Population dynamics of normal human blood inferred from somatic mutations

Henry Lee-Six, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, Brian J. P. Huntly, Inigo Martincorena, Elizabeth Anderson, Laura O’Neill, Michael R. Stratton, Elisa Laurenti, Anthony R. Green, David G. Kent, Peter J. Campbell

Haematopoietic stem cells drive blood production, but their population size and lifetime dynamics have not been quantified directly in humans. Here we identified 129,582 spontaneous, genome-wide somatic mutations in 140 single-cell–derived haematopoietic stem and progenitor colonies from a healthy 59–year-old man and applied population-genetics approaches to reconstruct clonal dynamics. Cell divisions from early embryogenesis were evident in the phylogenetic tree; all blood cells were derived from a common ancestor that preceeded gastrulation. The size of the stem cell population grew steadily in early life, reaching a stable plateau by adolescence. We estimate the numbers of haematopoietic stem cells that are actively making white blood cells at any one time to be in the range of 50,000–200,000. We observed adult haematopoietic stem cell clones that generate multilineage outputs, including granulocytes and B lymphocytes. Harnessing naturally occurring mutations to report the clonal architecture of an organ enables the high-resolution reconstruction of somatic cell dynamics in humans.

Human haematopoiesis balances the production and destruction of hundreds of billions of specialized blood cells every day. This process relies on a multilayered hierarchy of progressively more differentiated and more populous cells, at the top of which sits the pool of stem cells. First described functionally in the 1960s, haematopoietic stem cells are defined by their ability to establish long-term, stable contributions to multiple lineages of blood cells, including myeloid, T and B cells. The numbers and dynamics of stem cells in homeostatic human haematopoiesis remain poorly defined, despite their routine use in therapeutic transplantation for haematological diseases.

Historical studies in animals quantified haematopoiesis either by labelling cells in vitro and transplanting them into a recipient animal or by modelling X chromosome inactivation patterns. More recently, studies tracking the clonal contributions of cells labelled directly in vivo have suggested that long-term homeostatic haematopoiesis is driven by many thousands of cells that do not function as classical stem cells in transplantation assays.

Approaches to measuring the dynamics and potential of stem cells in humans have been less direct. Studies of X chromosome inactivation patterns in haematological malignancies have demonstrated their clonal origin in stem cells that had multilineage potential. In patients who received gene therapy, hundreds to thousands of clones contribute to lymphoid and myeloid lineages more than a year after transplantation. Studies of unperturbed haematopoiesis in humans have relied on ex vivo cellular assays or modelling of telomere lengths and X chromosome inactivation patterns. These analyses have suggested that the numbers of stem cells increase through childhood and adolescence, reaching a plateau in adulthood, with some shift in lineage potential.

Spontaneous mutations used to measure haematopoiesis

Mutations accumulate in somatic cells throughout life. A mutation arising in a cell is inherited by its descendant cells, a feature that has enabled the reconstruction of clonal structures in cancer and normal development. In normal blood stem cells, the burden of somatic mutations increases linearly with age, suggesting that they represent an accurate molecular clock.

We hypothesized that spontaneous somatic mutations could act as clonal markers, thus enabling the quantification of the number, activity and longevity of human blood stem cells during normal haematopoiesis. Analogous to capture–recapture experiments in ecology, our design followed two phases (Fig. 1). First, in the ‘capture’ phase, we isolated single haematopoietic stem and progenitor cells from a bone marrow aspirate and peripheral blood draw from a 59-year-old male with normal blood counts and no past history of blood disorders (Extended Data Fig. 1). These were expanded in single-cell liquid cultures or colony-forming cell assays. We performed whole-genome sequencing on 198 colonies, each to around 15× depth (Supplementary Table 1), and identified somatic mutations. Second, in the ‘recapture’ phase, we isolated bulk populations of mature peripheral blood cells from the same individual: granulocytes at three time points after the bone marrow aspirate, together with B and T lymphocytes, both from one time point. We performed deep targeted sequencing on these bulk populations for mutations discovered in the capture phase.

Bringing together stem cell biology and population genetics creates a risk of lexical confusion. We reserve the term ‘clone’ for the in vivo descendants of a single ancestral cell, and use ‘colony’ to describe the cells derived in vitro from a single stem or progenitor cell. We use ‘lineage’ to denote a specific functional group of blood cells, such as granulocytes; and ‘line-of-descent’ for the set of cells that are direct antecedents/descendants of the cell in question (see Supplementary Information).

Mutation burden and spectrum

In total, 140 colonies had variant allele fractions that were distributed around 50%, confirming that they did in fact derive from a single cell, whereas 58 of the colonies had lower allele fractions (Extended Data Fig. 2 and Supplementary Table 1); this was probably caused by colonies

© 2018 Springer Nature Limited. All rights reserved.
growing into each other in the methylcellulose. These polyclonal colonies were excluded from further analyses. It proved more difficult to derive clonal colonies from some progenitor types than others, such that our final set of 140 colonies was composed of 89 immunophenotypic haematopoietic stem cells, 38 megakaryocyte–erythocyte progenitors, eight granulocyte–macrophage progenitors and five common myeloid progenitors.

We assessed whether variants were acquired during in vitro expansion. Any mutation on the X chromosome in the original colony-forming cell should be present at allele fractions near 100%. The mean percentage of X chromosome mutations found on <50% reads was only 5.6% per colony, suggesting that variants acquired in vitro were infrequent.

Mutation burden was consistent across colonies, with a mean of 1,023 substitutions (range, 815–1,210) and 20 small insertion/deletion (indel) events (range 2–37) (Extended Data Fig. 3). No somatically acquired structural variants were identified. The spectrum of mutations was dominated by C > T and T > C transitions, as described in myeloid cancers20 and age-related clonal haematopoiesis20 (Extended Data Fig. 4).

Driver mutations in myeloid cancer genes occur in some older individuals with normal blood counts26–28. In the colonies from our subject, we found no known myeloid driver mutations. Furthermore, the ratio of non-synonymous to synonymous mutations, a metric that can detect positive or negative clonal selection across genes29, was exactly as expected for the background mutation spectrum (dN/dS = 1.001; 95% confidence interval, 0.889–1.127; with dN/dS = 1 representing neutrality). Finally, driver mutation hotspots were included in the bait set for the recapture phase (Supplementary Table 2), and no such drivers were detected. Thus, haematopoietic cells in this subject have undergone selectively neutral accumulation of somatic mutations.

**Somatic mutations acquired in embryo development**

To explore clonal relationships among the 140 colonies, we constructed a phylogenetic tree (Fig. 2 and Extended Data Figs. 5, 6). Of the 129,582 somatic mutations of colonies on the tree, 8,676 were seen in more than one colony. At the top of the tree, two mutations completely partitioned the colonies, one found in 52 and the other in 88 colonies, with every colony possessing one or other mutation, and no colony possessing both (Fig. 2b). These same two mutations were found in a buccal swab taken from the patient, and at clonal contributions that suggested the same two thirds to one third split. This indicates that the most recent common ancestor of all blood cells in this subject was also the most recent common ancestor of buccal epithelial cells. Because blood derives from the mesoderm and the buccal epithelium from the ectoderm, this common ancestor must have predated gastrulation. It is likely that this most recent common ancestor of both blood and buccal epithelium was in fact the fertilized egg; the two observed mutations would have occurred during its first cell division, one to each daughter cell. These two daughter cells then contributed unequally to adult somatic tissues, as previously observed22,30,31.

After the first division, a cascade of further mutations provides even finer partitions of the colonies (Fig. 2b), consistent with data in mice that adult haematopoiesis is a mosaic of embryonic clones32. By 10 mutations of molecular time, 33 lines of descent were created, which required at least 5 generations of cell doublings. Embryonic lineages that are lost or unobserved would suggest more than 5 generations, so average mutation rates in early embryogenesis could be around 2 per division at most. Of the 32 cell divisions by 10 mutations in molecular time, at least 10 were associated with no mutations (Fig. 2c), noting that unobserved embryonic lineages can convert polynomials (multiway splits) to dichotomies. This provides an estimate of the mutation rate of 1.2 per division (Supplementary Methods). Thus, mutation rates in embryonic cells that ultimately contribute to somatogenesis fall in the range of 1–2 per cell division, similar to estimates from human neural progenitor clones33 and de novo germline mutations33.

**Clonal relationships of haematopoietic cells**

Clonal relationships among the 140 colonies were evident beyond the early embryonic mutations as a series of branch points scattered down the vertical axis of the phylogenetic tree (Fig. 2a). Mutations shared between two colonies indicate that they derive from a common ancestor; the ratio of shared to unique mutations is a measure of the age at which the two lines of descent split. Long lines of descent with few branches suggest that we have sequenced only a small fraction of the stem cell pool.

The distribution of different cell types across the tree provides information on population stratification among haematopoietic stem and progenitor cells. Stem cells from the bone marrow aspirate, taken from the right iliac crest, were no more clustered together on the phylogenetic tree than those derived from peripheral blood (P = 0.14; Fig. 2 and Extended Data Fig. 3a). This suggests that stem cells recirculate and redistribute sufficiently often that the population in the iliac crest is a random sample of the whole-body stem cell pool. Similarly, progenitors were not more clustered on the tree than stem cells (P = 0.12; Extended Data Fig. 3b, c), suggesting that the progenitors that we sequenced are not drawn from a more restricted set of historic lines of descent than the stem cells.

The interspersed distribution of stem and progenitor cells on the tree and relative scarcity of recent branch points indicates that the phylogeny is dominated by events that occurred in stem cells. As a progenitor is short-lived, only the most recent mutations in its line of descent will have occurred while it was a progenitor. Therefore, a branch point hundreds of mutations ago represents an ancient symmetrical division in which one stem cell gave rise to two stem cells, since descendents of both daughters persist decades later as haematopoietic cells. In traditional experimental models, long-term self-renewal of haematopoietic stem cells is established prospectively (through serial transplantation assays), whereas here we infer it retrospectively through the decades-long persistence of lines of descent from an ancestral stem cell division. Therefore, for the analyses that follow, which rely...
Fig. 2 | The phylogeny of cells, showing the relationship between cell types and embryological cell divisions. a, Phylogeny of 140 single haematopoietic stem and progenitor cells showing the relationship between cell types. At each tip of the tree is a colony. Branches connect colonies to each other to form a family tree. Branch lengths are proportional to the number of somatic mutations. Branches are coloured according to the phenotype of their descendants. Branches ancestral to haematopoietic progenitor cells (HPCs) are coloured red; branches ancestral to bone marrow–derived haematopoietic stem cells (BM HSCs) are blue; branches ancestral to peripheral blood-derived haematopoietic stem cells (PB HSCs) are green; branches ancestral to both stem and progenitor cells are coloured black. b, The same phylogeny as in a, but showing only the first 10 mutations of molecular time. c, The number of descendants of each node for the first 10 mutations of molecular time, used to estimate the embryonic mutation rate.

Estimates of stem cell number and generation time
We designed a hybridization bait set for 7,116 mutations that were identified in the colonies and assigned to the phylogenetic tree, choosing 6,317 mutations that were shared by more than one colony and 799 that were unique to single colonies. We performed targeted sequencing of three peripheral blood granulocyte samples taken 4 months (mean coverage, 776 ×), 9 months (mean coverage, 4,669 ×) and 14 months (mean coverage, 268 ×) after the bone marrow aspirate, together with control cord blood from two individuals (mean coverage, 5,305 ×) (Supplementary Table 3). We used Bayesian generalized linear mixed models to estimate the fraction of reads that were derived from true mutant alleles versus sequencing errors. Estimated allele fractions were stable across the three time points and steadily decreased down individual lines-of-descent (Extended Data Fig. 8).

Three observations emerge (Fig. 4). First, the majority of mutations in the bait set were not detectable (horizontal grey ticks in Fig. 4). This suggests that the number of active stem cells in our subject must be much higher than a few thousand, since our detection threshold was ancestral to bone marrow–derived haematopoietic stem cells (BM HSCs) are blue; branches ancestral to peripheral blood–derived haematopoietic stem cells (PB HSCs) are green; branches ancestral to both stem and progenitor cells are coloured black.

Lifetime trajectory of the population size of stem cells
The relative timings of branch points, or more formally ‘coalescences’, in the phylogenetic tree inform on historical population dynamics. In brief, under neutral drift, the rate at which lines-of-descent coalesce is related to both the effective population size and the generation time (here, the time between symmetrical stem cell divisions). Methods to infer historic population dynamics from coalescences17,18 were applied to the tree built from the blood cells of our subject. This revealed a rapid population expansion of haematopoietic stem cells during early life, reaching a relatively stable plateau by late childhood or early adolescence (Fig. 3), consistent with previous inferences17,18. The stable population size during adulthood suggests that symmetric self-renewal divisions, during which one stem cell divides into two stem cells, are balanced by stem cell death, senescence and symmetric divisions into committed progenitors.

The same broad coalescence structure of the phylogenetic tree that arises with a given population size and number of generations can be generated by a population ten times larger going through ten times as many generations. Therefore, from the structure of the tree alone, without knowing how many generations the cells have been through, we cannot directly estimate the absolute number of stem cells. We therefore performed deep-sequencing of blood cells in the recapture phase of our study.

Fig. 3 | Population size trajectory of the stem cell pool. Phylodynamic methods reveal changes in the effective population size of stem cells over life based on the timing of coalescences (branch points) in our observed phylogeny. Shading illustrates different credibility intervals. The y axis is shown in units of ‘population size multiplied by generation time’ (cell years), because the same distribution of coalescences can be generated from a population of 10 times the size with 10 times as many generations.
Approximate Bayesian computation of the number of stem cells

Typically approximately 1/2,000. Second, most branches at the top of the phylogenetic tree have mutations that can be detected in granulocytes. This indicates that many stem cells that are not closely related to each other actively contribute to haematopoiesis during the investigated time period. If only one or a few stem cells produce granulocytes, which are replaced over time as they become exhausted, we would expect only the branches of the currently active stem cells to be detectable in granulocytes. Third, some branches contribute more to haematopoiesis than most, as indicated by a handful of branches with mutations that were detectable much further down the tree than average (although still representing a relatively small proportion: <0.6% of granulocytes). Assuming that no undiscovered driver mutations are present, this would be a consequence of genetic drift.

We developed an approximate Bayesian computation framework to quantify key properties of the stem cell compartment (Supplementary Methods and Supplementary Information). In brief, we generated 200,000 simulations of neutral haematopoiesis in adulthood (Fig. 5), varying the number of stem cells and the time between successive stem cell renewal divisions. The following assumptions were used: all active stem cells have an equal probability of dividing per unit time; the size of this pool is constant over adulthood; all active stem cells produce similar numbers of granulocytes; and stem cells accumulate somatic mutations stochastically. We then recapitulated our experimental design on each simulation in silico. Using informative summary statistics, we compared the properties of the simulations to the observed data, looking for the combinations of stem cell numbers and symmetrical cell division rate that could most closely replicate the observed data.

Under our model, the 90% credibility interval for the number of stem cells actively contributing to circulating granulocytes at one time was 44,000–215,000 (Fig. 5 and Extended Data Fig. 7). Furthermore, the estimated time between successive self-renewal stem cell divisions is most credibly in the range of 2–20 months. Simulations within this range (Extended Data Fig. 7n) most closely resembled our observed data (Extended Data Fig. 7m), whereas simulations from outside the plausible ridge showed marked differences (Extended Data Fig. 7o, p).

We note the uncertainties in these estimates, which could be improved by investigating more subjects; more stratified sampling (cells from spleen, thymus and other regions of the bone marrow); sequencing by investigating more subjects; more stratified sampling (cells from spleen, thymus and other regions of the bone marrow); sequencing more colonies in the capture phase (we now know that our 140 colonies represent only one-thousandth of the active stem cell pool); and increasing the sensitivity of mutation detection in the recapture phase.

**Clonal contributions to granulocytes and lymphocytes**

One of the defining features of a stem cell is its contribution to multiple cell types. Alongside granulocytes, we deep-sequenced samples of peripheral blood B and T lymphocytes (Fig. 6 and Extended

---

**Fig. 4** Recapture of mutations by targeted sequencing. The phylogenetic tree of cells is shown as in Fig. 2, but information from targeted sequencing of peripheral blood granulocytes from the nine-month time point is overlaid. This is shown more clearly in the inset, which zooms in on a portion of the tree. The underlying structure of the tree is shown in grey. Horizontal bars indicate the locations of every mutation in the bait set for targeted sequencing. Bars are coloured according to the proportion of cells in the sample that carry the mutations (obtained by multiplying the variant allele fraction (VAF) for autosomal mutations by two). Undetectable mutations are coloured grey and are shown as smaller bars. Mutations have been spaced evenly along a branch according to their mean variant allele fraction from the combined targeted sequencing of all granulocyte and lymphocyte time points. A higher density of baits were designed for branches that were shared by more than one colony. On these branches, the mutations are so close together that they appear as one continuous bar.

---

**Fig. 5** Approximate Bayesian computation of the number of stem cells and their replication rate. a, A contour plot of the most likely values for stem cell numbers and time between symmetrical cell divisions over the sample space that was simulated. It shows the stem cell number and generation times of the 500 simulations that produced summary statistics that were most similar to the summary statistics extracted from the observed data. b, The prior distribution for the number of stem cells contributing to granulocytes for the approximate Bayesian computation. 90% credibility interval: 1,696–368,457. c, The distribution of stem cell numbers for the 1,000 simulations that produced summary statistics most similar to the observed summary statistics. 90% credibility interval: 15,995–295,189. d, The posterior distribution of a neural network regression run on these 1,000 simulations. 90% credibility interval: 44,511–215,417.
Data Fig. 9). Most of the early mutations, at the top of the phylogenetic tree, were detectable in all cell types that were investigated (black branches in Fig. 6), suggesting that there is a shared common ancestry of lymphocytes and granulocytes during development. Beyond 100 mutations in molecular time, when population size reached a plateau (Fig. 3), 464 mutations distributed across 39 branches could be detected, of which 217 on 12 branches were detected in more than one cell type.

Some adult stem cell clones contributed to detectable numbers of granulocytes and B lymphocytes, whereas their mutations were not detectable in T lymphocytes (Extended Data Fig. 9b–f). We had equivalent sequencing coverage in B and T lymphocytes, so the discrepancy is not technical. This finding suggests that descendants of these particular stem cell lines contribute a higher fraction of currently produced granulocytes and B cells than T cells. Owing to the low fraction of stem cells that were sequenced in the capture phase, we cannot exclude the existence of stem cell lines that currently contribute to all three cell types, as found in age-related clonal haematopoiesis.

Conclusions

The mosaic contribution of many embryonic clones to haematopoiesis and the large number of clones active in adulthood agrees well with lineage-marking studies in mice. We estimate that the number of stem cells contributing to unperturbed human haematopoiesis at any given time is most probably in the hundreds of thousands and that the time between symmetric stem cell divisions falls in the range of 2–20 months; the latter estimate is similar to previous inferences.

We demonstrate the existence of adult human stem cell clones with multilineage output in vivo. Branches with mutations detectable in granulocytes and B lymphocytes exist beyond 300 mutations of molecular time, which would represent our subject’s late teens or early twenties. The similar allele fractions in B cells and granulocytes suggest ongoing contribution of these stem cell lines to the two cell types. The key principle here is that under neutral dynamics, the drift of these branches to an appreciable proportion of granulopoiesis would have been gradual. If myeloid and B lymphocyte lineages separated early in life and underwent genetic drift independently, then we would not expect the exact same lines-of-descent to become enriched in both populations. Instead, the parsimonious explanation is that the pool of shallower lymphocyte sequencing dataset and pink for mutations that were only detected in granulocytes, but at such a low allele fraction (<1 in 2,000 reads) that if they had been present in lymphocytes at this allele fraction they would not have been detected. Arrows indicate adult clones with multilineage output. B, B lymphocytes; G, granulocytes; T, T lymphocytes; VAF, variant allele fraction.

By contrast, we do not observe shared mutations between granulocytes and T lymphocytes much beyond 100 mutations of molecular time. Several possible explanations for this include a deeper separation between myeloid and T cell production than between myeloid and B cells; a large and long-lived pool of T lymphocytes that dilute any ongoing contribution to T lymphopoiesis from recent stem cells; or a lower number of stem cells actively contributing to T lymphopoiesis, requiring more than the 140 colonies that were sequenced here to uncover their contribution.

The 40 trillion cells in the human body all trace their ancestry back through a series of cell divisions to the fertilised egg. All of these cells can be visualized on a single phylogenetic tree with the fertilized egg at its root. Establishing the exact position of any given cell on this tree requires a unique and permanent mark stamped on each cell with each division. Somatic mutations provide such a mark. Layering phenotypic information—be it transcriptional state, lineage output or functional resilience—on the phylogeny will enable estimation of the heritability from mother cell to daughter cell of germline somatic phenotypes. Humans can be directly researched, as the ubiquity and permanence of somatic mutations enables studies through youth and old age, in steady state and after perturbation, across blood and other organs, and in health and disease.

Reporting summary

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

Code availability

Custom R scripts and their input data central to this analysis are available on GitHub at https://github.com/HLee-Six/HSC_phylodynamics. All other codes are available from the corresponding authors upon reasonable request.

Data availability

Whole-genome and targeted sequencing data have been deposited in the European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/). Whole-genome sequencing data have been deposited with EGA accession...
number EGAD00001004086 and targeted sequencing data with accession number EGAD00001004087. Substitution calls have been deposited on Mendeleev Data (Population dynamics of human blood inferred from spontaneous somatic mutations; https://doi.org/10.17632/zyjw2zkx71). Simulated datasets (from the approximate Bayesian computation) are available from the corresponding authors upon reasonable request.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0497-0.

Received: 28 February 2018; Accepted: 1 August 2018; Published online 5 September 2018.

1. Till, J. E. & McCulloch, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. Radiat. Res. 14, 213–222 (1961).
2. Becker, A. J., McCulloch, E. A. & Till, J. E. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. Nature 197, 452–454 (1963).
3. Lemschik, J. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. Cell 45, 917–927 (1986).
4. Kim, S. et al. Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study. Cell Stem Cell 14, 473–485 (2014).
5. Naik, S. H. et al. Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature 496, 229–232 (2013).
6. Koelle, S. J. et al. Quantitative stability of hematopoietic stem and progenitor cell output in rhesus macaques receiving transplants. Blood 129, 1448–1457 (2017).
7. Ackbowski, J. L., Catlin, S. N. & Guttorp, P. Evidence that hematopoiesis may be a stochastic process in vivo. Nat. Med. 2, 190–197 (1996).
8. Sun, J. et al. Clonal dynamics of native haematopoietics. Nature 514, 322–327 (2014).
9. Busch, K. et al. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. Nature 518, 542–546 (2015).
10. Sawai, C. M. et al. Hematopoietic stem cells are the major source of multilineage hematopoiesis in adult animals. Immunity 45, 597–609 (2016).
11. Bystrykh, L. V., Verovskaya, E., Zwart, E., Broekhuis, M. & Haan, G. Counting stem cells: methodological constraints. Nat. Methods 9, 567–574 (2012).
12. Fialkow, P. J., Gartler, S. M. & Yoshida, A. Clonal origin of chronic myelocytic leukemia in man. Proc. Natl Acad. Sci. USA 58, 1468–1471 (1967).
13. Fialkow, P. J., Jacobson, R. J. & Papayannopoulou, T. Chronic myelocytic leukemia: clonal in a stem cell common to the granulocyte, erythrocyte, platelet and monocyte/macrophage. Am. J. Med. 63, 125–130 (1977).
14. Carter, N. et al. Hematopoietic stem cell gene therapy with a lentiviral vector in a linearized vector in X-linked adrenoleukodystrophy. Science 326, 818–823 (2009).
15. Bisaccio, L. et al. In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. Cell Stem Cell 17, 119–131 (2015).
16. Notta, F. et al. Isolation of single human hematopoietic stem cells capable of reconstituting human hematopoiesis in vivo. Nat. Med. 13, 818–823 (2007).
17. Werner, B. et al. Reconstructing the in vivo dynamics of hematopoietic stem cell clonal output in rhesus macaques receiving transplants. Blood 117, 4460–4466 (2011).
18. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264 (2016).
19. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukaemia. Cell 150, 264–278 (2012).
20. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 1004–1012 (2012).
21. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature 513, 422–425 (2014).
22. Bae, I. et al. Different mutational rates and mechanisms in human cells at preagulation and neurogenesis. Science 359, 550–555 (2018).
23. Notta, F. et al. Isolation of single human hematopoietic stem cells capable of long-term multilineage engraftment. Science 333, 218–221 (2011).
24. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood 130, 742–752 (2017).
25. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. 371, 2488–2498 (2014).
26. Xie, M. et al. Age-related clonal hematopoietic expansion and malignancies. Nat. Med. 20, 1472–1478 (2014).
27. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hematopoiesis. Cell Rep. 10, 1239–1245 (2015).
28. Notta, F. et al. Universal patterns of selection in cancer and somatic tissues. Cell 171, 1029–1041 (2017).
29. Pluta, B. et al. The first cleavage of the mouse zygote predicts the blastocyst axis. Nature 434, 391–395 (2005).
30. Xu, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. Nature 543, 714–718 (2017).
31. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature 548, 456–460 (2017).
32. Rubbani, R. et al. Timing, rates and spectra of human germline mutation. Nat. Genet. 48, 126–133 (2016).
33. Lan, S., Palacios, J. A., Karcher, M., Minin, V. N. & Shahbaba, B. An efficient Bayesian inference framework for coalescent-based nonparametric phylogenetics. Bioinformatics 31, 3282–3289 (2015).
34. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal hematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun. 7, 12484 (2016).
35. Schiemer, H. et al. Fate mapping reveals separate origins of T cells and myeloid lineages in the thymus. Immunity 32, 426–436 (2010).

Acknowledgements
This work was supported by the Leukemia Lymphoma Society, the WBF Foundation and the Wellcome Trust. P.J.C. is a Wellcome Trust Senior Clinical Fellow (WT088340MA). H.L.-S. is a recipient of a Wellcome Trust PhD studentship. N.F.O. is a recipient of a Danish Lundbeck Fellowship (2016–17) and M.S.S. is the recipient of a BBSRC CASE Industrial PhD Studentship. Work in the D.G.K. laboratory is supported by a Bloodwise Bernstein Fellowship (15008), a European Research Council Starting Grant (ERC-2016-STG–715371) and a European Hematology Association Non-Clinical Advanced Research Fellowship. Work in the A.R.G. laboratory is supported by the Wellcome Trust Bloodwise, Cancer Research UK, the Kay Kendall Leukemia Fund, and the Leukaemia and Lymphoma Society of America. Work in E.L. laboratory is supported by a Wellcome Trust Sir Henry Dale Fellowship, BBSRC and a European Haematology Association Non-Clinical Advanced Research Fellowship. The D.G.K. and A.R.G. laboratories are supported by a core support grant from the Wellcome Trust and Medical Research Council to the Cambridge Stem Cell Institute. We acknowledge further assistance from the National Institute for Health Research Cambridge Biomedical Research Centre and the Cambridge Experimental Cancer Medicine Centre. We thank R.J.O. and P.J.C. separated signal from noise in the phylogeny. I.M. tested for positive selection. M.R.S. oversaw the analysis extraction and library preparation. H.L.-S. performed most of the data collection and statistical analysis. S.G. helped to call mutations and construct the phylogeny. U.M. tested for positive selection. M.R.S. oversaw the analysis of mutational signatures. R.J.O. and P.J.C. separated signal from noise in targeted sequencing analysis. K.D. helped to design the approximate Bayesian computation along with P.J.C. and H.L.-S. and co-wrote the supplementary results on inference of population size. H.L.-S., P.J.C., D.G.K. and A.R.G. wrote the manuscript, with contributions from all authors.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0497-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0497-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.R.G., D.G.K. and P.J.C.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Extended Data Fig. 1 | Cell sorting strategy. a, Sorting of stem and progenitor cells. Human bone marrow (BM) and peripheral blood (PB) mononuclear cells (time point 1) were stained with anti-CD34, anti-CD38, anti-CD45RA, anti-CD90, anti-CD10 and anti-CD135 antibodies. After exclusion of debris and doublets, gatings on CD34, CD38 and CD90 were used to separate CD34+CD38−CD90+ HSCs. The CD34+CD38+ compartment was gated for CD10− cells before gating on CD135 (also known as FLT3) and CD45RA to separate progenitor compartments: CD135+CD45RA− common myeloid progenitor (CMPs), CD135+CD45RA+ granulocyte–macrophage progenitors (GMPs) and CD135−CD45RA− megakaryocyte–erythrocyte progenitors (MEPs). b, Sorting of B and T lymphocytes. Peripheral blood mononuclear cells (time point, after 4 months) were stained with anti-CD4, anti-CD8 and anti-CD19 antibodies. After exclusion of debris and doublets, the CD4+CD8+CD19− gate was used to isolate T cells, while the CD4−CD8−CD19+ gate was used to isolate B cells. n = 20,000 events.
Extended Data Fig. 2 | Quality control of colonies from single-cell derived clones. Example histograms of the variant allele fraction (VAF— the proportion of sequencing reads that report the mutation) of mutations in single colonies. a, The VAF of all mutations on autosomes in a typical clonal colony. Because there are two copies of each autosome, and each mutation occurs on only one of them, in a clonal sample the VAF of autosomal mutations is binomially distributed with a mean of 0.5. b, The VAF of all mutations on the X chromosome in the same clonal colony. Because the subject is male, there is only one copy of the X chromosome, and so true mutations here must have a VAF of 1. Occasionally, lower VAFs are seen when a mutation is not detected on a read, when a read from another locus is aberrantly mapped to the locus in question and thus lowering the apparent coverage, or when a mutation is acquired in vitro. c, d, The VAF of autosomal and X chromosome mutations, respectively, in a typical colony seeded by more than one cell. As not all the reads come from the same cell, and most mutations are private to a given cell, a lower proportion of DNA molecules carry the mutation in a polyclonal colony than in a clonal colony, resulting in a leftward shift of the peak of the VAF histogram. These histograms suggest that the number of mutations acquired by the colonies after a few weeks of in vitro expansion is a small fraction of those acquired in vivo over 60 years of life.
Extended Data Fig. 3 | Mutation burden of colonies. a, A histogram of substitution (left) and indel (right) burden per colony. b, The location around the genome of substitutions from all clones combined is shown as a circos plot. The outermost ring of the circos plot depicts the karyotypic ideogram. Moving inwards, base substitutions are shown as rainfall plots in which the height of the dot in the substitution ring is proportional to the log₁₀ of the distance to the next mutation and with the colour of the dot illustrating the base change, as shown in the key. c, A comparison of the substitution burden between stem cells and progenitor cells. There were not significantly more mutations in progenitors than stem cells (P = 0.14, Wilcoxon rank-sum test).
Extended Data Fig. 4 | Trinucleotide context of mutations in normal blood colonies. a, The trinucleotide context of substitutions for all colonies combined. Substitutions can be classed according to the base change (referred to by the pyrimidine of the mutated base pair), and the bases 5’ and 3’ of the mutated one, into 96 categories. The counts in each of these categories are shown. b, Comparison with pooled acute myeloid leukaemia genomes, excluding genomes with >1,500 mutations, and publicly available data on normal tissues that have been whole-genome sequenced so far. The ordering of bars is the same as in a, and the same figure as in a is provided again at the same resolution of the other data for ease of comparison. Please note that these samples have been sequenced on different platforms using different systems, which is likely to result in small differences. Normal liver, normal colon and normal small intestine data were obtained from whole-genome sequencing of single-cell-derived organoids19, whereas normal neurons were derived from single cells that had undergone whole-genome amplification37. c, Example trinucleotide substitution plots for a selection of individual colonies derived from either stem cells (which have the prefix BMH) or progenitor cells (which have the prefix BMP). The ordering of bars is the same as in a.

37. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559 (2018).
Extended Data Fig. 5 | Construction of the phylogeny using different methods. a, The phylogeny of cells as presented in Figs. 2, 4, 6, but with the addition of P values next to every node, derived by bootstrapping the substitution matrix 1,000 times, building a tree using SCITE for each replicate, and counting the proportion of the bootstrapped trees that support each node. b–f, Phylogenies constructed using different datasets and methods. In each case the phylogeny was constructed using 100 bootstraps of the data, and the P value for each node shown underneath it. Branches are coloured by whether a branch ancestral to exactly the same descendants is also present in the SCITE tree, and are drawn with a thicker line if the branch is recovered in ≥70% of bootstrap replicates.

b, Substitution and indel datasets combined, building the tree by maximum parsimony. c, Substitution, indel and neighbour-joining datasets combined, building the tree by neighbour joining.

d, Substitutions, tree build by maximum parsimony. e, Indels, tree built by maximum parsimony. f, Short tandem repeats, tree built by neighbour joining.
Extended Data Fig. 6 | Relationship between cell types in the phylogeny. a, The phylogeny showing different stem and progenitor cell types. b, The phylogeny is shown as in a, but with the labels underneath coloured according to which cell types are being compared. The first row of labels has stem cells from bone marrow in red, progenitor cells from bone marrow in grey and stem cells from peripheral blood in black. The second row of labels has stem cells in red and bone marrow progenitors in black. The third row of labels has MEPs in red, CMPs in black, GMPs in blue and stem cells in grey. c–e, Analysis of molecular variance is used to test for clustering on the phylogeny for stem cells derived from peripheral blood versus bone marrow cells (c), stem cells versus progenitors (d) and different progenitor types (e). In each panel, a histogram of the null distribution of the statistic used to detect clustering is shown. Distributions were obtained by randomly permuting which cells were assigned to which category. Comparisons are only between cell types not shown in grey in b. The observed value of the statistic is shown as a red vertical line.
Extended Data Fig. 7 | Approximate Bayesian computations. a, The joint prior distribution for stem cell numbers (HSCs) and the generation time for the first approximate Bayesian computation (ABC). b, The location in sample space of the 10% of simulations that produced summary statistics (using only the ltt summary statistics; Supplementary Methods and Supplementary Information) most similar to the observed summary statistics (using only the ltt summary statistics; Supplementary Methods). c, The joint prior distribution for the second ABC, in the area of sample space indicated to be plausible by the first set of simulations. d, The joint posterior distribution of the best 500 simulations from the second ABC, as shown in Fig. 5 for ease of reference. e–i, Cross-validation of the model to choose the number of accepted simulations and the weighting applied to the ltt summary statistics (Supplementary Methods and Supplementary Information). j, For illustrative purposes, five simulations were sampled for each of three population sizes along the plausible diagonal of sample space indicated in b. One set of summary statistics are shown for these simulations in k, k. The red line indicates a simulation coming from the area of sample space indicated by a red point in j; and similarly for blue and green lines. The black dotted line indicates the observed values for these summary statistics. These summary statistics provide a count—for the different numbers of samples (x axis)—of how many of the 3,952 mutations that we considered (y axis) are in this many samples with two or more reads, using error model 1 (which simulates errors according to the error rate in control DNA (Supplementary Methods)). The same summary statistics were calculated for different mutant read number cut-offs. l, For each of the 1,000 simulations that produce summary statistics that were the most similar to the observed data, the Euclidean distance from the observed data (y axis) is plotted against the number of stem cells in that simulation (x axis). This information is used by the neural network regression step to define the most likely value for the number of stem cells. The most similar values are seen at around 100,000 stem cells, which was the location of the median of the posterior distribution from neural network regression. m, The observed phylogeny, with branch points indicated by asterisks. n–p, Phylogenies drawn from simulations that occur at the points in sample space indicated in d, n. A relatively plausible simulation, since the pattern of branch points is not dissimilar from the pattern of the observed phylogeny (m). Simulations with smaller stem cell populations and faster stem cell turnover rates resulted in phylogenies in which the stem cells were very closely related to each other (o), whereas those with larger populations and slower turnover result in phylogenies in which the stem cells only share an embryonic common ancestor, and no branches are seen through the tree (p).
Extended Data Fig. 8 | Targeted sequencing data. a, Correlations between the VAFs of all sequenced samples, shown on a log scale. Note that samples that were sequenced to a lower depth cannot have VAFs as small as samples sequenced to higher depths. b, Targeted sequencing information with no error correction. The data are shown as in Fig. 4 for all analysed samples, but focusing on only the first 350 mutations of molecular time. To allow a better comparison between samples that were sequenced at different depths, a higher detection threshold and different detection threshold are used relative to Fig. 4. c, Targeted sequencing information after using cord blood controls for sequencing error correction with the Bayesian generalized Poisson mixed-effects model. The colour scale is the same as in b. The data for the granulocytes at the nine-month time point are the same as in Fig. 4 (provided again for ease of comparison), but plotted with a different colour scale.
Extended Data Fig. 9 | Multilineage clonal output. a. The phylogeny with targeted sequencing information in different blood fractions overlaid as in Fig. 6, shown again here for ease of reference. The colouring of mutations reflects in which peripheral blood cell fractions they could be detected, as indicated by the colour key. Arrows indicate adult clones with multilineage output, with letters corresponding to panels b–f. B, B lymphocytes; G, granulocytes; G low VAF, granulocytes, allele fraction too low to be detected in lymphocytes; T, T lymphocytes. b–f, VAFs of all mutations on branches (indicated by arrows in a) with mutations beyond molecular time 100 that are detectable in granulocytes and B lymphocytes but not in T lymphocytes.
Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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 |
|     | An indication of 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 statistics 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 |
|     | Clearly defined error bars |
|     | State explicitly what error bars represent (e.g. SD, SE, CI) |

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection Variant calling algorithms are all cited in the manuscript and code is fully available for download.

Data analysis We have released code used on GitHub - https://github.com/HLee-Six/HSC_phyldynamics

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Whole genome sequencing data have been released on the European Genome-Phenome Archive (EGAD00001004086). Targeted sequencing data are also released on EGA (EGAD00001004087). Substitution calls have been deposited on Mendeley Data ("Population dynamics of human blood inferred from spontaneous somatic mutations": http://dx.doi.org/10.17632/yjzjw2stk7f.1).
Field-specific reporting

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

☑️ Life sciences  ☐ Behavioural & social sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences

Study design

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

Sample size  Sample size (number of colonies sequenced) was chosen to capture clonal relationships among polyclonal blood cells.

Data exclusions  58 colonies were excluded from the main data analysis because they were not single cell-derived, as described in the manuscript.

Replication  No replication in other individuals has yet been attempted.

Randomization  Not applicable

Blinding  Not applicable

Materials & experimental systems

Policy information about availability of materials

n/a

Involved in the study

☑️ Unique materials
☑️ Antibodies
☑️ Eukaryotic cell lines
☑️ Research animals
☐ Human research participants

Human research participants

Policy information about studies involving human research participants

Population characteristics  One normal 59-year old male was analysed.

Method-specific reporting

n/a

Involved in the study

☑️ ChIP-seq
☐ Flow cytometry
☐ Magnetic resonance imaging

Flow Cytometry

Plots

Confirm that:

☑️ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
☑️ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
☑️ All plots are contour plots with outliers or pseudocolor plots.
☑️ A numerical value for number of cells or percentage (with statistics) is provided.
### Methodology

#### Sample preparation

Mononuclear cells (MNCs) from BM and PB were isolated by density gradient centrifugation (Lymphoprep; Axis-Shield, Oslo, Norway), and enriched for CD34 positive cells (EasySep Human CD34+ enrichment kit, STEMCELL Technologies, Vancouver, Canada (STEMCELL)) as per the manufacturer’s guidelines except that only one round of depletion in the magnet was performed.

#### Instrument

Influx sorter (BD), equipped with the following lasers; 405nm, 488nm, 561nm, and 640nm, and filter sets; 530/40 (for FITC), 710/50 (for PerCPcy5.5), 750LP (for APC-Cy7), 670/30 (for APC), 460/50 (for Violet450), and 585/29 (for PE).

#### Software

BD Influx software

#### Cell population abundance

As shown in Extended Figure 1.

#### Gating strategy

Bulk HSCs and CD34+CD38+CD90-CD135+CD45RA- (common myeloid progenitors, CMPs), CD34+CD38+CD90-FLK2+CD45RA+ (granulocyte-macrophage progenitors, GMPs) and CD34+CD38+CD90-FLK2-CD45RA- (megakaryocyte-erythroid progenitors, MEPs)1,2 from BM and PB were isolated.

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