Clonal memory of cell division in humans diverges between healthy haematopoiesis and acute myeloid leukaemia

*Donada A[1**], Hermange G[2,3], Tocci T[1], Midoun A[1], Prevedello G[4,5], Hadj Abed L[1,6], Dupré D[6], Sun W[1], Milo I[1], Tenreira Bento S[1], Pospori C[7,8], Innes A[9], Willekens C[10], Vargaftig J[11], Michonneau D[12,13], Lo Celso C[7,8,9], Servant N[6,14], Duffy KR[15], Isambert H[1], Cournède PH[2], Laplane L[16,17] and Perié L[1,6,18*]*

1. Institut Curie, PSL Research University, Sorbonne University, CNRS UMR168, Physics of Cells and Cancer lab, Paris, France

2. Laboratory of Mathematics and Informatics (MICS), Université Paris-Saclay, CentraleSupélec, Gif-sur-Yvette, France

3. IHU-National PRecISion Medicine Center in Oncology, Villejuif, France

4. Sony Computer Science Laboratories, Paris, France

5. Enrico Fermi's Research Center (CREF), via Panisperna 89A, 00184, Rome, Italy

6. Single Cell Initiative, Institut Curie, Paris, France

7. The Francis Crick Institute, London, UK

8. Department of Life Sciences, Imperial College, London UK

9. Centre for Haematology, Imperial College, London UK

10. Department of Haematology, Gustave Roussy Cancer Campus, Villejuif, France

11. Department of Haematology, Institut Curie, Saint-Cloud, France

12. Haematology and Transplantation Unit, Saint Louis Hospital, AP-HP, Paris, France

13. INSERM U1342, Saint Louis Research Institute, IHU Leukemia Institute Paris Saint Louis, SIRIC InSitu, Université Paris Cité, Paris, France

14. Institut Curie, INSERM U900, Mines Paris Tech, PSL Research University, Paris, France

15. Northeastern University, Boston, MA, USA

16. Hematopoietic Tissue Aging, INSERM UMR 1287, Gustave Roussy Cancer Campus, Villejuif, France

17. Institut d'Histoire et Philosophie des Sciences et des Techniques, CNRS UMR 8590, University Paris I Pantheon-Sorbonne, Paris, France

18. Lead Contact

* Correspondence : leila.perie@curie.fr

** Correspondence : alessandro.donada@curie.fr

Summary

Clonal memory, a cellular property inherited across at least two divisions, has emerged as a key driver of cell heterogeneity. To uncover its roles in human haematopoiesis, we developed high-resolution *ex vivo* tools that track both division and fate commitment of individual primary human haematopoietic stem and progenitor cells (HSPCs). We show that human HSPCs display a clonal memory of division, as cells descending from the same ancestor cell divide synchronously over multiple generations. In parallel, HSPCs inherit a clonal memory of fate commitment, independently of lineage identity. Both forms of clonal memory persist over at least two divisions, across different HSPC commitment stages and cell culture conditions. In contrast, malignant haematopoiesis exhibits lower synchronicity, revealing a disruption of clonal memory in leukemic cells. Epigenetic remodelling using a bromodomain inhibitor partially restores the clonal memory in division in leukemic HSPCs, highlighting the plasticity of this trait and its potential for therapeutic modulation. Our findings position clonal memory as a key regulator of human haematopoietic stem cell behaviour. Demonstrating that clonal memory can be modulated opens new avenues for tuning cell heterogeneity in healthy and pathological tissues.

Introduction

Stem cells, and how they proliferate, play a key role in tissue development, homeostasis, and regeneration. However, not all stem cells divide or die at the same rate[1]. These proliferation differences impact the cellular composition of tissues and thereby tissue formation and renewal over time. In cancer, cell division changes directly affect the fitness of cancer cells, leading to their clonal expansion, increased malignant phenotypes and poorer prognosis[2]. Currently, we still know little about what drives differences in stem cell proliferation in healthy or diseased tissues.

Recently, clonal memory has emerged as a driver of stem cell heterogeneity[3–7]. Clonal memory is defined as a cellular property inherited over at least two successive rounds of cell division. Despite its importance in stem cell heterogeneity, the role of clonal memory in proliferation heterogeneity is poorly understood. Additionally, fundamental questions about clonal memory remain, including how long clonal memory persists over time (e.g. hours[8], days[9], months[10]) and its long-term functional impact. Although clonal memory is an intrinsic property, whether and how extrinsic factors (e.g. microenvironment, soluble factors) could modulate clonal memory remains unknown. Modulating clonal memory is crucial as it could promote the stability required for homeostasis by transmitting cellular traits across multiple cell generations. However, this stability could come at the expense of adaptability to environment changes. Thus, understanding how clonal memory contributes to heterogeneity in stem cell division is critical to determining its role in balancing stability and adaptability in both homeostasis and disease.

Haematopoiesis is a good model to address the role of clonal memory. It is a hierarchically organized differentiating system that generates over 13 different cell types across tens of

divisions[11]. HSPCs are a heterogeneous population with diverse division and differentiation abilities[12]. In mice, clonal memory of differentiation partly contributes to this heterogeneity. Indeed, HSPCs show heterogeneous differentiation biases that are maintained and transmitted throughout cell division and differentiation[13], as shown by single-cell lineage tracing[3] and serial transplantations[10]. However, most clonal memory studies have focused on differentiation, while its influence on HSPC proliferation[14], in particular in human cell, has remained unexplored. Moreover, as differentiation tends to be coupled with division[15,16] (with some exceptions), and both properties are central to stem cell biology; thus the clonal memory of both processes needs to be studied together. Critical for this aim is the ability to simultaneously track division and differentiation within individual cell lineages, using lineage tracing tools such as the MultiGen assay[14,17–19].

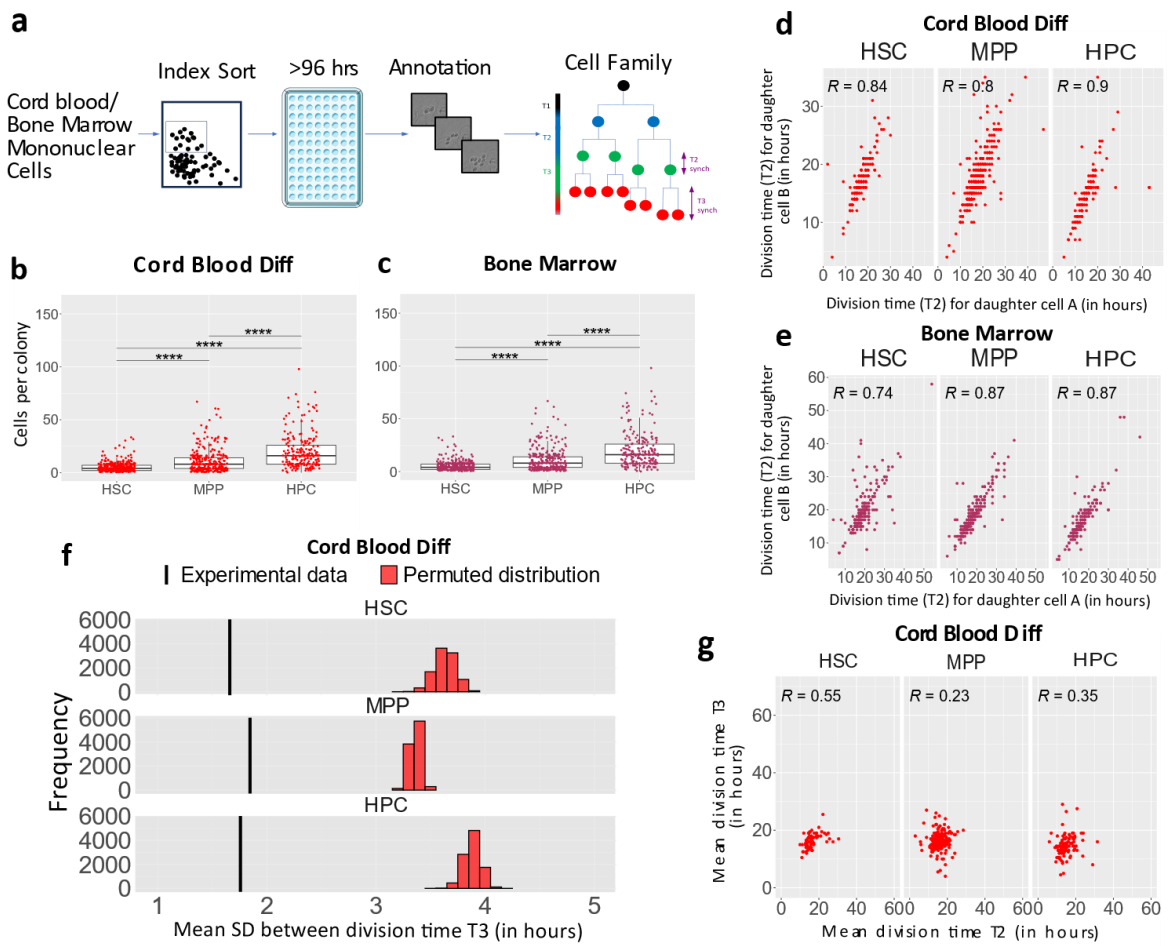
In pathological human haematopoiesis, evidence for clonal memory of division and fate in HSPCs remains largely indirect. While stem cell bone marrow transplantation induces cell proliferation to repopulate the recipient, the observation that stem cells accumulate mutations at similar rates in both recipient and donor support the hypothesis that stem cell proliferation is an intrinsic properties[20–22]. In clonal haematopoiesis, known genetic drivers account only for approximately 20% of all age-related clonal expansions, indicating that while genetic mutations contribute to proliferation heterogeneity, other intrinsic and extrinsic factors are at play[23]. In humanized mouse models, cellular memory was demonstrated to impact the heterogeneity in clonal fitness for otherwise genetically identical acute myeloid leukaemia (AML) stem cells, shaping their clonal evolution[3] and leading to differences in treatment resistance[24,25]. However, whether clonal memory can be modulated in leukaemia and leveraged for therapeutic benefit remains unexplored.

Here, we address whether clonal memory plays a role in human haematopoiesis, by studying clonal memory in primary human HSPCs during healthy and malignant haematopoiesis, focusing on division. We developed several tools to follow division, differentiation, and lineage commitment of individual human HSPC *ex vivo*. Using live-cell microscopy, we establish that human HSPCs display a clonal memory of cell division over at least two division and three cell generations across different HSPCs commitment stages and several culture conditions. Further, we demonstrate that there are two distinct inherited clonal memory properties, one for division and one for fate commitment. Counterintuitively, we find that HSPCs in malignant AML haematopoiesis show altered clonal memory in division, with a decreased level of synchronicity within cell families. By manipulating clonal memory of division through inhibition of epigenetic remodelling in AML cells, we observe an increased clonal memory but no changes in their total number of divisions, suggesting that clonal memory is regulated independently of the number of divisions. Altogether, we demonstrate the existence of distinct clonal memories that govern division and lineage commitment in healthy human primary HSPCs. Furthermore, we show that while the clonal memory of division is altered in AML, it remains amenable to modulation. This ability to modulate clonal memory opens new avenues for tuning cell heterogeneity.

Results

Clonal memory of division in human HSPC families

To measure the heterogeneity in division in human HSPCs and their descendant cells, we analysed three types of HSPCs, via single-cell live-cell imaging. We isolated hundreds of haematopoietic stem cells (HSCs; $Lin^{-}CD34^{+}CD38^{-}CD90^{+}CD45RA^{-}$), multipotent progenitors (MPPs; $Lin^{-}CD34^{+}CD38^{-}CD90^{-}CD45RA^{-}$), and haematopoietic progenitor cells (HPCs; $Lin^{-}CD34^{+}CD38^{+}$) from umbilical cord blood and adult bone marrow from unrelated donors **(Extended Data Fig. 1a).** We then cultured single HSPCs for 96 hours in two serum-free media: one to promote cell division and differentiation towards the myeloid and erythroid lineages (Diff)[26], and one to maintain the immature $CD34^{+}$ phenotype (gene therapy medium, GT)[27]. Analysing 1973 cell families, each composed of the descendants of a single HSPC, we manually extracted the time necessary to complete the first three divisions **(Fig. 1a)**. First, we observed differences of proliferation between cell types, with HPCs dividing faster and forming bigger colonies at 96 hours than HSCs or MPPs, for multiple donors in both culture media **(Fig. 1b-c, Extended Data Fig. 2m)**. Past the first division **(Extended Data Fig. 2a-b)**, the time to divide shortened not only for HSCs and MPPs **(Extended Data Fig. 2d-e, 2g-h),** as previously described[28], but also for HPCs. Therefore, cellular quiescence alone was not responsible for this shortened division time, as HPCs are not known to be quiescent[29,30] **(Extended Data Fig. 2j-k)**. Additionally, cells of the same type displayed heterogenous division times, as previously reported[31]. This heterogeneity was of similar amplitude between the three cell types (i.e., HSCs, MPPs, and HPCs) and between the culture media, although adult bone marrow HSPCs **(Extended Data Fig. 2c, 2f, 2i, 2l)** displayed more heterogeneity than foetal cord blood HSPCs. To assess the existence of clonal memory of division, we next explored whether this heterogeneity in division times between cells of the same type was inherited within a cell family. Cells within a family divided synchronously **(Fig. 1d-e),** as sister cells were highly correlated for all tested progenitors (HSCs, MPPs, and HPCs) and both culture media **(Extended Data Fig. 3a)**. Furthermore, cousin cells also divided synchronously **(Fig. 1f)**, with the four division for a given set of cousin cells displaying a lower standard deviation compared to permuted families composed of unrelated cells. This synchronicity in division was present in all HSPC types, irrespective of their origin **(Extended Data Fig. 3b)** or culture media **(Extended Data Fig. 3c)**, suggesting a clonal memory of division across commitment stages. However, the synchronicity was not an inherited trait, as cell division time across generations correlated weakly **(Fig. 1g),** similar to other cell systems. This was similar for each HSPC type, irrespective of culture media **(Extended Data Fig. 3d-e, 3g-h, 3j)** or cell origin **(Extended Data Fig. 3f, 3i, 3k)**. In sum, we observed a clonal memory of division in human HSPCs that is present in at least three cell generations, over different HSPC commitment stages, and across different culture conditions.



150

151 ***Figure 1. Division in synchronicity within cell family.*** ***a)*** *Overview of the live cell imaging experimental*
152 *workflow: individual progenitors were index sorting via FACS in 96-well tissue plates and kept in culture for at*
153 *least 96 hours. Division dynamics were manually annotated.* ***b)*** *Colony size, measured as number of light-*
154 *refringent cells in the well at T = 96 hours, for foetal cord blood HSPCs expanded in Diff medium, in 7 independent*
155 *experiments (n = 166 for HSC, n = 369 for MPP and n = 157 for HPC).* ***c)*** *Colony size, measured as number of light-*
156 *refringent cells in the well at T = 96 hours, for bone marrow HSPCs, in 8 independent experiments (n = 356 for*
157 *HSCs, n = 328 for MPPs and n = 222 for HPCs). Dunn post-hoc test with Hochberg correction, ****p < 1x10^{-4}.* ***d-***
158 ***e)*** *Correlation in division time (expressed in hours) between two daughter cells in generation 1, for foetal HSPCs*
159 *isolated from cord blood and expanded in Diff medium (7 independent experiment;, n = 149 for HSC, n = 335 for*
160 *MPP and n = 144 for HPC), and* ***e)*** *adult bone marrow cells (8 independent experiments, n = 255 for HSC, n = 269*
161 *for MPP and n = 205 for HPC). Spearman's rank correlation coefficient R.* ***f)*** *Mean SD of third division time (T3) of*
162 *cells, derived from the same ancestor cells (black line), and permuted families (red histograms), for HSPCs from*
163 *foetal cord blood and expanded in Diff medium. Only families with the four recorded T3 divisions were used.*
164 *Seven independent experiments; n = 65 for HSC, n = 170 for MPP and n = 99 for HPC), p values < 1x10^{-4}.* ***g)***
165 *Correlation in division time (expressed in hours) between the mean division time T2 and the mean division time*
166 *T3, for HSPCs isolated from foetal cord blood and expanded in Diff medium. Spearman's rank correlation*
167 *coefficient R.*

168

Clonal memory of lineage commitment in human HSPCs

To determine if human HSPCs display a clonal memory of lineage commitment, we performed the MultiGen assay[14,17,18] to track simultaneously the cell family, the number of cell divisions and the commitment phenotype of each cell produced by individual HSPCs **(Fig. 2a)**. We tested 1634 individual HSC- and MPP-derived families, from human cord blood, after 72 hours and 96 hours and using both Diff and GT media. We assigned the commitment phenotype of the progeny using an antibody panel that identifies six different HSPC subtypes **(Extended Data Fig. 1b)**. At the population level, we observed that MPP families divided and differentiated more often than HSC families **(Fig. 2b-c, Extended Data Fig. 4a-b)**, as expected from their positions in the haematopoietic hierarchy. Consistent with earlier observations[14], cells that divided more were more likely to be more differentiated than the ones that divided less **(Extended Data Fig. 4c-j)**. The large majority of immature HSPCs (i.e., HSCs and MPPs) were associated with fewer divisions, while committed progenitors and precursors emerged with more divisions **(Fig. 2d-e)**. Finally, we observed substantial heterogeneity in lineage commitments and divisions between families of the same progenitor cell type **(Fig. 2c, Extended Data Fig. 4k-r)**. Therefore, our experimental set-up mimics the heterogeneity of the lineage commitment and division found in human HSPCs.

Next, we explored the heterogeneity in lineage commitment within cell families. Cells from the same family consistently co-localized on the UMAP representation of the HSPC subsets **(Fig. 2f).** Most of the families displayed only one subset **(Fig. 2g),** despite the presence of multiple cell types in our culture system and the multipotent nature[32] of HSC and MPP. Permutation testing revealed significant intrafamilial similarities in lineage commitment, more than expected if each cell were committing independently of the sister and cousin cells. This observation was independent of the ancestor cell type, the timing of the analysis, and the culture condition **(Extended Data Fig. 5a-c)**. We concluded that a clonal memory of lineage commitment in human HSPCs is present over different HSPC commitment stages and across culture conditions.

We also observed clonal memory of division with the MultiGen assay, consistent with our results using live-imaging microscopy **(Fig. 1)**. Most dividing families displayed high concordance in division: for 1183 HSC or 1634 MPP families, all cells within the same family had the same division number (range = 0) after 72 and 96 hours. Only 38% of families were spread over more than one generation (range > 1) **(Fig 2h, Extended Data Fig. 5d-f)**. As cell death and sampling effects could potentially make families look more concordant in division, we developed an improved version of the previously published mathematical model[14] to account for sampling and cell death. We did not detect cell death during the first 96 h, using live-imaging data, and we estimated a recovery rate of 0.6 using beads. We fitted the model to the family distribution over range, to estimate the correlation in division progression decisions among cells within a family. High correlation coefficients (70–85%) **(Fig. 2h)** resulted in the best fit to the measured familial ranges, demonstrating that cell families progress through division together. We used permutation testing to assess if families sharing the same microwell will eventually influence each other, for cell division and differentiation. We did not see a microwell effect on these properties, as families sharing the same well retained both different division times and cell type composition **(Extended Data Table 1).** These results

suggest the existence of a cell-intrinsic clonal memory of division and lineage commitment over several divisions.

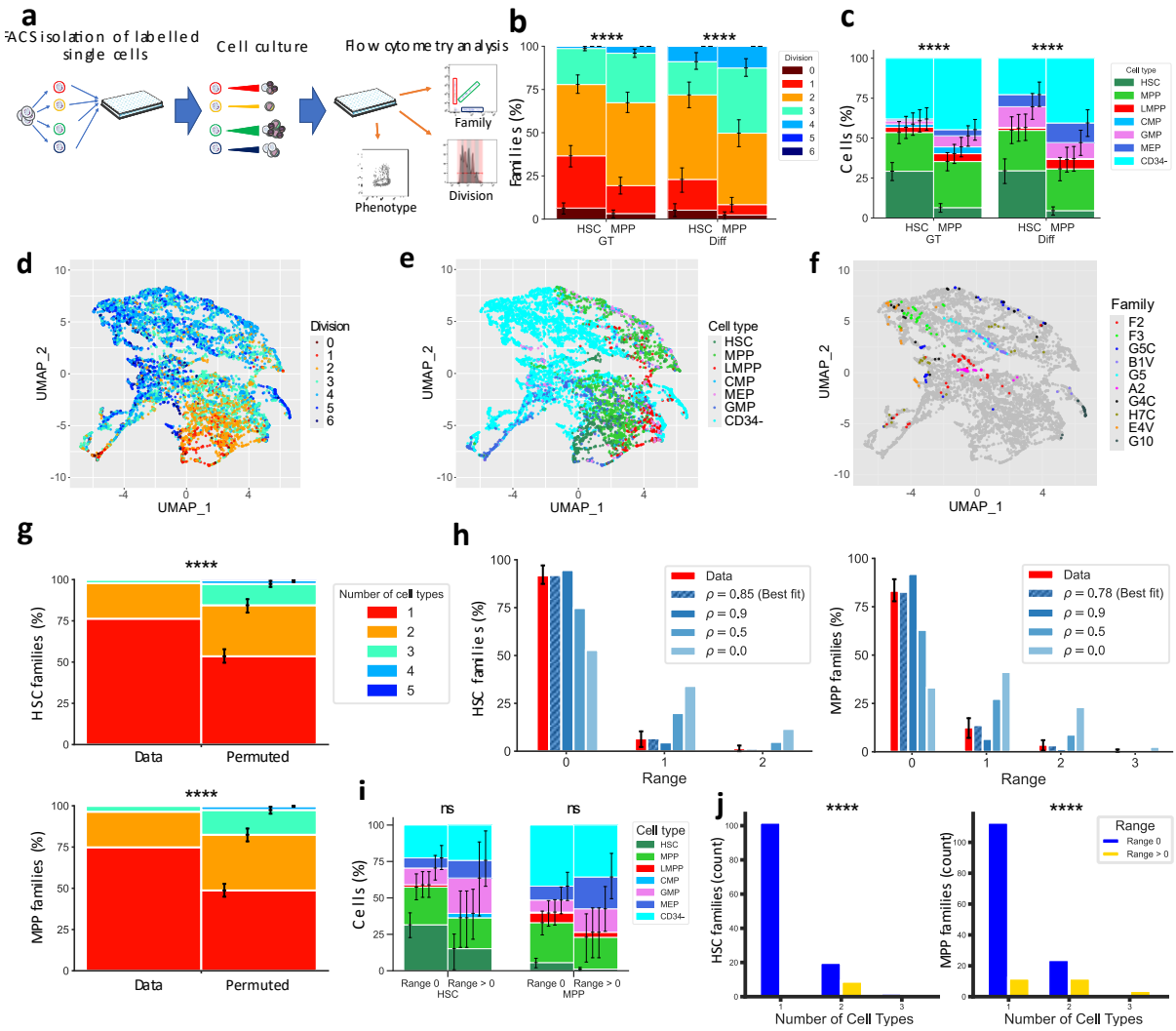


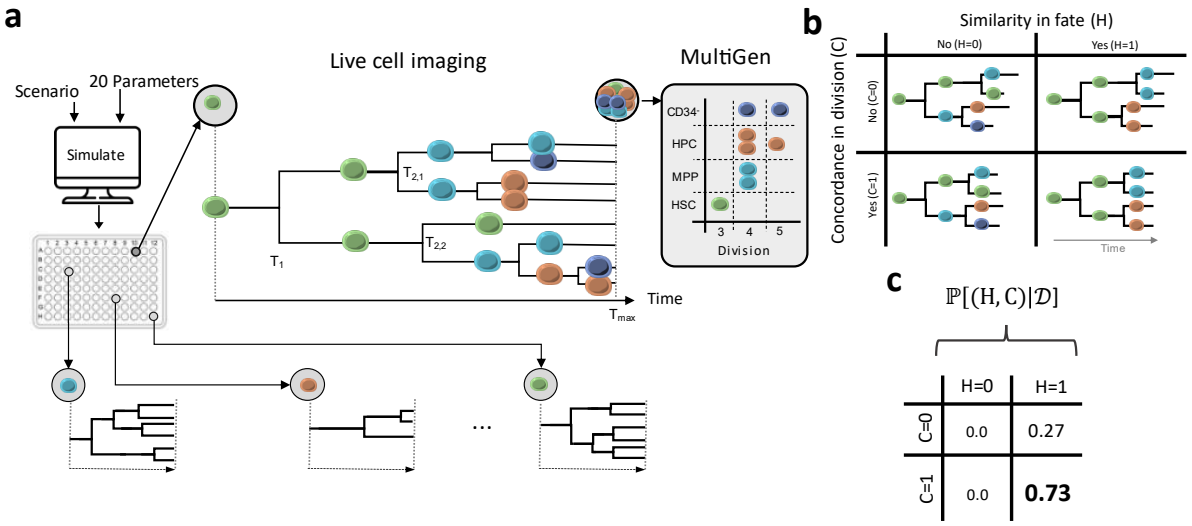
Figure 2. Clonal memory of fate and division in human HSPCs ***a)*** *Overview of the MultiGen experimental workflow: HSPCs stained with four distinct combinations of CFSE and CTV were individually sorted in 96-wells tissue plates, with four cells per well, each one from a different CTV/CFSE combination. After cell culture (72 and 96 hours respectively), plates were analysed by flow cytometry.* ***b)*** *Distribution of the maximum division reached by individual families, derived from HSCs and MPPs and expanded in two different media (Diff and GT) after 72 hours of cell culture(7 independent experiments, n = 235 for HSC GT, n = 248 for MPP GT, n = 135 for HSC Diff, n = 167 for MPP Diff).* ***c)*** *Proportions of cell types per families derived from HSC and MPP and cultured in the GT and Diff media for 72 hours.* ***d-f)****, UMAP projection of 6148 individual cells from 13 independent experiments, each cell coded by* ***d)*** *division (based on CTV/CFSE fluorescence),* ***e)*** *phenotype (based on cell surface markers fluorescence) and* ***f)*** *family (displaying the ten largest families).* ***g)*** *Distribution of the number of cell types per family, compared with the distribution derived from 250,000 permutations of the experimental data (HSCs and MPPs in Diff medium, 72 hours ex vivo).* ***h)*** *Percentage of families per division range, defined as the difference between divisions within each family (range = 0 means that all the cell in the family divided the same number of times). The red bar corresponds to the experimental data (135 and 167 families for HSC and MPP respectively, in Diff medium for 72 hours). The dashed blue bar corresponds to the best fit value to the experimental data of the correlation coefficient in division r between sister cells. The filled blue bars result from simulations, parametrized by different values of the correlation coefficient r.* ***i)*** *Proportions of cell types per families that have a division range = 0 or > 0, derived from HSC or MPP cultured in the Diff media for 72 hours.* ***j)***

*Number of progenitor subsets per family, for families with range = 0 or > 0. For the panels **b, c, g, i** and **j** the error bars indicate 95% confidence intervals calculated via basic bootstrap. The p values are calculated using the permutation test, **** $p < 1x10^{-4}$.*

Two distinct clonal memories for division and lineage commitment within cell families

While we observed a clonal memory in division and lineage commitment within HSPC families, we also observed that lineage commitment progresses with division. Therefore, the clonal memory in division could be due to the clonal memory in lineage commitment, and vice-versa. For example, while concordant and discordant families showed the same repertoire of cell subsets **(Fig. 2i, Extended Data Fig. 5g-i)**, individual discordant families (range > 0) displayed more progenitor subsets than concordant families **(Fig. 2j, Extended Data Fig. 5j-l)**. We next tested whether the presence of two distinct clonal memories of division and lineage commitment would best explain the similarity in lineage commitment and concordance in division. To this end, we applied a stochastic modelling framework with approximate Bayesian computation (ABC)[33–35] to the live-cell imaging and the MultiGen data combined. Briefly, we built a generic model of the HSPCs division and differentiation over 96 h **(Fig. 3a)** starting from one cell of a known phenotype in a well and considering four possible phenotypes HSC, MPP, HPC, or $CD34^-$. In the model cells divided at random times, according to log-normal probability distributions[36,37] for each cell type. Differentiation only happened after a division[38]. Asymmetric and symmetric division were allowed, but de-differentiation was not allowed. We tested four distinct modelling scenarios, whereby cells from the same family could be similar (or not) in lineage commitment and/or concordant in division **(Fig. 3b)**. To compare how the different models fit the entire dataset, we used a distance that aggregated the derivation of many summary statistics. We ran 10,982,900 simulations with different parametrizations of the four models and then compared the posterior probability of the models. We found that the models lacking similarity in commitment were inconsistent with the data **(Fig. 3c)**. The best-fitting model displayed both similarity in commitment and concordance in division for cells from the same family. The model in which cells from the same family were only similar in commitment but not concordant in division had a lower posterior probability **(Fig. 3c).** Overall, we demonstrated that two distinct clonal memories for division and lineage commitment underlie the similarity in lineage commitment and concordance in division.

Figure 3. Two distinct clonal memories of division and lineage commitment ***a)*** *Overview of the modeling strategy: proliferation and differentiation for each well were simulated by agent-based modeling for different scenarios. For each simulation, simulated data were extracted to be compared with both the live-cell and MultiGen datasets. From the live-cell imaging datasets, the experimental parameters used were the values of time of 1st (T_1), 2nd ($T_{2,1}$ and $T_{2,2}$) and 3rd divisions ($T_{3,1}$ to $T_{3,4}$) and the colony size at T_{max}. From the MultiGen datasets, the relative percentages of cell types and the associated cell divisions were used. Green, cyan, orange and dark blue correspond to HSC, MPP, HPC, and $CD34^-$ types, respectively.* ***b)*** *The four distinct modelling scenarios tested, based on whether there is similarity in fate (H=1) or not (H=0), and whether there is concordance in division (C=1) or not (C=0).* ***c)*** *Value of the posterior probability of each scenario, given the experimental observations, after running the ABC-SMC (Approximate Bayesian Computation – Sequential Monte-Carlo) procedure.*



Clonal transcriptional programs of clonal memory in human HSPCs

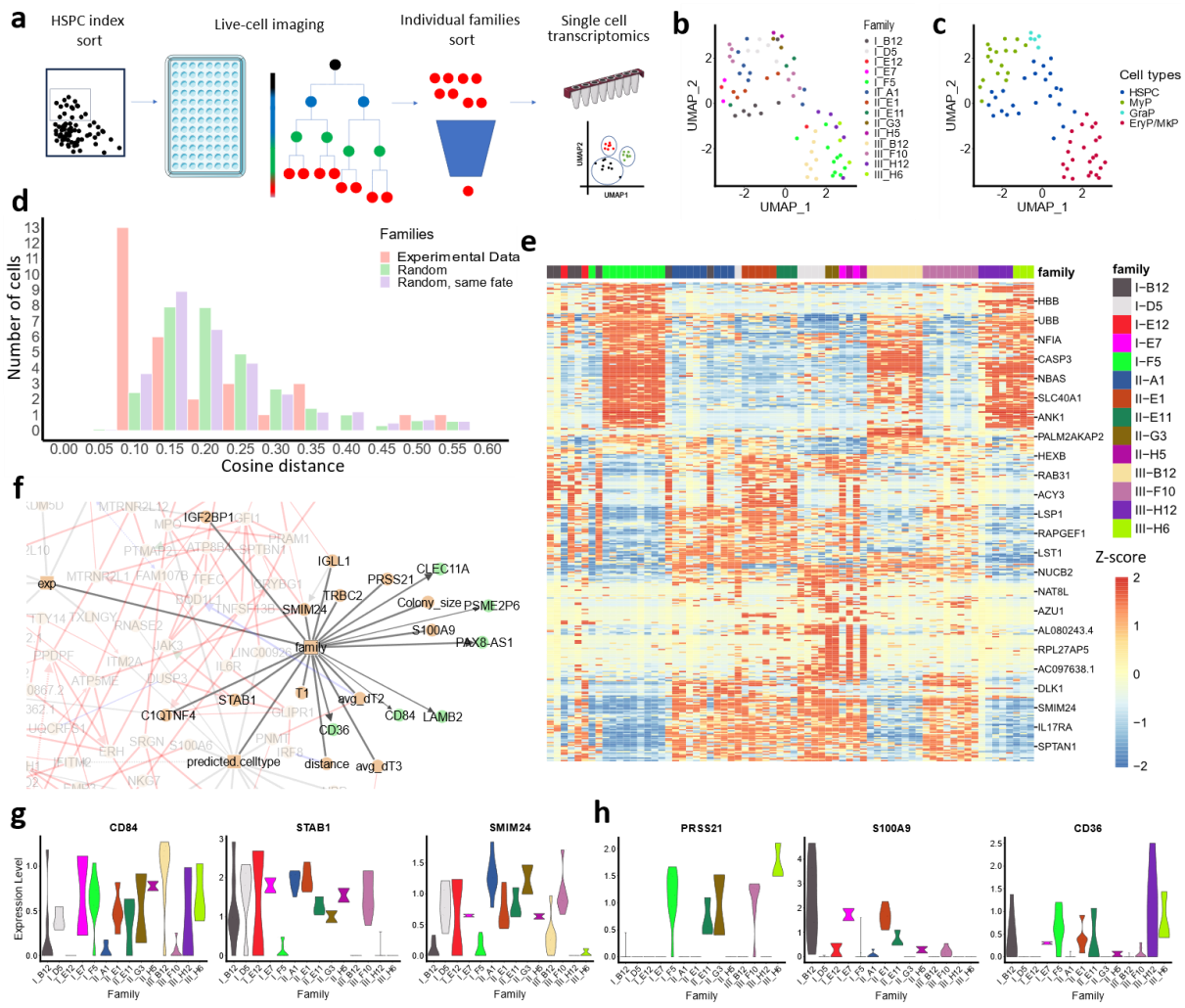
To explore the transcriptional program involved in the clonal memory of human HSPCs, we developed an experimental protocol integrating lineage tracing and molecular profiling with the division history of each cell. Briefly, we FACS-sorted individual MPPs from two cord blood samples and cultured them *ex vivo* in the Diff media while recording their division history by live-cell imaging over 90 hours. We then re-isolated individual cells from randomly selected wells, using low throughput, multi-fluorescence image sorting[39], and profiled each cell with high-coverage single-cell RNA sequencing (scRNA-seq) via SMART-seq3[40] **(Fig. 4a)**. We obtained two multidimensional datasets containing information on division, family, and transcriptome for 142 individual cells from 30 families after QC and filtering **(Fig. 4b, Extended Data Fig. 6d)**. After supervised annotation using known marker genes, we identified several cell subsets: immature progenitors (HSPCs), erythroid-megakaryocytic progenitors (EryP/MkP) and myelomonocytic committed progenitors (MyP and GraP) **(Fig. 4c, Extended Data Fig. 6e-f)**. These subsets globally matched the subsets we identified via CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing) profiling of 8,580 cord blood HSPCs after 4 days in Diff media (**Extended Data Fig. 6a-c**)

We next explored the similarity of cell families in transcriptomics space. We observed that cell families co-localized within the same UMAP clusters, suggesting that cell from the same family share a similar transcriptome. These clusters were more associated with the annotated cell type than with the number of divisions performed **(Extended Data Fig. 6f)**. To quantify the transcriptomic similarity between cells from the same family, we compared the median cosine distance[41] distributions between the observed cell families and artificial families made of permutated cells. Cell families had lower median cosine distance than permutated families **(Fig. 4d)**, demonstrating that cells within a family are more similar transcriptomically than cells between families and in line with the MultiGen results. Moreover, this result held after permutating cells within the same subsets **(Fig. 4d),** suggesting that transcripts associated with cell subsets alone were insufficient to explain this similarity. Thus, biological processes other than those associated with lineage commitment contributed to the similarities between the cells within a family.

To characterize the transcriptional programs associated with the familial similarity[9,42], and therefore with clonal memory, we extracted the genes that varied the least within a family and the most between families **(Fig. 4e, Extended Data Fig. 6g)**. We could then identify transcriptional fingerprints specific for each family, contributing to the unique intrafamilial similarities beyond fate commitment. These signatures contained genes associated to many biological processes, including membrane trafficking (e.g., *RAB31, CAVIN2, VAMP8*), cell proliferation (e.g., *CASP3, IGFBP2*), cytoskeleton structure (e.g., *LESP1, ANK1, MYH10)*, cell signalling (e.g., *IL17RA, IL1RL1, LAT2*), and metabolism and biosynthesis (e.g., *DLK1, CALR, XBP1*) **(Fig. 4e; Extended Data Table 2)**. To identify common genes between families that contributed to clonal memory, we ran the MIIC causal network reconstruction algorithm[43] on both transcriptomic datasets using the transcriptomes, the family and the division information from live imaging of each cell **(Fig. 4f; Extended Data Fig. 7a, Extended Data Table 3)**. The MIIC algorithm is a multivariate information-based network reconstruction method[44,45] that reconstruct causal and non-causal graphical models from heterogeneous data. The MIIC analysis identified the variables "family" (sharing the same ancestry), "predicted cell type" and "distance" (transcriptomic similarity) as key hubs in the networks **(Extended Data Fig. 7b)**. Gene expression of *IGLL1* and surface markers like *CD36* and *CD84*, amongst others, were strongly associated with the "family" hub **(Fig. 4f)**. Notably, the transmembrane protein STAB1, which directly connected with the "family" hub, emerged as a central hub within the gene network, orchestrating interactions with a broad array of other genes **(Extended Data Fig. 7c)**. In contrast, variables associated with division properties, such as division time and colony size, were only directly connected to less than 5 genes, and were strongly related to the "family" variable **(Extended Data Fig. 7d**), as all these families were synchronous in division **(Fig. 1d-f)**. The limited number of genes associated with division-related properties suggests that the transcriptome is not a suitable proxy for these properties, or that the relevant genes are expressed at low levels and thus remain undetected. Finally, the genes associated to the variable "family" were not equally expressed by all families **(Fig. 4g-h, Extended Data Fig. 7e-f)**, suggesting the existence of different transcriptomic program between families. Some genes were expressed in all families but at different levels **(Fig. 4g)**, while others were specific to some families **(Fig. 4h)**. In summary, human HSPC families displayed heritable transcriptomes linked to several cellular processes, including cell commitment, that contributed to clonal memory.

Figure 4. Clonal transcriptional programs of clonal memory in human HSPCs ***a)*** *UMAP representation of CITE-seq data from 8580 foetal HSPCs, cultured for 96 hours in Diff condition, color-coded by the annotated cell types.* ***b)*** *Individual progenitors (MPP) were index sorted via FACS in 96-well plates and live-imaged for at least 90 hours (Diff medium). Then, individual cells from 30 randomly selected families from three 96-well plates were further sorted using an image-based cell sorter, distributed into 384-wells plates and processed accordingly to the SMART-Seq3 protocol.* ***c)*** *UMAP representations of a single cell transcriptomic dataset (70 cells), color-coded accordingly to family code (left) and annotated cell type (right), one experiment representative of two.* ***d)*** *Distribution of the median cosine distance in gene expression for actual cell families (red) and two permutation-generated distributions, either retaining (purple) or discarding (green) cell types of the permuted cells. Wilcoxon rank-sum test, p value < 0.0035 for actual family/permutation distribution discarding cell type and p value < 0.02 for actual family/permutation distributions retaining cell type.* ***e)*** *Heatmap displaying the z-score of the 371 most variant genes (rows) between families, with 24 randomly selected genes were annotated on the right side. Cells (columns) are color-coded according to their original family.* ***f)*** *Snapshot*

of the MIIC network generated using the transcriptomic data and the divisional data measured via live-cell imaging. Cellular and gene features directly connected to Family are highlighted, variables in orange are causal regarding Family, and those in green are consequential. Edges represent an association between two variables, while edges with an arrow represent a causal relationship. Red edges highlight positive association, blue edges represent negative associations, and grey edges represent categorical association with no sign. A green arrow indicates genuine causality, while a red or blue or grey arrow stands for putative causality (i.e. the presence of an unobserved common cause between the two variables cannot be ruled out). Full network with detailed legend is shown in Extended Data Figure 8a. ***g-h)*** *Expression in the dataset A of a selection of "family"-associated genes accordingly to the MIIC analysis, displayed per individual family and organised accordingly to* ***g)*** *genes expressed in almost every family and* ***h)*** *genes expressed only in a subset of families*



The clonal memory of division is disrupted in leukemic cells

AML is a proliferative disease with a differentiation blockade, resulting in a disproportionate production of blast cells. To test if this disruption affects the cellular clonal memory, we analysed primary CD34+ blasts from AML patients **(Extended Data Table 4),** that had been isolated from bone marrow or peripheral blood and cultured for at least 72 hours in Diff media supplemented with IL-1β[46], using live-cell imaging and MultiGen. By live-cell imaging, the intragenerational correlations in the division time between daughter and cousin cells were diminished **(Fig. 5a, Extended Data Fig. 8a).** The MultiGen assay showed that $CD34^+$ AML cells possessed a lower clonal memory of division than healthy HSPCs, as the percentage

372 of families in range 0 was lower **(Fig. 5b)** than for cells from healthy cord blood **(Fig. 2h).**
373 Stratifying the analysis per ancestor cell types (i.e., HSCs, MPPs, HPCs), we found that the
374 clonal memory of division reduction was the same for all cell types **(Extended Data Fig. 8b,**
375 **8i-j),** even if we did not observe significant changes in the differentiation profile **(Extended**
376 **Data Fig. 8g)** or in the number of divisions over the 72 hours **(Extended Data Fig. 8h)**. As blasts
377 displayed aberrant division time and proliferation potential, the reduction in synchronicity
378 and concordance in division was not simply due to fewer divisions **(Extended Data Fig. 8i)**.
379 This revealed that primary CD34+ blasts from AML patients had a lower synchronicity of
380 division within families.

381 We observed some variability in the synchronicity in division between patient samples **(Fig.**
382 **5c),** with best-fitted correlation coefficients of division within families varying between 53%
383 and 72%. Nevertheless, this variability was uncorrelated to the patient mutation profile,
384 clinical history **(Extended Data Table 4)** or the ability of the blasts to expand *ex vivo* **(Extended**
385 **Data Fig. 8k-n)**. These findings demonstrated a reduction in the synchronicity of primary AML
386 blast divisions, suggesting a disruption of the clonal memory that orchestrates cell division.
387 Decreasing clonal memory could be a strategy for AML blasts to burst their heterogeneity in
388 division and promote adaptability over stability.

389 *Clonal memory of division within AML HSPC families can be modulated*

390 Finally, we assessed if the clonal memory in division was a plastic trait that can be modulated.
391 As we observed that the clonal memory was independent of genetic mutations, we
392 hypothesized an epigenetic mechanism as central to clonal memory[3,7,9,42]. To this end, we
393 tested the effects of JQ1, a pan-bromodomain and extra-terminal domain (BET) inhibitor[47] on
394 the clonal memory of AML cells.

395 First, we used a leukaemia stem cell (LSCs) model of differentiating AML[48,49], the OCI_8227,
396 to test the effect of JQ1 on clonal memory **(Extended Data Fig. 9a)**. Without JQ1, the
397 OCI_8227 cells displayed a decreased clonal memory in division, similarly to primary CD34+
398 blasts, with fewer number of families in range 0 and low correlation coefficients of division
399 within families **(Fig. 5d).** After exposure to sublethal concentrations of JQ-1 (5 nM), the cell
400 families that divided simultaneously increased by approximately 10%, and the best fitted
401 correlation coefficient by 1.5-fold **(Fig. 5d).** We concluded that the clonal memory of cell
402 division could be modulated, in this case increased with JQ-1. As the JQ-1 treatment affected
403 the differentiation profile **(Extended Data Fig. 9b)** but not the total number of divisions
404 **(Extended Data Fig. 9c),** we concluded that JQ-1 inhibition modulated only the division clonal
405 memory but not the division capacity of cells, suggesting that these two processes are
406 regulated by different mechanisms[50].

407 We obtained similar results in primary $CD34^+$ blasts from AML patients. JQ-1 treatment
408 increased the synchronicity between sister cells during their second divisions **(Fig. 5e)** without
409 significantly modifying their division time **(Extended Data Fig. 9d-g).** Using the MultiGen
410 assay, JQ-1 treatment also increased the percentage of families in range 0 and the correlation
411 coefficients of division within families **(Fig. 5f)** as compared to vehicle, without changing
412 either the family progression through divisions **(Fig. 5g)** or their surface phenotype **(Extended**

Data Fig. 9h). These results were confirmed using samples from two other patients **(Extended Data Fig. 9i-p)**. Thus, as the clonal memory in division increased *ex vivo* in primary AML CD34+ blasts when exposed to an agent acting on the cellular epigenetic landscape, we concluded that clonal memory can be modulated.

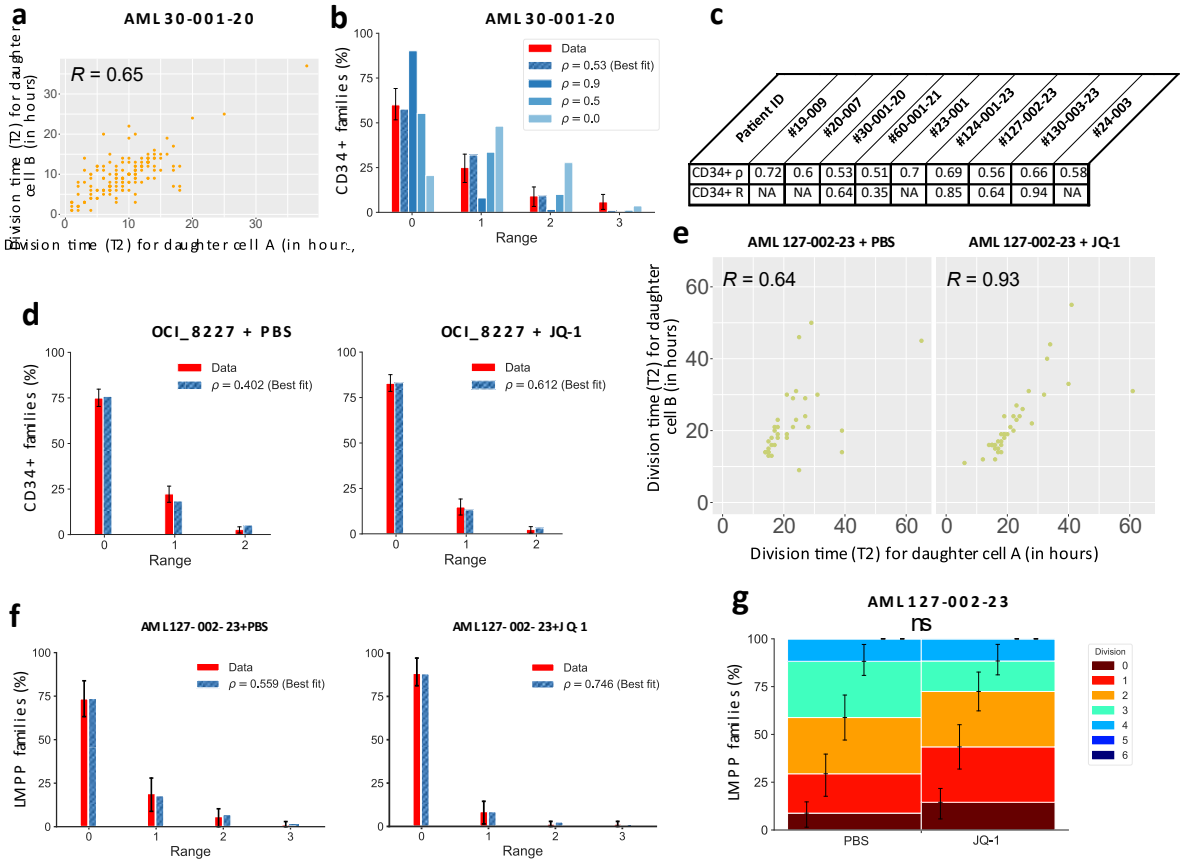


Figure 5. Reduced concordance in division of leukemic families can be restored ***a)*** *Correlation in division time (expressed in hours) between two daughter cells in feneration 1, for CD34⁺-derived families from AML patient #30-001-20. Spearman's rank correlation coefficient R (n = 154).* ***b)*** *Percentage of CD34⁺-derived families per division range, for an AML patient (#30-001-20), after 72 hours of cell culture. The red bar represents the experimental data, the dashed blue bar represents the best fit of the correlation coefficient r (likelihood for a given cell to divide), the filled blue bars represent simulations for different parametrization of r, n = 120 families.* ***c)*** *Summary of correlation coefficients as in a) and b) for the 9 AML samples analysed with live-cell imaging, MultiGen or both assays. NA for not available.* ***d)*** *Percentage of families per division range for OCI_8227 treated for 72h with sublethal-doses of JQ-1 (treatment) or PBS (control). The red bars correspond to the experimental data (n = 250 and 327 respectively, four independent experiments), the dashed blue bar represents the best fit of the correlation coefficient ρ. We used a recovery rate r of 0.6 and a r_surv of 0.95.* ***e)*** *Correlation in division time (expressed in hours) between two daughter cells in generation 1, for LMPP-derived families for AML patient #127-002-23 exposed to PBS and JQ-1 for 96 hours (n = 32 for both). Spearman's rank correlation coefficient R.* ***f)*** *Percentage of families per division range for LMPP-derived families from AML patient #127-002-23, treated for 72h with sublethal doses of JQ-1 or PBS (n = 69 and n = 68 respectively). The red bars correspond to the experimental data, the dashed blue bar represents the best fit of the correlation coefficient r.* ***g)*** *Distribution of the maximum division reached by individual families treated for 72h with sublethal doses of JQ-1 and PBS (n = 69 and 68 respectively). The error bars indicate 95% confidence intervals calculated via basic bootstrap. The p values are calculated using permutation testing, ns for p > 0.05.*

Discussion

Our study revealed that human HSPCs display two distinct forms of clonal memory— one for division and one for fate commitment. These clonal memories were consistently observed across multiple progenitor types, for more than two successive divisions, and diverse culture conditions, suggesting that clonal memory is a fundamental and conserved feature of early haematopoiesis. To our knowledge, this is the first demonstration of both forms of clonal memory in human HSPCs. Clonal memory of differentiation has been previously described in murine hematopoiesis[51,52], and we previously identified clonal memory of division in mouse haematopoiesis[14]. They also align with indirect *in vivo* evidence of proliferation inheritance in twin patients developing the same neoplasms[53,54] and in donor and recipient of bone marrow transplants [20–22]. Although clonal memory of division has been reported in other cell systems[55–57], such memory has not yet been described in differentiating systems.

While differentiation and division are linked, we demonstrated that there are two distinct forms of memory for division and differentiation, raising the possibility that they are governed by separate regulatory mechanisms. This echoes findings in B cell differentiation, where division and differentiation are independently clonally inherited[58]. Whether these two memory types are independent or coordinated remains an open question. We also identified that transcriptional fingerprints, gene modules shared by each member of that family, extend beyond differentiation states[42], implicating clonal identity in shaping gene expression. Proliferative parameters such as clone size and division timing are poorly captured at the transcriptomic level, indicating that these features may be regulated through epigenetic or post-transcriptional mechanisms. Our results confirm that clonal analysis by transcriptomics[3,4,55–57,59] is an effective method when the cells analysed display a strong clonal memory (cells from the same family display high transcriptional similarity). This should be confirmed beforehand, as we could observe a non-negligible fraction of discordant, more heterogeneous families, which could bias the clonal analysis by transcriptomics.

The existence of clonal memory in healthy haematopoiesis represents a form of non-genetic inheritance under non-pathological conditions. For this reason, we use the term *family* rather than *clone* to describe cell genealogies independently of the underlying cellular genetics. Our work focused on early haematopoiesis, the stage at which initial commitment decisions are made. Further studies are needed to determine how long clonal memory persists through subsequent stages of hematopoietic differentiation.

Functionally, clonal memory of division may serve to maintain proliferative heterogeneity among haematopoietic stem cells, a feature critical for homeostasis and stress-induced haematopoiesis. Such heterogeneity could enable distinct cell families to contribute to steady-state haematopoiesis versus responses to environmental fluctuations (e.g. inflammation, blood loss). Unexpectedly, we observed a marked reduction in clonal memory of division in AML, which could be partially restored by treatment with the epigenetic modifier JQ-1. This demonstrates that clonal memory is not only regulated, but can also be tuned pharmacologically. While AML is composed of genetically distinct clones, our data reveal that proliferation rates vary even within genetically defined families. This intra-clonal variability may reflect a strategy by which leukemic cells reduce memory to enhance

480 adaptability. Whether this reduction is a dynamic, reversible process or a fixed trait of
481 selected clones remains to be determined.

482 The ability to modulate clonal memory pharmacologically opens new therapeutic avenues. By
483 shifting the balance from plasticity toward stability, it may be possible to reduce intra-tumour
484 heterogeneity and sensitize malignant cells to treatment. More broadly, our findings position
485 clonal memory as a key regulatory layer in stem cell biology, with implications for both normal
486 development and disease.

487 **Limitations of study**

488 To enhance the reproducibility of our findings, we employed several culture conditions,
489 including at least two distinct media and multiple timepoints, to ensure the reproducibility of
490 our observations. However, future studies should aim to better capture the influence of the
491 microenvironment, which we could only partially addressed here.

492 We prioritized the analysis of many individual families per sample, particularly for the
493 pathological ones, and the inclusion of a representative selection of AML samples.
494 Nonetheless, our study faced the classical technical challenges with primary AML samples,
495 notably: (1) the limited capacity for *ex vivo* expansion of primary leukaemia samples,
496 especially in the absence of supportive stromal cells, and (2) the inherent constraints of single-
497 cell culture systems. Overcoming these limitations will require the development and
498 implementation of advanced culture platforms to assess the proliferative and differentiation
499 properties of AML cells more accurately.

500

501 **Material and Methods**

502 **Ethical approval, banking, and processing of human samples**

503 De-identified umbilical cord blood units were collected accordingly to the guidelines defined
504 by the Saint-Louis Hospital Umbilical Cord Blood biobank (authorization AC-2016-2759) and
505 in accordance with the Declaration of Helsinki.

506 Human Bone Marrow samples were collected accordingly to the guidelines defined by the
507 Saint-Louis Hospital Institutional Review board (IRB 00003888), authorization 21-799.

508 Human AML primary samples (bone marrow or peripheral blood) were obtained from
509 patients with informed consent for sample collection and use in research were obtained
510 accordingly to the IRB-approved procedures from Institut Curie (Saint-Cloud, France), Gustave
511 Roussy (Department of Clinical Haematology and Drug Development Department DITEP,
512 Villejuif, France) and Hammersmith Hospital (London, UK). A summary of the patient samples
513 used for this study can be found in **Extended Data Table 4**.

514 Mononuclear Cells (MNC) were isolated via Ficoll density gradient. Human bone marrow cells
515 were systematically processed fresh, while umbilical cord blood units and human AML
516 primary samples underwent CD34-enrichment using immunomagnetic microbeads (MACS

517 Miltenyi). Total MNCs or enriched CD34$^+$ fractions were immediately processed or frozen in
518 Fetal Bovine Serum (FBS, Eurobio Scientific) supplemented with 10% dimethyl sulfoxide
519 (DMSO, Sigma) for further analysis.

520

521 **Fluorescence Activated Cell-Sorting (FACS) and single-cell isolation**

522 Total MNCs or enriched CD34$^+$ fractions were prepared for immunostaining as previously
523 described[17]. The list of antibodies used with the respective fluorochromes can be found in
524 Extended Data Table 3. Single-cell FACS was performed using BD Aria III or BD Fusion
525 instruments (Becton Dickson), in 96-wells U-bottom tissue culture vessels (Falcon). Cells were
526 isolated using single cell sorting precision when individually isolated or Purity precision when
527 isolated in bulk. The sorted cells were indexed for all the fluorescence parameters, using the
528 Index sorting function, to record the fluorescence levels for each marker. Data analysis was
529 performed using the software FlowJo v10.1 (BD Biosciences).

530

531 **MultiGen cell dye labelling**

532 The cell dye labelling step was performed as previously described[17]. Briefly, cells were stained
533 with four combinations of 5-(and 6)-carboxyfluorescein diacetate succinimidyl ester (CFSE,
534 ThermoFisher Scientific) and CellTrace Violet (CTV, ThermoFisher Scientific): CFSE only,
535 CFSEhigh + CTVlow, CFSElow + CTVhigh and CTV only.

536 For Umbilical Cord Blood HSPCs, the labelling wasperformed in two sequential steps, starting
537 with the CFSE staining and concluding with the CTV staining. The final concentration used for
538 the labelling are respectively CFSE 5 µM (CFSE only and CFSEhigh CTVlow), CFSE 2.5 µM
539 (CFSElow CTVhigh), CTV 5 µM (CTV only and CFSElow CTVhigh) and CTV 2.5 µM (CFSEhigh
540 CTVlow).

541 For 8227 cells, the labelling was performed in a single step, preparing 2X solutions for each of
542 the cell dye combination. The final concentration used for the labelling are respectively CFSE
543 2.5 µM (CFSE only and CFSEhigh CTVlow), CFSE 1.25 µM (CFSElow CTVhigh), CTV 2.5 µM (CTV
544 only and CFSElow CTVhigh) and CTV 1.25 µM (CFSEhigh CTVlow).

545 For primary human AML cells, the labelling was performed in a single step, preparing 2X
546 solutions for each of the cell dye combination. The final concentration used for the labelling
547 are respectively CFSE 5 µM (CFSE only and CFSEhigh CTVlow), CFSE 2.5 µM (CFSElow CTVhigh),
548 CTV 5 µM (CTV only and CFSElow CTVhigh) and CTV 2.5 µM (CFSEhigh CTVlow).

549

550 **Short-term liquid culture of primary cells**

551 Progenitors isolated from umbilical cord blood or healthy bone marrow, either as single cells
552 or in bulk, have been cultured in two types of cell culture medium, named Gene Therapy
553 medium (GT)[27] and Differentiation medium (Diff)[26]. Both media are based on StemPro-34 SFM
554 (SP34, ThermoFisher Scientific) supplemented with Penicillin – Streptomycin 1X (Pen/Strep,

555 ThermoFisher Scientific), L-Glutamine 2 mM (ThermoFisher Scientific) and human Low-
556 Density Lipoprotein 500 ng/mL (LDL, Stem Cells). For GT medium we added the following
557 cytokines: SCF 300 ng/mL (Peprotech), FLT3-ligand 300 ng/mL (FLT3l, Peprotech), TPO 100
558 ng/mL (Peprotech), IL-3 60 ng/mL (Peprotech). For Diff medium: SCF 100 ng/mL, TPO 100
559 ng/mL, IL-6 50 ng/mL (Peprotech), GM-CSF (Miltenyi) 25 ng/mL, FLT3l 20 ng/mL, EPO 1 U/mL
560 (Janssen) and IL-3 10 ng/mL. Progenitors isolated from AML samples, both as single cells
561 orbulks, were cultured in a slightly modified version of the Diff medium, containing IL-1b 50
562 ng/mL (Peprotech) in addition. A complete list of references for cytokines and small molecules
563 is available in **Extended Data Table 5**.

564

565 **MultiGen flow cytometry assessment of cell division, cell differentiation and cell family**

566 Flow cytometry experiments, both using umbilical cord blood progenitors and 8227 cells,
567 were performed as previously described[17]. The list of antibodies used with the respective
568 fluorochromes can be found in **Extended Data Table 6**. Data analysis was performed using
569 the software FlowJo v10.1 (BD Biosciences) and custom Python scripts.

570

571 **Time-lapse imaging**

572 Haematopoietic progenitors from umbilical cord blood or adult bone marrow were, as
573 previously mentioned, individually sorted via FACS in 96-wells U bottom plates (Falcon) and
574 then continuously imaged for 96 hours, using an Incucyte Zoom system (Sartorius). Cell
575 culture were performed using the same culture media previously described. Time-lapse
576 experiments were conducted at 37 °C, 5% CO_2, with time intervals of bright field image
577 acquisition of 1-2 hours. Cell viability and cell division were assessed manually: high
578 refractility, membrane integrity and cell size were the criteria used to assess viability, and cell
579 death time was measured as the first image when visible blebbing could be observed,
580 together with cell fragmentation, loss of circularity and lack of cell movement. The first cell
581 division time was measured as the time between the beginning of the experiment and the
582 time point at which two cells could be clearly distinguished. For the second and third divisions,
583 the time was measured as the interval between the end of the former division and the time
584 point at which two novel cells could be clearly distinguished and were adjacent in the same
585 region. We stopped to measure the time for cell division after the third division, due to cell
586 motility making impossible to reliably track each individual cell manually.

587

588 **Cell line culture and *ex vivo* treatments**

589 A total of 8227 cells were maintained in culture for no more than six weeks at the time. Cells
590 were passaged every 5 days in T-25 or T-75 cell culture-grade flasks, at a cell concentration of
591 1.5-3 x 10^5 cells/mL. Cells were cultured in StemSpan II (Stem Cells) supplemented with
592 Pen/Strep 1x, SCF 50 ng/mL, TPO 50 ng/mL, FLT3l 50 ng/mL, G-CSF 10 ng/mL, IL-6 10 ng/mL
593 and IL-3 10 ng/mL. 8227 cells were treated *ex vivo* using sublethal doses of JQ-1 (Sigma): the

working solutions were prepared accordingly to the manufacturer's instructions, while the used doses were defined as the lowest dose inducing a cell viability of 90%.

Confidence Intervals for MultiGen analysis

The confidence intervals at 95% level were calculated via basic bootstrap[60] with 250,000 bootstrap datasets. Following this procedure, each bootstrap dataset was constructed by sampling with replacement as many cellular families as were in the original data. The distribution of the statistics, each calculated from one bootstrap dataset, then provided a reference from which the confidence interval was derived. Formally, given the statistics ϑ^ calculated from the original data, and $\vartheta^*_{(0.025)}$ and $\vartheta^*_{(0.975)}$ percentiles, respectively derived from the bootstrapped distribution, the confidence interval of ϑ^ at 95% level was calculated as follows:

$$\mathrm{CI}(95\%) = (2\vartheta - \vartheta^*_{(0.975)},\ 2\vartheta - \vartheta^*_{(0.025)})$$

Cell division range analysis

To quantify how correlated were the cell decisions to divide, for cells from the same generation and descending from the same ancestor, we employed a previously used stochastic model[14,18,19], upgraded to account for cell death. Biologically, each cell reaching a certain generation have a chance of surviving that is independent from other cells. All surviving cells can then divide or stop dividing, with a given probability that is correlated between cells from the same family and generation. When sampling, cells that survive but divide less tend to be also less detected, simply because they are less abundant. Thus, it is important to take both survival and sampling into account when measuring the range of division within a family. In this model, cells have a certain probability to survive before dividing, as defined by the survival rate r_{surv}. Cells have also a certain probability to be measured, defined by the recovery rate r. The correlation coefficient that links the cell division within the same family was fitted to the data and was used to evaluate the degree of correlation in division within families, otherwise called concordance in division. The model was directly parameterized by the data, apart from one variable that encapsulated the correlation in decision-making that fitted to the data. With n being the maximum generation recorded, let $p_i \in [0, 1]$ for $i = 0, \ldots, n$ denoted the proportion of cells that divided from generation i to the next, which was determined from the data as follows. Set z_i the total number of cells recovered in generation I, the p_i was estimated by

$$p_i = \frac{\sum_{j=i+1}^{n} z_j\, 2^{-j}}{\sum_{j=i}^{n} z_j\, 2^{-j}}$$

For $i = 0, \ldots, n$-1 and by p^$_n = 0$.

In this model, given k_i the number of cells from a particular family that reach generation i, the number of cells that continued dividing to $i + 1$ followed a beta-binomial distribution with a

number of trials that was a random variable K_i (independent of all the others) and parameters $a_i = p_i(1-\rho)/\rho$, and $b_i = (1-p_i)(1-\rho)/\rho$, namely $\beta(K_i, a_i, b_i)$, where $\rho \in [0, 1]$ was a free parameter.

To account for cell death, $K_i = B(k_i, r_{surv})$ followed a binomial distribution with k_i that reached generation i and survived to divide. r_{surv} was estimated from time-lapse microscopy experiments for each cell type and specified in the results section. Finally, one family was generated recursively by setting $k_o = 1$ and defining $k_{i+1} = 2\beta(K_i, a_i, b_i)$. The recovery rate r was determined experimentally using microbeads and quantified as r = 0.6. In the model, we accounted therefore for the sampling effect by sub-sampling each cell from every artificially generated family with a probability of 0.6 and independently of all other cells. The beta-binomial model interpolated between surviving cells that decided to divide independently of each other ($\rho = 0$) and cells that were perfectly aligned in their decision to divide ($\rho = 1$). A value between 0 and 1 reflected the concordance level within each family in division, but, by design irrespective of the values of p_i, that determined the population distribution among generations. Defining the family range as the difference between maximum and minimum generation for all recovered surviving cells per family, the best fit p^ of p was determined to be the value from 0 to 1 that maximized the likelihood of recapitulating the family range distribution in the data. The 95% confidence interval for p^ was calculated following the basic bootstrap procedure previously described, that was by sampling with replacement the cellular families and setting the maximum-likelihood estimation as p^ the statistic to be calculated on the bootstrapped datasets. To assess the significance of p^, the statistical hypothesis $H_0 : \rho = 0$ was tested against the alternative $H_1 : 0 < \rho \leq 1$ using the likelihood-ratio statistic

$$\lambda_{LR} = -2\,(L(0) - L(\hat{\rho}))$$

where $L(\rho)$ was the log-likelihood of the cellular family ranges under the beta-binomial model with parameter p. The p-value for the test was derived from the asymptotic distribution of λ_{LR} that was a χ^2 with 1 degree of freedom.

Mathematical modelling and statistical inference

The detailed description of the mathematical model, the parameter estimation and model selection procedure, and supplementary results are available in the Supplementary Information (SI).

Model assumptions: we considered a stochastic model, at the scale of a family, continuous in time (denoted by t), over a short period ($T_{max} = 96$h), and where each cell was characterized by a phenotype p, with four different possibilities: $p \in \mathcal{P} =$ (HSC, MPP, HPC, CD34-), where $\mathcal{P}$ was naturally provided with an order relation (e.g., an MPP being more differentiated than an HSC). Therefore, we confounded $\mathcal{P}$ with the set [1,2,3,4]. p = 3 would then correspond to an HPC. At t = 0, the process begun with a single cell of a known phenotype p_1 **(Fig. 3a)**. The initial cell division occurred at a time T_1, randomly drawn from the empirical distribution of observed first division times (of cells of type p_1) (for further details, see **SI – B.3.1**). We treated

separately the first division from the subsequent ones since we observed that the initial division typically exhibits a longer duration **(Extended Data Figure 2j-l)**. After division, two daughter cells were considered. No cell death was considered in our model, and each of the daughter cells was assigned a cell phenotype at least as differentiated as the mother cell (we assume no de-differentiation). The probability for a daughter cell to be of type $p = i$, given its mother's type $p_m = a$, was denoted by $p_{a\to i} = \mathbb{P}[\, p = i \mid p_m = a]$. Then, the cells pursued the proliferation process, with division times randomly sampled from Log-Normal distributions ($\mathcal{LN}$) where the distribution parameters (μ_p, σ_p) varied according to the dividing cell's phenotype: $T \sim \mathcal{LN}(\mu_p, \sigma_p)$.

Possible scenarios: The generic model of HSPC proliferation and differentiation could be declined into several scenarios to explore different biological hypotheses **(Fig. 3b)**. These scenarios modify the interactions between sister cells, to reflect distinct biological hypotheses that we evaluated for their agreement with our experimental data. Scenarios were defined based on two criteria: homogeneity in fate (H) and concordance in division timing (C). Homogeneity (H) assesses whether sister cells tended to share the same fate (H=1) or not (H=0). Concordance (C) evaluated whether sister cells tended to divide simultaneously, with C = 1 for synchronised divisions and C = 0 for independent timing. Each scenario was identified by a pair of values for (H, C), creating a matrix of possibilities. The default scenario, characterized by H = 0 and C = 0, posited that sister cells were fully independent from each other (conditioning on their mother). This scenario served as a baseline, against which the implications of fate homogeneity and division concordance were assessed. When we hypothesized a homogeneity in fate (H = 1), we considered in our model a higher probability to have a symmetrical division compared to the scenario where H = 0. To do that, we introduced the parameter $\rho_H \in [0,1]$ such that the probability of having two daughter cells of the same type *b*, given their mother's type *a*, was equal to: $\rho_H\, p_{a\to b} + (1-\rho_H) p_{a\to b}^2$. In the limiting case where $\rho_H = 1$, all divisions were symmetrical. The case $\rho_H = 0$ corresponded to the hypothesis H = 0 of absence of homogeneity in fate. When we hypothesised a concordance in division (C = 1), we considered that the division times of two sister cells were not independent anymore but correlated. Let us denote by T_i and T_j the division times of two daughter cells of type p_i and p_j, respectively. Then, in our model, the division times of the two sister cells followed a bivariate log-normal distribution:

$$\left(T_i, T_j\right)^t \sim \mathcal{LN}\left((\mu_{p_i}, \mu_{p_j})^t, \Sigma\right),$$

where the covariance matrix was expressed by: $\Sigma = \begin{pmatrix} \sigma_{p_i}^2 & \rho_C\, \sigma_{p_i}\sigma_{p_j} \\ \rho_C\, \sigma_{p_j}\sigma_{p_i} & \sigma_{p_j}^2 \end{pmatrix}$.

With parameter $\rho_C \in [0,1]$ quantifying to which extent sister cells were concordant.

The limiting case $\rho_C = 0$ would correspond to the hypothesis C = 0, which was absence of concordance in division.

In addition to our four main scenarios, we also studied more complex ones, by considering that ρ_H would depend on the mother cell type (SI – B.2.3), or that there could be concordance

in division between cousin cells, or that the concordance between sister cells would depend on their types (SI – B.3.4).

We denote by:

$\theta = (p_{1\to1}, p_{1\to2}, p_{1\to3}, p_{2\to2}, p_{2\to3}, p_{3\to3}, \mu_1, \sigma_1, \mu_2, \sigma_2, \mu_3, \sigma_3, \mu_4, \sigma_4, \rho_H, \rho_C)^t$
our parameter vector.

Comparing the model simulation with the experimental observations: To evaluate which scenario is the most likely, we compared the output of our model simulations to the experimental observations, defining a distance between the data and simulations. We denoted by $\mathcal{D}$ our dataset (for further details about its definition, the inclusion and pre-treatment step, and the uncertainties in the observations, see **SI – A**). $\mathcal{D}$ is of size 1,740, meaning that we had the observations for 1740 different families, either from the MultiGen or the single-cell live imaging assay, for different initial conditions, and for different observation times. Because of the large size of our dataset, and the heterogeneity in the observations, defining a proper distance was challenging. Our methodology involved the derivation of an extensive set of summary statistics, denoted as S_i, where *i* indexes the specific summary statistic within the set. Each summary statistic can be computed from experimental observations, in that case we noted: $\hat{S}_i = S_i(\mathcal{D})$.

Given a parameter vector θ, we simulated the model to generate a virtual dataset of equivalent size to the number of experimental observations, from which we calculated each theoretical (i.e., given by our model) summary statistic $S_i(\theta)$. Summary statistics synthesised the information from the observations (or from the simulations) making comparisons possible. This was accompanied by a potential loss of information, which we first minimised by defining 340 summary statistics. Ideally, we aimed to find parameter values such that the simulations of the model lead to computed summary statistics as close as possible to the experimental ones, that is, find θ^* such that, $\forall i \in \{1, \dots, 340\}$, $S_i(\theta^*) \approx \hat{S}_i$.

However, given the number of summary statistics considered, and the potential redundancy and high correlation among them, we refined our approach using Principal Component Analysis (PCA), thus reducing the dimensionality of our observation space. We ended-up with 20 principal components, accounting for 80% of the variance, denoted as $\left(S_{p,j}\right)_{j=1}^{20}$ where each $S_{p,j}$ was a linear combination of the original 340 summary statistics.

The distance d(θ) was then defined as a mean squared error (MSE) between the simulated and observed values of these principal summary statistics:

$$d(\theta) = \sum_{j=1}^{20} \left(S_{p,j}(\theta) - \hat{S}_{p,j}\right)^2$$

This approach provided a practical balance between simulation accuracy and computational cost, making it well-suited for methods like Approximate Bayesian Computation (ABC).

The ABC method: to compare the experimental observations to the output of our model, through the construction of a relevant distance d(θ), we ran the ABC method to determine

the posterior probability of each scenario, given our experimental data. First, we defined our prior distributions. We considered *a priori* our four scenarios to be equiprobable. Then, we defined the prior distributions of our model parameters. Except for ρ_C and ρ_H, all the parameters did not depend, *a priori*, on the scenario. We considered them uniformly distributed (see **SI – B.1 and B.4** for further information on the possible range of values of the different parameters). For our parameter associated with the hypothesis of homogeneity in fate, we had: $\mathbb{P}[\,\rho_H = 0\,|H = 0] = 1$ and $\rho_H|\,H = 1\ \sim \mathcal{U}([0.3, 1])$ ($\mathcal{U}$ denotes a uniform distribution) such that we enforced a clear distinction between our scenarios of absence (H=0) or presence (H=1) of homogeneity in fate. We did similarly for ρ_C. In a Bayesian framework, the objective was to update our prior, through the integration of the data, to get the posterior distribution. Given the complexity of our model, the likelihood was intractable, hence the need to use an approximate Bayesian Computation method where the posterior distribution might be approximated. Instead of using the expression of a likelihood, we relied on the definition of our distance d(θ) between the model and the observations.

The ABC-rejection sampling method: For $i \in \{1, \cdots, N\}$, we sampled *a* modelling scenario (H_i, C_i) and *a* parameter vector θ_i from the prior distributions. We ran as many simulations as observations (1,740), computed the principal summary statistics $S_{p,i}(\theta_i)$ and then the distance to the observation $d_i = d(\theta_i)$. We kept only the samples $i' \in I$ such that $d_{i'} < \varepsilon$, where ε corresponded to the 0.1 percentile computed over the $(d_i)_{1 \le i \le N}$. That is, we kept the 0.1% best samples. These samples were used to approximate our posterior distribution **(Extended Data Fig. 10a)**. Finally, the posterior distribution of each of our four scenarios – when using the ABC-rejection algorithm - was approximated by:

$$\mathbb{P}[H = i, C = j\,|\mathcal{D}] \;\approx\; \frac{1}{0.001N}\sum_{i' \in I}\{H_{i'} = i\} \cap \{C_{i'} = j\}$$

That is, we counted how many times we selected each scenario among the 0.1% best samples. With N=1,733,106, the results of this procedure led us to select the scenario (H=1, C=1) as the most likely **(SI - D.1)**. Using the ABC-rejection sampling algorithm to explore how more complex modelling scenarios performed, and testing nearly 11 million of different parametrisations, we found no evidence that more complex scenarios would be more appropriate to describe our experimental observations **(SI - E)**.

The ABC-SMC method: The ABC rejection algorithm requires to sample parameter vectors a high number of times to converge. Then, to refine our results, and ensure their robustness and the convergence of the algorithm, we implemented the ABC-SMC sampler from Toni and colleagues[61] (details of the implementation in **SI - D.2**). With this method, we considered N = 2000 parameter vectors called particles, to approximate the posterior distribution. The ABC-SMC algorithm is sequential, so we iteratively sampled the N particles from intermediate distributions, starting from the prior distribution to the posterior distribution. At each iteration i, we sampled a parameter vector from the previous intermediate distribution (sampling the previously selected particles, according to some weights, and using pre-defined perturbation kernels) and ran the simulation to compute the distance to the observations. As for the ABC-rejection algorithm, particles were accepted if they are below a given threshold,

786 discarded otherwise. In the ABC-SMC method, there was a decreasing sequence of tolerances.
787 The ABC-SMC method ran for many iterations, until convergence, that was, until the
788 distributions of the distances did not evolve for new iterations **(Extended Data Fig. 10b)**. Over
789 the iterations, those distributions evolved until reaching a stable value, which corresponded
790 to the approximated posterior distributions of the modelling scenarios **(Extended Data Fig.**
791 **10c).** To ensure the robustness of the ABC-SMC algorithm that we used, also regarding the
792 choice of the perturbation kernels used, we ran different configurations, leading to the same
793 results **(Extended Data Fig. 10e)**. ABC-SMC approach refined the results from the ABC-
794 rejection sampling approach, from iteration 9, allowing to better explore the parameter
795 space, progressively focusing on the region where the parameter vectors were more likely to
796 have generated the experimental data.

797 We focused on the discrimination between modelling scenarios, rather than the estimation
798 of the parameters of the most likely model. Yet, we also got the intermediate posterior
799 distributions of the model parameters and saw how each iteration allowed to be more
800 informative on the different parameter values, and how the parameters for which the model
801 converged faster than other ones **(Extended Data Fig. 10f** and **SI - D.4**). Some parameters
802 stayed undetermined (or weakly determined) at this stage, meaning that their posteriors did
803 not move away from the prior. When comparing the summary statistics computed from the
804 models – sampling from the posterior – to the experimental values, we observed for most of
805 them already a good agreement **(Extended Data Fig. 10d** and **SI - D.5**)

806

807 **CITE-seq profiling of short-term liquid culture**

808 *Antibody labelling:* magnetic sorted $CD34^+$ cells from umbilical cord blood were plated at a
809 cell density of $1x10^5$ cells/mL, using the same culture condition described for the condition
810 Diff. Cells were harvested after 96 hours, counted with a haemocytometer, and resuspended
811 in cold PBS 1X (ThermoFisher Scientific) containing 2% Bovine Serum Albumin (BSA,
812 Euromedex) and 0.01% Tween 20 (Sigma), to a final concentration of 10^6 cells/mL. 100 µL of
813 this cell suspension was incubated with a mix of previously prepared CITE-seq antibodies.
814 Cells were incubated for 30 minutes at 4 °C, washed three times with the same solution used
815 for the incubation and resuspended in 16.5 µL, at a final concentration of 1000 cells/mL.

816 *Library sequencing:* approximately $1.65x10^4$ cells were loaded, to obtain a final yield of $1x10^4$
817 recovered cells. Cells were processed accordingly to the Chromium Single Cell 3' library v3.1
818 (10x Genomics). Libraries were sequenced using a Novaseq s1 (Illumina), allocating at least
819 20,000 reads per cell for the mRNA library, and 5,000 reads for the Antibody Derived Tags
820 (ADT) library **(Extended Data Table 7)**.

821 *Computation methods:* Raw sequencing reads were processed using Cellranger v3.1.0. To
822 obtain a reads/cell/gene count table, reads were mapped to the human GRCh38 reference
823 genome. Cells with less than 1000 genes per cell and with a high percentage (> 30%) of
824 mitochondrial genes were removed from downstream analyses. Following our filtering
825 procedures, the average UMI count per cell was 4406. The median number of genes detected
826 per cell was 1662, 2.2% mapped to mitochondrial genes. Data normalization and integration

827 were performed using the default Seurat v4 approach FindIntegrationAnchors() followed by
828 IntegrateData(), and differentially expressed genes were determined using a logistic
829 regression in Seurat on the non-integrated data using the FindConservedMarkers() function.
830 Pathway based analyses were performed using the enrichR package. Unsupervised clustering
831 was performed on the significant variable genes using the 10 first PCA followed by the
832 nonlinear dimensionality reduction technique UMAP. Annotation of the data was obtained by
833 mapping published signatures using the AddModuleScore() method of Seurat.

834

835 **Single-cell re-distribution**

836 Single cell distribution of individual cell families was performed using the cellenONE platform
837 (Cellenion). Briefly, we selected individual wells of 96-wells plates displaying at least four cells,
838 harvested the entire well, centrifugated for 8' at 300g and resuspended in 6 µL of PBS 1X
839 (Sigma). Then, the whole volume was loaded into the machine, but due to the dead volume,
840 effectively only 4 µL of this cell suspension could be redistributed. Individual cells were each
841 sorted directly into single wells of 96-wells plates (90 cells per plate) at 4 °C, each well
842 containing 3 µL of lysis buffer. The lysis buffer contains 0.03 µL Triton X100 (ThermoFisher
843 Scientific), 0.0375 µL Recombinant RNAse Inhibitor (Takara Bio), 0.2 µL dNTP 10mM
844 (ThermoFisher Scientific), 0.5 µL PEG 40% (Sigma), 0.04 µL SMARTseq3 oligo-dT 100 µM
845 (/5BiosG/ACGAGCATCAGCAGCATACGATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN; IDT) and
846 ultrapure RNAse water (Invitrogen).

847

848 **SMART-Seq3 library preparation and sequencing**

849 Reverse transcription (RT) and the subsequent steps were performed accordingly to the
850 SMART-seq3 technique[40]. Briefly, lysates in each well underwent RT (42 °C for 90 mins, 10
851 cycles at 50 °C 2 mins + 42 °C 2 mins, and finally 85 °C for 5 mins) on ice, in presence of
852 template switch oligonucleotide at 100 µM
853 (/5BiosG/AGAGACAGATTGCGCAATGNNNNNNNNrGrGrGIDT), 0.04 µL Maxima H minus
854 reverse transcriptase (ThermoFisher Scientific), TRIS-HCl 1M PH 8.3 (Invitrogen), 0.12µL NaCl
855 1M (Invitrogen), 0.04 µL GTP 100mM (ThermoFisher Scientific), 0,32 µL DTT 100mM
856 (ThermoFisher Scientific), 0.1 µL MgCl2 50mM (ThermoFisher Scientific) and 0.05 µL
857 Recombinant RNAse Inhibitor (Takara Bio) and water. The following PCR step was performed
858 immediately after, using the Kapa Hifi HotStart Ready Mix (Roche), and forward (100 µM
859 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTGCGCAATG) and reverse (25 µM
860 ACGAGCATCAGCAGCATACGA) primers (IDT technologies) for 25 cycles. After DNA purification
861 using AMPure XP microbeads (ratio 0.8:1) (Beckman Coulter, A63881), the cDNA was
862 tagmented in a final volume of 2 µL, using the Nextera XT DNA Library preparation kit
863 (Illumina). The following library amplification was performed adding 1.5 µL of UDI primers
864 and 2,5 µL of PCR mix including the Phusion High Fidelity DNA polymerase (ThermoFisher
865 Scientific). After tagmentation, the different samples were pooled, purified using AMPure XP

microbeads (ratio 0.9:1) and sequenced at 100-bp paired end on a high-output flow cell, using a NovaSeq SP, 800M reads (Illumina).

Single-cell 3' mRNA sequencing pre-processing

FASTQ files for each single cell were generated using bcl2fastq (version 2.20.0), with default parameters. The Read 1 contain the TSO (constant tag sequence and a UMI) followed by cDNA or only cDNA. The Read 2 contain cDNA only. Index1 and index2 reads correspond to cell-specific barcode. The scRNA-SmartSeq3 pipeline v1.2.1 (https://doi.org/10.5281/zenodo.15056215) was used to generate UMI count tables from FASTQ files. Briefly, reads were aligned to the human genome (hg38) using STAR (version 2.7.8a) and assigned to genes using FeatureCounts (deeptools version 3.5.0). UMIs were extracted, deduplicated and counted for each gene using umiTools (version 1.1.1). Low-quality cells were filtered out based on a minimum of 500 genes detected and a percentage of mitochondrial RNA lower than 20. Families composed of only one cell were also removed. Counts were normalised using SCTransform function with default parameters from R Seurat package (version 4.3.0).

Dimensionality reduction and clustering

PCA was performed using the RunPCA function from the Seurat R package. Using the elbowPlot from the same package, the first 10 PCs were selected, as able to cover most of the variation in the data. To perform the clustering analysis, a K-nearest neighbour (KNN) graph was constructed based on the Euclidean distance in the PCA space, to find neighbour cells having similar feature expression patterns. Finally, Uniformed Manifold Approximation and Projection (UMAP) for dimension reduction was performed, using the RunUMAP function with the top ten PCA dimensions and default parameters.

Cosine distance

The variable "distance" was a measure of how cells of the same family tended to be transcriptomically similar, regardless of their cell fate. The cosine distance between two cells was defined as follows[41]:

$$cosine\ distance\ (A, B) = 1 - \frac{A \cdot B}{\|A\| \|\ B\|}$$

where A represented the gene expression vector of the first cell and B represented the gene expression vector of the second cell. For each family, the cosine distance was computed for each pair of cells. The cosine distance distributions were computed on the log-normalized UMIs for real families and both types of random families.

903 **Identification of Variably Expressed Genes Across Families**

904 Single-cell RNA sequencing (scRNA-seq) data from Smart-seq3 was analysed to investigate
905 gene expression variability across families. The pre-processed and normalized data were used
906 for the analysis.

907 Cells were grouped into families based on well information. Families with fewer than two cells
908 were excluded. An analysis of variance (ANOVA) was performed for each gene to assess
909 differential expression across families:

910 $$x_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

911 where x_{ij} represents the expression of gene ii in family j, μ is the overall mean expression, α_j
912 represents family-specific effects, and ϵ_{ij} denotes residual error.

913 The total variance was decomposed into between-family variance (σ_B^2) and within-family
914 variance (σ_W^2), and an F-test was applied to determine whether the between-group variance
915 was significantly greater than the within-group variance:

916 $$F = \frac{\sigma_B^2}{\sigma_W^2}$$

917 where a high F-value indicates greater variance between families than within families. To
918 control for multiple hypothesis testing, p-values were adjusted using the Benjamini-Hochberg
919 procedure to limit the false discovery rate (FDR). Genes with adjusted p-values (FDR < 0.05)
920 were considered significantly homogeneous within families. To visualize gene expression
921 patterns across families and clusters, a heatmap was generated using pheatmap R package.
922 The expression matrix was prepared by including the significantly family conserved genes. The
923 samples were ordered by family information, while genes were clustered separately for cells
924 and genes using hierarchical clustering with Euclidean distance and the complete linkage
925 method. The genes were clustered into modules by cutting the hierarchical cluster tree.

926

927 **MIIC analysis**

928 MIIC was used to investigate which genes were significantly associated with specific families,
929 their recorded division times, colony size, predicted cell types, and transcriptomic dispersion
930 "distance", using transcriptomics data of 142 cells obtained via SMART-Seq3[40]. In the
931 reconstructed graphical networks, the variables were represented as nodes and the edges
932 encoded for dependency. An edge with an arrow stands for causality, an edge without an
933 arrow stands for association, a bi-directed dotted edge represented an unobserved common
934 cause between two connected variables[44]. The green arrows represented genuine causality
935 while the non-green arrows represented putative causality. In the case of putative causality,
936 it was not possible to exclude the presence of an unobserved common cause. The link-colours
937 between two nodes were assigned based on the Spearman's correlation coefficient: red stand
938 for positive correlation and blue for negative correlation. Gray was used for categorical
939 variables without natural order, such as the family IDs. The networks were reconstructed from

940 two scRNA-seq datasets, corresponding to two different experiments: 70 cells belonging to
941 14 families and 28,750 genes for the first, 72 cells belonging to 16 families and 30,682
942 sequenced genes for the second. The metadata for each dataset was the following: family of
943 origin, times of division, colony size at 90 hours. The network was reconstructed from the 142
944 cells with their respective raw UMI counts of selected genes and additional features of
945 interest. The gene selection was performed based on their high mutual information with the
946 cellular features of interest: "family", "T1", "avg dT2", "avg dT3", "predicted.celltype",
947 "Colony size", "exp", "distance". The genes selected to reconstruct the networks shared the
948 highest mutual information with these properties of interest (i.e. we retained all genes above
949 some tuneable mutual information threshold). Mitochondrial, Ribosomal Protein Large
950 subunit (RPL), Ribosomal Protein Small subunit (RPS), Large Subunit Mitochondrial Ribosomal
951 Proteins (MRPL) and Small Subunit Mitochondrial Ribosomal Proteins (MRPS) genes were
952 filtered out from the analysis. A confidence parameter of 1% was used. This parameter sets
953 the probability upper bound that this edge could be spurious due to sampling noise. Hence,
954 the confidence parameter allowed to filter the least robust edges[43].

955 Two networks were obtained either including or excluding the "Family" variable:

956 • A network with the "family" variable and a total of 1035 variables (including 151 regulatory
957 genes or variables and 884 consequence genes):

958 https://miic.curie.fr/job_results.php?id=t4n5bx95nzm6jExIDvdG

959 • A network without the "family" variable and a total of 759 variables (including 117
960 regulatory genes or variables and 642 consequence genes):

961 https://miic.curie.fr/job_results.php?id=PVPRcJsspyU75PGQuMtB

962

963 **Statistical analysis**

964 Statistical analyses are detailed in figure legends and performed using R (version 4.3) or
965 Python (version 3.10). The number of independent replicates, experiments and donors is
966 stated for each figure in the legend.

967

968 **Data availability**

969 Raw and pre-processed single-cell RNA-sequencing dataset are deposited here:

970 CITE-Seq: https://zenodo.org/records/15720816

971 SMART-Seq3: https://zenodo.org/records/15721710 -
972 https://zenodo.org/records/15722093 - https://zenodo.org/records/15722300 -
973 https://zenodo.org/records/15722418 - https://zenodo.org/records/15722505
974

975 Additional data for reanalysis and reproducibility effort is available upon request to the lead
976 author.

Code availability

All the scripts used in this paper are freely accessible on GitHub:

Live-cell imaging analysis: https://github.com/TeamPerie/Live-cell_imaging.git

MultiGen analysis: https://github.com/TeamPerie/MultiGen.git

Mathematical model of division and differentiation: https://gitlab-research.centralesupelec.fr/2012hermangeg/HSPC_proliferation

CITE-seq analysis: https://github.com/TeamPerie/CITEseq_analysis_pipeline.git

Smartseq3 pre-processing:

https://github.com/bioinfo-pf-curie/scRNA-SmartSeq3/releases/tag/v1.2.1

Smartseq3 quality control: https://github.com/TeamPerie/smartseq3_familly_analysis

Single-cell 3' mRNA sequencing pre-processing and analysis:

https://github.com/tizianatocci/Donada2025

https://github.com/TeamPerie/Donada-et-al2025-family-ANOVA.git

Acknowledgments

We would like to thank all the members of the Perié lab for helpful discussion, particularly Anais Lamoureux, Jason Cosgrove, Lucie Hustin. We would like to thank the members of the Institut Curie Flow Facility for their help with setting up the flow cytometry experiments. We thank the members of the Institut Curie Imaging Platform for their assistance with the live-cell imaging experiments. We thank the Saint Louis hospital cord blood biobank for providing the biological resources necessary for this project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 847718. The study was supported by grants from the Labex CelTisPhyBio (ANR-10-LBX-0038) (to L.P. and A.D.), the Canceropole INCA Emergence (2021-1- EMERG-54b-ICR-1, to L.P.), and the ITMO MIIC grant (21CM044, to L.P. and P.H.C.). As well as funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program ERC StG 758170-Microbar (to L.P.), A.D. was supported by a fellowship from Fondation de France (00119150 / WB-2021-36197).

Authors Contributions

A.D. and A.M. performed, analysed and interpreted the live-cell imaging data. A.D., I.M. and C.P. performed, analysed and interpreted the MultiGen assay. G.P. developed the statistical framework for the MultiGen analysis, supervised directly by K.D. and with feedback from A.D. and L.P. G.H. developed the mathematical framework, under the supervision of P.H.C. and with feedback from A.D. and L.P. A.D., S.T.B. and D.D. performed the transcriptomics

experiments. L.H.A., W.S. and T.T. analysed and interpreted the transcriptomics data, supervised directly by H.I. and L.P. with feedback from N.S. and A.D. T.T. performed and interpreted the network analysis, supervised directly by H.I. C.W., J.V., D.M., C.P. and C.L.C. provided the biological and clinical samples. A.D. prepared the paper final figures, with contributions from T.T., A.M. and W.S. A.D. and L.P. wrote the manuscript, with substantial contribution from L.L. and all the co-authors.

Declaration of interests

D.M. received research funding from Novartis, Sanofi and CSL Behring, and consulting fee from Incyte, Novartis, Sanofi, CSL Behring, Jazz pharmaceuticals and Mallinckrodt. None of this COI are directly linked to this research.

The other authors have nothing to declare.

References

1. Till, J. E., McCulloch, E. a. & Siminovitch, L. A Stochastic Model of Stem Cell Proliferation, Based on the Growth of Spleen Colony-Forming Cells*. *Proc. Natl. Acad. Sci. U. S. A.* **51**, 29–36 (1964).
2. Laplane, L. & Maley, C. C. The evolutionary theory of cancer: challenges and potential solutions. *Nat. Rev. Cancer* **24**, 718–733 (2024).
3. Fennell, K. A. *et al.* Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature* **601**, 125–131 (2022).
4. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (80-.).* **367**, eaaw3381 (2020).
5. Marsolier, J. *et al.* H3K27me3 conditions chemotolerance in triple-negative breast cancer. *Nat. Genet.* **54**, 459–468 (2022).
6. Reyes Hueros, R. A., Gier, R. A. & Shaffer, S. M. Non-genetic differences underlie variability in proliferation among esophageal epithelial clones. *PLOS Comput. Biol.* **20**, e1012360 (2024).
7. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
8. Sigal, A. *et al.* Variability and memory of protein levels in human cells. *Nature* **444**, 643–646 (2006).
9. Shaffer, S. M. *et al.* Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors. *Cell* **182**, 947-959.e17 (2020).

1048 10. Sieburg, H. B. *et al.* The hematopoietic stem compartment consists of a limited
1049 number of discrete stem cell subsets. *Blood* **107**, 2311–2316 (2006).

1050 11. Cosgrove, J., Hustin, L. S. P., de Boer, R. J. & Perié, L. Hematopoiesis in numbers.
1051 *Trends in Immunology* vol. 42 1100–1112 (2021).

1052 12. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation
1053 landscapes. *Nature* **553**, 418–426 (2018).

1054 13. Dykstra, B. *et al.* Long-Term Propagation of Distinct Hematopoietic Differentiation
1055 Programs In Vivo. *Cell Stem Cell* **1**, 218–229 (2007).

1056 14. Tak, T. *et al.* HSPCs display within-family homogeneity in differentiation and
1057 proliferation despite population heterogeneity. *Elife* **10**, (2021).

1058 15. Kueh, H. Y. *et al.* Positive feedback between PU.1 and the cell cycle controls myeloid
1059 differentiation. *Science* **341**, 670–3 (2013).

1060 16. Treichel, S. & Filippi, M. D. Linking cell cycle to hematopoietic stem cell fate decisions.
1061 *Frontiers in Cell and Developmental Biology* vol. 11 (2023).

1062 17. Donada, A. *et al.* Simultaneous Assessment of Kinship, Division Number, and
1063 Phenotype via Flow Cytometry for Hematopoietic Stem and Progenitor Cells. *J. Vis.*
1064 *Exp.* 1–26 (2023) doi:10.3791/64918.

1065 18. Horton, M. B. *et al.* Multiplexed Division Tracking Dyes for Proliferation-Based Clonal
1066 Lineage Tracing. *J. Immunol.* **201**, 1097–1103 (2018).

1067 19. Marchingo, J. M. *et al.* T-cell stimuli independently sum to regulate an inherited
1068 clonal division fate. *Nat. Commun.* **7**, 13540 (2016).

1069 20. Spencer Chapman, M. *et al.* Clonal dynamics after allogeneic haematopoietic cell
1070 transplantation. *Nature* (2024) doi:10.1038/s41586-024-08128-y.

1071 21. Oshima, M. U. *et al.* Characterization of clonal dynamics using duplex sequencing in
1072 donor-recipient pairs decades after hematopoietic cell transplantation. *Sci. Transl.*
1073 *Med.* **16**, eado5108 (2024).

1074 22. Aldoss, I., Clark, M., Marcucci, G. & Forman, S. J. Donor derived leukemia in allogeneic
1075 transplantation. *Leuk. Lymphoma* **62**, 2823–2830 (2021).

1076 23. Calabria, A. *et al.* Long-term lineage commitment in haematopoietic stem cell gene
1077 therapy. *Nature* **636**, 162–171 (2024).

1078 24. Oliveira, E. *et al.* Epigenetic heritability of cell plasticity drives cancer drug resistance
1079 through one-to-many genotype to phenotype mapping. *bioRxiv* 2023.11.15.567140
1080 (2023) doi:10.1101/2023.11.15.567140.

1081 25. Bell, C. C. *et al.* Targeting enhancer switching overcomes non-genetic drug resistance
1082 in acute myeloid leukaemia. *Nat. Commun.* **10**, 1–15 (2019).

1083 26. Laurenti, E. *et al.* CDK6 levels regulate quiescence exit in human hematopoietic stem
1084 cells. *Cell Stem Cell* **16**, 302–313 (2015).

1085 27. Aiuti, A. *et al.* Lentiviral Hematopoietic Stem Cell Gene Therapy in Patients with

1086 Wiskott-Aldrich Syndrome. *Science (80-.).* **341**, 1233151 (2013).

1087 28. Knapp, D. J. H. F. *et al.* Dissociation of Survival, Proliferation, and State Control in
1088 Human Hematopoietic Stem Cells. *Stem Cell Reports* **8**, 152–162 (2017).

1089 29. Passegué, E., Wagers, A. J., Giuriato, S., Anderson, W. C. & Weissman, I. L. Global
1090 analysis of proliferation and cell cycle gene expression in the regulation of
1091 hematopoietic stem and progenitor cell fates. *J. Exp. Med.* **202**, 1599–1611 (2005).

1092 30. Johnson, C., Belluschi, S. & Laurenti, E. Beyond "to divide or not to divide": Kinetics
1093 matters in hematopoietic stem cells. *Experimental Hematology* vol. 92 1-10.e2 (2020).

1094 31. Hammond, C. A., Wu, S. W., Wang, F., MacAldaz, M. E. & Eaves, C. J. Aging alters the
1095 cell cycle control and mitogenic signaling responses of human hematopoietic stem
1096 cells. *Blood* **141**, 1990–2002 (2023).

1097 32. Majeti, R., Park, C. Y. & Weissman, I. L. Identification of a Hierarchy of Multipotent
1098 Hematopoietic Progenitors in Human Cord Blood. *Cell Stem Cell* **1**, 635–645 (2007).

1099 33. Grelaud, A., Roberty, C. P., Marinz, J. M., Rodolphe, F. & Talys, J. F. ABC likelihood-
1100 free methods for model choice in Gibbs random fields. *Bayesian Anal.* **4**, 317–335
1101 (2009).

1102 34. Toni, T. & Stumpf, M. P. H. Simulation-based model selection for dynamical systems
1103 in systems and population biology. *Bioinformatics* **26**, 104–110 (2009).

1104 35. Sunnåker, M. *et al.* Approximate Bayesian Computation. *PLOS Comput. Biol.* **9**,
1105 e1002803 (2013).

1106 36. Dowling, M. R., Milutinović, D. & Hodgkin, P. D. Modelling cell lifespan and
1107 proliferation: is likelihood to die or to divide independent of age? *J. R. Soc. Interface*
1108 **2**, 517–526 (2005).

1109 37. Hawkins, E. D., Turner, M. L., Dowling, M. R., Van Gend, C. & Hodgkin, P. D. A model
1110 of immune regulation as a consequence of randomized lymphocyte division and
1111 death times. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 5032–5037 (2007).

1112 38. Upadhaya, S. *et al.* Kinetics of adult hematopoietic stem cell differentiation in vivo. *J.*
1113 *Exp. Med.* **215**, 2815–2832 (2018).

1114 39. Chen, T. N., Gupta, A., Zalavadia, M. D. & Streets, A. µCB-seq: microfluidic cell
1115 barcoding and sequencing for high-resolution imaging and sequencing of single cells.
1116 *Lab Chip* **20**, 3899–3913 (2020).

1117 40. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution
1118 using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).

1119 41. Smets, T. *et al.* Evaluation of Distance Metrics and Spatial Autocorrelation in Uniform
1120 Manifold Approximation and Projection Applied to Mass Spectrometry Imaging Data.
1121 *Anal. Chem.* **91**, 5706–5714 (2019).

1122 42. Mold, J. E. *et al.* Clonally heritable gene expression imparts a layer of diversity within
1123 cell types. *Cell Syst.* **15**, 149-165.e10 (2024).

1124 43. Verny, L., Sella, N., Affeldt, S., Singh, P. P. & Isambert, H. Learning causal networks
1125 with latent variables from multivariate information in genomic data. *PLoS Comput.*
1126 *Biol.* **13**, 1–25 (2017).

1127 44. Ribeiro-Dantas, M. da C. *et al.* Learning interpretable causal networks from very large
1128 datasets, application to 400,000 medical records of breast cancer patients. *iScience*
1129 **27**, (2024).

1130 45. Cabeli, V. *et al.* Learning clinical networks from medical records based on information
1131 estimates in mixed-type data. *PLOS Comput. Biol.* **16**, e1007866 (2020).

1132 46. Carey, A. *et al.* Identification of Interleukin-1 by Functional Screening as a Key
1133 Mediator of Cellular Expansion and Disease Progression in Acute Myeloid Leukemia.
1134 *Cell Rep.* **18**, 3204–3218 (2017).

1135 47. Filippakopoulos, P. *et al.* Selective inhibition of BET bromodomains. *Nature* **468**,
1136 1067–1073 (2010).

1137 48. Lechman, E. R. *et al.* MiR-126 Regulates Distinct Self-Renewal Outcomes in Normal
1138 and Malignant Hematopoietic Stem Cells. *Cancer Cell* **29**, 214–228 (2016).

1139 49. Schoof, E. M. *et al.* Quantitative single-cell proteomics as a tool to characterize
1140 cellular hierarchies. *Nat. Commun.* **12**, 1–15 (2021).

1141 50. Horton, M. B. *et al.* Lineage tracing reveals B cell antibody class switching is
1142 stochastic, cell-autonomous, and tuneable. *Immunity* **55**, 1843-1855.e6 (2022).

1143 51. Müller-Sieburg, C. E., Cho, R. H., Thoman, M., Adkins, B. & Sieburg, H. B. Deterministic
1144 regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* **100**,
1145 1302–1309 (2002).

1146 52. Sieburg, H. B., Rezner, B. D. & Muller-Sieburg, C. E. Predicting clonal self-renewal and
1147 extinction of hematopoietic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4370–4375
1148 (2011).

1149 53. Valdés-Mas, R. *et al.* Transplacental transfer of essential thrombocythemia in
1150 monozygotic twins. *Blood* **128**, 1894–1896 (2016).

1151 54. Sousos, N. *et al.* In utero origin of myelofibrosis presenting in adult monozygotic
1152 twins. *Nat. Med.* **28**, 1207–1211 (2022).

1153 55. Singh, I., Fernandez-Perez, D., Sanchez, P. S. & Rodriguez-Fraticelli, A. E. Pre-existing
1154 stem cell heterogeneity dictates clonal responses to the acquisition of leukemic driver
1155 mutations. *Cell Stem Cell* **32**, 564-580.e6 (2025).

1156 56. Schönberger, K. *et al.* Multilayer omics analysis reveals a non-classical retinoic acid
1157 signaling axis that regulates hematopoietic stem cell identity. *Cell Stem Cell* **29**, 131-
1158 148.e10 (2022).

1159 57. Bonora, M. *et al.* A mitochondrial NADPH-cholesterol axis regulates extracellular
1160 vesicle biogenesis to support hematopoietic stem cell fate. *Cell Stem Cell* **31**, 359-
1161 377.e10 (2024).

1162 58. Duffy, K. R. *et al.* Activation-induced B cell fates are selected by intracellular

1163 stochastic competition. *Science (80-.).* **335**, 338–341 (2012).

1164 59. Eisele, A. S., Tarbier, M., Dormann, A. A., Pelechano, V. & Suter, D. M. Gene-
1165 expression memory-based prediction of cell lineages from scRNA-seq datasets. *Nat.*
1166 *Commun.* **15**, 1–16 (2024).

1167 60. Davison, A. C. & Hinkley, D. V. Bootstrap Methods and their Application. *Bootstrap*
1168 *Methods their Appl.* (1997) doi:10.1017/CBO9780511802843.

1169 61. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. Approximate Bayesian
1170 computation scheme for parameter inference and model selection in dynamical
1171 systems. *J. R. Soc. Interface* **6**, 187–202 (2009).

1172