Article

Clonal dynamics in early human embryogenesis inferred from somatic mutation

https://doi.org/10.1038/s41586-021-03786-8

Received: 22 November 2020

Accepted: 29 June 2021

Published online: 25 August 2021

Check for updates

Seongyeol Park[1,2,10], Nanda Maya Mali[3,10], Ryul Kim[1,10], Jeong-Woo Choi[3,4], Junehawk Lee[5], Joonoh Lim[1], Jung Min Park[3,4], Jung Woo Park[5], Donghyun Kim[3,4], Taewoo Kim[1], Kijong Yi[1], June Hyug Choi[3], Seong Gyu Kwon[3], Joo Hee Hong[3], Jeonghwan Youk[1], Yohan An[1], Su Yeon Kim[1], Soo A Oh[1], Youngoh Kwon[2], Dongwan Hong[6], Moonkyu Kim[7], Dong Sun Kim[3], Ji Young Park[8], Ji Won Oh[3,4,9,11] ✉ & Young Seok Ju[1,2,11] ✉

Cellular dynamics and fate decision in early human embryogenesis remain largely unknown owing to the challenges of performing studies in human embryos[1]. Here, we explored whole-genomes of 334 single-cell colonies and targeted deep sequences of 379 bulk tissues obtained from various anatomical locations of seven recently deceased adult human donors. Using somatic mutations as an intrinsic barcode, we reconstructed early cellular phylogenies that demonstrate (1) an endogenous mutational rate that is higher in the first cell division but decreases to approximately one per cell per cell division later in life; (2) universal unequal contribution of early cells to embryo proper, resulting from early cellular bottlenecks that stochastically set aside epiblast cells within the embryo; (3) examples of varying degrees of early clonal imbalances between tissues on the left and right sides of the body, different germ layers and specific anatomical parts and organs; (4) emergence of a few ancestral cells that will substantially contribute to adult cell pools in blood and liver; and (5) presence of mitochondrial DNA heteroplasmy in the fertilized egg. Our approach also provides insights into the age-related mutational processes and loss of sex chromosomes in normal somatic cells. In sum, this study provides a foundation for future studies to complete cellular phylogenies in human embryogenesis.

An adult human body consists of trillions of cells of more than 200 types[2]. These cells can be traced back to the fertilized egg, which continues to divide and eventually organizes into an individual body. Most of our knowledge about human development originates from rare direct observation of human fetal specimens, or extrapolation from model organisms despite the interspecies divergence[3–6]. Although extensive studies have been carried out[1,7], our understanding of the fundamental aspects of early human embryogenesis remains limited.

Genomic mutations spontaneously accumulate in somatic cells throughout life[8,9], presumably to a substantial rate beginning with the first cell division[10]. These mutations can be used as cellular barcodes to reconstruct developmental phylogenies of somatic cells[11], as used in tracking the clonal composition of tumours[12], blood[13] and brain tissues[14]. To this end, we applied the 'capture–recapture' strategy[13] for detecting and tracing early embryonic mutations (EEMs) in human individuals (Fig. 1a).

Capture–recapture of embryonic mutations

First, in the capture phase, we isolated viable single cells from various anatomical locations during autopsies of seven individuals (Extended Data Fig. 1a, Supplementary Table 1, Supplementary Discussion 1). The single cells were expanded in vitro to 1,000–10,000 cells, which provided sufficient genomic DNA for whole-genome sequencing (WGS) without error-prone whole-genome amplification. Overall, we obtained high-quality somatic mutations from WGS (mean coverage approximately 25×) of 334 single-cell colonies (hereafter referred to as clones) (Extended Data Fig. 1b, Supplementary Table 2). We then extracted EEMs from the mutation list as described below.

Second, in the recapture phase, small bulk tissues (n = 379), dissected from various organs and locations of the donated bodies, were deep sequenced (mean coverage approximately 2,900×) targeting EEMs discovered in the capture phase (Extended Data Fig. 1a, Supplementary Table 3). Of note, targeted deep sequencing was not possible from two donors (DB2

[1]Graduate School of Medical Science and Engineering (GSMSE), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea. [2]GENOME INSIGHT Inc, Daejeon, Republic of Korea. [3]Department of Anatomy, School of Medicine, Kyungpook National University, Daegu, Republic of Korea. [4]Immune Square Inc, Daegu, Republic of Korea. [5]Korea Institute of Science and Technology Information (KISTI), Daejeon, Republic of Korea. [6]Department of Medical Informatics, College of Medicine, Catholic University of Korea, Seoul, Republic of Korea. [7]Department of Immunology, School of Medicine, Kyungpook National University, Daegu, Republic of Korea. [8]Department of Pathology, Kyungpook National University Chilgok Hospital, School of Medicine, Kyungpook National University, Daegu, Republic of Korea. [9]Biomedical Research Institute, Kyungpook National University Hospital, Daegu, Republic of Korea. [10]These authors contributed equally: Seongyeol Park, Nanda Maya Mali, Ryul Kim. [11]These authors jointly supervised: Ji Won Oh, Young Seok Ju. ✉e-mail: ohjiwon@knu.ac.kr; ysju@kaist.ac.kr

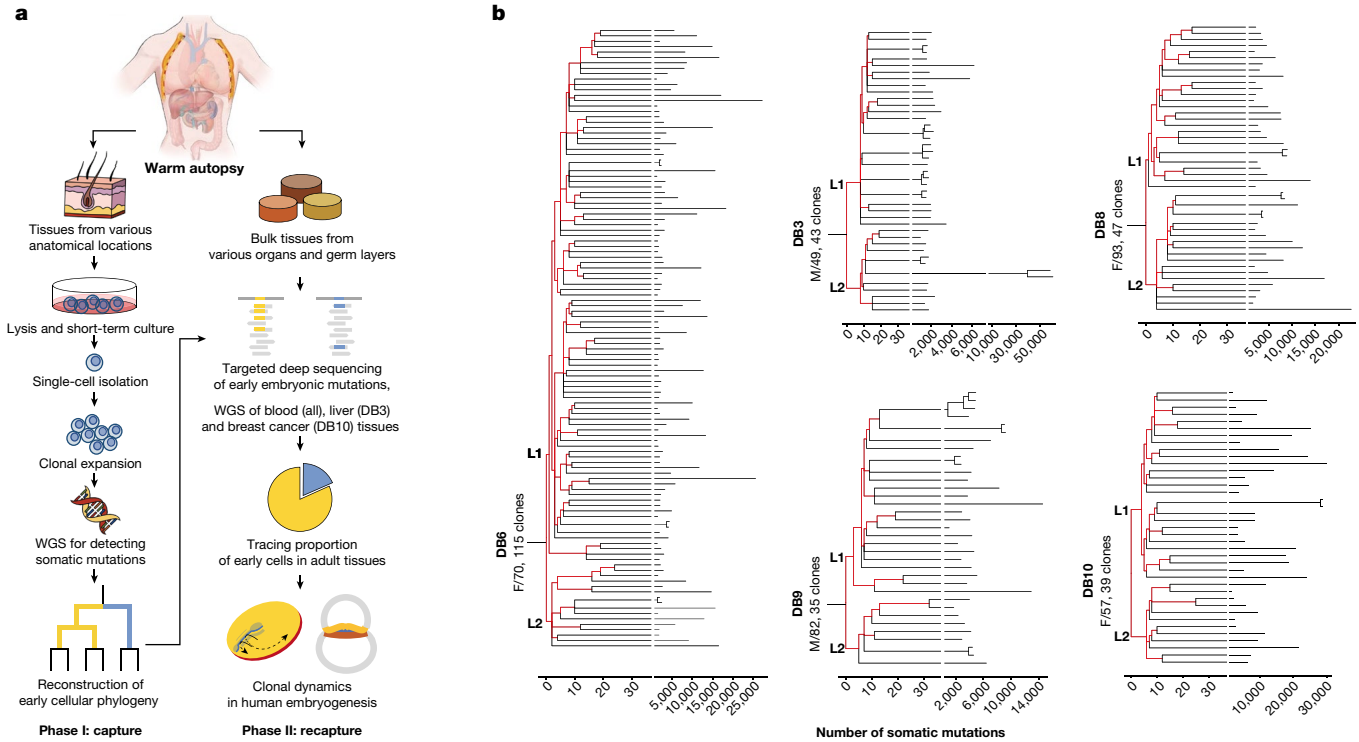


Fig. 1 | Tracing early cellular phylogenies using somatic mutations. **a**, Experimental design with two phases: a capture phase (left), in which adult viable single cells are collected and expanded in vitro, and analysed by WGS to detect genome-wide somatic mutations, and a recapture phase (right), in which bulk tissues were deep-sequenced for tracing VAFs of EEMs in the tissues. **b**, Cellular phylogenies of the 279 clones from five donors (sex (M, male or F, female) and age of donors are shown). The horizontal axes show molecular time as measured by the number of somatic mutations. Branches with recognized early mutations are coloured in red. L1 and L2 are major and minor lineages at first branching, respectively.

and DB5) owing to the limited amount of tissues available. In addition to the targeted sequencing, EEMs in polyclonal blood tissues (from each of the individuals), 22 microdissected hepatic lobules (DB3) and a breast cancer tissue (DB10) were traced by WGS of these tissues (Supplementary Table 4).

Somatic mutations and early phylogenies

From the capture phase with the 279 clones from the five donated bodies, we detected 1,532,625 somatic single nucleotide variants (SNVs) and 35,257 insertions and deletions (indels) (Supplementary Table 2). Multiple lines of evidence indicated that most of the SNVs and indels were truly somatic, rather than culture-mediated artefacts (Supplementary Discussion 2). Of these, 117,340 SNVs and 2,405 indels were shared by at least 2 clones, and were therefore informative for the phylogeny reconstruction (Extended Data Fig. 1c, d). Notably, only a few mutations completely dichotomized all the clones of a donor (referred to as lineages L1 and L2) (Fig. 1b). For example, 8 mutations were present in 30 of the 43 clones from DB3 (L1), whereas another set of 8 mutations were present in all the remaining clones (L2) (Extended Data Fig. 2a). We found such mutually exclusive mutation sets defining L1 and L2 in all five donors (Extended Data Fig. 1d). In the recapture, the two mutation sets showed variant allele fraction (VAF) of approximately 50% when aggregated (Extended Data Fig. 2b). These data collectively suggest that the two mutation sets are cellular barcodes for the two earliest ancestral cells of all cells in an individual, which were potentially, but not necessarily, the two cells at the two-cell-stage embryo.

After the first cell branching, clones in a specific lineage were further partitioned by additional mutations (Fig. 1b, Extended Data Fig. 1d, Supplementary Table 5). Except for a few outliers, most clone pairs diverged from each other before 35 mutations of molecular time, which defined the latest embryonic period that it was possible to explore in this study. Overall, we detected 537 EEMs (488 SNVs and 49 indels) from the five phylogenies (Supplementary Table 6). The number of detected EEMs per individual showed a positive correlation with the number of clones established from a donor (Extended Data Fig. 2c). The mutational spectra[15] of the EEMs directly indicate that intrinsic processes generate mutations mainly early in life, as previously suggested[10] (Extended Data Fig. 3, Supplementary Discussion 3).

Our phylogenies indicate that anatomically adjacent cells are generally remote developmentally (overall 92% of informative cases), although clones of specific lineages are occasionally observed (Extended Data Figs. 2a, 4).

Features of developmental phylogenies

We observed two common characteristics from the five phylogenies as well as three others in the accompanying paper[16]. First, unlike absolute bifurcations (dichotomies) in the first branchings, multifurcations (polytomies) are often observed from the second branchings (Fig. 1b, Extended Data Fig. 5a). This implies that the mutation rate is higher in the first cell branching and that mutation-based tracking of early cellular doublings is not always possible from the second branchings owing to the stochastic absence of intrinsic mutations (Extended Data Figs. 5b, c).

Second, as observed previously[10], the two earliest ancestral cells (as well as two daughter cells in later branchings) showed an unequal contribution in the phylogenies (Fig. 2a, Extended Data Fig. 5d). For instance, 112 early lineages in DB6 split at a ratio of 6.5:1 at the first branching, substantially skewed from the 1:1 expectation ($P < 0.001$). The imbalance was not caused by sampling bias of clones in the capture phase, because VAFs of L1 mutations expected in the phylogeny are corroborated by VAFs in the recapture phase (Fig. 2a, Extended Data Figs. 5d, e). Such unequal contributions were consistently observed in the other phylogenies, but imbalance ratios were variable across individuals (from 1.4:1 to 6.5:1), suggesting stochasticity of clonal segregation in human development, unlike the fully deterministic embryogenesis of *C. elegans*[17] (Extended Data Fig. 5f).

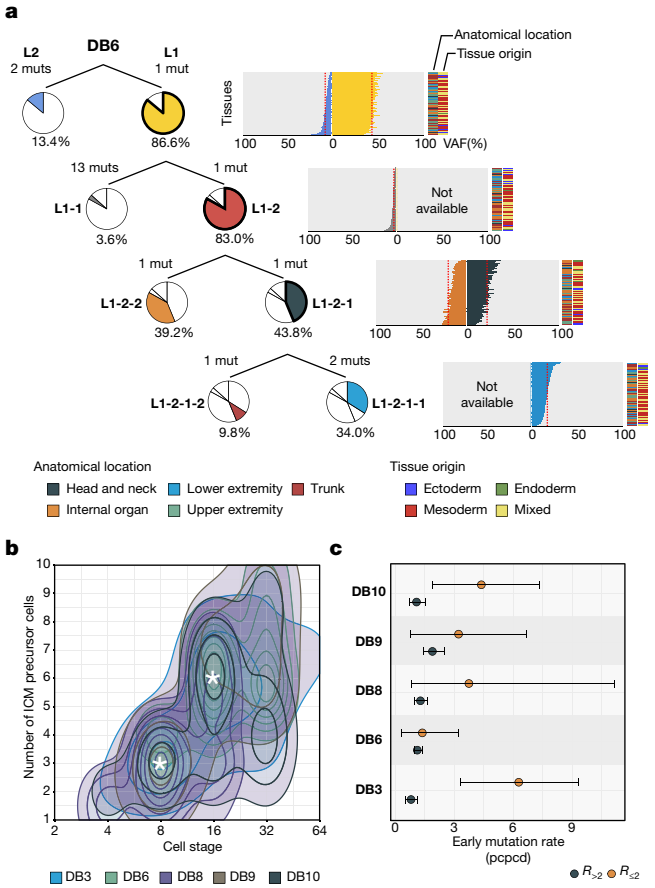


Fig. 2 | Unequal contribution of early lineages to human bodies and early mutation rate. a, The unequal contribution of two daughter lineages in body formation. The imbalances from the 1st to the 4th branchings of DB6 are shown. Pie graphs represent the proportion of each lineage counted by clones in the phylogeny. Horizontal bar graphs show the VAFs of early mutations in targeted deep-sequencing. Expected VAFs from the phylogeny are shown by red dashed lines. Mut, mutation. **b**, A contour plot showing the simulation for estimating the number of epiblast-contributing cells at given cell-stage embryos. White asterisks indicate the maximum-likelihood values. **c**, Estimates of early embryonic mutation rates with 95% confidence intervals from simulation (n = 500,000). $R_{\leq 2}$ and $R_{>2}$ are mutation rates up to and after 2-cell stage, respectively.

For more quantitative investigation of the phylogenies, we simulated a cell genealogy scenario in which a few cells in a fully bifurcating tree are stochastically set to epiblast precursor cells through a cellular bottleneck[18,19] at a certain cell stage (Extended Data Fig. 6, Supplementary Discussion 4). Maximum-likelihood simulations suggested that three cells in the 8 cell-stage embryos, or 6 cells in the 16 cell-stage embryos, are contributing mainly to the epiblast (Fig. 2b), consistent with previous studies in mouse and human[20–22].

By accommodating missing cell divisions in the phylogenies (Supplementary Discussion 4), our simulations determined early mutation rate per absolute cell division of 3.8 (range 1.4 to 6.3) mutations per cell per cell division (pcpcd) for the earliest cell divisions ($R_{\leq 2}$) and 1.2 mutations pcpcd (range 0.8 to 1.9) from 4-cell stages ($R_{>2}$), with a substantial interindividual variability (Fig. 2c; Supplementary Table 1; Supplementary Discussion 5). We speculate that the higher mutation rate in the earlier period may be, in part, owing to the lack of mature DNA repair processes before zygotic genome activation[23,24]. Notably, these donor-specific EEM rates and the number of detected EEMs determined the latest developmental time that it was possible to explore for each individual (ranging from around the 11th cell generation (CG) in DB8 to around the 17th CG in DB6; 95th percentile molecular time).

Tracing the early embryonic cells

We next traced the VAF differences of EEMs in various adult tissues (the recapture phase). First, left- and right-side tissues exhibited substantial differences in all five individuals. Despite slight variations in the level and timing between donors, the left–right VAF differences emerged from EEMs acquired at 1st–4th CGs (Fig. 3a, b, Extended Data Fig. 7a, b). Notably, these EEMs occur before gastrulation, which can be partially explained by the primitive streak separating the epiblast layer in pre-gastrulation embryos[25] (Extended Data Fig. 7c, Supplementary Table 3). Tissues from internal organs (such as lung, kidney and liver) of DB6 also showed concordant orientation (Fig. 3b, Supplementary Discussion 6). By contrast, such early imbalances were not evident across the cranio-caudal axis (Extended Data Fig. 7d).

In DB6, EEMs acquired at around the 9th CG_{DB6} begin to collectively differentiate into different body parts (for example, right upper extremity, 6th–9th CG_{DB6}) and particular organs (for example, stomach, 9th–17th CG_{DB6}), suggesting the timing of the anatomical restriction of early embryonic cells (Fig. 3b).

Notably, from around the 3rd to 7th CGs, the VAFs of EEMs were consistently lower in ectodermal tissues (epidermis) than in mesodermal tissues (such as dermis or muscle) or endodermal tissues (liver) in all available cases (Fig. 3c, Extended Data Fig. 8a). This implies that a few lineages contributing more to ectodermal tissues diverged from lineages predominantly contributing to mesodermal and endodermal tissues (Extended Data Fig. 8b). Mesodermal and endodermal tissues showed more similar composition of early cells compared with ectodermal tissues (Fig. 3c, Supplementary Discussion 7).

Next, we investigated early founder lineages in specific cell types. In DB9, for example, a lineage branched at a molecular time of around 15 mutations (approximately 8th CG_{DB9}) contributed to about 50% of the adult blood cell pool (approximately 25% VAF), 15-fold higher than expected in the phylogeny (Fig. 3d). We consistently found these blood-enriched cell lineages in all informative individuals at the similar molecular time of around 9–17 mutations (8th–12th CGs; Extended Data Fig. 8c, d, Supplementary Table 4). Whole-genome sequences of the blood tissues further showed 0–22 mutations that were not found in the capture phase (Extended Data Fig 8e). These mutations suggest that the timing of the lineage expansion is slightly later than the branching time observed in the phylogenies. The blood-enriched early cells have been found in model organisms[26,27]. Importantly, it should be distinguished from age-related clonal haematopoiesis[28].

Similarly, lineages with a higher contribution to hepatocytes were traced in DB3 (Fig. 3e, Supplementary Discussion 7). Whole genome data from microdissected hepatic lobules suggested that 4 out of 22 lobules (18%) were predominantly composed of cells originating from a lineage (L1-4) (Fig. 3e), about sixfold higher than expected. The read-depth-adjusted aggregate of 22 WGSs, further encompassing minor contributing lineages, showed similar clonal dynamics. The number of mutations shared between the lobules and the phylogeny (9–15 mutations) suggests that lineage expansion in the liver occurred around the 5th–13th CG_{DB3}.

Finally, we also compared somatic mutations in the breast cancer tissue of DB10 to its phylogeny. The tumour whole genome shared eight mutations with a skin fibroblast clone established from the right knee, suggesting the ancestral lineage of the cancer was branched out from lineage L2-2-1 at around the 5th CG_{DB10} (Extended Data Fig. 8f).

In sum, the clonal dynamics in early human embryogenesis can be conceptually summarized in Extended Data Fig. 8g.

Heteroplasmy of mitochondrial DNA

From the mitochondrial DNA (mtDNA) sequences in WGS of 279 clones, we identified somatic-like mtDNA variants (485 SNVs and 48 indels). Most of them (84.4%) were single-clone specific, exhibiting various levels of heteroplasmy. However, a fraction of mutations (96 out of 533, 18.0%)

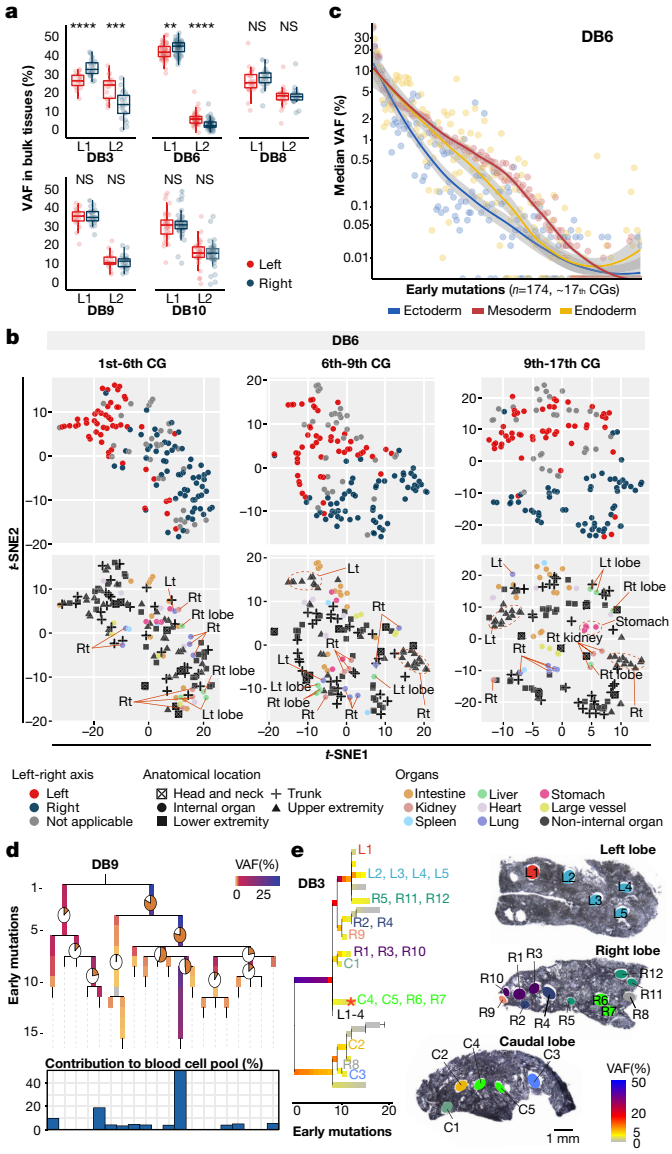


Fig. 3 | Timing of fate determination of early cells. a, Box plots illustrating median VAFs with interquartile ranges (IQRs) with whiskers (1.5 × IQRs) of L1 and L2 mutations in left- or right-side tissues. Two-sided Wilcoxon test; $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$; NS, not significant. Exact *P*-values are 1.2×10^{-5} (L1) and 3×10^{-4} (L2) for DB3, and 0.0041 (L1) and 1.5×10^{-7} for DB6. **b**, *t*-SNE clustering of DB6 bulk tissues using the VAFs of the embryonic mutations occurred at three different times of early embryogenesis. The separation between left (lt) and right (rt) tissues is apparent from the earliest molecular time (≤6th CG, top row). This is followed by the separation of anatomical regions and organs (bottom row). **c**, Median VAFs of the early mutations in the bulk tissues according to their dominant germ layers in DB6. The horizontal axis shows early mutations sorted by the averaged VAFs in bulk tissues in descending order. Tissues with mixed germ layers are excluded. Fitted curves by locally estimated scatter plot smoothing are shown. **d**, The contribution of early cells to adult blood tissues. The phylogenetic tree is coloured by VAFs of EEMs in the blood. Pie and bar graphs represent the estimated contribution of each lineage to blood tissues. **e**, The contribution of early cells to adult hepatocytes in DB3. The phylogenetic tree is coloured by VAFs of EEMs from hepatic lobules (left). The anatomical locations of the hepatic lobules and their dominant lineages are shown (right).

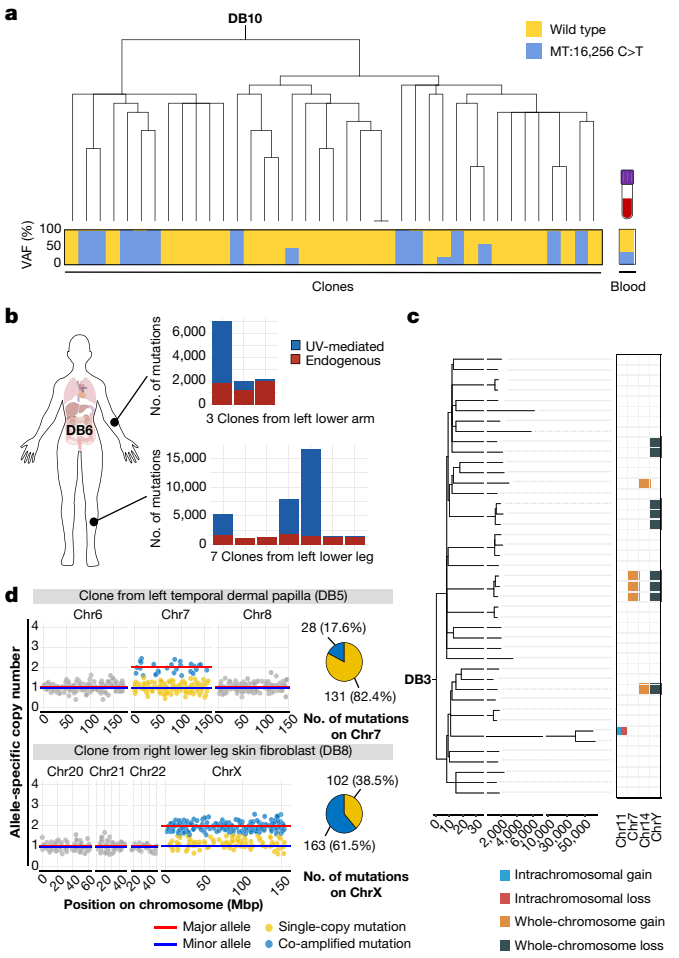


Fig. 4 | mtDNA heteroplasmy in fertilized egg and late-stage mutations. **a**, An example of mtDNA mutation (MT:16,256 C>T) recurrently found in clones (left) and polyclonal blood (right) of DB10. Heteroplasmy levels (VAFs) are shown in bar plots (bottom). **b**, Examples of highly variable burdens of UV-mediated mutations among clones established from the close anatomical locations. **c**, Large-segmental copy-number changes (>10 Mb) in DB3 correlated with the phylogeny. Chr, chromosome. **d**, Examples of the whole-chromosomal gains that occurred in the third and tenth decades of life (top and bottom, respectively), estimated by the number of co-amplified mutations.

were repeatedly observed in clones established from an individual. For example, mtDNA:16,256 C>T was found exclusively in 14 clones of DB10 (36.8% out of a total 38 clones), frequently in homoplasmy (approximately 100% VAF) and in the polyclonal blood (approximately 35.0% VAF) (Fig. 4a, Extended Data Fig. 9a). Its distribution in the phylogeny indicated that the first node, probably the fertilized egg, harboured the mtDNA variant with heteroplasmy of about 29.8% (95% confidence interval 25.0–34.0%; Extended Data Fig. 9b-d), despite the mtDNA-purification mechanism in the maternal germline[29]. The bimodal distribution of its VAFs among somatic clones (0% and 100% VAFs in 24 and 10 clones, respectively) may reflect bottlenecks during a progressive mtDNA copy number reduction in the cleavage[30,31] followed by random-walk mtDNA segregations during mitoses for a lifetime[32,33] (Extended Data Fig. 9e). Albeit at lower levels, we found similar mtDNA heteroplasmy in the first cells of the other phylogenies (Extended Data Fig. 9f).

Late-stage mutations and copy-number changes

Most somatic mutations identified from the 334 clones from all seven donors (1,647,501 SNVs and 40,967 indels) were not attributed to the early branches, and were therefore regarded as late-stage mutations. Most of the mutations were single-clone specific (92.5%) or shared by clone pairs that diverged later than embryogenesis (7.5%) (Fig. 1b, Supplementary Table 5). Independently acquired recurrent mutations were also rarely observed ($n = 619$ when collapsed; Extended Data Fig. 10a, Supplementary Table 2).

The vast majority of the late-stage mutations were attributable to a few known mutational processes, including UV-mediated DNA damage (COSMIC single base substitution (SBS) 7a–7d) and endogenous clock-like mutagenesis[15] (COSMIC SBS 5 and 1) (Extended Data Fig. 10b, c, Supplementary Discussion 3). The spectrum of the recurrent mutations was very similar to SBS 7a and 7b (Extended Data Fig. 10d), suggesting that they shared UV-mediated mutagenesis as their main origin. There was a large variation in the burden of UV-mediated mutations even in anatomically adjacent clones (Fig 4b; Extended Data Fig. 10e), which could be caused by heterogeneous DNA repair capabilities across cell lineages[34]. Two corneal epithelial cell clones in DB3 harboured extremely large numbers of UV-mediated mutations (58,858 and 61,146). After adjusting UV-mediated mutations (Supplementary Discussion 8), we concluded that around 24 mutations endogenously accumulate in skin fibroblasts per year (Extended Data Fig. 10f).

Although most of the clones exhibited diploid genomes, 10%–23% of the clones in each individual exhibited somatic DNA copy-number alterations (CNAs) more than 10 Mb in size. Some of these variations were shared by clones (Fig. 4c), confirming their in vivo origin. Approximately 70% (38 out of 54) of the CNAs were chromosome-level CNAs, and frequently in sex chromosomes (19 clones) (Extended Data Figs. 11a, b). Because these chromosomes are transcriptionally less active than autosomes, such whole-chromosome changes would be more tolerated in cells[35].

The timing of large copy-number gains was estimated using the clock-like property of endogenous somatic mutations[36]. Of the 21 long-segmental amplifications (larger than 50 Mb), 7 (33%) were dated to have occurred in the first 4 decades of life, implying stochastic acquisition of these events in normal cells (Fig. 4d, Extended Data Fig. 11c).

Discussion

Several complementary approaches have been used for detection of somatic mutations in single cells, such as whole-genome amplification[37], laser-capture microdissection[16,38–41], sequencing of small biopsies[42–45] and massive in vitro clonal expansion of single cells[13,46]. Although the clonal expansion method is labour-intensive and only applicable to dividing cells, it provides the most sensitive and precise somatic mutation data at the absolute single-cell level as shown in this study. This study indicates that human embryogenesis can be further dissected to the later cell stages by scaling up our approach together with a large number of clones and tissues from additional organs. As the cost of genome sequencing falls and with further technological innovations, advanced studies should soon be deliverable.

Although this study focused mainly on common features, human embryogenesis is, in principle, a personal and variable event. Many developmental diseases may be caused by catastrophic mutations and misregulation of clonal dynamics in early embryogenesis. Similar analyses on more individuals, particularly for diseases with unknown aetiologies, are likely to yield the clinical impact of early human embryogenesis to an unprecedented resolution in the future.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03786-8.

1. Wamaitha, S. E. & Niakan, K. K. Human pre-gastrulation development. *Curr Top Dev Biol* **128**, 295–338 (2018).
2. Sender, R., Fuchs, S. & Milo, R. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*. **14**, e1002533 (2016).
3. Gasser, R. F., Cork, R. J., Stillwell, B. J. & McWilliams, D. T. Rebirth of human embryology. *Dev. Dyn*. **243**, 621–628 (2014).
4. Nakamura, T. et al. A developmental coordinate of pluripotency among mice, monkeys and humans. *Nature* **537**, 57–62 (2016).
5. Rossant, J. Mouse and human blastocyst-derived stem cells: vive les differences. *Development* **142**, 9–12 (2015).
6. Xiang, L. et al. A developmental landscape of 3D-cultured human pre-gastrulation embryos. *Nature* **577**, 537–542 (2020).
7. Shahbazi, M. N. & Zernicka-Goetz, M. Deconstructing and reconstructing the mouse and human early embryo. *Nat. Cell Biol*. **20**, 878–887 (2018).
8. Samuels, M. E. & Friedman, J. M. Genetic mosaics and the germ line lineage. *Genes (Basel)* **6**, 216–237 (2015).
9. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
10. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
11. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
12. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med*. **366**, 883–892 (2012).
13. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
14. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
15. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
16. Coorens, T. H. H. et al. Extensive phylogenies of human development reveal variable embryonic patterns. *Nature* https://doi.org/10.1038/s41586-021-0379-y (2021).
17. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol*. **100**, 64–119 (1983).
18. Hardy, K., Handyside, A. H. & Winston, R. M. The human blastocyst: cell number, death and allocation during late preimplantation development in vitro. *Development* **107**, 597–604 (1989).
19. Sancho, M. et al. Competitive interactions eliminate unfit embryonic stem cells at the onset of differentiation. *Dev. Cell* **26**, 19–30 (2013).
20. Biggins, J. S., Royer, C., Watanabe, T. & Srinivas, S. Towards understanding the roles of position and geometry on cell fate decisions during preimplantation development. *Semin. Cell Dev. Biol*. **47-48**, 74–79 (2015).
21. Spencer Chapman, M. et al. Lineage tracing of human development through somatic mutations. *Nature* **595**, 85-90 (2021).
22. Wennekamp, S., Mesecke, S., Nédélec, F. & Hiiragi, T. A self-organization framework for symmetry breaking in the mammalian embryo. *Nat. Rev. Mol. Cell Biol*. **14**, 452–459 (2013).
23. Blakeley, P. et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151–3165 (2015).
24. Schulz, K. N. & Harrison, M. M. Mechanisms regulating zygotic genome activation. *Nat. Rev. Genet*. **20**, 221–234 (2019).
25. Gardner, R. L. Normal bias in the direction of fetal rotation depends on blastomere composition during early cleavage in the mouse. *PLoS ONE* **5**, e9610 (2010).
26. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
27. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
28. Shlush, L. I. Age-related clonal hematopoiesis. *Blood* **131**, 496–504 (2018).
29. Wai, T., Teoli, D. & Shoubridge, E. A. The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. *Nat. Genet*. **40**, 1484–1488 (2008).
30. Cummins, J. M. The role of maternal mitochondria during oogenesis, fertilization and embryogenesis. *Reprod. Biomed. Online* **4**, 176–182 (2002).
31. Floros, V. I. et al. Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol*. **20**, 144–151 (2018).
32. Coller, H. A. et al. High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet*. **28**, 147–150 (2001).
33. Wonnapinij, P., Chinnery, P. F. & Samuels, D. C. The distribution of mitochondrial DNA heteroplasmy due to random genetic drift. Am. *J. Hum. Genet*. **83**, 582–593 (2008).
34. Sanders, M. A. et al. Life without mismatch repair. Preprint at https://doi.org/10.1101/2021.04.14.437578 (2021).
35. Thompson, D. J. et al. Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
36. Lee, J. J.-K. et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell* **177**, 1842–1857.e21 (2019).
37. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
38. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
39. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
40. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
41. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* https://doi.org/10.1038/s41586-021-03822-7 (2021).
42. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
43. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
44. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
45. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell* **177**, 608–621.e12 (2019).
46. Fasching et al. Early developmental asymmetries in cell lineage trees in living individuals? *Science* **371**, 1245–1248 (2021).

Methods

Warm-autopsy and tissue sampling

A total of seven donors transitioning to imminent death were included in post-mortem tissue donation for this study (Supplementary Table 1). The objectives and procedure of this study were explained to the patients or nearest kin, and permission for autopsy was obtained before their death. From June 2016 to January 2019, 10 immediate autopsies had been performed at the Kyungpook National University Hospital and Department of Anatomy at Kyungpook National University, School of Medicine. In seven of them, primary culture of tissues was successful. When a participant died, we collected 20–30 ml of blood in a purple-topped, heparinized tube. The whole body was sterilized with 70% ethanol before tissue sample collection. The skin, subcutaneous tissues, muscles and hair follicles from several randomly chosen anatomical sites and corneas were taken and immersed in transportation media as soon as possible (Extended Data Fig. 1a). The internal organs were exposed through an anterior Y-shaped incision extending from the shoulders to the pubis and dissected from their posterior attachments to the body. Tissues were collected as fresh tissue, snap-frozen tissue using liquid nitrogen (stored at −70 °C), and formalin-fixed. Fresh tissues were transported to the laboratory for tissue culture experiments. All tissues were collected within 36 h of death and processed on the day of the collection according to the protocol. All the procedures in this study were approved by Institutional Review Board of Kyungpook National University (approval number: KNU-2018-0088, KNU-2019-0151) and KAIST (approval number: KH2020-029). This study was conducted in accordance with the Declaration of Helsinki provisions. No statistical methods were used to predetermine the sample size.

Primary culture of tissues

We obtained a total of 165 tissues, approximately $1 \times 1 \times 1\ cm^3$ in size, for a single-cell expansion experiment from the seven autopsies (Extended Data Fig. 1a, Supplementary Table 2, Supplementary Discussion 1). All tissues were processed as previously described with minor modifications (detailed description below). Growth medium contained Dulbecco's modified Eagle's medium (DMEM) low glucose, 20% fetal bovine serum, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin, 2 mM L-glutamine, and 1 μg ml^{-1} fungizone. All components were obtained from Gibco (Thermo Fisher Scientific, Waltham, MA, USA). The detailed protocols for each tissue type are described below.

Skin culture

The dermal skin fibroblasts were cultured by the explant method described previously with minor modification[47–49]. In brief, the skin samples were washed 2–3 times with buffer saline (1X PBS; Gibco, Thermo Fisher Scientific). The adipose and blood vessels associated tissues were removed. Subsequently, the tissues were cut into 1–2 mm^2 pieces and put in 1 mg ml^{-1} collagenase/dispase solution (Roche Diagnostics) at 37 °C for 1 h. After the treatment, the epidermis layer was separated from the dermis layer. The dermis tissue was washed with DMEM medium containing 20% FBS (Gibco) to inhibit collagenase/dispase activity. Then, the dermis tissue was minced into small pieces and cultured in collagen I-coated 24-well plates (Corning BioCoat) with 200 μl medium in a humidified incubator at 37 °C and 5% CO_2 concentration.

Muscle culture

The skeletal muscle and other smooth muscle were used for muscle culture with a previously described protocol with minor modification[50]. The muscle tissue was washed in 1× PBS (Gibco) and then 10% FBS-containing medium. Next, the connective tissue and adipose tissues were removed. The tissue was chopped into small pieces of 1–2 mm^2 and digested for 45–60 min at 37 °C with 1mg ml^{-1} collagenase/dispase solution with frequent agitation. The tissue slurry was washed with FBS containing growth medium and centrifuged at 1,400 rpm for 5 min at room temperature. The pellet was resuspended in a fresh DMEM medium containing 20% FBS. A small piece of tissue was taken out from a slurry and minced with sharp blades and cultured in collagen I-coated 24-well plates with 200 μl medium in a humidified incubator at 37 °C and 5% CO_2.

Corneal tissue culture

The corneal tissue was washed with 1× PBS (Gibco). Then, the cornea was cut diagonally into 16 pieces, and each piece was transferred into a well of a collagen I-coated 24-well plate. The fragments of cornea tissue were chopped in a culture plate with 20% FBS containing DMEM medium. For corneal epithelium, the medium was replaced with keratinocyte-specific CNT-PR medium (CELLnTEC Advanced Cell System) after 24 h of tissue culture.

Hair follicle culture

Hair follicle samples were collected from the scalp and other anatomical sites of individuals. Excess hair shafts were trimmed with scissors and washed with 1× PBS. The single hair follicles were isolated and cultured using the previously described methods[51] after the determination of hair follicle stages. The anagen hair follicles were used for dermal papilla (DP) and dermal sheath (DS) culture. DP and DS were isolated from the bulbs of micro-dissected hair follicles. Each DP and DS were separately transferred to collagen I-coated culture dishes and cultured with 20% FBS containing growth medium.

Cell outgrowth and passage

The medium was changed every four days after culture to remove cellular debris in suspension. Usually, the cells were proliferated for two weeks until passages, depending on tissue and cell types. After reaching sub-confluency, the cells were maintained in DMEM supplemented with 10% FBS in 60-mm or 100-mm dishes to adjust the microenvironment of cell growth.

Single-cell cloning and clonal expansion

The subconfluent cells were washed in ice-cold 1× PBS and collected using 1× trypsin (TrypLE, Gibco) in a 15-ml conical tube (BD Falcon) with sufficient amount of fresh medium to neutralize the trypsin activity. The cell pellet was resuspended in 1 ml growth medium, and the total cell number was counted by trypan blue using a disposable haematocytometer slide in an automatic cell counter (EVE automatic cell counter, NanoEnTek). After the calculation of the total cell number, a serial-dilution method was applied to dilute the cells up to 10 cells per ml. Then, 100 μl cell suspension medium was seeded in each well of 96-well plates (BD Biosciences) and incubated at 37 °C with 5% CO_2. Each well contained a single cell. Within 12–15 h, the cell was manually checked under a light inverted microscope (Olympus) at 10× magnification, and wells containing a single cell were selected. Single cells were proliferated up to 80–90% confluency; then, the individual cell clones were expanded into 24-well plates. The confluent clones were collected for DNA extraction and WGS.

Library preparation and WGS

We extracted genomic DNA materials from clonally expanded cells and matched peripheral blood using DNeasy Blood and Tissue kits (Qiagen) according to the manufacturer's protocol. DNA libraries for WGS were generated by an Accel-NGS 2S Plus DNA Library Kit (Swift Biosciences) from 1 μg of genomic DNA material. WGS was performed on either the Illumina HiSeq X Ten platform or the NovaSeq 6000 platform to generate mean coverage of 25.2× for 374 clonally expanded cells and 94.8× for 7 matched blood tissues. In addition, 22 hepatic lobules dissected from DB3 (213× for a lobule, and 22× on average for the remaining 21 lobules) and 1 breast cancer tissue (26×) from DB10 were also analysed by WGS.

Variant calling and filtering of WGS data

Sequenced reads were mapped to the human reference genome (GRCh37) using the BWA-MEM algorithm[52]. The duplicated reads were removed by Picard (available at http://broadinstitute.github.io/picard), and indel realignment and base quality score recalibration were performed by GATK[53]. Initially, we identified base substitution and short indels by using HaplotypeCaller[54] and VarScan2[55]. Variants called by both of the tools were included for future analysis. For every variant in any of the single-cell-derived clones, we evaluated the number of mutant and wild type reads. To establish the high-confidence somatic variant sets, we applied additional filtering processes for these variants. In brief, we removed variants with the following features: low mapping quality (<40), high proportion of indels or clipping (>90%), 5 or more mismatched bases in the variant reads, 1% or more VAF in the panel of normals, and frequent existence of error reads in other clones. Then, we visually inspected the most shared variants using the Integrative Genomics Viewer[56]. The detailed method and performance of each step is available in Supplementary Discussion 9 and Supplementary Table 7.

Quality control of samples

Among a total of 374 clonally expanded cells (colonies), 18 with low depth of coverage (mean depth <10) were excluded. Based on the VAFs of somatic mutations and established phylogenetic trees, we removed additional 19 colonies thought to be multiclonal without a single dominant clone: they have variants in mutually exclusive lineages simultaneously (four-gamete test violation), and/or low peak VAFs (Extended Data Fig. 1b). In addition, three colonies that show unexplainable high peak VAFs were excluded for data integrity. The final 334 colonies were mostly single-cell derived, or having a dominant clone for unambiguously reconstructing phylogenetic trees.

Detection of culture-associated mutations

Since we performed two times of in vitro culture, culture-associated mutations could be integrated in our mutation list. First, the mutations which were generated in short-term culture before single cell isolation would exist in all cells in clones (clonal). Second, the mutations which occurred after single cell isolation would exist in part of cells in clones (subclonal). The latter mutations seem to be rare, because additional VAF peaks < 50% were not detected in most 334 single-cell derived clones and in the merged VAF distribution of all somatic mutations (n = 1,688,468; Supplementary Discussion 2). To approximate the amount of the culture-associated mutations before single cell isolation, we conducted serial single-cell clonal expansions followed by WGS of both clones in three samples (Supplementary Discussion 2). In these experiments, mutations, which are clonal in second clones but subclonal or not detected in first clones, are culture-associated mutations that occurred during first single cell cloning. To discriminate clonal mutations and subclonal mutations, we calculated mutant-cell fraction (MCF) by the same method estimating cancer cell fraction in the previous studies[36,57]. Mutations, which showed MCF < 0.25 in the first clone and MCF > 0.25 in the second clone simultaneously, were defined as culture-associated mutations.

Reconstruction of the phylogenetic trees and defining early mutations

The developmental phylogenetic tree of an individual was inferred from multi-sample genotype information. Let $S = [s_1, s_2, \ldots, s_n]$ be the set of n samples, and $G = [g_1, g_2, \ldots, g_m]$ be the union set of mutations detected in one or more samples from the same individual. We then build a matrix M with rows labeled $g_1, g_2, \ldots, g_m$, and columns labeled $s_1, s_2, \ldots, s_n$. If the VAF of somatic mutation g_i in sample s_j is determined to be true (Supplementary Discussion 9), M_{ij} was assigned to 1, while others were assigned to 0. Mutations shared in all samples were considered to be germline variants. After removing these germline variants, we grouped all mutations with the same profile into a mutation group according to the sharing pattern between samples (Extended Data Fig. 1c, d). Over m' distinct mutation groups, mutation matrix $M_{m' \times n}$ is defined such that each column represents a sample and each row represents a mutation group. From the mutation matrix $M_{m' \times n}$, we reconstructed a phylogenetic tree by using graphic theoretic methods[58]. Mutations shared by two or more samples, which violate the four-gamete test thus cannot be rooted to a most recent common ancestor in the phylogenetic tree, were considered to occur independently in those samples, and therefore were not employed to reconstruct a phylogenetic tree. A phylogenetic tree for the mutation matrix $M_{m' \times n}$ is a rooted tree where each sample is attached to exactly one terminal node of the tree, and each of the m' mutation group is associated with exactly one branch of the tree, with the number of mutations in the corresponding mutation group being the length of the branch.

Reconstructed phylogenetic trees were manually converted into Newick format. The trees were drawn in R using 'ggtree' package[59]. Most samples shared the small number of mutations (<35) except for some samples which shared more than hundreds of mutations. Only the mutations shared in small amount (<35) among samples were defined as early mutations. To validate the early mutations, five biological replicates (culturing the sample twice from three samples, and separating the cultured cell pool into two from two samples) and two technical replicates (sequencing the DNA pool twice) were made. All the detected early mutations were validated in the replicates.

Estimating the number of inner cell mass precursor cells and mutation rate in early embryogenesis

Several studies implicate the presence of a developmental bias among blastomeres at an early stage of embryogenesis[10,11,60]. However, little is known regarding to the exact details of this critical event in human embryogenesis, and both the precise timing and number of cells contributing to epiblast have yet to be clarified. We have attempted to answer these parameters, using the framework of an approximate Bayesian computation (ABC)[61]. Our simplified model of early human embryogenesis has four parameters: the number of blastomeres contributing to epiblast (n), the stage at which the precursors of epiblast are chosen (s), the mutation rate before zygotic genome activation (ZGA) ($R_{\leq 2}$), and the mutation rate after ZGA ($R_{>2}$) (Extended Data Fig. 6a). ZGA was assumed to occur at the 4-cell stage of early embryogenesis[23,62]. From a fertilized egg, cells are doubled at each stage, and somatic mutations accumulate in the nuclear genome with a mutation rate $R_{\leq 2}$ before ZGA and $R_{>2}$ after ZGA. At a stage s, n_{preEPI} cells are committed towards precursor of epiblast, and the cells divide 15 more times forming $n_{\text{preEPI}} \times 2_{15}$ cells. Of these simulated cells, n_{samples} cells (the number of clones established in each individual) are randomly selected for reconstructing a phylogenetic tree by using the somatic mutations accumulated in each cell. The major assumptions of this model are as follows:

1. All the cells in an embryo are simultaneously doubled at a stage s (that is, all the cells have the same division rate).
2. All the cells at a stage s are equally likely to be epiblast precursors.
3. All the cells randomly distribute over the embryo before precursor selection.
4. ZGA begins right after the 4-cell stage.
5. Both $R_{\leq 2}$ and $R_{>2}$ are constant during the simulation and follow the Poisson distribution.
6. Cell losses do not occur other than selection of n_{preEPI} and n_{samples} during the simulation.

This model was simulated 500,000 times for each individual, varying the parameters drawn from a uniform distribution with the following ranges: [2,7] for n_{preEPI}; [1, $2^{n_{\text{preEPI}}-1}$] for s; [2.0, 10.0] for $R_{\leq 2}$; and [0.5, 2.0] for $R_{>2}$. For each simulation and observed data from the five phylogenetic trees (DB3, DB6, DB8, DB9 and DB10), the following set of summary statistics was extracted.

1. The number of shared mutations between more than two samples ($\sum_{branches} m$, where m represents the number of mutations assigned to each branch)
2. The number of sample groups split right after fertilization ($n_{1st\ lineage}$).
3. The number of mutations assigned to each first branch.
4. Multifurcation score is defined as $\frac{\sum_{nodes} n_{branches}}{n_{nodes}}$, where $n_{branches}$ and n_{nodes} denote the number of branches for each node and the total number of nodes, respectively.

The summary statistics of each simulation are compared to the observed summary statistics. The parameters of simulations that produce summary statistics that are highly similar to the observed statistics are selected. To estimate the posterior distribution of the $R_{\leq 2}$ and $R_{>2}$, we used the neural network regression algorithm implemented in the 'abc' package in R[63].

Separation of skin tissues into epidermis and dermis

We removed adipose tissue and blood vessels from the skin tissue before treatment. After washing with 1× PBS, 2–3 times, we cut the tissue into small pieces. The section was placed into a collagenase/dispase solution (1 mg ml^{-1}) such that the epidermis layer is at the bottom. The tissue should be well covered with a collagenase/dispase solution and incubated at 4 °C for overnight. Then, we took out the plate and peeled off the epidermis from the dermis using a pair of fine forceps. Epidermis and dermis tissues were transferred into the fresh medium and washed with 1× PBS.

Laser-capture microdissection of hepatic lobules

Freshly frozen liver tissues from DB3 were incubated in xylene for 2.5 min 2 times. Then, the tissues were incubated in 100%, 90%, 80% and 70% ethanol for 10–20 s each. The tissues were stained with haematoxylin and moved to slide glasses. Hepatic lobules were microdissected with the PALM laser-capture system and collected in an adhesive cap (Carl Zeiss). Genomic DNA was extracted using the QIAamp micro DNA kit laser-capture microdissection tissue protocol (Qiagen). The library for WGS was constructed using NEBNext Ultra II DNA library kit (NEB) and sequenced using the NovaSeq 6000 platform (Illumina).

Targeted deep sequencing for bulk tissues

To estimate the contribution of early embryogenic cell lineages in adult tissues, we performed deep targeted sequencing on bulk tissues from various organs of the five individuals. Whenever possible, skin tissues were separated into the epidermis and dermis. Custom DNA bait sets were designed according to the manufacturer's guidelines to include early embryonic substitutions, which were confirmed before designing the baits, and several randomly-chosen germline mutations (SureSelect^{XT} Custom DNA Target Enrichment Probes, Agilent). Of the 441 EEMs targeted, 411 mutations could have high-quality baits designed for them. DNA libraries were prepared by SureSelect^{XT} Library Prep Kit (Agilent), hybridized to the appropriate capture panel, multiplexed on flow cells, and subjected to paired-end sequencing (151-bp reads) on the NovaSeq 6000 platform (Illumina) with a mean ~2,900× depth of coverage for the early mutations. Sequence reads were trimmed and mapped to the human reference genome (GRCh37) using the BWA-MEM algorithm[52]. Reads with mapping quality ≥20 and base quality ≥20 were included for analysis. Read counts of all mutated positions in the bait-set were counted with a custom script.

Subgrouping of early mutations in each branch using data from bulk tissues

Selection processes such as epiblast selection and random sampling of single cells can integrate multiple divisions into a branch in phylogenetic trees (Extended Data Fig. 6b). If a missing branch contributed to the embryo, the mutations that occurred before the hidden branching would have higher VAFs in bulk tissues than the other mutations in the same branch that occurred later. We performed the paired t-test on all possible pairs of EEMs in each branch. Mutations without significant differences in the test were clustered into a subgroup.

Analysis of clonal imbalance on germ layers, anatomical regions, and organs

To investigate the distribution of early embryonic cells in each germ layer, we classified bulk tissues into several germ layer groups: ectoderm, mesoderm, endoderm, and mixed (Supplementary Table 3). For example, fresh frozen skin tissues were physically separated into the epidermis and the dermis, which were considered as ectodermal and mesodermal origins. VAFs of early mutations were compared among these germ layer groups using the two-sided Wilcoxon test.

We also classified bulk tissues according to the left-right, cranio-caudal axis, and anatomical groups (Supplementary Table 3). Tissues with the specified anatomical locations such as skin, muscle and fat tissues were used in the analysis. A mid-sagittal plane and xiphoid process were adopted for distinguishing left-right and cranio-caudal tissues, respectively.

To compare the amount and timing of forming clonal imbalance between the left and right tissues, we used two approaches. First, we compared the mean VAF of early mutations in each subgroup of mutations in each branch by two-sided Wilcoxon test (Fig. 3a, Extended Data Fig. 7a). The timing of occurrence of imbalance was defined as the minimum value of mutation (x-axis) in the phylogenetic tree which shows $P < 0.05$. The number of mutations was converted to the stage of CGs using the estimated mutation rate ($R_{\leq 2}$ and $R_{>2}$; Supplementary Table 1).

Second, we divided early mutations of DB6 into three groups using two thresholds of median VAF of 5% and 0.5% computed from all bulk tissues. Since VAFs of early mutations decrease rapidly as cells undergo divisions, early mutations generally show higher VAFs than late mutations. Thus, early mutations with VAF > 5% indicate the intrinsic barcodes of the earliest cells. We also inferred the stage of CGs of each using the estimated mutation rate. Using the VAF by sample matrix of each group, we performed t-distributed stochastic neighbour embedding (t-SNE) analysis using the 'Rtsne' package in R[64]. In the analysis, we could identify the presence of clusters of anatomical regions and their relative timing of emergence (Fig. 3b). Since the number of informative early mutations was less than for DB6, the other individuals had been grouped into one (DB3 and DB10) or two (DB8 and DB9) groups for t-SNE presentation (Extended Data Fig. 7b).

Tracing the dominant lineage of blood or liver precursor cells

By calculating VAFs of early mutations in high-depth WGS of blood, we estimated the contribution of early cells to the blood cell pool. Since we observed the correlation between lineage proportion in phylogenetic trees and cellular proportion in bulk tissues (Extended Data Fig. 5e), we compared blood VAFs (observed VAF) to the expected VAFs from phylogenetic trees. The observed/expected VAF ratio was plotted by molecular time of mutations in subgroups of mutations in each branch (Extended Data Fig. 8d). The molecular time of the peak ratio was defined as the emergence time of dominant lineage contributing to haematopoietic tissues. We also used the observed/expected VAF ratio to find the major lineage of each hepatic lobule.

We extracted blood-specific mutations to identify the single-cell-derived subclones and their timing of emergence. Variants were called by HaplotypeCaller[54] and included if satisfied all of the following criteria: (1) median mapping quality of variant reads >25, (2) the distance of median mapping quality between reference reads and variant reads <10, (3) minimum mapping quality of reference reads >0, (4) mean base quality of variant reads >20, (5) the number of variant reads >2, (5) the proportion of clipped reads <0.95, (6) the median number of mismatched bases in variant reads <5, (7) the

number of single-cell clones having the same variant read <3. All the mutations constituting the phylogenetic tree of each case were also removed. Thus, the number of final blood-specific mutations reflects the molecular time after diverging from the phylogenetic tree. In this analysis, DB10 was excluded due to the prevalent contamination of tumour cells in the blood.

Discovery of variants in mitochondrial genome

Because the reads sequenced from an inserted mitochondrial sequence in nuclear DNA (the nuclear mtDNA transfer, NUMT) can be misaligned to the mitochondrial reference genome (chrM), we included reads mapped to chrM only if both paired reads (1) mapped to chrM, (2) mapped properly in pair, (3) had read length ≥100 bp, and (4) had no chimeric alignment.

We filtered mitochondrial variants according to the following criteria:

We assigned a VAF cut-off to each mitochondrial variant by considering the VAF distribution locus-by-locus over all the samples. A VAF cut-off of 0.5% appeared to be optimal for the majority of variants (Extended Data Fig. 9a). For each sample, variants with VAF lower than the cutoff were discarded.

The average variant position in supporting reads relative to read length should be between 10% to 90%. Variants that fall outside of this range were filtered out.

Because of the highly repetitive nature of the mitochondrial genome, indel mutations appeared to be more subject to false positives. Therefore, we applied more stringent filtering criteria to indel mutations. For indel mutations to be included, more than half of the supporting reads of an indel mutation should have no additional mutations within 10 base pairs.

Variants that fell in the following regions, which have an extensive level of homopolymers or sequencing error in the reference mtDNA genome (3107N), were explicitly removed:

1. Misalignment due to ACCCCCCCTCCCCC (rCRS 302–315)
2. Misalignment due to GCACACACACACC (rCRS 513–525)
3. Misalignment due to 3107N in rCRS (rCRS 3105–3109)

Simulating mitochondrial heteroplasmy in maternal line

The mtDNA is present in many copies per cell and is inherited through the maternal germline[65]. In addition, a single cell has different mtDNA variants, and the level of heterogeneity (mitochondrial heteroplasmy) can vary considerably between cells. During the production of primary oocytes, a selected number of mtDNA molecules are transferred into each oocyte. Following the rapid replication of the mtDNA population during oocyte maturation, each oocyte has a variable number of heteroplasmic mitochondrial clones[66]. We attempted to infer the number of clones and levels of mitochondrial heteroplasmy in an oocyte using a simplified random-drift segregation model (Extended Data Fig. 9b). This model assumes that (1) a cell has 1,000 copies of mtDNA, and (2) they are partitioned to two daughter cells with an equal amount (500 for each daughter cell), after which (3) all the mtDNA are doubled to reach 1,000 copies of mtDNA in a daughter cell, (4) f% of mtDNA in a fertilized egg has the same variant (MT-mtDNA), (5) the variant is neutral, and therefore MT-mtDNA replicates at the same rate as wild type mtDNA (WT-mtDNA). After fifteen generations of cell division, n_{samples} of cells (the number of clones established in each individual) are randomly selected. Two summary statistics were drawn from the n_{samples} of cells: the proportion of samples harbouring MT-mtDNA (p), the median heteroplasmic level of MT-mtDNA (h). This model simulated 500,000 times for each individual, varying the parameter f drawn from a uniform distribution with a range of [0.001, 1.000]. We compared the summary statistics (p, h) of each simulation to the observed summary statistics, and estimated the posterior distribution of f using the neural network regression algorithm of an approximate Bayesian computation[67].

Mutational signature analysis

We estimated contributions of mutational signatures to an observed mutational spectrum in each sample (that is, the presumed amount of exposure to corresponding mutational processes)[68]. This can be achieved by solving the following constrained optimization problem:

$$\underset{h}{\operatorname{argmin}}\, v - W h_2^2$$

where $v \in \mathbb{R}_{0+}^{m\times 1}$, $W \in \mathbb{R}_{0+}^{m\times k}$, and $h \in \mathbb{R}_{0+}^{k\times 1}$ (m is the number of mutation types and k is the number of mutational signatures). For each sample, given the observed counts of each mutation type $\boldsymbol{v}$ from a sample and a pre-trained mutational signature matrix W, we calculated exposure h. We used an R package (pracma) that internally uses an active-set method to solve the above problem[69].

The number of mutational signatures is a parameter to be determined for each sample. We started with all known mutational signatures available from the version 3 of the COSMIC mutational signature catalogue and examined the results to sort out the most likely set of signatures (available at https://cancer.sanger.ac.uk/cosmic/signatures). To avoid overfitting, we narrowed them down to a handful of signatures by excluding signatures that do not add much to explaining the mutational spectrum up to a certain fitting error (cosine similarity >approximately 0.9). Finally, SBS1, SBS5, and SBS18 were used for the analysis of early mutations, and additionally SBS 7a to 7d for the late mutations. For indels, ID1, ID2, ID4, and ID10 for early mutations and ID5, ID8 and ID9 for late mutations were considered.

We found consistent positive correlation between the number of SBS5 mutations (n_{SBS5}) and SBS7a–d mutations (n_{SBS7}) in individuals. We made a linear model estimating n_{SBS7} (dependent variable) using n_{SBS7} and age ($n_{\text{SBS5}} \sim n_{\text{SBS7}} + \text{age}$). Both n_{SBS7} and age were significant as explanatory variables (both $P < 2\times10^{-16}$), and the estimate for n_{SBS7} (e) was 0.11. We also estimated the number of SBS18 mutations (n_{SBS18}) using the same variables ($n_{\text{SBS18}} \sim n_{\text{SBS7}} + \text{age}$). In this analysis, n_{SBS7} and age were also significant ($P < 2\times10^{-16}$ and 0.012), and the estimate for n_{SBS18} (e') was −0.01. Using the above relationships, we calculated the adjusted number of endogenous mutations (n_{adjENDO}) as below.

$$n_{\text{adjSBS5}} = n_{\text{SBS5}} - n_{\text{SBS7}} \times e$$

$$n_{\text{adjSBS18}} = n_{\text{SBS18}} - n_{\text{SBS7}} \times e'$$

$$n_{\text{adjENDO}} = n_{\text{SBS1}} + n_{\text{adjSBS5}} + n_{\text{adjSBS18}}$$

Linear model estimating n_{adjENDO} by age was more significant than model estimating the number of unadjusted mutations ($n_{\text{SBS1}} + n_{\text{SBS5}} + n_{\text{SBS18}}$) by age (adjusted R^2: 0.46 and 0.17, respectively), reflecting the effectiveness of the adjustment.

Copy number variations and genomic rearrangements

Segmented copy-number profiles were estimated for the whole genome sequenced samples by Sequenza[70] algorithms using matched blood samples as controls. Copy number changes with length >10 Mb were included for further analysis. Subclonal copy-number changes, of which depth ratios were not properly fitted to the integer values of the absolute copy numbers, were excluded manually.

We identified somatic genomic rearrangements of the WGS samples using Delly[71] with an unmatched blood sample as a control not to miss early mutations existed in a matched blood sample. We filtered out the somatic rearrangements in the similar way as our previous report[36]. In brief, we discarded rearrangement events that were present in the all clonal samples or in panel of normal (that is, genomic rearrangements detected in any normal blood samples in our previous study[36]). Rearrangements with poor mapping quality ($Q < 25$), with an insufficient number of supporting reads ($n < 5$), or those with many discordant

reads in matched blood samples were considered to be false positives and were removed. Additionally, we removed short-sized deletions and duplications (<1 kb) without soft-clipped reads and unbalanced inversions (<5 kb) without sufficient supporting reads ($n < 5$), which are mostly DNA library artefacts. To remove remaining false-positive events and to rescue false-negative events located nearby the breakpoints, we visually inspected all the rearrangements passed the filtering process using the Integrative Genomics Viewer[56]. The detailed method and performance of filtering step are available in Supplementary Discussion 9 and Supplementary Table 7.

Timing estimation of copy-number gains

Copy number gains with length > 50 Mb were included in the timing analysis. We estimated mutation copy number (n_{mut}) by a previously described formula shown below[57]:

$$n_{mut} = f_s \frac{1}{\rho}[\rho n^t_{locus} + n^n_{locus}(1-\rho)]$$

where f_s, ρ, n^t_{locus}, and n^n_{locus} denote VAF, tumour cellularity, absolute copy number in tumour and absolute copy number in normal cells, respectively. In our experimental design, the formula will be simplified as follows because $\rho = 1$ (single-cell derived clone):

$$n_{mut} = f_s n_{locus}$$

Here, n_{locus} is given by

$$n_{locus} = 2 \times \frac{RD_{locus}/Cov_{sample}}{RD_{blood}/Cov_{blood}}$$

where RD_{locus} and RD_{blood} indicate the read depth of the locus of interest in a sample (mean coverage of Cov_{sample}) and paired bulk blood sample (mean coverage of Cov_{blood}), respectively.

To assess the temporal relationship between a specific somatic point mutation and chromosomal gains, we compared the mutation copy number (n_{mut}) with the allele-specific major copy number (n_{major})of the chromosomal segment. Mutations with $n_{mut} \geq n_{major}$ were classified as a pre-amplification event, while the others as post-amplification or minor allele mutations. The probabilities of pre-amplification (P_{pre}), post-amplification (P_{post}) were evaluated with the following binomial distributions:

$$B_{pre} = \binom{DP}{VC}\left(\frac{n_{major}}{n_{locus}}\right)^{VC}\left(1-\frac{n_{major}}{n_{locus}}\right)^{DP-VC}$$

$$B_{post} = \sum_{i=1}^{n_{major}-1}\binom{DP}{VC}\left(\frac{i}{n_{locus}}\right)^{VC}\left(1-\frac{i}{n_{locus}}\right)^{DP-VC}$$

where DP and VC indicate the total read depth and variant read count, respectively. Finally, P_{pre} and P_{post} are given by:

$$P_{pre} = \frac{B_{pre}}{B_{pre} + B_{post}}$$

$$P_{post} = \frac{B_{post}}{B_{pre} + B_{post}}$$

In order to estimate the timing of amplification events, we analysed the expected number of pre-amplification substitutions (E_{pre}) by summation of the P_{pre} values of all substitutions in amplified segments larger than 50 Mb. The 95% confidence interval of E_{pre} was evaluated from z-scores using the sum of $P_{pre} \times (1 - P_{pre})$ as a variance. The expected number of post-amplification substitutions (E_{post}) was also calculated in the same way. Then, age multiplied by the proportion of E_{pre} among clonal mutations ($E_{pre} + E_{post}$) were used as chronological timing of copy-number gains.

Reporting summary

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

Data availability

Whole-genome and targeted sequencing data are deposited in the European Genome-phenome Archive (EGA) with accession EGAS00001004824 and are available for general research use.

Code availability

The information of sequenced clones and tissues, detected early mutations, and their anatomical tracking can be browsed through Somatic Clone Viewer (https://julab.kaist.ac.kr/somatic-clone-viewer). In-house scripts for genomic analyses and simulation studies are available on GitHub (https://github.com/seongyeol-park/Human_Lineage_Tracing and https://github.com/chrono0707/Human_Lineage_Tracing).

47. Jones, G. E. & Wise, C. J. Establishment, maintenance, and cloning of human dermal fibroblasts. *Methods Mol. Biol.* **75**, 13–21 (1997).
48. Rittié, L. & Fisher, G. J. Isolation and culture of skin fibroblasts. *Methods Mol. Med.* **117**, 83–98 (2005).
49. Vangipuram, M., Ting, D., Kim, S., Diaz, R. & Schüle, B. Skin punch biopsy explant culture for derivation of primary human fibroblasts. *J. Vis. Exp.* **(77)**, e3779 (2013). https://doi.org/10.3791/3779.
50. Spinazzola, J. M. & Gussoni, E. Isolation of primary human skeletal muscle cells. *Bio Protoc.* **7**, e2591 (2017).
51. Oh, J. W. et al. A guide to studying human hair follicle cycling in vivo. *J. Invest. Dermatol.* **136**, 34–44 (2016).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
54. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**, 11.10.1–11.10.33 (2013).
55. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
56. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
57. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
58. Gusfield, D. Efficient algorithms for inferring evolutionary trees. *Networks* **21**, 19–28 (1991).
59. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. GGTREE: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* (2016).
60. Strnad, P. et al. Inverted light-sheet microscope for imaging mouse pre-implantation development. *Nat. Methods* **13**, 139–142 (2016).
61. Bertorelle, G., Benazzo, A. & Mona, S. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol. Ecol.* **19**, 2609–2625 (2010).
62. Zernicka-Goetz, M., Morris, S. A. & Bruce, A. W. Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. *Nat. Rev. Genet.* **10**, 467–477 (2009).
63. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
64. Krijthe, J. H. Rtsne: *t*-distributed stochastic neighbor embedding using a Barnes–Hut implementation (R package, 2015).
65. van den Ameele, J., Li, A. Y. Z., Ma, H. & Chinnery, P. F. Mitochondrial heteroplasmy beyond the oocyte bottleneck. *Semin. Cell Dev. Biol.* **97**, 156–166 (2020).
66. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.* **6**, 389–402 (2005).
67. Hickerson, M. J., Stahl, E. & Takebayashi, N. msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. *BMC Bioinformatics* **8**, 268 (2007).
68. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
69. Borchers, H. W. pracma: practical numerical math functions (R package, 2019).
70. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
71. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).

Acknowledgements We are deeply indebted to the individuals who donated their cells and tissues for this study. We thank M. R. Stratton, P. J. Campbell, T. H. H. Coorens, R. Rahbari, L. Moore, A. Cagan and M. V. Plikus for their fruitful comments and discussions. We thank J.-Y. Shin, M. S. Jun, H. Jung, J. H. Lee, H. S. Lee, J. Y. Jeon, J. H. Jeon, S. Cho and J. S. Lee, for their methodological advice and technical assistance. This work was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health and Welfare of Korea (HI17C1836 to Y.S.J.), the Suh Kyungbae Foundation (SUHF-18010082 to Y.S.J.), a National Research Foundation (NRF) of Korea funded by the Korean Government (NRF-2020R1A3B2078973 to Y.S.J.; NRF-2019R1I1A3A01060675, NRF-2020R1A5A2017323 and NRF-2021R1C1C1014425 to J.W.O.; NRF-2020R1A6A3A01100621 to S.P.; and NRF-2019H1D3A2A02061168 to S.Y.K.).

Author contributions Y.S.J., J.W.O. and S.P. conceived the study; N.M.M. designed the warm autopsies and developed the entire protocol of the clonal expansion and bulk tissue preparation with help from J.W.O.; J.W.O., N.M.M., J.-W.C, J.M.P., D.K., J.H.C., S.G.K., J.H.H., M.K., D.S.K., J.Y.P., K.Y., T.K., J.Y., and Y.A. conducted autopsies, tissue sampling and clonal expansions. S.A.O. conducted DNA work. S.P. and R.K. conducted most of the genome and statistical analyses with a contribution from S.Y.K. and Y.S.J.; J. Lee and J.W.P. contributed to large-scale genome data management. J. Lim conducted mutational signature analysis. Y.K. and D.H. constructed the web tool (SCV). Y.S.J., S.P., R.K., N.M.M and J.W.O. wrote the manuscript with contributions from all the authors. Y.S.J. and J.W.O. supervised the study.

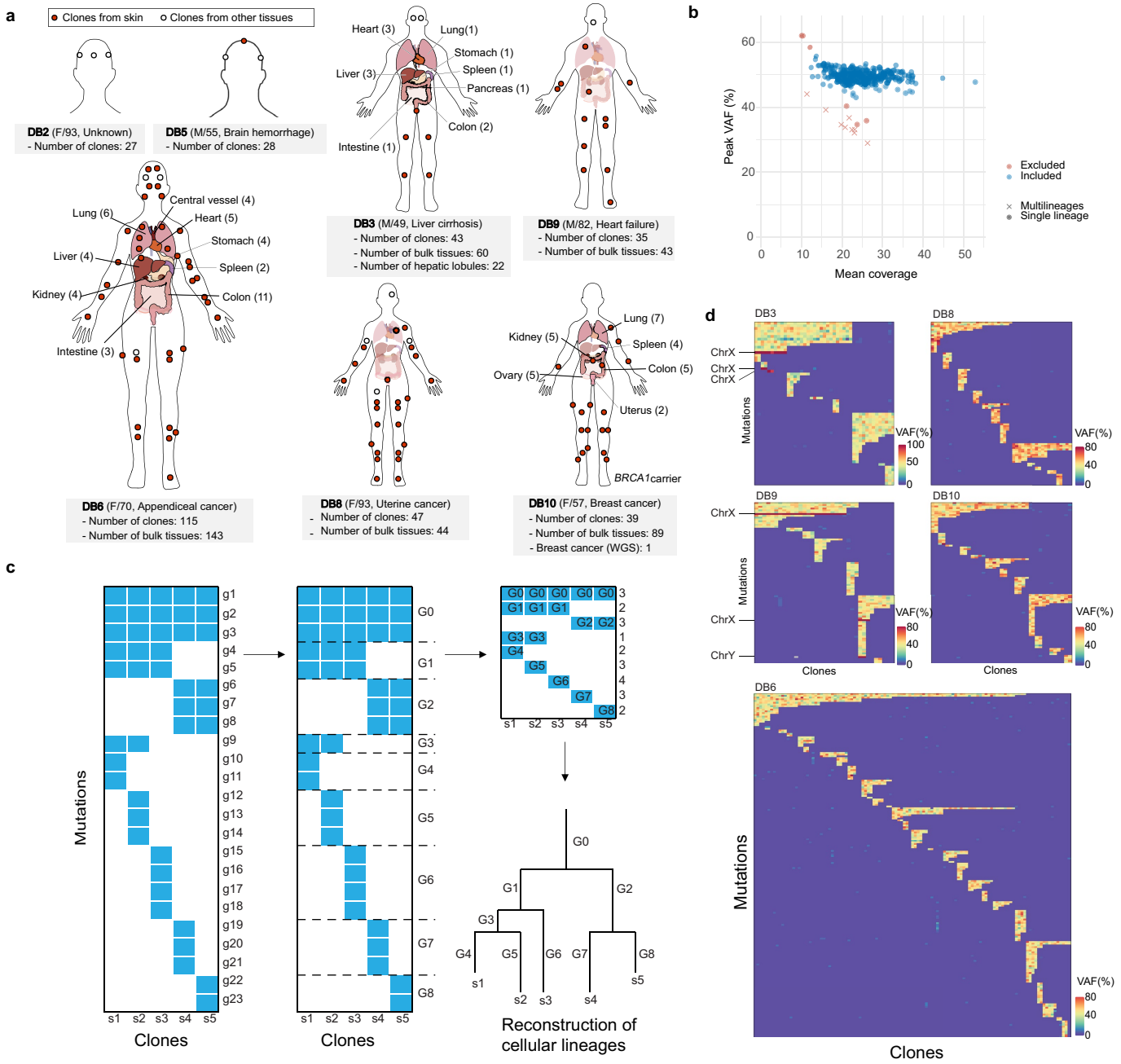
Competing interests Y.S.J. is a founder and chief executive officer of GENOME INSIGHT Inc. J.W.O. is a founder and chief executive officer of Immune Square Inc.

Additional information

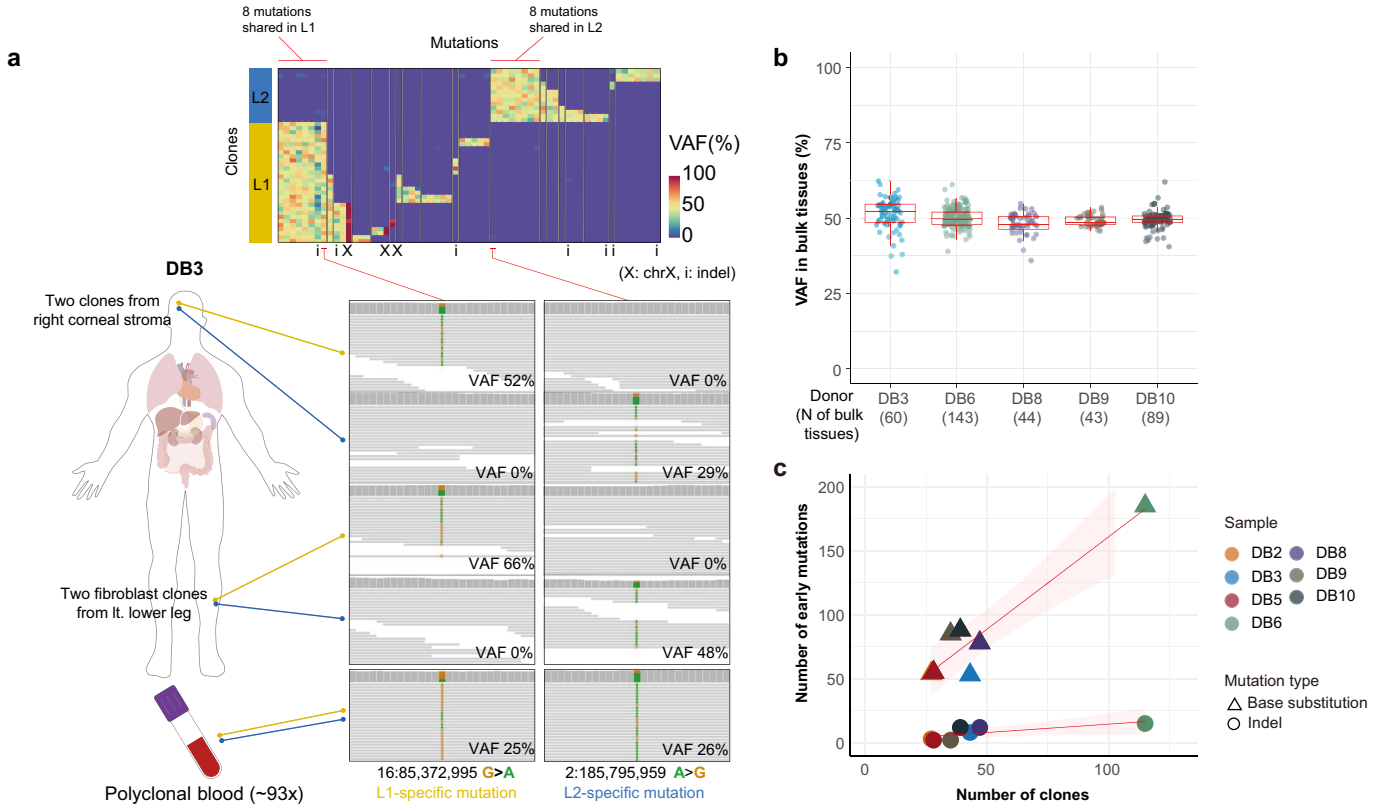
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03786-8.

Correspondence and requests for materials should be addressed to J.W.O. or Y.S.J.

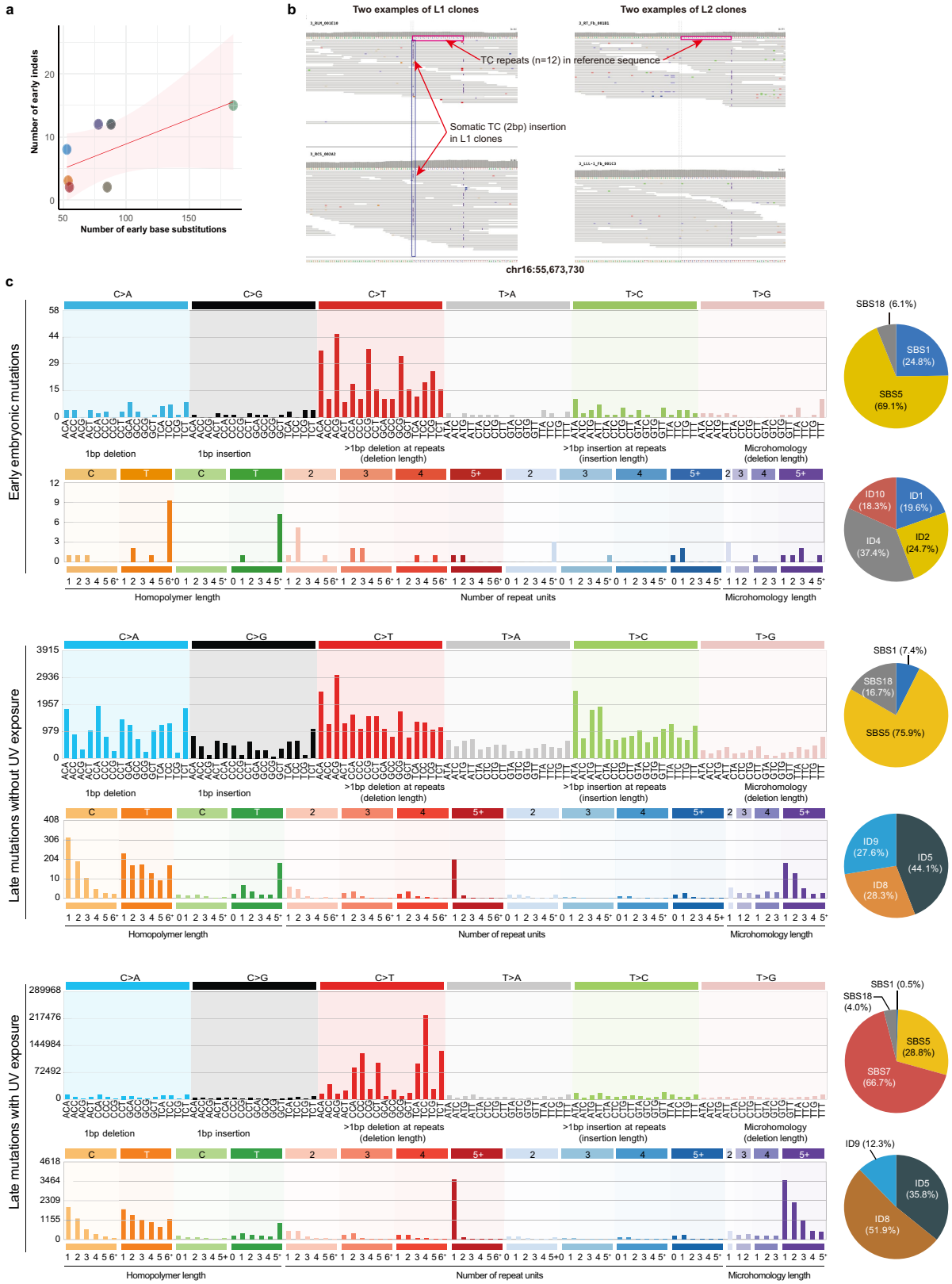
Peer review information *Nature* thanks Chloé Baron, Aaron Mckenna and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



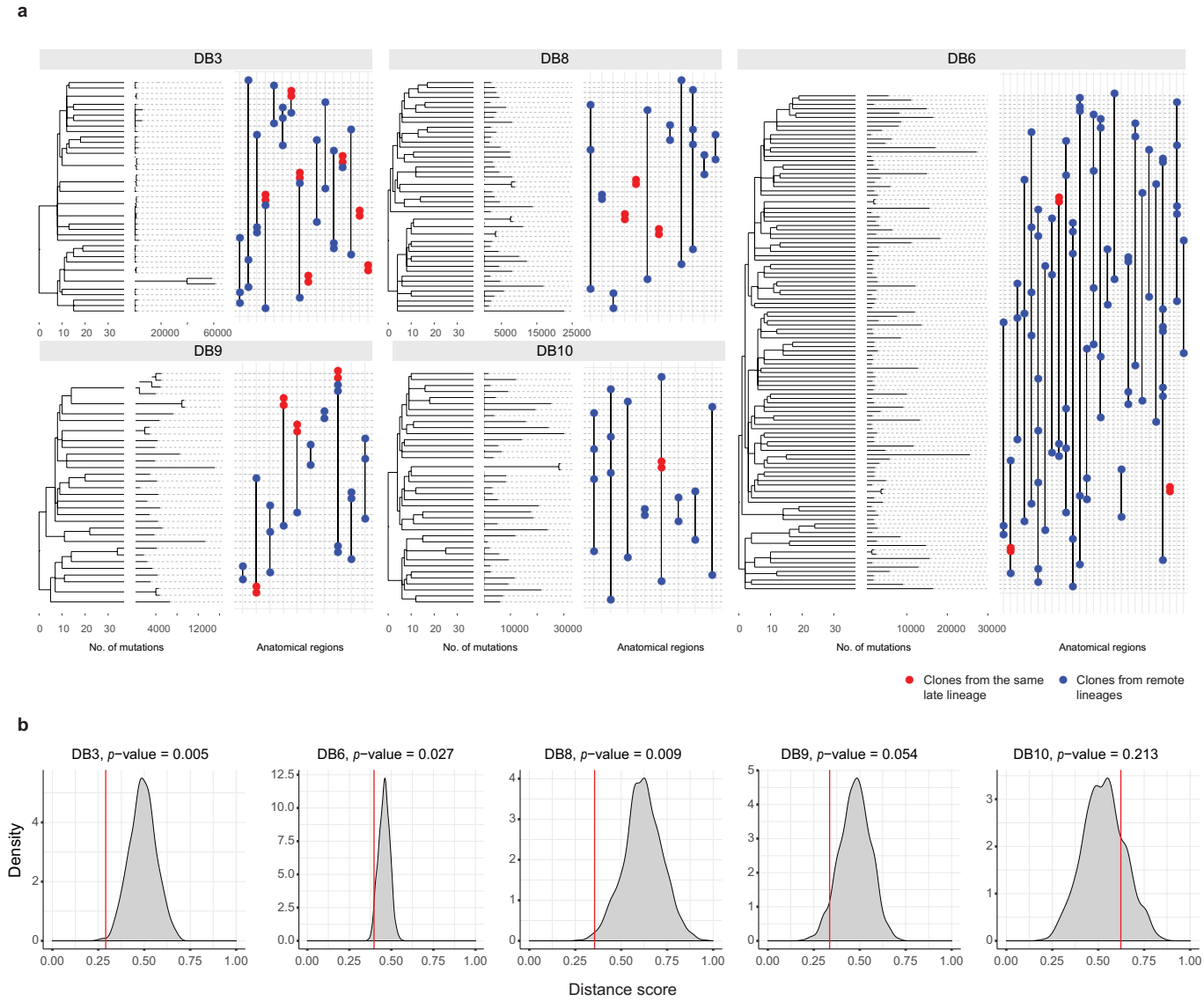
Extended Data Fig. 1 | Sample collection and phylogeny reconstruction. **a**, A summary of the seven warm autopsies, in which clones (dots) and bulk tissues were collected. Internal organs contributing to bulk tissues are annotated (black line). The information can also be browsed in the Somatic Clone Viewer (https://julab.kaist.ac.kr/somatic-clone-viewer). The detailed information for each sample is shown in Supplementary Discussion 1, Supplementary Tables 2 and 3, for clones and bulk-tissues, respectively. **b**, A scatter plot showing peak VAF and mean coverage of the WGS of all the clones established in this study. Excluded clones, due to their multiclonal origins and/or atypical VAF peaks, are coloured in red. Finally, 334 clones are included for the downstream analyses. **c**, A schematic illustration demonstrating our approach reconstructing a developmental phylogenetic tree. Let $S=[s_1, s_2, \ldots, s_5]$ be the set of 5 clones, and $G=[g_1, g_2, \ldots, g_{23}]$ be the union set of mutations detected in one or more clones from the same individual. We then build a matrix M with rows labeled $g_1, g_2, \ldots, g_{23}$, and columns labeled $s_1, s_2, \ldots, s_5$. If the VAF of somatic mutation g_i in clone s_j is determined to be true, M_{ij} was assigned to 1 (blue-coloured tile), while others to 0 (white-coloured tile). After removing germline variants (G0), we grouped all mutations with the same profile into a mutation group according to the sharing pattern between clones. Over 8 distinct mutation groups in this example, mutation matrix $M_{8\times5}$ is defined such that each column represents a clone and each row represents a mutation group. From the mutation matrix $M_{8\times5}$, we reconstructed a phylogenetic tree. **d**, Mutation matrices constructed from all clones and all the detected embryonic mutations of the five individuals.



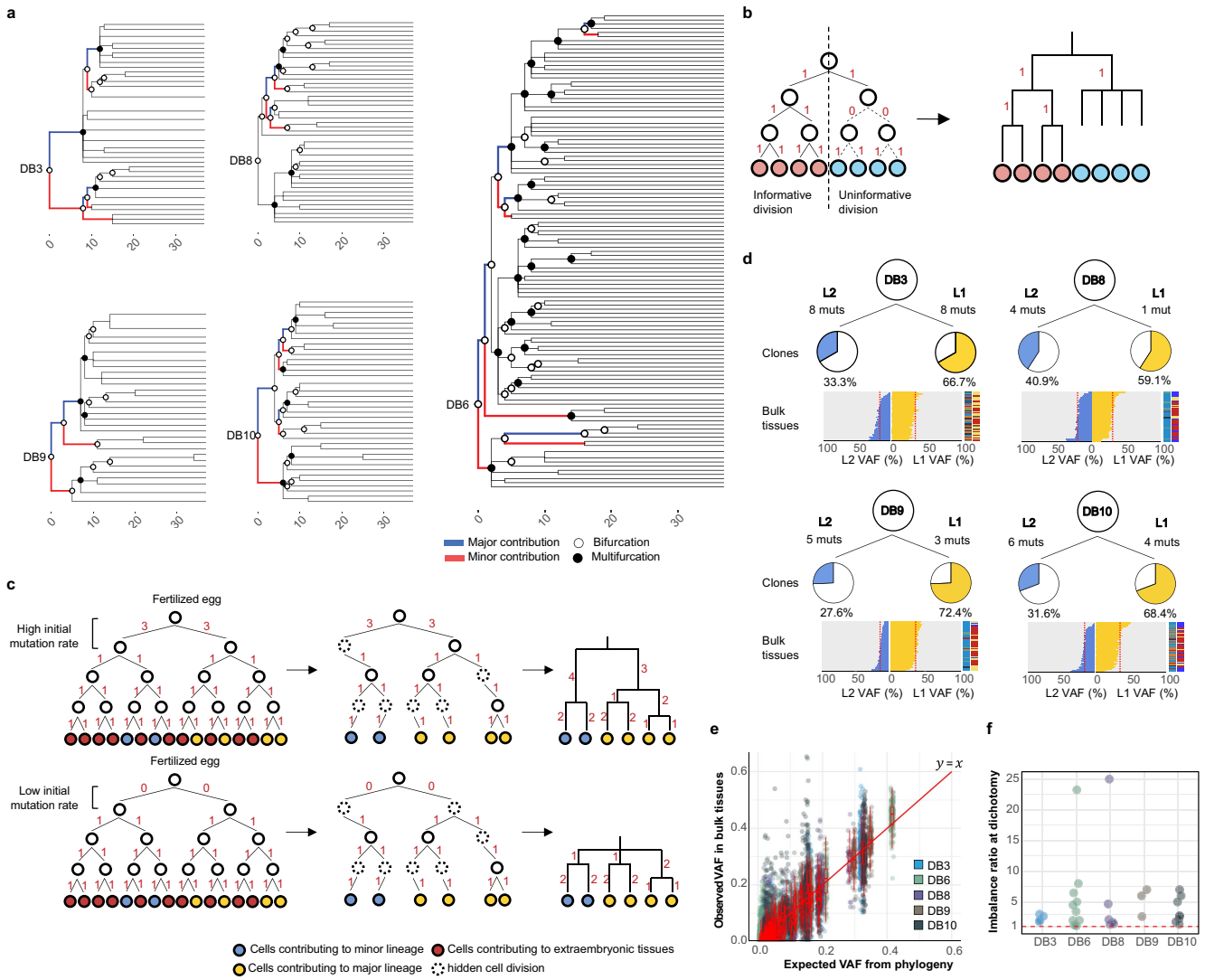
Extended Data Fig. 2 | Detecting EEMs. a, Examples of embryonic mutations found in DB3. A heatmap in the upper panel shows the VAFs of the early mutations detected in the capture phase. Integrative Genomics Viewer screenshots for two early mutations in WGS of four clones and a polyclonal blood are also shown in the lower panel. **b**, The aggregated VAFs for L1 and L2 mutation sets in bulk tissues (dot). Median and interquartile ranges (IQRs) are shown in boxplots with whiskers (1.5*IQRs). **c**, A scatter plot showing the number of clones established and the number of discovered early mutations. The number of base substitutions (triangle) and indels (circle) are shown separately. Red lines and shaded areas represent fitted lines from linear regression and 95% confidence intervals.



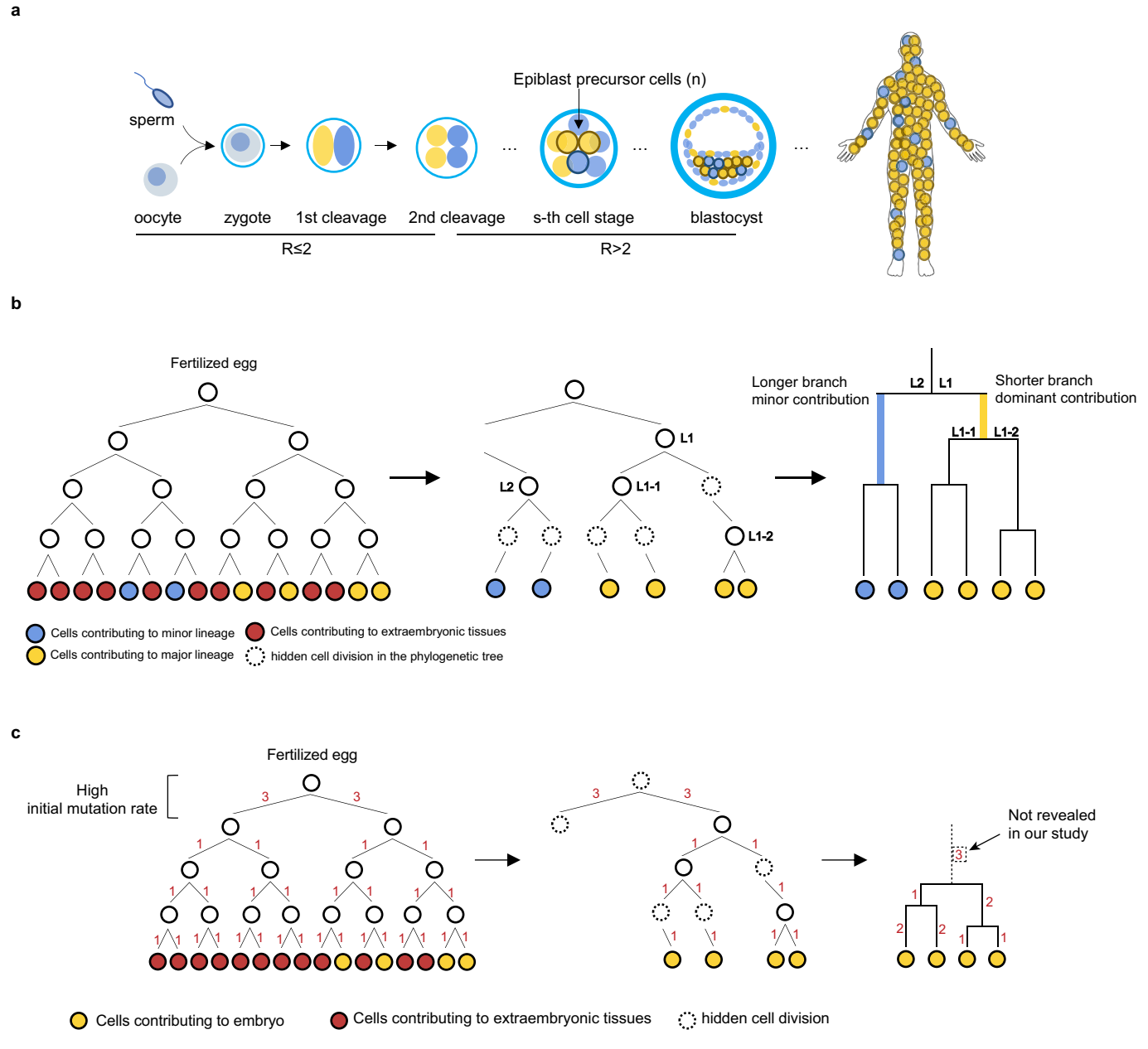
Extended Data Fig. 3 | Features of EEMs. a, The correlation between the number of early base substitutions and indels. A red line and shaded area represent fitted line from linear regression and 95% confidence intervals. **b**, An example of microsatellite length-changing mutation identified in the L1 branch of DB3. In this study, microsatellite regions were defined as 5 or more repeat of 1-6 nucleotides in the reference genome. **c**, Mutational spectrums of early and late mutations found in the study. The signatures of the early base substitutions (n=488) and indels (n=49) are delineated by version 3 COSMIC signatures (top). For late mutations, we categorized clones into two groups by the amount of ultraviolet (UV) light mediated mutations, (1) clones with prevalent UV-mediated mutations and (2) clones with lack of UV-mediated mutations, using the 5% cutoff for the proportion of the SBS7 mutations. The middle panel displays the mutational spectrums of base substitutions (n=74,824) and indels (n=3,404) in clones with SBS7 ≤5% (without UV exposure), while the bottom panel displays the spectrums of substitutions (n=1,457,489) and indels (n=31,805) in clones with SBS7 >5% (with UV exposure).



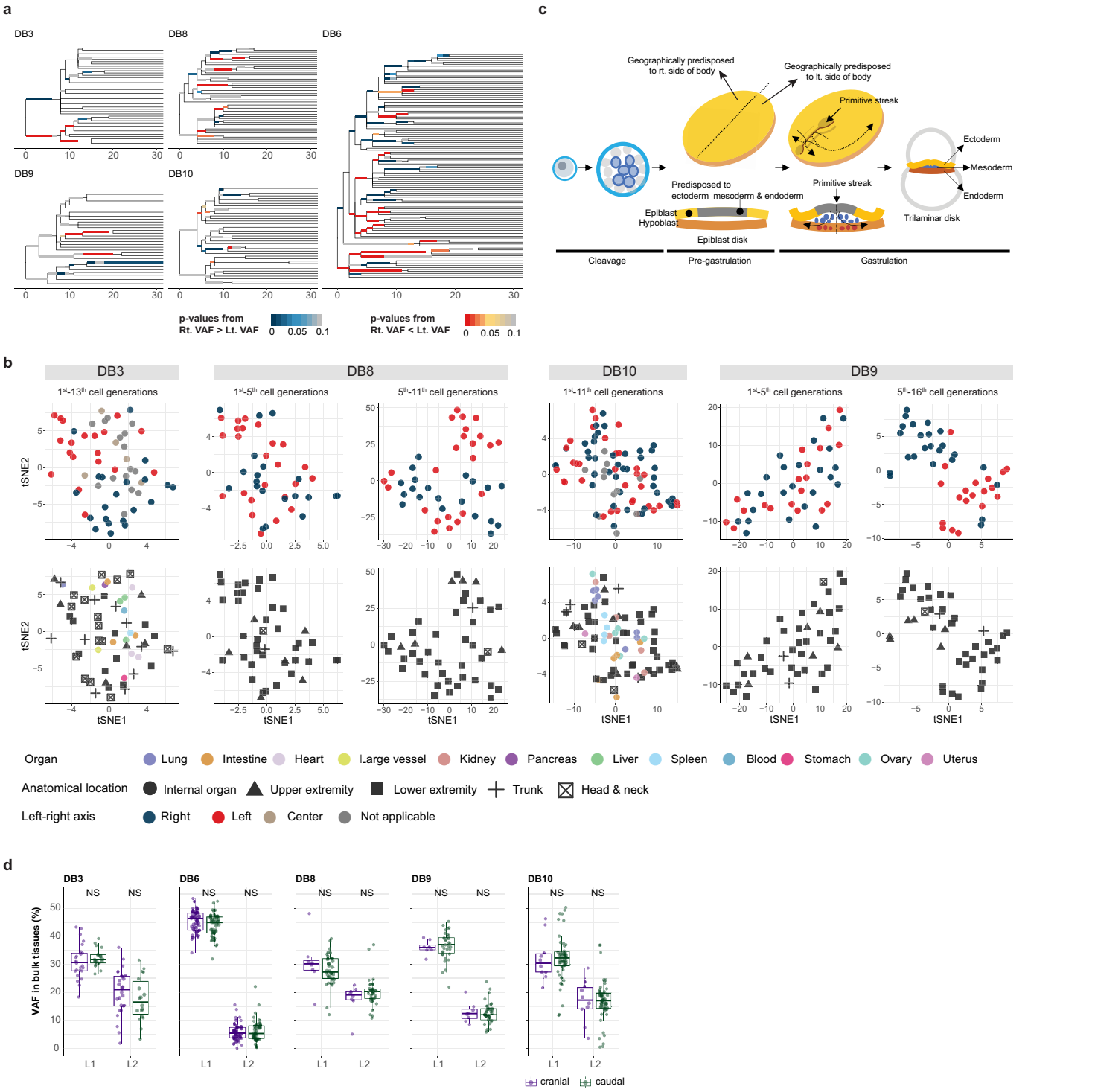
Extended Data Fig. 4 | Lineage relationship among physically adjacent clones. a, Lineages of physically adjacent clones established from <1cm of the distance are shown with early phylogenies. Clone pairs from the same late lineage are coloured in red. **b**, The distribution of distance scores between clones from same anatomical region. Distance score was calculated by 1/(No. of shared mutations +1). The random distributions (density plots) were generated by randomly assigning clones to lineages on the established phylogenies. A red line represents the actual mean distance score the clones. Empirical p-values from simulation (n=1,000) are shown.



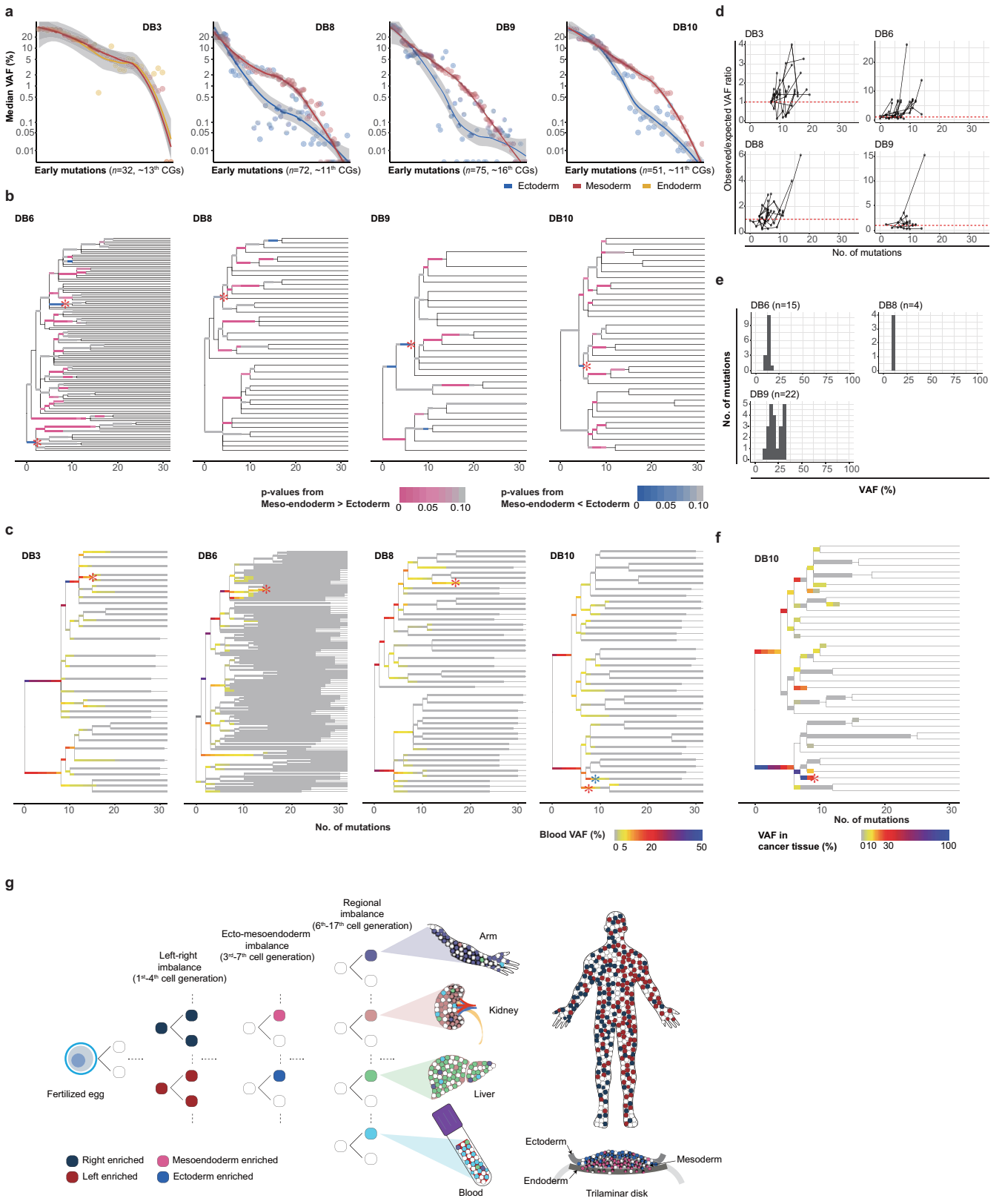
Extended Data Fig. 5 | The patterns of early phylogenies. a, Annotated phylogenetic trees of DB3, DB6, DB8, DB9, and DB10. Dichotomy (bifurcation) and polytomy (multifurcation) nodes are indicated by black-filled and hollow circles, respectively. At bifurcation nodes, two daughter lineages are then coloured in red (major) or blue (minor), according to their relative contribution in phylogenies. **b**, A schematic illustration demonstrating informative and uninformative cell divisions. In contrast to a cell division accompanying spontaneous mutations (informative division, left of dashed line), a cell division without intrinsic mutation cannot be reflected in our phylogenetic tree (uninformative division, right of dashed line) due to a lack of 'cellular barcodes'. **c**, A schematic illustration showing the effect of initial mutation rate on the pattern of the trees. **d**, The unequal contribution of the two earliest branches (L1 and L2) consistently found in phylogenies and bulk tissues. Pie graphs represent the proportion of each lineage counted in the phylogenetic trees. Horizontal bar graphs show the VAFs of the lineage-specific mutations in targeted sequencing of bulk tissues. Expected VAFs from the phylogenies are shown by red dashed lines. **e**, A correlation between VAFs of the early mutations expected in the phylogenies (*x*-axis) and observed in bulk tissues (*y*-axis). Median and interquartile ranges (IQRs) are shown in boxplots with whiskers (1.5*IQRs). A red line drawing shows *y*=*x* for comparison. **f**, The imbalance ratio at bifurcating nodes, which is the ratio between the numbers of the clones of major and minor lineages. The late-branched clones from the same lineage were counted as a single clone.



Extended Data Fig. 6 | Cellular bottleneck and phylogenetic tree. **a**, A developmental model showing the lineage imbalance in epiblast as an origin of global unequal L1 and L2 contribution. This model assumes the number of cells (*n*) is selected for epiblast at s-cell stage. We presume two different mutation rates ($R_{\leq 2}$ and $R_{>2}$) in early embryogenesis. $R_{\leq 2}$, $R_{>2}$ are mutation rates until and after 2-cell stage, respectively. **b**, A cellular genealogy scenario that can explain the features of early phylogenetic trees. Assuming that the mutation rate is constant, a longer branch results from the lineage that contribute less to the embryo. **c**, Impact of stochastic cellular segregation during embryogenesis on early developmental phylogeny. This illustration shows the consequence of biased selection on cellular phylogeny. In a case that all of the epiblast cells are derived from one cell in two-cell stage, mutations accumulated at first division are shared among all embryonic cells. High initial mutation rate could be masked in this situation.

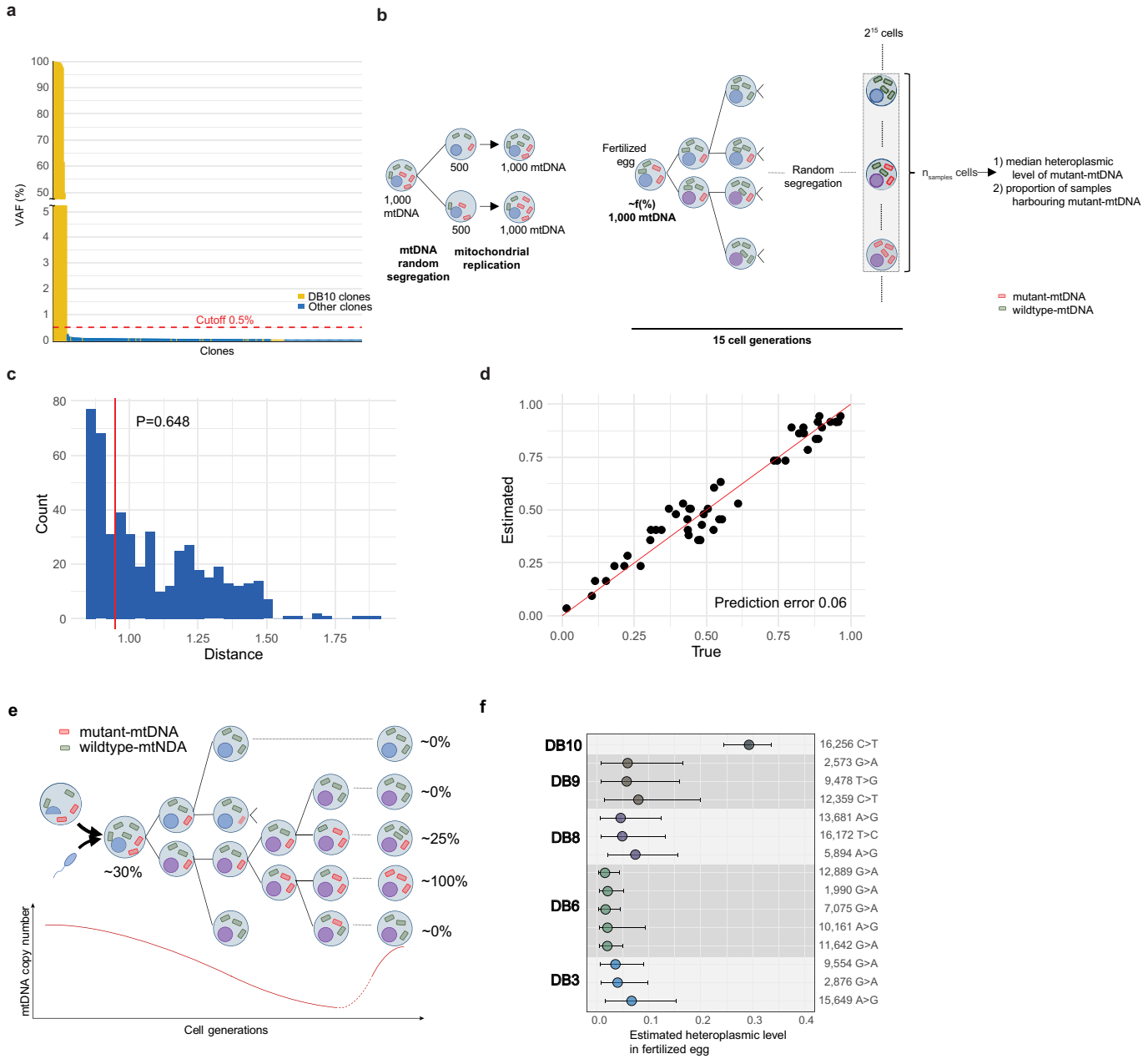


Extended Data Fig. 7 | Imbalanced distribution of early embryonic cells in adult body axes. a, The phylogenetic trees of five individuals coloured by the significance levels of imbalances between the left and right tissues. Mutations with higher VAF in left tissues are coloured in red, while with higher VAF in right tissues are coloured in blue. The left-right clonal imbalances are frequently seen from very early lineages (<10 early mutations) in all the individuals. **b**, t-SNE clustering of bulk tissues using the VAFs of the embryonic mutations for the four individuals (DB6 is available in Fig. 3b). The separation of left and right tissues is shown in top panels. Other information, such as anatomical regions and organs, is shown in bottom panels. **c**, A schematic illustration showing the emergence of primitive streak and formation of left-right axis in early embryogenesis. **d**, Box plots illustrating the median VAFs and interquartile ranges (IQRs) with whiskers (1.5*IQRs) of L1 and L2 mutations in cranial or caudal tissues. There was no significant VAF difference between cranial- and caudal-side tissues (two-sided Wilcoxon test). NS, not significant.

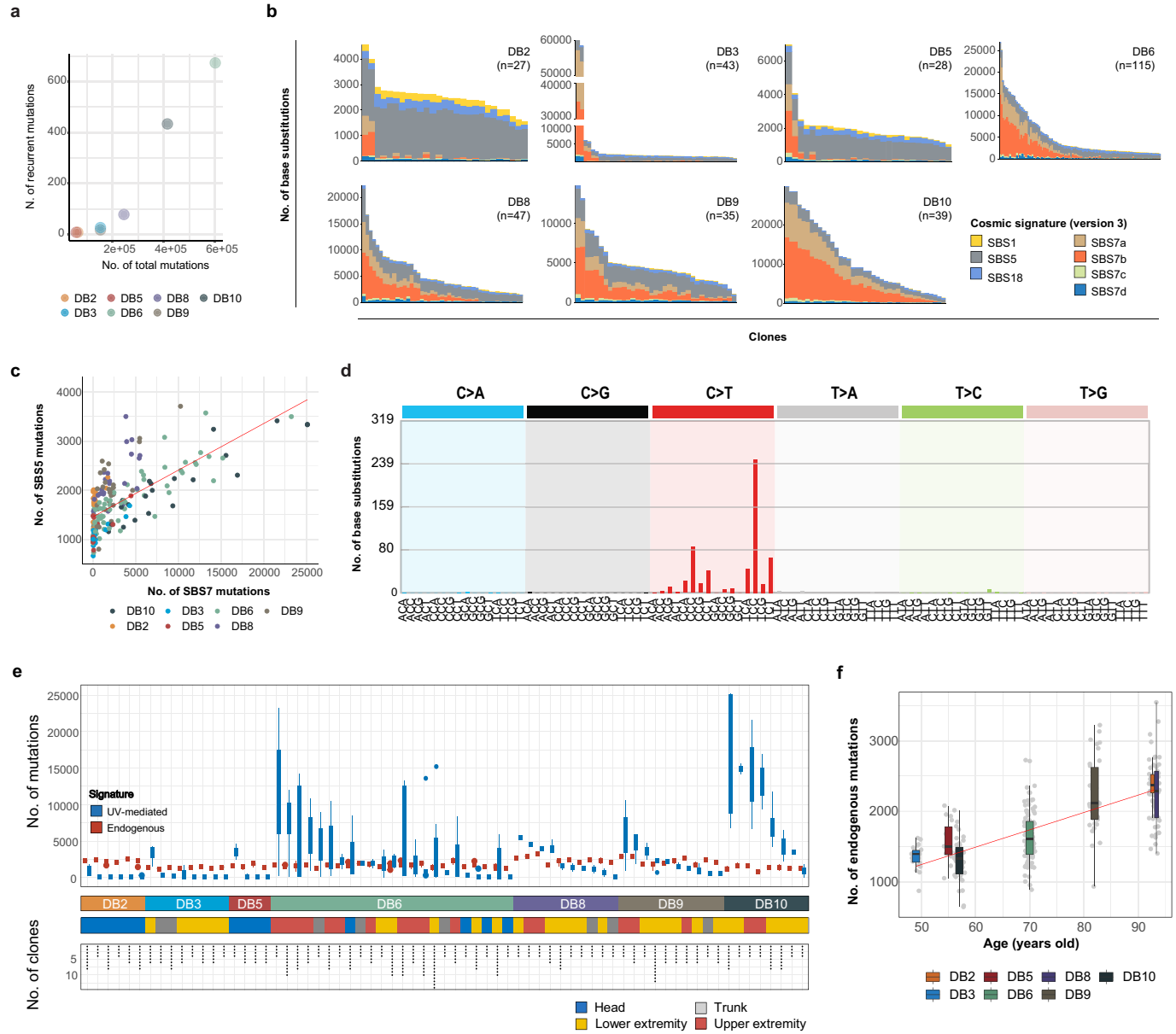


Extended Data Fig. 8 | See next page for caption.

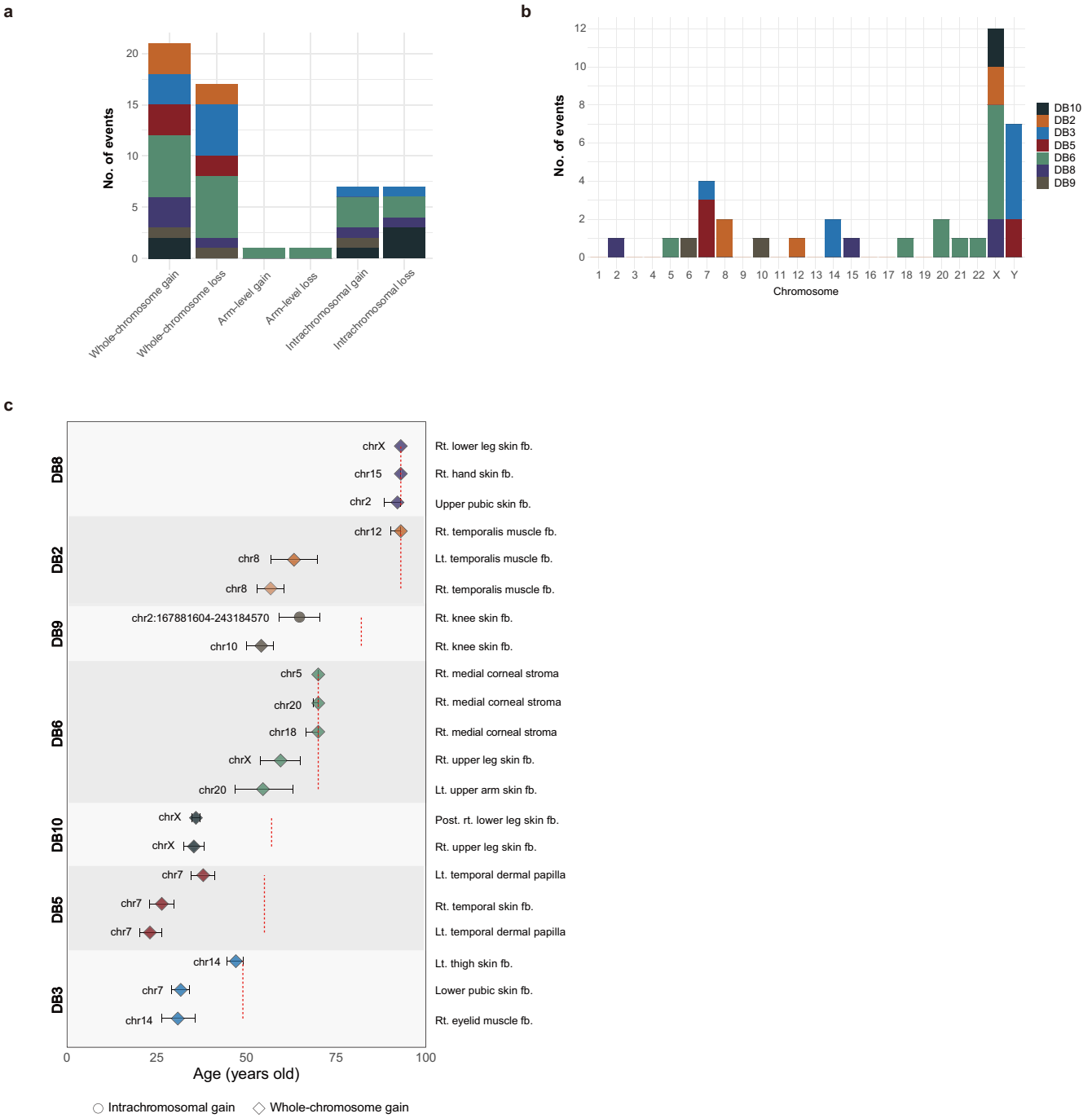
Extended Data Fig. 8 | Imbalanced distribution of early embryonic cells in adult anatomical regions. a, Median VAFs of the early mutations in the bulk tissues according to their dominant germ layers. The horizontal axis shows early mutations sorted by the averaged VAFs in bulk tissues in descending order, approximately from earlier to later mutations. Tissues with mixed germ layers are excluded in this figure. The lines are fitted curves by locally estimated scatterplot smoothing (LOESS) methods. **b**, The phylogenetic trees coloured by the significance of imbalances between ectoderm and meso-endodermal tissues. DB3 was not suitable for the analysis due to a lack of ectodermal tissues sequenced. Comparisons were performed by two-sided Wilcoxon tests. Red asterisks indicate the estimated point of the branching of the ectoderm-dominant lineage. **c**, the phylogenetic trees of four individuals (DBs 3, 6, 8, and 10) coloured by the VAF in blood tissues. An equivalent figure for DB9 is shown in Fig. 3d. Red asterisks indicate the estimated point of the branching of the blood-enriched lineages. Blue asterisk in DB10 indicate the major lineage of contaminated tumour cells in blood. **d**, Ratios of VAF for embryonic mutations between observed in blood tissues and expected from phylogenetic trees. The molecular time of the embryogenesis is shown on the *x*-axis by the number of mutations. Dots in the direct lineages are linked by lines. **e**, Histograms for showing the number of the blood-specific mutations (absent in phylogenetic trees). Blood-specific mutation was not found in DB3 (n=0). DB10 is excluded due to tumour contamination in the blood. **f**, The phylogenetic tree of DB10 coloured by VAF in cancer tissue with the branching point of the ancestral cell of breast cancer (shown by a red asterisk). **g**, Schematic representation demonstrating the clonal imbalance and their timing in early embryogenesis.



Extended Data Fig. 9 | Heteroplasmic mtDNA variants in fertilized eggs. **a**, VAFs of MT:16,256 C>T substitution (frequently found in DB10 clones) in WGS of 279 clones explored in the study. Applying a VAF cutoff of 0.5%, the variant was detected only in the 14 clones all established from DB10. **b**, a developmental model for inferring the heteroplasmic level of a mitochondrial variant in a fertilized egg. We assumed that *f*% of mtDNA in a fertilized egg has a functionally neutral mtDNA variant (mutant-mtDNA), which randomly segregates to daughter cells in successive cell divisions. Two summary statistics were drawn from this model: 1) the proportion of samples harbouring mutant-mtDNA (*p*), and 2) the median heteroplasmic level of mutant-mtDNA (*h*). We compared the summary statistics (*p*, *h*) of each simulation to the observed summary statistics, and constructed the posterior distribution of *f* using the neural network regression algorithm of an approximate Bayesian computation. For detail, see the Methods section. **c**, a histogram of the null distribution of the statistic for the goodness of fit test assuming our model. **d**, Cross-validation to access the accuracy of parameter inference. **e**, a possible scenario underlying the recurrent mtDNA mutation. Mitochondrial bottleneck during the cleavage and random segregation of mtDNA during mitosis may underlie the early mtDNA variant. **f**, mtDNA variants and their heteroplasmic levels with 95% confidence intervals estimated by simulation (n=500,000) to be harboured in fertilized egg.



Extended Data Fig. 10 | Features of late mutations. a, Scatter plot showing the correlation between the number of total mutations and recurrent mutations. **b**, Total number and signature of somatic mutations in each of the 334 clones. Horizontal axis represents each clone in decreasing order of total mutation numbers. **c**, Linear correlation between the numbers of SBS7 (UV-mediated) and SBS5 (an endogenous, clock-like) mutations in skin fibroblast clones. Approximately one additional SBS5 mutation is acquired per ten SBS7 mutations. A red line represents the result of linear regression. **d**, The mutational spectrum of late recurrent base substitutions (n=619). **e**, A massive heterogeneity of UV-mediated mutational burden among clones established in the close anatomical location (inter-clonal distance < ~1cm; top). The number of clones in each location is illustrated at the bottom. **f**, The rate of purely endogenous mutations in skin fibroblasts showing a linear correlation with age (24.3 substitutions per year). A red line represents the result of linear regression. The median number of endogenous substitutions with interquartile ranges (IQRs) from clones are drawn by boxplots (whiskers=1.5*IQRs) and scatter plots.



Extended Data Fig. 11 | Somatic copy-number changes in normal cells. a, Bar plot showing the frequency of large-scale copy-number alterations (> 10 Mb) per segment type detected in clones. **b**, Bar plot showing the counts of whole-chromosomal copy-number changes per chromosome. **c**, Timing estimation of the copy-number gains (> 50 Mb) observed in the clones. Ages at death are shown in red dashed lines. fb., fibroblast.

nature portfolio

Corresponding author(s): Young Seok Ju, Ji Won Oh

Last updated by author(s): Jun 19, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	All the sequences were generated by Illumina sequencer (Hiseq X Ten or Novaseq 6000)
Data analysis	Sequenced reads were mapped to the human reference genome (GRCh37) using the BWA-MEM (v0.7.17-r1188) algorithm. The duplicated reads were removed by Picard (v2.1.0), and indel realignment and base quality score recalibration were performed by GATK (v4.0). Initially, we identified base substitution and short indels by using HaplotypeCaller (v4.0) and VarScan2 (v2.3.9). We identified somatic genomic rearrangements of the whole genome sequenced samples using Delly (v0.7.6). Segmented copy number profiles were estimated for the whole genome sequenced samples by Sequenza (v3.0.0). Detected variants were inspected using Integrative Genomics Viewer (IGV v2.8.7). Custom codes were written by Python (v3.7.6) and R (v3.6.0). Following packages were used in R: ggtree (v1.16.6), abc (v2.1), Rtsne (v0.15), and pracma (v2.2.9). All custom scripts are available at GitHub (https://github.com/seongyeol-park/Human_Lineage_Tracing and https://github.com/chrono0707/Human_Lineage_Tracing).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Whole-genome and targeted sequencing data have been deposited in the European Genome-phenome Archive (EGA) with accession id: EGAS00001004824.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine the sample size. We acquired various number of single-cell clones from individuals, and we estimated the latest developmental time which can be revealed by our dataset.
Data exclusions	Among a total of 374 clonally expanded cells, 18 with low depth of coverage (mean depth < 10) were excluded. Based on the VAFs of somatic mutations and established phylogenetic trees, we removed additional 19 samples thought to be multiclonal: they have variants in mutually exclusive lineages simultaneously, and/or low peak VAFs. In addition, three samples which show unexplainable high peak VAFs were excluded for data integrity. Finally, we analyzed the remaining 334 whole-genome sequencing data from single-cell derived clones to reconstruct phylogenetic trees. In targeted-deep sequencing, we only used the mutations with the sufficient read depth for the accurate analysis. The exclusion criteria were not pre-established. We determined the criteria based on the mutation sharing pattern and certainty of lineage assignment.
Replication	Five biological replicates (culturing the sample twice from three samples, and separating cultured cell pool into two from two samples) and two technical replicates (sequencing the DNA pool twice) were made. All attempts at replication were successful, and all the detected early mutations were validated in the replicates.
Randomization	Not applicable since there was no predetermined group of samples. All possible samples were used.
Blinding	Not applicable since this is a descriptive study. Based on the assumption that all the sample information (e.g. anatomical position) is accurate, we traced early embryogenesis using both genomic data and sample information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	We analyzed cells and tissues from seven donors. Age, gender, and cause of death were described at Supplementary Table 1.

Recruitment	From June 2016 to January 2019, ten immediate autopsies had been performed at the Kyungpook National University Hospital and Department of Anatomy at Kyungpook National University, School of Medicine after informed consent. In seven of them, primary culture of tissues was successful and the donors were enrolled in this study. Since most donors were hospitalized before death, cause of death in our cases could be biased.
Ethics oversight	All the procedures related to warm-autopsy and tissue sampling were approved by Institutional Review Board of Kyungpook National University (KNU-2018-0088 and KNU-2019-0151) and KAIST (KH2020-029).

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