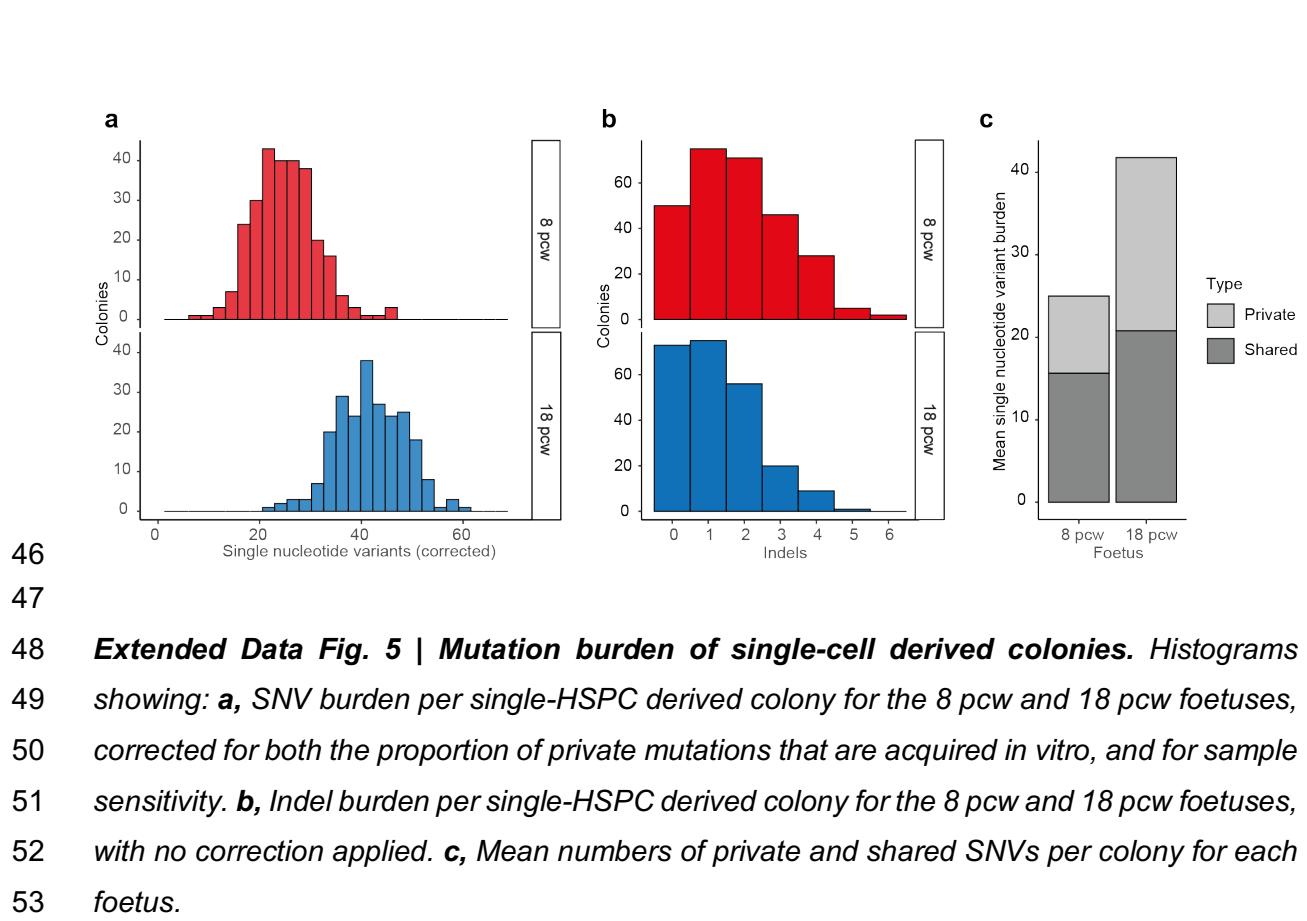
27 **Extended Data Fig. 3 | Mutation validation.** Histograms showing variant allele fraction (VAF)  
 28 of **a**, shared mutations with targeted sequencing depth  $\geq 8\times$  in the 8 pcw foetus, **b**, shared  
 29 mutations with targeted sequencing depth  $\geq 8\times$  in the 18 pcw foetus, **c**, private mutations with  
 30 targeted sequencing depth  $\geq 8\times$  in the 8 pcw foetus, **d**, private mutations with a high read depth  
 31  $\geq 40\times$  in the 8 pcw foetus, **e**, private mutations with depth  $\geq 8\times$  in the 18 pcw foetus, **f**, private  
 32 mutations with a high read depth  $\geq 40\times$  in the 18 pcw foetus. Almost all shared mutations fitted  
 33 a single clonal distribution. Private mutation distributions were best explained by a mixture of

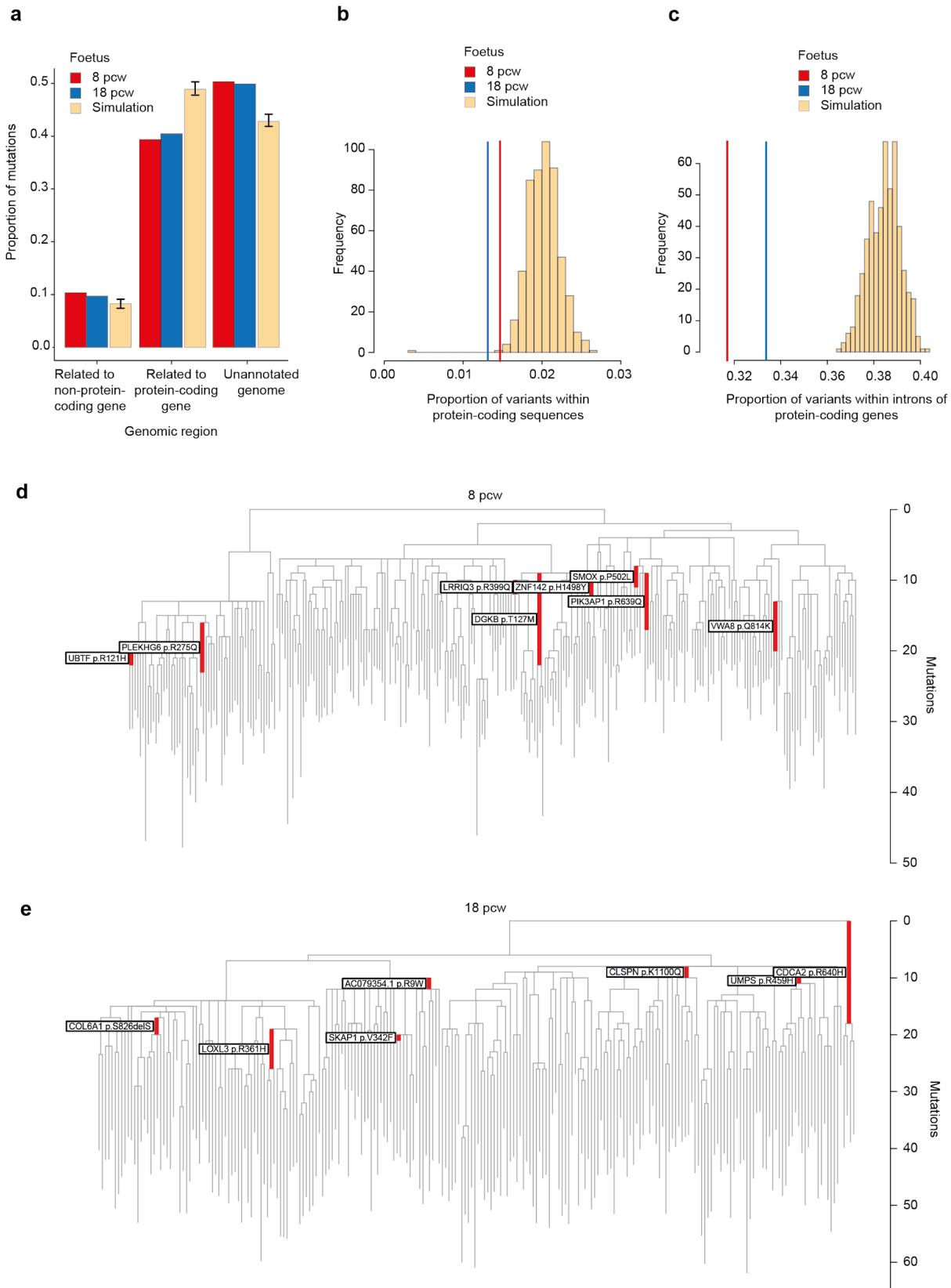
34 *clonal (blue line, probability ~0.50) and one or more sub-clonal binomial distributions (red and*  
35 *yellow lines).*



36  
37

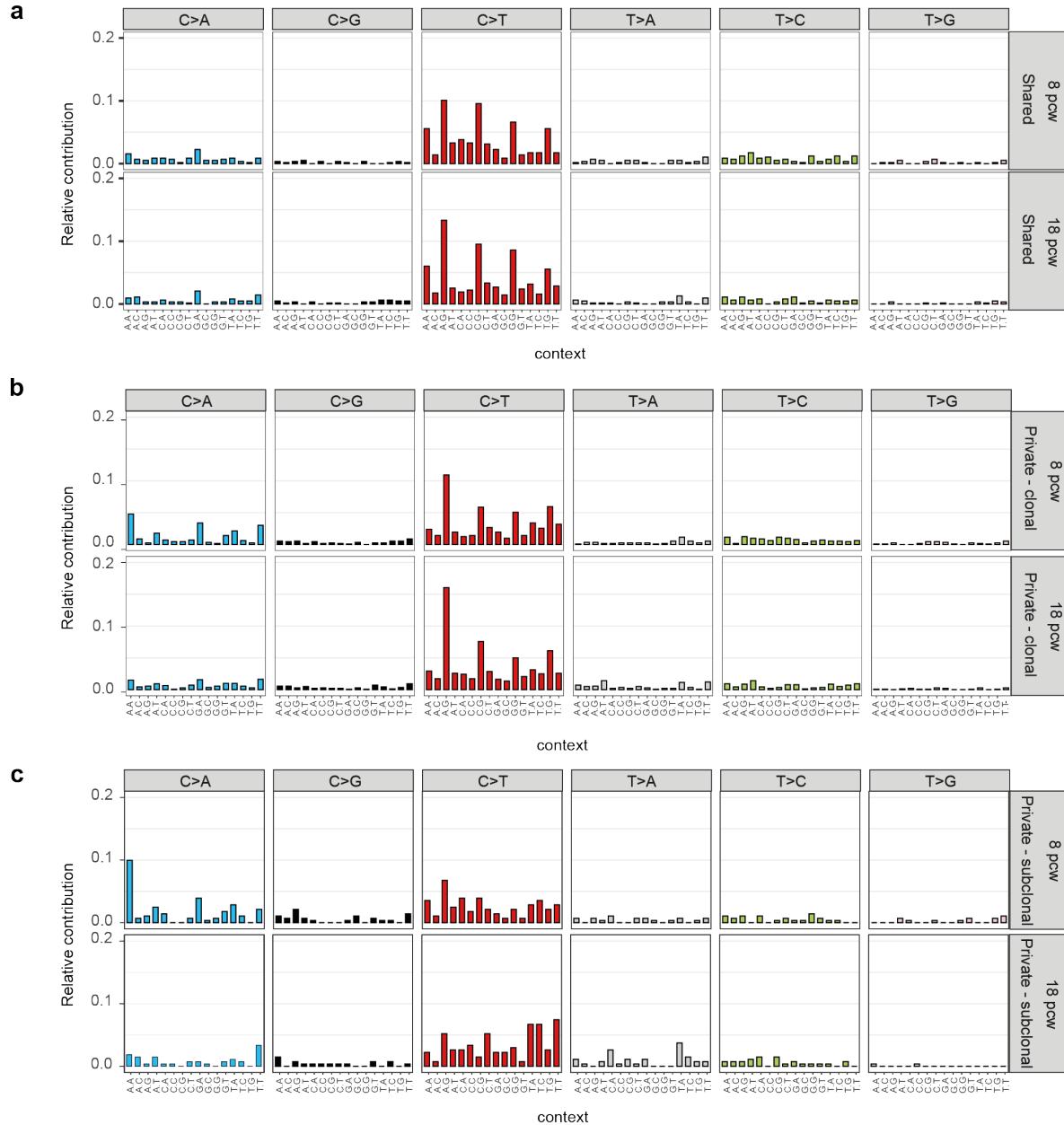
**Extended Data Fig. 4 | Laser-capture microdissected tissues.** Representative histological slides of tissue structures with different developmental origins. Slides were stained with haematoxylin and eosin prior to microdissection. **a**, Structures microdissected from the 8 pcw sample included: mesodermal core and syncytiotrophoblast from the placenta, blood circulating in the heart, muscle from the heart, tubules from the kidney, epithelium from the gut, epidermis from the skin, vertebral disc from the vertebrae. **b**, Structures microdissected from the 18 pcw sample included: epidermis and peripheral nerves from the skin, muscle from the heart and glomeruli from the kidney.





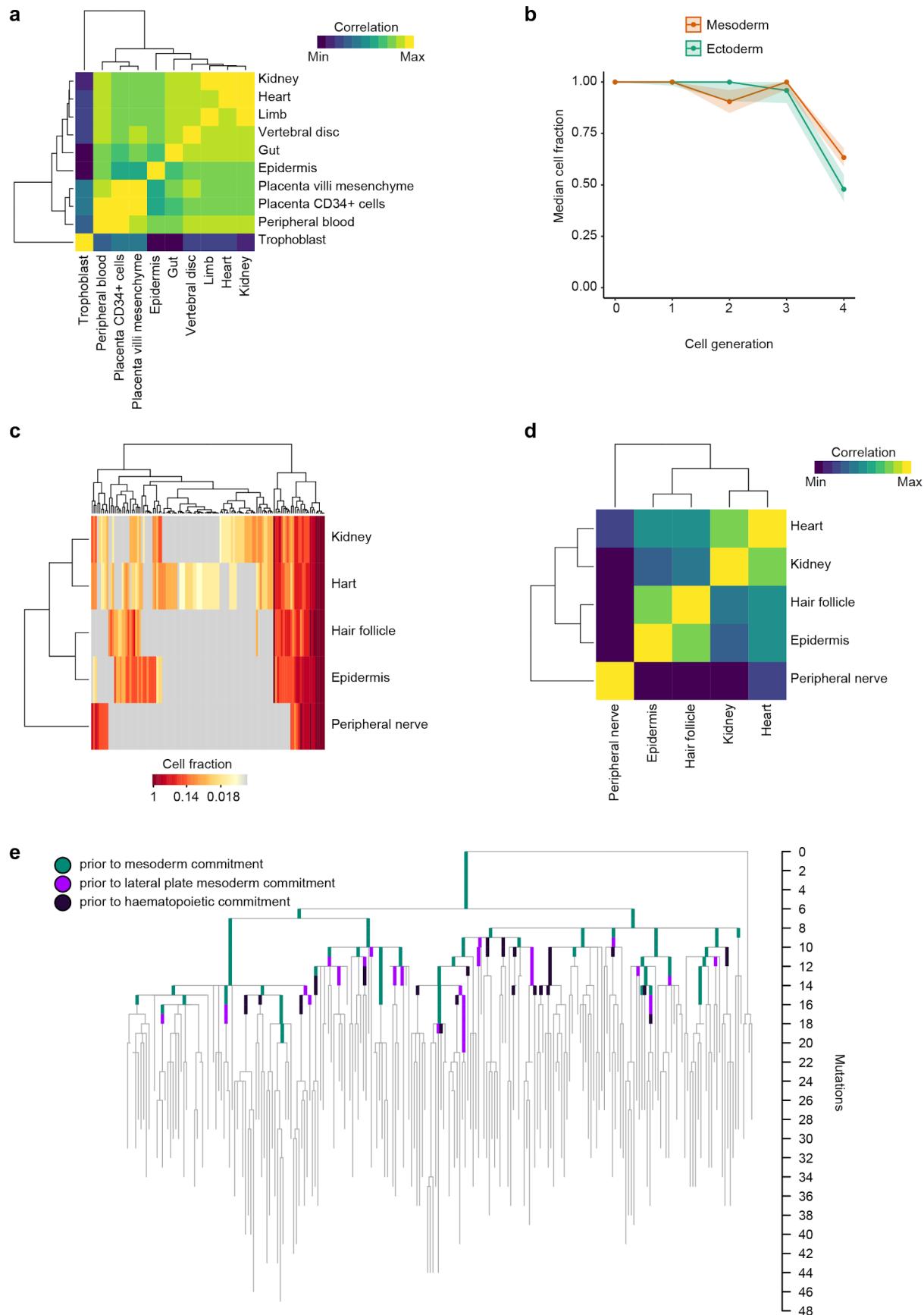
54  
55  
56 **Extended Data Fig. 6 Genomic distribution of identified somatic mutations.** **a**, Bar plot  
57 showing the proportion of variants mapping to different genomic features. **b**, Plot showing the

58 *proportion of variants within protein-coding sequences. The histograms show the distribution*  
59 *obtained by 500 simulations of random acquisition of mutations across the genome. c, Plot*  
60 *showing the proportion of variants within introns of protein-coding genes. The histograms*  
61 *show the distribution obtained by 500 simulations of random acquisition of mutations across*  
62 *the genome. d, Mutations in protein-coding sequences in the 18 pcw foetus, mapped on the*  
63 *branch of acquisition. Only those found in  $\geq 2$  colonies are shown. e, Mutations in protein-*  
64 *coding sequences in the 8 pcw foetus, mapped on the branch of acquisition. Only those found*  
65 *in  $\geq 2$  colonies are shown.*

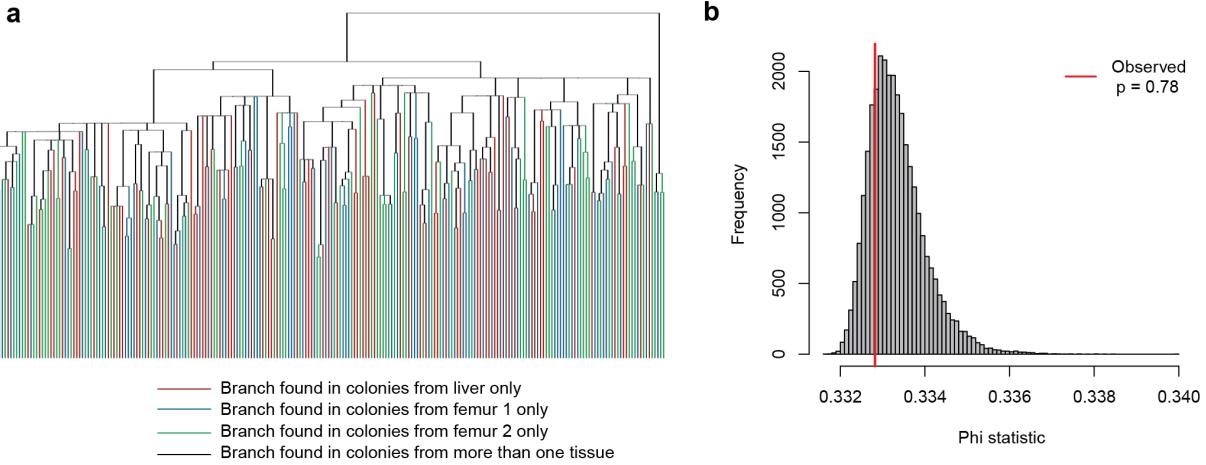


66

67 **Extended Data Fig. 7 | Mutational signatures.** Mutational signatures incorporating the  
 68 trinucleotide context of **a**, Shared mutations. **b**, Private mutations assigned to the clonal peak  
 69 in the binomial mixture model, i.e. likely *in vivo* acquired mutations. **c**, Private mutations  
 70 assigned to sub-clonal peaks in the binomial mixture, i.e. likely *in vitro* acquired mutations.



73 **Extended Data Fig. 8 | a**, Heatmap representation of the targeted sequencing data from the  
74 8 pcw foetus. Each column and each row represents a single tissue. The colour shows the  
75 level of correlation. **b**, Line plot showing lineage loss of microdissected tissues in the 18 pcw  
76 foetus at different times (represented as cell generations from the zygote). Shaded areas  
77 represent confidence intervals. **c**, Heatmap representation of the targeted sequencing data  
78 from the 18 pcw foetus. Each column represents an individual mutation, and each row a single  
79 tissue. The colour shows the VAF of the mutation in that tissue with grey indicating that it was  
80 not detected. **d**, Heatmap representation of the targeted sequencing data from the 18 pcw  
81 foetus. Each column and each row represents a single tissue. The colour shows the level of  
82 correlation. **e**, Phylogenetic tree showing the clonal relationships of 18 pcw HSPCs. Mutations  
83 identified in HSPCs by WGS that were also detected in non-haematopoietic tissues are  
84 coloured on the tree according to the earliest diverging tissue in which the mutation was  
85 reliably detected. Branch lengths are proportional to the number of mutations accumulated.  
86 The tree only contains those SNVs included in the bait set for targeted sequencing. ICM =  
87 inner cell mass. In all heatmaps of this figure the tissues are clustered using soft cosine  
88 similarity (see methods).



89

90 **Extended Data Fig. 9 | High level of intermixing of 18 pcw HSPCs among fetal liver and**  
91 **two sites of bone marrow. a, Ultrametric phylogenetic tree of 18 pcw HSPCs with branches**  
92 *coloured by the tissue from which HSPCs bearing the mutations were isolated. Black branches*  
93 *indicate that cells were found in more than one tissue, coloured branches indicate cells were*  
94 *unique to one specific tissue. b, Analysis of molecular variance used to formally test for*  
95 *clustering on the phylogeny of HSPCs isolated from the same tissue. The histogram shows*  
96 *the null distribution used to detect clustering. Distributions were obtained by randomly*  
97 *permuting which cells were assigned to which tissue. The observed value of the phi statistic*  
98 *is shown as a red line, with the p-value indicating no statistically significant clustering by tissue.*

99

Antibody	Clone	Manufacturer
CD3 AF700	OKT3	Biolegend
CD8 AF700	SK1	Biolegend
CD11b AF700	ICRF44	Biolegend
CD14 AF700	61D3	Thermo Fisher
CD19 AF700	HIB19	Biolegend
CD56 AF700	B159	BD Biosciences
CD34 APC-Cy7	581	Biolegend
CD38 PE-Cy7	HB7	BD Biosciences
CD45RA BV785	HI100	Biolegend
CD90 PE	5E10	BD Biosciences
CD49f PE-Cy5	GoH3	BD Biosciences
CD10 BUV737	HI10a	BD Biosciences
CD7 FITC	MT701	BD Biosciences
CD123 BV421	9F5	BD Biosciences

100

101 **Extended Data Table 1 | Antibodies list.** Antibodies used for the study, including clone and  
102 manufacturer.

103

104

Tissue (8 pcw foetus)	Number of biopsies undergoing sequencing	Mean aggregate depth across biopsies (x)
Epidermis	6	314
Gut epithelium	12	841
Heart	9	449
Intervertebral disc	4	184
Kidney	10	662
Limb	6	390
Mesenchymal core of placenta	5	152
Peripheral blood	5	268
Trophoblast	10	652
Bulk FACS-sorted placental HSPCs	1	78

105

Tissue (18 pcw foetus)	Number of biopsies undergoing sequencing	Mean aggregate depth across biopsies (x)
Epidermis	10	551
Hair follicle	4	239
Heart	10	1246
Kidney	13	701
Peripheral nerve	1	32

106

107 **Extended Data Table 2 | Laser capture microdissection biopsies.** The number of biopsies  
108 undergoing targeted sequencing from each tissue for each foetus. The mean aggregated  
109 depth is the mean depth across all sites included in the custom capture panel after aggregating  
110 reads across biopsies from a given tissue.

111 **References**

112

113 1. OpenStax. *Anatomy and Physiology*. (2016).

114 2. Luckett, W. P. Origin and differentiation of the yolk sac and extraembryonic mesoderm  
115 in presomite human and rhesus monkey embryos. *Am. J. Anat.* **152**, 59–97 (1978).

116 3. Palis, J. & Yoder, M. C. Yolk-sac hematopoiesis: the first blood cells of mouse and man.  
117 *Exp. Hematol.* **29**, 927–936 (2001).

118 4. Ivanovs, A. *et al.* Human haematopoietic stem cell development: from the embryo to the  
119 dish. *Development* **144**, 2323–2337 (2017).

120 5. Silver, L. & Palis, J. Initiation of murine embryonic erythropoiesis: a spatial analysis.  
121 *Blood* **89**, 1154–1164 (1997).

122 6. Arnold, S. J. & Robertson, E. J. Making a commitment: cell lineage allocation and axis  
123 patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10**, 91–103 (2009).

124 7. Rossant, J. & Tam, P. P. L. New Insights into Early Human Development: Lessons for  
125 Stem Cell Derivation and Differentiation. *Cell Stem Cell* **20**, 18–28 (2017).

126 8. Enders, A. C. & King, B. F. Formation and differentiation of extraembryonic mesoderm  
127 in the rhesus monkey. *Am. J. Anat.* **181**, 327–340 (1988).

128 9. Kelemen, E., Calvo, W. & Fliedner, T. M. Atlas of Human Hemopoietic Development.  
129 (1979) doi:10.1007/978-3-662-11193-2.

130 10. Charbord, P., Tavian, M., Humeau, L. & Péault, B. Early ontogeny of the human marrow  
131 from long bones: an immunohistochemical study of hematopoiesis and its  
132 microenvironment. *Blood* **87**, 4109–4119 (1996).

133 11. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during  
134 life. *Nature* **538**, 260–264 (2016).

135 12. Baron, C. S. & van Oudenaarden, A. Unravelling cellular relationships during  
136 development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.*  
137 **20**, 753–765 (2019).

138 13. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic

139 mutations. *Nature* **561**, 473–478 (2018).

140 14. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian  
141 genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).

142 15. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma.  
143 *Nature* **471**, 467–472 (2011).

144 16. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on  
145 regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

146 17. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat.*  
147 *Genet.* **47**, 1402–1407 (2015).

148 18. Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early  
149 human embryo. *Nature* **543**, 714–718 (2017).

150 19. Ye, A. Y. *et al.* A model for postzygotic mosaicism quantifies the allele fraction drift,  
151 mutation rate, and contribution to de novo mutations. *Genome Res.* **28**, 943–951  
152 (2018).

153 20. Schulz, K. N. & Harrison, M. M. Mechanisms regulating zygotic genome activation. *Nat.*  
154 *Rev. Genet.* **20**, 221–234 (2019).

155 21. Molè, M. A., Weberling, A. & Zernicka-Goetz, M. Comparative analysis of human and  
156 mouse development: From zygote to pre-gastrulation. *Curr. Top. Dev. Biol.* **136**, 113–  
157 138 (2020).

158 22. Xiang, L. *et al.* A developmental landscape of 3D-cultured human pre-gastrulation  
159 embryos. *Nature* **577**, 537–542 (2020).

160 23. Kuruppumullage Don, P., Ananda, G., Chiaromonte, F. & Makova, K. D. Segmenting the  
161 human genome based on states of neutral genetic divergence. *Proc. Natl. Acad. Sci. U.*  
162 *S. A.* **110**, 14699–14704 (2013).

163 24. Wu, J. *et al.* Chromatin analysis in human early development reveals epigenetic  
164 transition during ZGA. *Nature* **557**, 256–260 (2018).

165 25. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous  
166 process. *Nat. Cell Biol.* **19**, 271–281 (2017).

167 26. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells.  
168 *Nature* **574**, 532–537 (2019).

169 27. Conway, T. *et al.* Xenome--a tool for classifying reads from xenograft samples.  
170 *Bioinformatics* **28**, i172–8 (2012).

171 28. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion  
172 Events from Paired End Sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12  
173 (2015).

174 29. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to  
175 Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinformatics*  
176 **56**, 15.10.1–15.10.18 (2016).

177 30. Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and  
178 bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).

179 31. Kleinman, J. C. Proportions with Extraneous Variance: Single and Independent Sample.  
180 *Journal of the American Statistical Association* vol. 68 46 (1973).