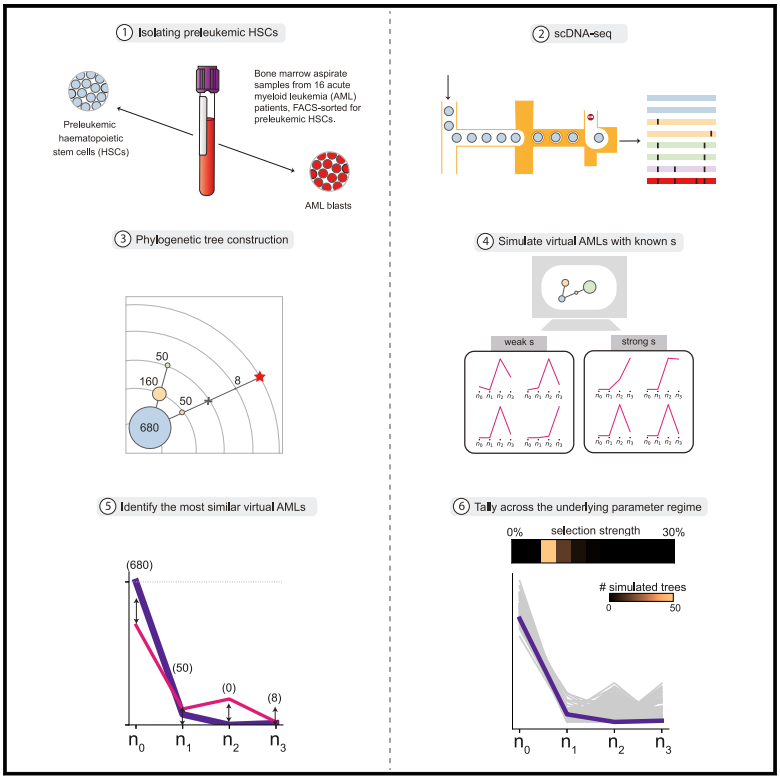
Cell Genomics

Article

Single-cell DNA sequencing reveals pervasive positive selection throughout preleukemic evolution

Graphical abstract



Highlights

- Performed scDNA-seq of pHSCs from bone marrow of 16 patients with AML
- Constructed phylogenetic trees based on major drivers for each patient
- Developed computational framework to infer levels of positive selection
- Revealed pervasive selection (9%–24% per year) during preleukemic evolution

Authors

Gladys Poon, Aditi Vedi, Mathijs Sanders, Elisa Laurenti, Peter Valk, Jamie R. Blundell

Correspondence

ypgp2@hku.hk (G.P.),
jrb75@cam.ac.uk (J.R.B.)

In brief

The representation of driver mutations in preleukemic HSCs provides a window into the somatic evolution that precedes AML. Here, Poon et al. isolate pHSCs from the bone marrow of 16 patients with AML and perform scDNA sequencing on thousands of cells to reconstruct phylogenetic trees of the major driver clones in each patient. Poon et al. develop a computational framework and use it to find that the highly variable structures of preleukemic trees emerge naturally from somatic evolution with pervasive positive selection.

Poon et al., 2025, Cell Genomics 5, 100744
February 12, 2025
https://doi.org/10.1016/j.xgen.2024.100744

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Single-cell DNA sequencing reveals pervasive positive selection throughout preleukemic evolution

Gladys Poon,[1,*] Aditi Vedi,[2,3] Mathijs Sanders,[4] Elisa Laurenti,[2] Peter Valk,[4] and Jamie R. Blundell[1,5,*]
[1]Early Cancer Institute, University of Cambridge, Cambridge, UK
[2]Wellcome - MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK
[3]Cambridge University Hospital NHS Foundation Trust, Cambridge, UK
[4]Department of Hematology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands
[5]Lead contact
*Correspondence: ypgp2@hku.hk (G.P.), jrb75@cam.ac.uk (J.R.B.)
https://doi.org/10.1016/j.xgen.2024.100744

SUMMARY

The representation of driver mutations in preleukemic hematopoietic stem cells (pHSCs) provides a window into the somatic evolution that precedes acute myeloid leukemia (AML). Here, we isolate pHSCs from the bone marrow of 16 patients diagnosed with AML and perform single-cell DNA sequencing on thousands of cells to reconstruct phylogenetic trees of the major driver clones in each patient. We develop a computational framework that can infer levels of positive selection operating during preleukemic evolution from the statistical properties of these phylogenetic trees. Combining these data with 67 previously published phylogenetic trees, we find that the highly variable structures of preleukemic trees emerge naturally from a simple model of somatic evolution with pervasive positive selection typically in the range of 9%–24% per year. At these levels of positive selection, we show that the identification of early multiple-mutant clones could be used to identify individuals at risk of future AML.

INTRODUCTION

Large-scale cancer genome sequencing efforts over the last 20 years have revealed that most cancers harbor multiple clonal driver mutations at diagnosis.[1–3] These mutations are sometimes acquired decades before the emergence of the first malignant cell.[4,5] How the full complement of driver mutations is acquired in a single cell over a human lifespan, however, is not fully understood. Early multi-hit theories of cancer considering mutation acquisition as the key factor seemingly capture the age-incidence relationships across a range of cancers.[6] However, these theories struggle to provide plausible estimates for overall cancer incidence[7] using estimates of somatic mutation rates[8] and stem cell numbers.[9,10]

Work over the last decade using deep bulk sequencing has shown that there is positive selection acting on mutations in cancer-associated genes across a range of healthy tissues.[10,11–21] This suggests that positive selection may act throughout the entire multi-hit trajectory of cancer. However, because of the inherent challenges in measuring precancerous evolution, the predictions of this conceptual idea have not been quantitatively tested.

Single-cell sequencing from multiple cells in a population can provide a quantitative picture of the evolutionary history of the population by using somatic mutations as unique lineage markers for building phylogenetic trees.[5,22–24] The hematopoietic system provides an ideal model for using single-cell approaches as hematopoietic stem cells (HSCs) can be easily isolated and sequenced. These approaches have revealed that large numbers of HSCs maintain the blood in healthy individuals,[9,24] that the driver events of certain blood cancers can be acquired *in utero*,[5] and that there is extensive clonal heterogeneity in acute myeloid leukemia (AML).[25–28] An innovative approach to studying precancerous evolution is through the isolation of ostensibly healthy preleukemic HSCs (pHSCs) from patients with AML at diagnosis. By coupling this idea with single-cell sequencing, the stepwise nature of preleukemic evolution and how preleukemic clones can seed relapse post-therapy has been demonstrated previously.[29] However, the quantitative information encoded in the preleukemic phylogenetic trees have not been fully exploited to understand how levels of positive selection operate during preleukemic evolution. Here, we hypothesize that the statistical properties of phylogenetic trees across many patients with AML could reveal levels of positive selection operating during preleukemic evolution and how those levels vary among individuals.

RESULTS

Single-cell phylogenies from pHSCs

To construct phylogenetic trees from pHSCs, we obtained bone marrow aspirate samples from 16 patients with AML collected at

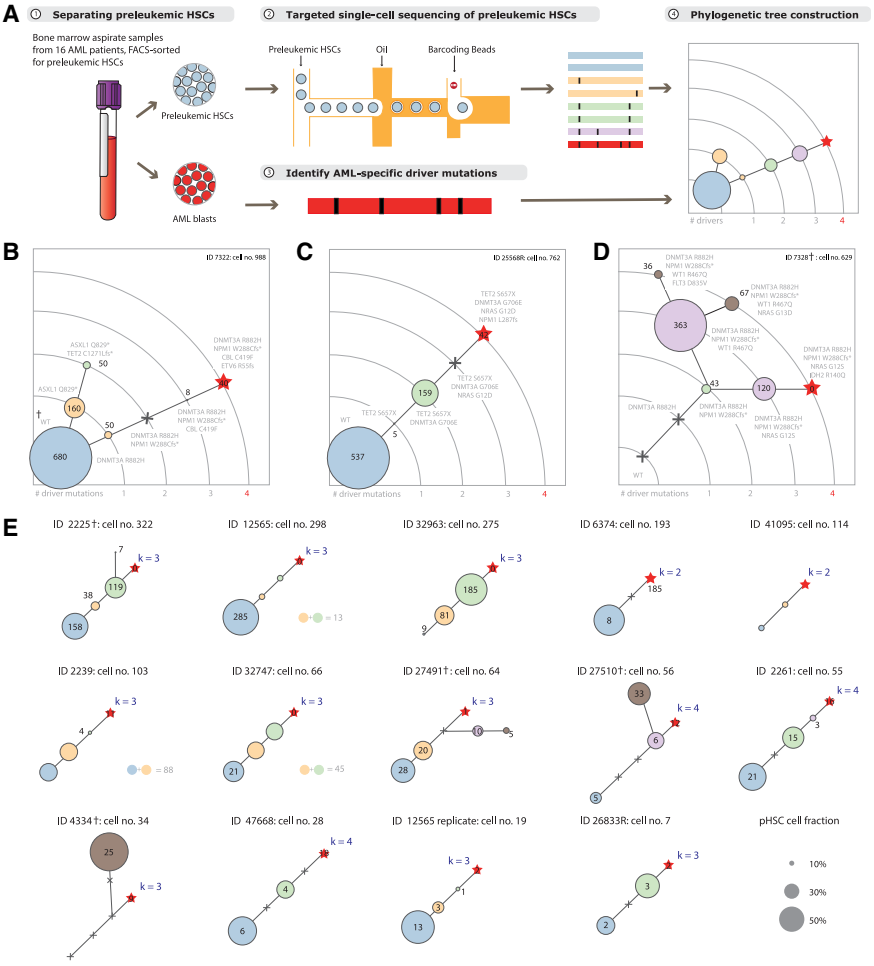


Figure 1. Construction of phylogenetic trees from preleukemic HSCs from patients with AML at diagnosis

(A) Preleukemic hematopoietic stem cells (HSCs) were sorted from bone marrow aspirate samples acquired from patients with AML at diagnosis. The sorted samples were then passed onto the single-cell DNA sequencing platform for genotype information on a targeted panel of 45 commonly mutated myeloid genes. Variant information at the single-cell level combined with bulk sequencing information of the same samples before sorting enables the assignment of individual cells to their respective clonal genotype. Phylogenetic trees were constructed for each patient based on the clonal architecture found in the preleukemic HSC population (see STAR Methods).

(B–D) Three 4-hit phylogenetic trees from three patients are shown. Clones are denoted by circles and ordered on ascending arcs based on the number of drivers detected. The number of cells assigned to each clone is labeled on the circle and replaced by a gray cross if zero. Clonal genotypes not fully ascertained are not assigned cell numbers—the summed total across indistinguishable clonal genotypes is shown in the bottom right corner instead. The genotype of a clone is also labeled and represented by a red star if it belongs to the clonal genotype with all high-confidence drivers.

(E) Across the 17 trees reconstructed for the 16 patients, 6 trees were identified as 4-hit events (including B–D), 9 trees were identified as 3-hit events (including the pair of technical replicates from patient ID12565), and 2 trees were identified as 2-hit events. Samples that contain multiple *k*-th mutants are marked with a †.

diagnosis from the multicenter HOVON-SAKK clinical trials (see sample selection in STAR Methods). To minimize inter-patient variation, we selected samples with a genetically restricted AML genotype (*DNMT3A*^{mut}/*NPM1*^{mut}). pHSCs were isolated by fluorescence-activated cell sorting (FACS) (Figure 1A; STAR Methods). We then performed single-cell DNA sequencing of these pHSCs targeting 45 myeloid genes using the Tapestri platform (Mission Bio). By combining these data with a high-confidence set of driver mutations identified by bulk sequencing of each bone marrow aspirate sample, we were able to genotype a total of 4,013 HSCs (range 7–988 per sample) for the reconstruction of phylogenetic trees. Individual cells from single-cell sequencing were then assigned to clones based on the presence or absence of these driver mutations, forming a phylogenetic tree based on driver mutations for each sample (see STAR Methods).

Phylogenies enabled us to reconstruct how AML evolves from initially healthy HSCs. Across the 16 individuals analyzed, we found that the AML clones (red stars, Figure 1) most commonly harbor 3 ($n = 8$) or 4 ($n = 6$) driver mutations (Figures 1B–1E). The most commonly co-occurring drivers alongside the *DNMT3A* and *NPM1* mutations were in *NRAS* ($n = 12$), *FLT3* ($n = 11$), *TET2* ($n = 8$), *IDH1*/*IDH2* ($n = 5$), *RAD21* ($n = 4$), and *WT1* ($n = 4$). In all of the samples where the single-cell-level data enabled us to determine the order of mutation acquisition, we found that *DNMT3A* preceded *NPM1*. *DNMT3A* is typically the first mutation acquired along the evolutionary branch to the AML clone, in agreement with previous findings.[30] Mutations in the RAS pathway (*FLT3*, *NRAS*), when observed in our data, are usually late driver events, either the 3rd or 4th "hit." Acquisition of *NPM1* is usually secondary to *DNMT3A* but earlier than *FLT3* or *NRAS*.[30] Mutations in *IDH1/2* and *TET2* were found to be mutually exclusive on the ancestral branch of the AML blasts, consistent with previous observations.[31,32]

The trees also revealed highly diverse patterns of preleukemic evolution. In approximately a third of cases (6/16), we observed evidence of branched evolution, indicating multiple evolutionary paths to higher fitness in HSCs. Our data suggest that these parallel competing lineages can coexist for a considerable time, as the coalescence events can occur early in a tree. In the remaining 10 samples, we observed patterns of linear evolution, which may indicate strong preferences for particular evolutionary paths in the fitness landscape. In addition to the qualitative differences among the trees, we also observed clear

CellPress
OPEN ACCESS

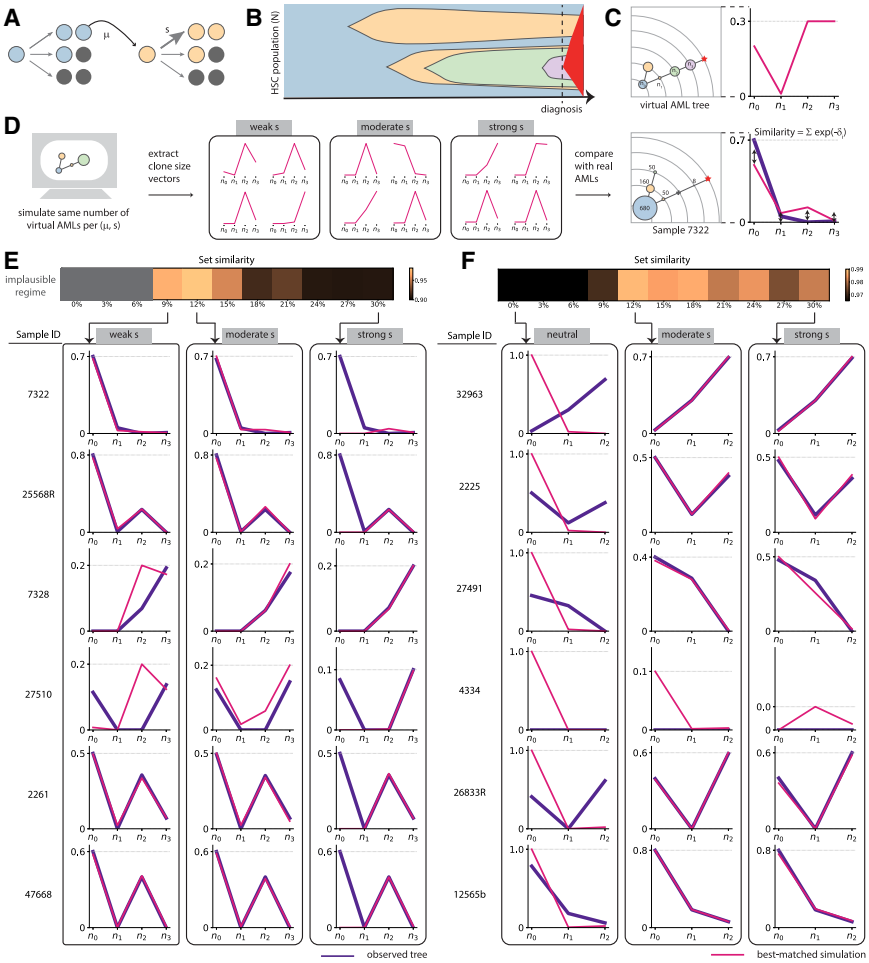


Figure 2. Statistics of preleukemic phylogenetic trees suggests pervasive positive selection operates during preleukemic evolution

(A) A model of HSC dynamics as a multi-class branching process where stem cells self-renew symmetrically at rate $1/\tau$ and acquire driver mutations stochastically at a rate of μ per year. Driver mutations confer a selective advantage s per year to the cell. Under the staircase model, every additional mutation confers an additional selective advantage to the lineage.

(B) An example of the evolutionary dynamics observed in this model where cancer diagnosis is defined as the time when the first lineage acquires k driver mutations with $k = 4$ in this example (red coloring).

(C) The phylogenetic tree constructed from driver mutations at diagnosis can be used to extract the clone size vector $*n = (n_0, n_1, n_2, n_3)$.

(D) Stochastic simulations produced four thousand 3-hit and 4,000 4-hit virtual AML events diagnosed before age 80 per parameter combination. The clone size vectors of these simulated virtual AMLs ($*n'$, pink lines) were compared against real trees ($*n$, purple line) from our single-cell experiments using a similarity metric (see STAR Methods).

(E and F) Set similarity (color scale) between observed trees (purple lines) and closest simulated trees (pink lines) for different selection coefficients across the 12 observed trees considered.

quantitative diversity at the level of clone sizes. In approximately a third of cases (6/16), we observed that wild-type HSCs—cells carrying no detectable driver mutations—remain the dominant clone in the HSC pool at the time of diagnosis (Figures 1B and 1C; e.g., ID12565 and ID2261 in Figure 1E). On the other hand, in 5 cases, we observed that the wild-type HSCs have largely been out-competed by later-occurring clones carrying multiple driver mutations (Figure 1D; e.g., ID4334 and ID32963 in Figures 1E). In the remaining samples, there appears to be coexistence of intermediate clones—harboring different numbers of driver mutations—at similar clone sizes (ID47668 and ID26833R in Figure 1E).

Inferring positive selection from phylogenies

The clone size statistics of the phylogenetic trees are the outcome of clonal competition, mutation, and stochasticity during preleukemic evolution. We reasoned that the diverse patterns observed in the AML trees may emerge naturally due to these three interacting elements in the evolutionary dynamics, independent of the specific details of the mutations involved. To test this hypothesis we developed a simplified model of preleukemic evolution, which was informed by experimental observations[9,10,24,33–36] (Figure 2A; STAR Methods).

To avoid overfitting to the data, we impose the simplifying assumption that each driver mutation is acquired at the same rate μ and confers the same advantage s to cells regardless of mutation identity, and the effects of driver mutations combine additively. Diagnosis occurs once k driver mutations occur in the same cell (red clone, $k = 4$ case in Figure 2B). This framework can be used to simulate phylogenies of HSCs at the time of diagnosis. In order to compare the quantitative patterns observed in simulated trees to those experimentally measured in the AML cases, we considered the clone size vector, $*n = (n_0, n_1, \ldots, n_{k-1})$, which quantifies the clonal frequencies of the wild-type, single-mutant, double-mutant, etc., clones along the main evolutionary branch ancestral to the AML clone at the time of diagnosis (Figure 2C). The statistical properties of the clone size vector, $*n$, reflect different levels of selection (s) operating during preleukemic evolution.

We used our model to perform in excess of a billion stochastic simulations across a range of fitness effects and mutation rates consistent with previously inferred values[9,10,13,24,34,37] (Figure 2D; STAR Methods). In these simulations, we considered a "virtual AML" to occur when a clone that carries $k = 3$ or $k = 4$ driver mutations emerges before 80 years, in line with the diagnosis ages in our cohort. We recorded $4,000$ $k = 3$ and $4,000$ $k = 4$ virtual AMLs for every selection strength considered and used them for a tree comparison with the observed clone size vectors.

Of the sixteen reconstructed trees from our single-cell experiments, twelve were unambiguously genotyped for all driver mutations. We identified the best-matched virtual AML for every one of these observed clone size vectors and calculated the total similarity over all the best-matched virtual AMLs across a range of selection strengths (Figures 2E and 2F; see STAR Methods). This provides an estimate for the likelihood that the set of observed trees is generated from preleukemic evolution driven by the assumed selection strength s—which produces more robust inferences than comparing only single trees (STAR Methods). Although individual experimental trees may appear to be under neutral evolution, we found that the statistical properties of the observed tree set are consistent with simulated tree patterns generated under moderate to strong selection ($s \geq 9\%$ per year). These selection strengths cause intermediate clones to expand to high cell fractions and lead to the coexistence of wild-type and multiple mutant clones in the HSC pool at diagnosis. When combined with stochasticity in mutation timings and clonal growth during somatic evolution, this leads to the observed tree patterns where a small fraction of cases are still dominated by the wild type. In contrast, neutral evolution ($s = 0\%$) produces highly homogeneous sets of virtual AML tree patterns where the preleukemic stem cell population is dominated by the wild-type clone in all cases. This is inconsistent with the set of our experimentally observed trees, suggesting that stochastic acquisition of mutations alone cannot account for the emergence of AML.

Validation using an independent single-cell dataset

To test how well these findings generalize to other AML cases, we considered 27 AML trees from an independent single-cell study.[27] We extracted clone size vectors, $*n$, from each of these cases and performed the same procedure for inferring levels of positive selection (STAR Methods). As was observed from our single-cell data, the statistics of the clone size vectors are highly variable, with some cases exhibiting the dominance of wild-type clones and others the coexistence of intermediate clones. Consistent with our previous findings, these trees cannot be explained by a model of neutral evolution (left, Figure 3). Models of evolution driven by moderate levels of positive selection ($9\% < s < 24\%$ per year) produce clone size vectors that are in close agreement with those observed in the 27 trees (middle, Figure 3). Increasing selection strengths even further ($s > 24\%$ per year) produces clone size vectors that become less consistent with the data (right, Figure 3). The selection strengths implied by these data are, therefore, consistent with the selection strengths inferred from our single-cell data. These findings generalize to cases that were driven by 4 driver mutations as well as trees that show considerable branching (see STAR Methods).

Variation in selection strengths across individuals

To evaluate whether there is evidence of variation in levels of positive selection across individuals, we extracted all $k = 3$ and $k = 4$ AML trees (67 in total) from Morita et al.[27] and combined these with our own experimentally observed trees. To estimate the most likely selection levels that drove the preleukemic evolution in each of these 80 trees, we identified the 100 most similar virtual AMLs for each case and considered the distribution of the selection strengths that produced them (Figure 4A).

The majority of trees are consistent with preleukemic evolution driven by selection strengths in the range of $s = 9\% - 24\%$ per year (Figure 4B). However, we observe clear variation in selection strengths across the 80 trees. Trees in which wild-type clones remain dominant at diagnosis (top rows, Figure 4B) are consistent with weaker selection strengths. In these cases, our framework is confident in the inferred levels of selection: the best-matched virtual AMLs consistently fall onto a narrow range of s. For example, our inference using this framework strongly suggests that the preleukemic evolution in ID12565 and ID2225 is consistent with being driven by weak selection strengths at $s \approx 3\% - 9\%$ per year. Trees with a diverse coexistence of multiple-mutant clones (bottom rows, Figure 4B) are inferred to have been driven by stronger levels of positive selection (ID32963 and ID7328, Figure 4B). The similarity of the best-matched virtual AMLs to the observed clone size vector is a measure of how typical such a vector is to be an outcome of the evolutionary model. For example, sample 32963 is a very typical tree resulting from our preleukemic evolutionary model (all best-matched virtual AMLs look alike), whereas sample 4334 (all entries of the clone size vector are 0) is atypical under our model choice.

To assess how finite sampling impacts our inferences, we trained neural networks on subsampled virtual AMLs to infer selection strengths s and mutation rates μ from a single tree (STAR Methods). Applying these neural networks to our experimentally observed trees, we found that the inferred selection levels are broadly consistent with our previous inferences based on clone size similarity (see STAR Methods). This indicates that subsampling introduced by finite numbers of pHSCs only influences the inferred levels of selection where cell numbers are small.

In silico modeling identifies features of future AML risk

Our finding that most observed AML trees appear to be consistent with moderate selection strengths suggests that preleukemic clones expand over timescales of decades and could be used to identify individuals at high risk of developing AML. However, a key challenge in identifying robust predictors of risk for rare cancers such as AML is in understanding how rare these features are in individuals not destined to develop the cancer. Using our *in silico* model, we are able to compare the evolutionary trajectories that precede virtual AMLs (Figure 5A) to the trajectories observed in hundreds of thousands of simulated individuals who never develop virtual AML. In an HSC population of $\sim 10^5$ stem cells[9,10] where driver mutations occur at rate of $\sim 10^{-5}$ per year,[37] clones carrying single driver mutations occur routinely in both future AML cases and controls (orange shading, Figures 5B and 5C). This is consistent with what is known about the prevalence of clonal hematopoiesis in healthy individuals[14,15] and suggests that the detection of clones carrying a single driver mutation is unlikely to be a robust predictor of AML. However, the emergence of double-mutant clones (i.e., cells harboring two driver mutations) is predicted to be rare in healthy individuals under the age of 30. In contrast, more than half of future virtual

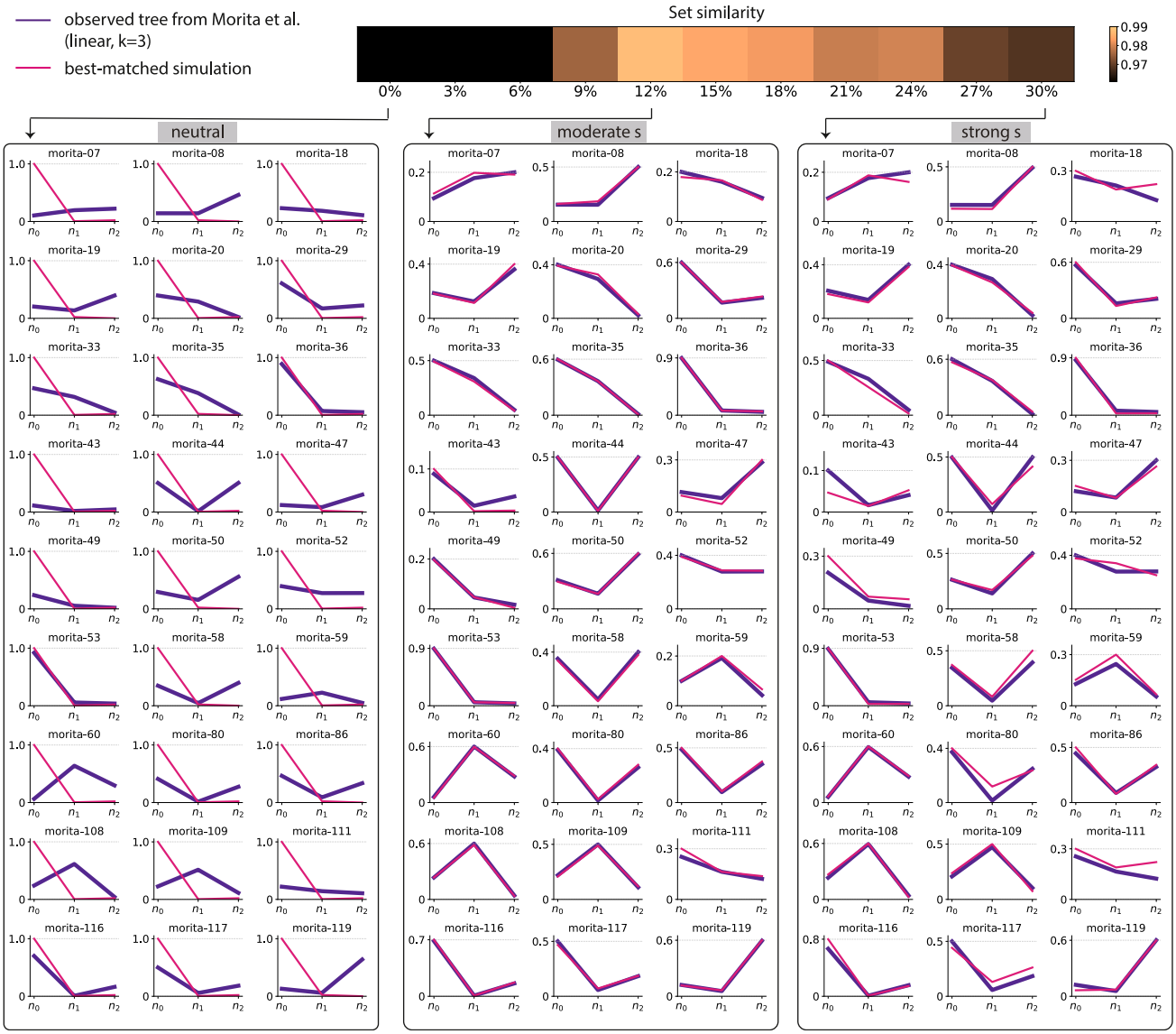


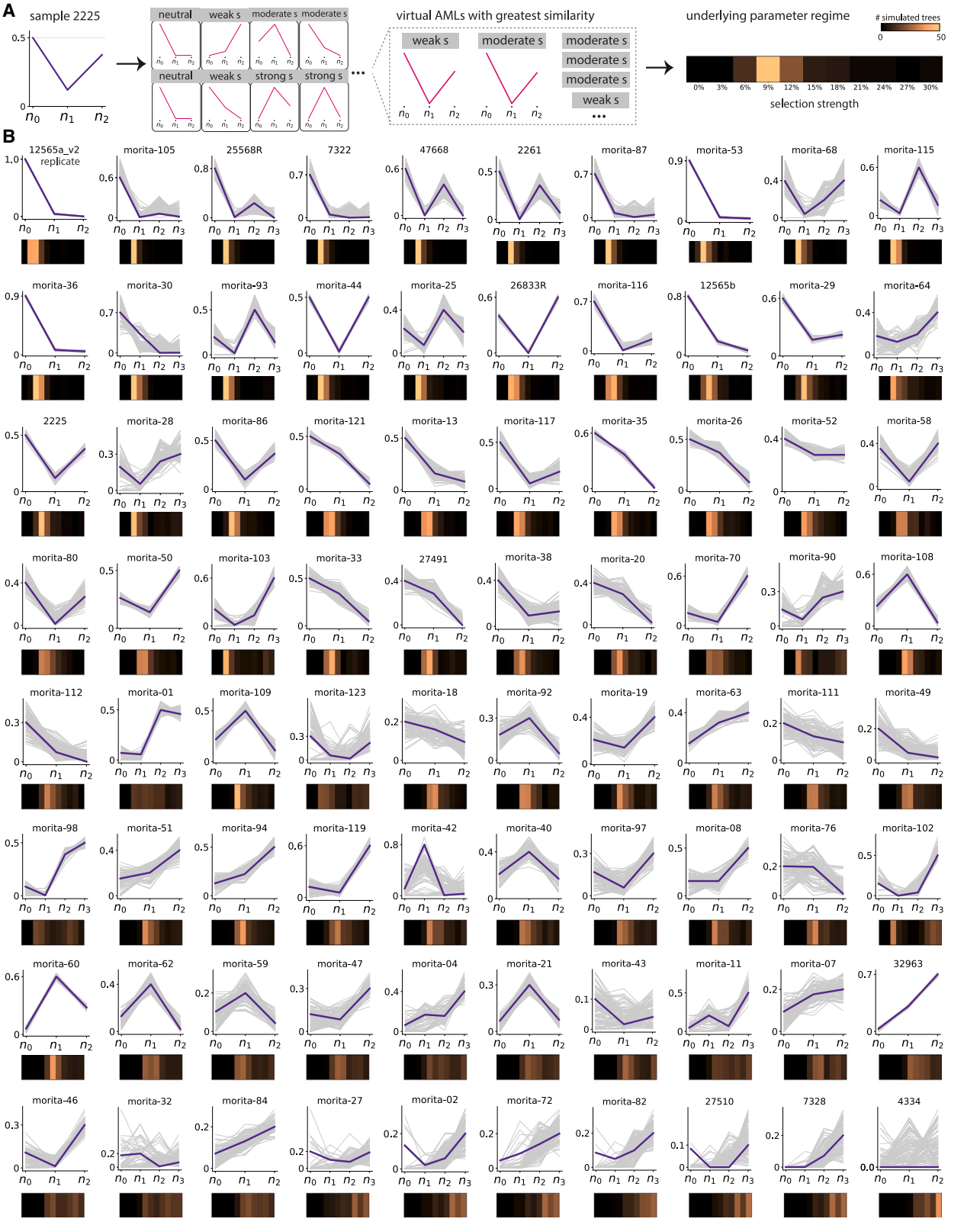
Figure 3. Independent AML single-cell dataset validates that moderate selection operates during preleukemic evolution

Clone size vectors from an independent dataset of 27 $k = 3$ single-cell AML trees from Morita et al.[27] (purple lines) compared to the closest virtual AML clone size vectors generated from our stochastic simulations (pink lines) for three different selection strengths (panels). The normalized set similarity (color scale) between all 27 trees and their closest simulated trees across the full set of selection strengths modeled is shown on the top (see STAR Methods).

AML cases have already acquired a double mutant by the same age (green shading, Figures 5B and 5C). While the relative risk conferred by the emergence of a double mutant is considerable, its absolute risk remains modest (Figure 5D) (see STAR Methods for $k = 3$). The emergence of a triple-mutant clone before the age of 50 confers both a large relative and absolute risk of being diagnosed with a 4-hit virtual AML before age 60. In summary, by simulating the evolutionary dynamics using parameters inferred from the experimental trees and examining large numbers of virtual AML cases and controls, one can learn the general features that identify future AML cases. Multiple-mutant clones (carrying >1 driver mutation in the same lineage) that arise anomalously early (e.g., in individuals <30 years of age) carry a considerable risk of transforming to AML.

DISCUSSION

We have shown that preleukemic clones, which harbor subsets of the driver mutations present in the AML clone, coexist in the HSC/multipotent progenitor (MPP) pool at the time of AML diagnosis. Single-cell sequencing of these clones enabled the construction of phylogenetic trees depicting the clonal evolution that occurs in HSCs antecedent to AML. We found that clone size patterns revealed by the phylogenetic trees were highly variable across individuals. Using evolutionary modeling, we



(legend on next page)

showed that these diverse patterns are representative outcomes of somatic evolution that is driven by pervasive positive selection. The major implication of these inferred selection strengths is that preleukemic clones expand from low frequencies to become detectable decades prior to the AML diagnosis. This provides a long window of opportunity for risk prediction and potential intervention. Our inferences are based on the assumption that selection strengths remain constant throughout life. However, we cannot rule out that selection pressures may vary in time due to extrinsic factors including aging,[34,38] cytotoxic therapies, and smoking.[37]

Our method is applicable to the early stage of cancer evolution when positive selection effects are weaker than during tumor evolution. Existing methods analyzing tree structures in cancer evolution[39–42] were designed with tumor evolution in mind, relying heavily on topological information, and rarely account for clonal sizes. Our approach takes advantage of the simplicity of preleukemic tree structures and directly utilizes clone size information by comparing the frequencies of preleukemic mutants with virtual AMLs, achieving higher resolution on the fitness landscape near neutrality ($s = 0$). Unlike existing methods, our approach is suited to revealing positive selection during precancerous evolution—when the stem cell population is dominated by a few coexisting mutant clones. This was the scenario observed in clonal hematopoiesis and precancerous evolution of healthy tissues,[10,19,43] as well as our own experimental preleukemic trees. By harnessing the clone size statistics of wild-type and early mutants, we showed that persistent positive selection acts throughout the multi-hit preleukemic trajectory toward AML.

Limitations of the study

There are a number of technical limitations in our work that need to be considered when interpreting the results. Chief among these is our premise that the sizes of preleukemic clones are determined only by the forces of selection, mutation, and drift operating during preleukemic evolution. One important possible technical effect is that clones sizes may also reflect biases introduced during the experimental process of isolating and sequencing HSCs. Specifically, if the immunophenotype of the HSC co-varies consistently with the presence of a particular driver mutation, then this could cause systematic biases in our inference of selection strengths. To evaluate the influence of this effect, we re-performed our analysis using a similarity metric that considers only the relative sizes of wild-type and first mutants. By focusing on their relative frequencies, this removes effects of potential blast contamination. We show that the inferred s values from this metric are highly consistent with our original inferences (see STAR Methods). Additionally, we performed a technical replicate of our experimental procedure on the bone marrow aspirates from one patient (ID12565; see STAR Methods) and confirmed that our model inferences were unaffected. The validation of our inferences on an independent single-cell dataset[27] generated from unsorted samples also suggests that the experimentally observed clone size vectors are not primarily dominated by systematic biases introduced during the isolation of pHSCs.

While the conceptual model presented here is clearly an oversimplification of the underlying biology (neglecting epistasis between driver mutations, extrinsic perturbations altering the fitness of clones,[34,44] explicit effects of aging,[38] and many other factors), it nevertheless highlights that the patterns observed in the phylogenetic trees can be explained by a far simpler model that has as its defining feature the requirement for positive selection on intermediate clones. In order to focus on the general behaviors that emerge from positive selection, the model intentionally ignores differences between driver mutations (i.e., all drivers confer the same selective advantage to HSCs and their effects combine additively in fitness effect). This simplified model highlights that the divergence between healthy and pre-malignant evolutionary dynamics in blood usually occurs due to the emergence of an anomalously early double-mutant clone. We propose that the detection of rare clones harboring two driver mutations within a single cell in young adults could provide a rational basis for identifying individuals at high risk of progression to AML. Robustly testing this prediction will require analysis of large single-cell datasets from healthy and pre-AML blood—an important area for future work. Our conceptual framework also provides a potential route to resolving the apparent discrepancies between models of oncogenesis based on stochastic mutation accumulation and bad luck,[6,45–47] the detection of driver mutations in healthy tissues,[9,10,14–17,48–50] and the observed lifetime risk of AML.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Jamie R. Blundell (jamieblundell@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All code used in this study is available on the Blundell lab GitHub page (https://doi.org/10.5281/zenodo.14257378). The raw sequencing data are deposited on the European Nucleotide Archive: project accession PRJEB81526.

ACKNOWLEDGMENTS

We thank Daniel Fisher for helpful discussions and the team at Mission Bio for their technical support throughout this project. We thank all members of the Blundell lab for input. G.P. and J.B. are funded by a UKRI Future Leaders Fellowship and by the CRUK Cambridge Cancer Centre. We thank the Cambridge NIHR BRC Cell Phenotyping Hub for flow cytometry sorting. E.L. is funded by a Wellcome-Royal Society Sir Henry Dale Fellowship (107630/Z/15/Z) and A.V. by a Gates Cambridge Trust PhD Scholarship (10350885) until 2022. Work in E.L.'s group is funded by BIRAX

Figure 4. Individual trees show variation in inferred selection levels

(A) The 100 virtual AML trees highest in similarity were identified for each observed tree. Sample 2225 is shown here. The inferred selection is visualized by the distribution of the number of virtual AMLs across the underlying parameter distribution.

(B) This procedure was carried out for 80 real trees, including 6 4-hit Poon trees, 7 3-hit Poon trees (containing a pair of technical replicates), 44 3-hit Morita trees, and 23 4-hit Morita trees.

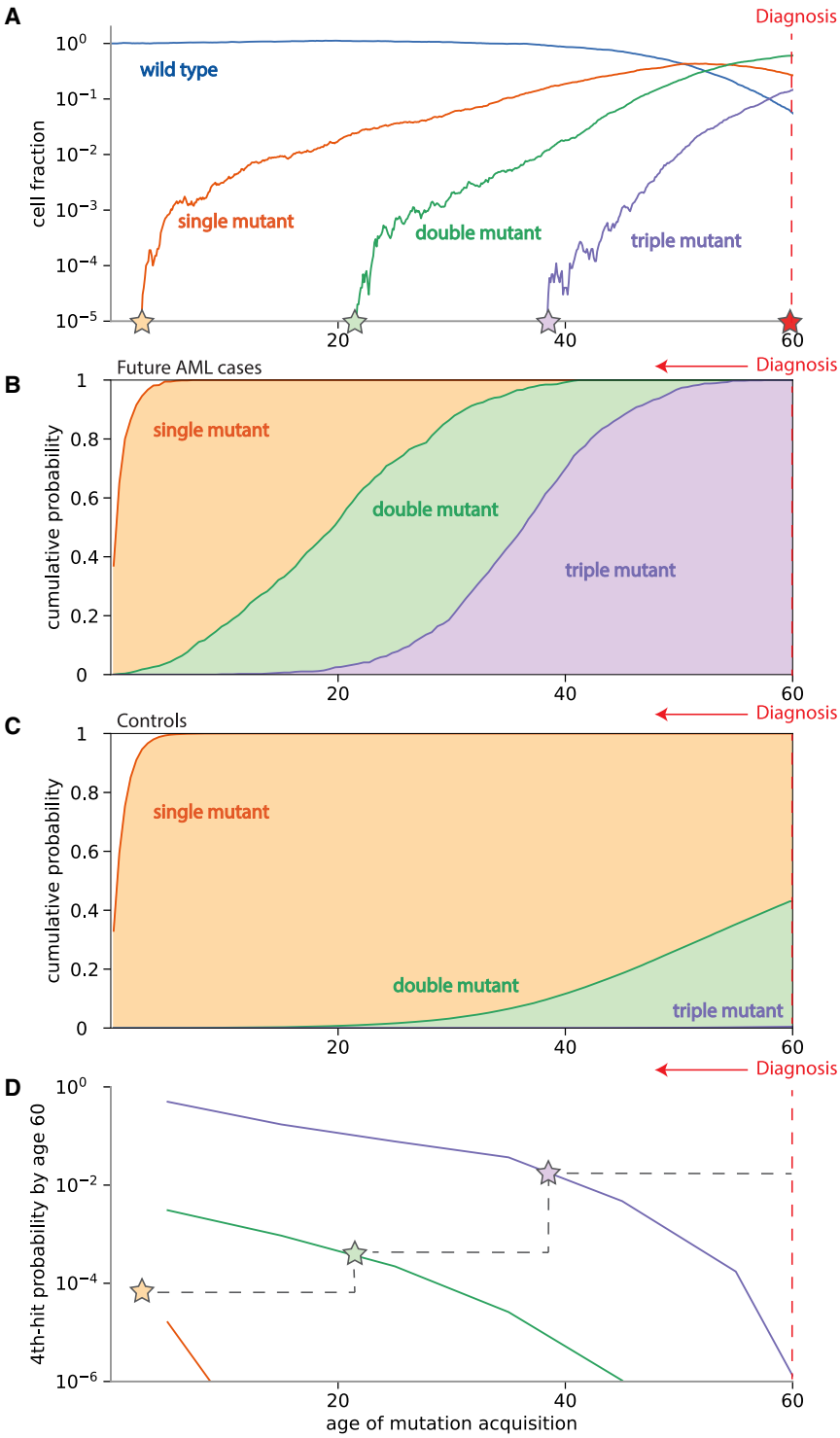


Figure 5. Early double-mutant and triple-mutant clones identify a high-risk group for AML

(A) An example of the preleukemic clonal history of a simulated individual who acquires an AML-defining 4th hit at age 60 where AML drivers confer moderate levels of positive selection.

(B) Distribution of mutation acquisition times in simulated individuals (under moderate selection and at medium mutation level) who acquired a 4th hit before age 60.

(C) Distribution of mutation acquisition times in unselected simulated individuals, showing marked differences compared to simulated individuals destined to develop AML.

(D) The absolute risk of AML diagnosis before age 60 is dependent on when single- (orange), double- (green) and triple- (purple) mutant clones are acquired. This is illustrated for the individual in (A), for whom the absolute risk jumps upward (gray dashed line) whenever the individual acquired a further mutation (star).

(47BX16ELLS) and core support grants from the Wellcome and Medical Research Council (MRC) to the Wellcome - MRC Cambridge Stem Cell Institute (203151/Z/16/Z) and the UKRI Medical Research Council (MC_PC_17230).

AUTHOR CONTRIBUTIONS

J.R.B. and G.P. conceived the project and wrote the manuscript, with input from E.L., A.V., P.V., and M.S. The single-cell experiments and all computational analyses were performed by G.P. with input from J.R.B. FACS of all samples was performed by A.V. Bone marrow aspirate samples were provided by M.S. and P.V.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Sample selection
  - AML blast genotyping and phenotyping
  - Isolation of pHSCs
  - Single-cell variant calling
  - Dropout clone analysis and missing genotypes
  - Construction of phylogenies
  - Stochastic simulation of 'virtual AMLs'
  - Tree similarity and set similarity
  - Benchmarking
  - Validation using published single-cell data
  - Extracting clone size vectors from stochastic simulations to account for subsampling
  - ML model for parameter inference

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100744.

Received: May 22, 2024
Revised: August 3, 2024
Accepted: December 26, 2024
Published: January 21, 2025

REFERENCES

1. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium; Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., and Ouellette, B.F.F. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
2. Cancer Genome Atlas Research Network; Ley, T.J., Miller, C., Ding, L., Raphael, B.J., Mungall, A.J., Robertson, A.G., Hoadley, K., Triche, T.J., and Laird, P.W. (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N. Engl. J. Med. *368*, 2059–2074. https://doi.org/10.1056/NEJMoa1301689.
3. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339. https://doi.org/10.1038/nature12634.
4. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122–128. https://doi.org/10.1038/s41586-019-1907-7.
5. Williams, N., Lee, J., Mitchell, E., Moore, L., Baxter, E.J., Hewinson, J., Dawson, K.J., Menzies, A., Godfrey, A.L., Green, A.R., et al. (2022). Life histories of myeloproliferative neoplasms inferred from phylogenies. Nature *602*, 162–168. https://doi.org/10.1038/s41586-021-04312-6.
6. Armitage, P., and Doll, R. (1954). The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. Br. J. Cancer *8*, 1–12. https://doi.org/10.1038/bjc.1954.1.
7. Durrett, R., and Moseley, S. (2010). Evolution of resistance and progression to disease during clonal expansion of cancer. Theor. Popul. Biol. *77*, 42–48. https://doi.org/10.1016/j.tpb.2009.10.008.
8. Araten, D.J., Golde, D.W., Zhang, R.H., Thaler, H.T., Gargiulo, L., Notaro, R., and Luzzatto, L. (2005). A quantitative measurement of the human somatic mutation rate. Cancer Res. *65*, 8111–8117. https://doi.org/10.1158/0008-5472.CAN-04-1198.
9. Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E., et al. (2018). Population dynamics of normal human blood inferred from somatic mutations. Nature *561*, 473. https://doi.org/10.1038/s41586-018-0497-0.
10. Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L., Druley, T.E., Fisher, D.S., and Blundell, J.R. (2020). The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science *367*, 1449–1454. https://doi.org/10.1126/science.aay9333.
11. Williams, M.J., Sottoriva, A., and Graham, T.A. (2019). Measuring Clonal Evolution in Cancer with Genomics. Annu. Rev. Genom. Hum. Genet. *20*, 309–329. https://doi.org/10.1146/annurev-genom-083117-021712.
12. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. Cell *171*, 1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.
13. Poon, G.Y.P., Watson, C.J., Fisher, D.S., and Blundell, J.R. (2021). Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. Nat. Genet. *53*, 1597–1605. https://doi.org/10.1038/s41588-021-00957-1.
14. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. en. N. Engl. J. Med. *371*, 2477–2487. https://doi.org/10.1056/NEJMoa1409405.
15. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498. https://doi.org/10.1056/NEJMoa1408617.
16. Loh, P.-R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A., Palamara, P.F., Birmann, B.M., Talkowski, M.E., Bakhoum, S.F., McCarroll, S.A., and Price, A.L. (2018). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature *559*, 350–355. https://doi.org/10.1038/s41586-018-0321-x.
17. Loh, P.-R., Genovese, G., and McCarroll, S.A. (2020). Monogenic and polygenic inheritance become instruments for clonal selection. Nature *584*, 136–141. https://doi.org/10.1038/s41586-020-2430-6.
18. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science (New York, N.Y.) *348*, 880–886. https://doi.org/10.1126/science.aaa6806.
19. Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R., et al. (2018). Somatic mutant clones colonize the human esophagus with age. Science *362*, 911–917. https://doi.org/10.1126/science.aau3879.
20. Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L., Latimer, C., Raine, K., Sanders, M.A., Warren, A.Y., Mahbubani, K.T.A., et al. (2020). Extensive heterogeneity in somatic mutation and selection in the human bladder. Science (New York, N.Y.) *370*, 75–82. https://doi.org/10.1126/science.aba8347.
21. Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P., Dentro, S.C., Dawson, K.J., Butler, T., Rahbari, R., Mitchell, T.J., et al. (2020). The mutational landscape of normal human endometrial epithelium. Nature *580*, 640–646. https://doi.org/10.1038/s41586-020-2214-z.
22. Coorens, T.H.H., Moore, L., Robinson, P.S., Sanghvi, R., Christopher, J., Hewinson, J., Przybilla, M.J., Lawson, A.R.J., Spencer Chapman, M., Cagan, A., et al. (2021). Extensive phylogenies of human development inferred from somatic mutations. Nature *597*, 387–392. https://doi.org/10.1038/s41586-021-03790-y.

23. Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. Nature *574*, 532–537. https://doi.org/10.1038/s41586-019-1672-7.

24. Mitchell, E., Spencer Chapman, M., Williams, N., Dawson, K.J., Mende, N., Calderbank, E.F., Jung, H., Mitchell, T., Coorens, T.H.H., Spencer, D.H., et al. (2022). Clonal dynamics of haematopoiesis across the human lifespan. Nature *606*, 343–350. https://doi.org/10.1038/s41586-022-04786-y.

25. Miles, L.A., Bowman, R.L., Merlinsky, T.R., Csete, I.S., Ooi, A.T., Durruthy-Durruthy, R., Bowman, M., Famulare, C., Patel, M.A., Mendez, P., et al. (2020). Single cell mutation analysis of clonal evolution in myeloid malignancies. Nature *587*, 477–482. https://doi.org/10.1038/s41586-020-2864-x.

26. Corces-Zimmerman, M.R., and Majeti, R. (2014). Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. Leukemia *28*, 2276–2282. https://doi.org/10.1038/leu.2014.211.

27. Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S.A., Yan, Y., et al. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nat. Commun. *11*, 5327. https://doi.org/10.1038/s41467-020-19119-8.

28. Ediriwickrema, A., Aleshin, A., Reiter, J.G., Corces, M.R., Köhnke, T., Stafford, M., Liedtke, M., Medeiros, B.C., and Majeti, R. (2020). Single-cell mutational profiling enhances the clinical evaluation of AML MRD. Blood Adv. *4*, 943–952. https://doi.org/10.1182/bloodadvances.2019001181.

29. Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L., Quake, S.R., and Majeti, R. (2012). Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Sci. Transl. Med. *4*, 149ra118. https://doi.org/10.1126/scitranslmed.3004315.

30. Majeti, R. (2023). Mutation order in acute myeloid leukemia identifies uncommon patterns of evolution and illuminates phenotypic heterogeneity, PREPRINT (Version 1). doi.org/10.21203/rs.3.rs-3516536/v1.

31. Figueroa, M.E., Abdel-Wahab, O., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F., et al. (2010). Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell *18*, 553–567. https://doi.org/10.1016/j.ccr.2010.11.015.

32. Gaidzik, V.I., Paschka, P., Späth, D., Habdank, M., Köhne, C.-H., Germing, U., Lilienfeld-Toal, M. von, Held, G., Horst, H.-A., Haase, D., et al. (2012). TET2 mutations in acute myeloid leukemia (AML): results from a comprehensive genetic and clinical analysis of the AML study group. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology *30*, 1350–1357. https://doi.org/10.1200/JCO.2011.39.2886.

33. Abelson, S., Collord, G., Ng, S.W.K., Weissbrod, O., Mendelson Cohen, N., Niemeyer, E., Barda, N., Zuzarte, P.C., Heisler, L., Sundaravadanam, Y., et al. (2018). Prediction of acute myeloid leukaemia risk in healthy individuals. Nature *559*, 400–404. https://doi.org/10.1038/s41586-018-0317-6.

34. Fabre, M.A., Almeida, J. G. de, Fiorillo, E., Mitchell, E., Damaskou, A., Rak, J., Orrù, V., Marongiu, M., Chapman, M.S., Vijayabaskar, M.S., et al. (2022). The longitudinal dynamics and natural history of clonal haematopoiesis. Nature *606*, 335–342. https://doi.org/10.1038/s41586-022-04785-z.

35. Weinstock, J.S., Gopakumar, J., Burugula, B.B., Uddin, M.M., Jahn, N., Belk, J.A., Bouzid, H., Daniel, B., Miao, Z., Ly, N., et al. (2023). Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. Nature *616*, 755–763. https://doi.org/10.1038/s41586-023-05806-1.

36. Robertson, N.A., Latorre-Crespo, E., Terradas-Terradas, M., Lemos-Portela, J., Purcell, A.C., Livesey, B.J., Hillary, R.F., Murphy, L., Fawkes, A., MacGillivray, L., et al. (2022). Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. Nat. Med. *28*, 1439–1446. https://doi.org/10.1038/s41591-022-01883-3.

37. Bolton, K.L., Ptashkin, R.N., Gao, T., Braunstein, L., Devlin, S.M., Kelly, D., Patel, M., Berthon, A., Syed, A., Yabe, M., et al. (2020). Cancer therapy shapes the fitness landscape of clonal hematopoiesis. Nat. Genet. *52*, 1219–1226. https://doi.org/10.1038/s41588-020-00710-0.

38. McKerrell, T., and Vassiliou, G.S. (2015). Aging as a driver of leukemogenesis. Sci. Transl. Med. *7*, 306fs38. https://doi.org/10.1126/scitranslmed.aac4428.

39. DiNardo, C.D., and Cortes, J.E. (2016). Mutations in AML: prognostic and therapeutic implications. Hematology: the American Society of Hematology Education Program *2016*, 348–355.

40. Satas, G., Zaccaria, S., Mon, G., and Raphael, B.J. (2020). SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses. Cell Syst. *10*, 323–332.e8. https://doi.org/10.1016/j.cels.2020.04.001.

41. Ciccolella, S., Bernardini, G., Denti, L., Bonizzoni, P., Previtali, M., and Della Vedova, G. (2021). Triplet-based similarity score for fully multilabeled trees with poly-occurring labels. Bioinformatics *37*, 178–184. https://doi.org/10.1093/bioinformatics/btaa676.

42. Skums, P., Tsyvina, V., and Zelikovsky, A. (2019). Inference of clonal selection in cancer populations using single-cell sequencing data. Bioinformatics (Oxford, England) *35*, i398–i407. https://doi.org/10.1093/bioinformatics/btz392.

43. Martincorena, I., Jones, P.H., and Campbell, P.J. (2016). Constrained positive selection on cancer mutations in normal skin. Proc. Natl. Acad. Sci. USA *113*, E1128–E1129. https://doi.org/10.1073/pnas.1600910113.

44. Bolton, K.L., Ptashkin, R.N., Gao, T., Braunstein, L., Devlin, S.M., Kelly, D., Patel, M., Berthon, A., Syed, A., Yabe, M., et al. (2019). Oncologic therapy shapes the fitness landscape of clonal hematopoiesis. Genetics. https://doi.org/10.1101/848739.

45. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer Genome Landscapes. Science (New York, N.Y.) *339*, 1546–1558. https://doi.org/10.1126/science.1235122.

46. Jassim, A., Rahrmann, E.P., Simons, B.D., and Gilbertson, R.J. (2023). Cancers make their own luck: theories of cancer origins. Nat. Rev. Cancer *23*, 710–724. https://doi.org/10.1038/s41568-023-00602-5.

47. Tomasetti, C., Li, L., and Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science *355*, 1330–1334. https://doi.org/10.1126/science.aaf9011.

48. Abascal, F., Harvey, L.M.R., Mitchell, E., Lawson, A.R.J., Lensing, S.V., Ellis, P., Russell, A.J.C., Alcantara, R.E., Baez-Ortega, A., Wang, Y., et al. (2021). Somatic mutation landscapes at single-molecule resolution. Nature *593*, 405–410. https://doi.org/10.1038/s41586-021-03477-4.

49. Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun. *7*, 12484. https://doi.org/10.1038/ncomms12484.

50. Young, A.L., Tong, R.S., Birmann, B.M., and Druley, T.E. (2019). Clonal haematopoiesis and risk of acute myeloid leukemia. Haematologica *104*, 2410–2417. https://doi.org/10.3324/haematol.2018.215269.

51. Jongen-Lavrencic, M., Grob, T., Hanekamp, D., Kavelaars, F.G., Al Hinai, A., Zeilemaker, A., Erpelinck-Verschueren, C.A.J., Gradowska, P.L., Meijer, R., Cloos, J., et al. (2018). Molecular Minimal Residual Disease in Acute Myeloid Leukemia. N. Engl. J. Med. *378*, 1189–1199. https://doi.org/10.1056/NEJMoa1716863.

52. Mission bio (2020). 2020, Tapestri Single-Cell DNA Sequencing Requirements Guide., accessed June 2020. en-US. Library Catalog: support.missionbio.com.

53. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10. https://doi.org/10.14806/ej.17.1.200.

54. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

55. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36. https://doi.org/10.1186/gb-2013-14-4-r36.

56. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

57. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

58. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498. https://doi.org/10.1038/ng.806.

59. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.

60. Schwede, M., Jahn, K., Kuipers, J., Miles, L.A., Bowman, R.L., Robinson, T., Furudate, K., Uryu, H., Tanaka, T., Sasaki, Y., et al. (2024). Mutation order in acute myeloid leukemia identifies uncommon patterns of evolution and illuminates phenotypic heterogeneity. Leukemia *38*, 1501–1510. https://doi.org/10.1038/s41375-024-02211-z.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Acute myeloid leukemia patients' bone marrow aspirates	Dutch-Belgian Cooperative Trial Group for Hematology-Oncology (HOVON): https://hovon.nl/en/about-hovon, or the Swiss Group for Clinical Cancer Research (SAKK) clinical trials: https://www.sakk.ch/en.	N/A
Critical commercial assays		
Tapestri Single-cell DNA Myeloid Panel	Mission Bio.	https://missionbio.com/products/panels/myeloid/
Deposited data		
Novel raw scDNA-seq data from 16 patients	On European Nucleotide Archive under project accession PRJEB81526.	N/A
scDNA-seq data from 67 AML patients	Morita et al.[27]	N/A
Software and algorithms		
Custom code	https://github.com/blundelllab/Preleukemic_evolution_at_single_cell_level; https://doi.org/10.5281/zenodo.14257378	N/A

METHOD DETAILS

Sample selection

Bone marrow aspirate samples containing mononuclear cells were collected between 1998 and 2017 from treatment-naive *de novo* AML patients either enrolled in the Dutch-Belgian Cooperative Trial Group for Hematology-Oncology (HOVON) or the Swiss Group for Clinical Cancer Research (SAKK) clinical trials. All of them had provided written informed consent in accordance with the Declaration of Helsinki. We processed 17 samples from 16 patients harboring both *DNMT3A* and *NPM1* hotspot mutations, two of which being biological replicates (ID12565a and 12565b). Patients' ages range between 37 – 75 at the time of sample collection (see STAR Methods).

AML blast genotyping and phenotyping

High molecular weight genomic DNA was isolated from bone marrow samples or peripheral blood. Bulk gene re-sequencing was performed on the diagnostic samples using a 54 gene panel containing the most frequently mutated genes in myeloid malignancies. Targeted next-generation sequencing (NGS) was carried out with the Illumina TruSight Myeloid Sequencing panel (HOVON-SAKK). The genomic libraries were sequenced on the Illumina platform (Illumina, San Diego, CA) with mean target coverage of at least 500x and an average sequencing depth of 2500x. Variant calling was carried out as described previously.[51] This procedure enables the identification of high-confidence driver mutations during preleukemic evolution. Variants called that were between 20 – 80% VAF (80 – 100% VAF if homozygous) and considered highly pathogenic were assigned to the high-confidence set of driver mutations for that sample.

Isolation of pHSCs

BM mononuclear cells samples listed in STAR Methods were thawed using 50% IMDM/50% FCS with 1:100 DNase and MNCs were counted manually using a haemacytometer. Positive selection of $CD34^+$ cells from the remaining sample was achieved by incubating with CD34 Micro Beads (30μL/ 10^8 cells, Miltenyi), FcR Blocking Reagent (30μL/ 10^8 cells) and PBS +3% FBS (90μL/ 10^8 cells) for 30 min at 4 ° C. Cells were then washed and resuspended in MACS buffer and applied to a prepared LS magnetic column for manual selection as per the Miltenyi user manual. Samples with 1-5x10^7 MNCs were selected using EasySep (StemCell Technologies).

EasySep was performed by incubating in the selection cocktail for 15min and with magnetic particles for 10 min at room temperature. Once in the magnet, supernatant containing $CD34^-$cells was poured out and frozen separately, while $CD34^+$ cells were separated for staining and selection.

MNCs were then stained for flow cytometry using the antibody panel in Table S2. We then performed fluorescence-activated cell sorting (FACS) on the $CD34^+$ cells in order to isolate preleukemic HSCs (pHSCs) in bulk. Here we define pHSCs based on being phenotypically-normal HSCs in AML samples. These pHSCs should only harbor early driver mutations, a subset of the leukemic-specific mutations. We aimed to isolate the CD19-/CD33-/CD34+/CD38-/CD45dim/CD45RA-compartment and gating requirements were relaxed in some cases to meet cell input requirements for the single-cell sequencing platform (see STAR Methods). All pHSC samples were at a minimum, CD19-/CD33-/CD34+, and some were also CD45dim. Any leukemic stem cells (red stars in relevant figures) leaked into the sorted population could be identified based on cell genotype and were excluded in downstream analysis. The sorted population size varies between 2500 – 68292 cells across the 17 samples as a result of variable sample immunophenotype on blasts, blast percentage and viable cell count after thawing.

The sorted pHSC population was then put through the commercial microfluidic-based encapsulation platform ('Tapestri', by Mission Bio) during which genomic DNA was barcoded and amplified within single-cell droplets. The panel used was the catalog myeloid panel which consists of 312 amplicons (65 kb) that target 45 commonly mutated myeloid genes.[52] Samples were then made into libraries and over-sequenced using standard paired-end (2×150 bp) sequencing on Illumina platforms (MiSeq and NovaSeq 6000 with the SP flowcell). Sequencing read output shows that >5% single cells were captured per sample but typically only ≈10 – 1000 single cells were well genotyped at all driver mutation positions (see STAR Methods).

Single-cell variant calling

FASTQ files generated were then uploaded to the Tapestri DNA pipeline v2 after downsizing for a target coverage of 50x per amplicon per cell. Adapter sequences were trimmed from sequenced reads using Cutadapt v2.3[53,54] and reads were mapped to the hg19 reference genome using the BWA-MEM algorithm.[55,56] Barcodes within a small Hamming or Levenshtein distance from corrupted barcodes were corrected. Genome Analysis Toolkit[57] with a joint calling approach that follows GATK Best Practices recommendations[58,59] was used to call single nucleotide variants and indels. A custom genotyping method was used to target internal tandem duplications in the *FLT3* gene in the Tapestri pipeline. However, *FLT3-ITD* was not detected in our samples due to their low variant allele frequencies even when present (ID 32747). A final.loom file was generated for each sample, which contains a matrix of values detailing the allele frequencies of variants called. The data was then visualized using the commercial software Tapestri Insight v2.2 provided by Mission Bio and analyzed quantitatively in Python 3.10.9.

Dropout clone analysis and missing genotypes

The dropout rates estimated based on known heterozygous variants ranges between 6 – 26% (average 6 – 26%) across 15 samples (estimates not available for the remaining two). This indicates that on average 14% of cells genotyped for a certain variant was wrongly labeled as either harboring a homozygous variant or the reference allele. Due to this effect, not all clonal genotypes detected are true and some may be dropout clones from true clones. These clones were marked clearly on the phylogenetic trees and only the total cell numbers combined across ambiguous clones were reported (ID32747, ID41095, ID12565a and ID2239) when the complete mutation order cannot be ascertained. In cases where the mutation orders were clear from the bulk mutation VAFs we resolved dropout clones by assigning cells of dropout clones back to their true clones noting that dropout clones are always considerably smaller than the true clones they belong to.

Construction of phylogenies

Clone sizes measured from single-cell DNA sequencing were used to construct phylogenetic trees for each sample. Cells were assigned to respective levels based on the number of high-confidence driver mutations they harbor. There are two reasons any high-confidence drivers may not be recovered during single-cell DNA sequencing. First it is possible that the panel does not cover the driver mutation or that the variant position was not genotyped. Bulk VAFs of these high-confidence driver mutations are marked with red triangle markers. They are not included in the phylogenetic tree unless they are *DNMT3A* or *NPM1* mutations. Samples harboring high-confidence driver mutations whose presence cannot be ascertained are labeled with 'hidden driver'. Second the high-confidence driver mutation may also be genuinely missing in the preleukemic HSC pool, in which case the cell fraction of the driver mutation (purple bars) is shown as zero. Clone size vectors were then extracted from the reconstructed phylogenetic trees where only the clonal sizes on the ancestral branch of the AML and of side mutants up to k-th hit (not including clonal genotype with all high-confidence drivers) were considered.

Sample 7328: *Patient age was 68 years*

(DNMT3A R882H - > NPM1 W288Cfs* - > NRAS G12S - > IDH2 R140Q) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. This sequence of mutation acquisition is consistent with observed clonal genotypes in the single cell sequencing results. We assigned cells to their respective clones based on mutation co-occurrence observed.

Sample 7322: *Patient age was 63 years*

(DNMT3A R882H, NPM1 W288Cfs*, CBL C419F, ETV6 R55fs, TET2 A295fs, RAD21 splicing) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. TET295fs was not genotyped in the single cell sequencing due

to poor amplicon performance and RAD21 splicing was not covered by the panel (hence sample marked with 'hidden driver'). The acquisition order of the remaining four high-confidence drivers was determined by observed clonal genotypes in the single cell sequencing results. We also identified an additional branch of evolution in the preleukemic stem cell pool ASXL1 Q829* - > TET2 C1271Lfs*. Both mutations are at much lower variant allele frequencies in the DNA sequencing results of the bulk sample.

Sample 32963: *Patient age was 62 years*

(DNMT3A W753C, NPM1 W288Cfs*, TET2 L371X, TET2 H724fs) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. TET2 H724fs was not covered by the panel (hence sample marked with 'hidden driver'). TET2 L371X was genotyped at very low cell fraction in the preleukemic stem pool due to sorting against blast phenotype, consistent with the mutation acquisition order DNMT3A W753C - > NPM1 W288Cfs*- > TET2 L371X.

Sample 6374: *Patient age was 55 years*

(DNMT3A R771X - > NPM1 W288fs*) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. This sequence of mutation acquisition is consistent with observed clonal genotypes in the single cell sequencing results.

Sample 2225: *Patient age was 68 years*

(DNMT3A R882H, NPM1 W288Cfs*, IDH1 R132H) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. The acquisition order of the high-confidence drivers was determined by observed clonal genotypes in the single cell sequencing results.

Sample 32747: *Patient age was 55 years*

(DNMT3A R882H, NPM1 W288Cfs*, IDH2 R140Q) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. NPM1 W288fs* was not genotyped in the single cell sequencing results due to poor amplicon performance (hence sample marked with 'hidden driver'). IDH2 R140Q was not detected, suggesting that it belongs to the blast genotype. Wild type cells were assigned under the assumption that DNMT3A R882H was the first mutation acquired.

Sample 41095: *Patient age was 56 years*

(DNMT3A E426X, NPM1 W288Cfs*, RAD21 N190fs) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. DNMT3A E426X was not covered by the panel and RAD21 N190fs was not genotyped in single-cell sequencing. Order of mutation acquisition cannot be ascertained.

Sample 12565a and b: *Patient age was 50 years*

Two technical replicates had been processed for this sample. (DNMT3A R882H, NPM1 W288Cfs*, IDH2 R140Q) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. This sequence of mutation acquisition was clear from observed clonal genotypes in the single cell sequencing results of sample 12565b. We assigned cells to their respective clones based on mutation co-occurrence observed. NPM1 W288Cfs* was not genotyped for sample 12565a and therefore cells belonging to the single and double mutant clones cannot be assigned with certainty.

Sample 25568R: *Patient age was 75 years*

(DNMT3A G706E, NPM1 W288Cfs*, NRAS G12D, TET2 S657X, RAD21 Q293fs) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. All drivers were well-genotyped in single-cell sequencing except for RAD21 Q293fs (hence sample marked with 'hidden driver'). We assigned cells to their respective clones based on mutation co-occurrence observed.

Sample 27491: *Patient age was 59 years*

(DNMT3A R882C, NPM1 W288Cfs*, PTPN11 E76G, PTPN11 T73I) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. The sequence of mutation acquisition is clear from observed clonal genotypes in single cell sequencing results.

Sample 2261: Patient age was 37 years.

(DNMT3A R882H, NPM1 W288Cfs*, FLT3 D835Y, SMC3 R661P) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. This sequence of mutation acquisition is clear from observed clonal genotypes in the single cell sequencing results. We assigned cells to their respective clones based on mutation co-occurrence observed.

Sample 4334: *Patient age was 54 years*

(DNMT3A delins, NPM1 W288Cfs*, NRAS G12V) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. We only found one single clone in the single cell sequencing results. It is not ancestral to the clonal genotype with all high-confidence drivers and diverged from the main branch earlier on. The clone size vector used for downstream analysis is (0,0,0).

Sample 2239: *Patient age was 48 years*

(DNMT3A M548I, NPM1 W288Cfs*, FLT3 D835Y) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. DNMT3A M548I was not genotyped in our single cell sequencing. Assuming DNMT3A mutation was the first driver, we assigned cells to double mutants.

Sample 27510: *Patient age was 56 years*

(DNMT3A R736H, NPM1 W288Cfs*, SMC3 R381Q, FLT3 D835V) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. The sequence of mutation acquisition can be inferred from observed clonal genotypes (with some ambiguity, as labeled) in single cell sequencing results.

Sample 47668: *Patient age was 71 years*
(DNMT3A P115S, NPM1 W288Cfs*, TP53 ins, TET2 V781fs, TET2 Q1539X, SRSF2 P95L) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. SRSF2 P95L and TET2 Q1539X were not genotyped in the single cell sequencing (hence sample marked with 'hidden driver'). The acquisition order of the remaining four high-confidence drivers was inferred (with some ambiguity, as labeled) from observed clonal genotypes.
Sample 25568R: *Patient age was 70 years*
(DNMT3A L584H, NPM1 W288Cfs*, WT1 R363fs, RAD21 A572fs) formed the set of high-confidence drivers for this patient, according to the bulk sequencing results. All drivers were well-genotyped in single cell sequencing except for RAD21 A572fs (hence sample marked with 'hidden driver'). We assigned cells to their respective clones based on mutation co-occurrence observed.

Stochastic simulation of 'virtual AMLs'

We simulated the preleukemic evolutionary process before AML diagnosis based on a multi-type branching process. Stem cells divide symmetrically at rate $1/\tau = 1$ per year and acquire driver mutations stochastically at a rate of μ per cell per year. Every driver mutation confers the same absolute selective advantage s to the lineage and their effects combine additively, i.e., lineages harboring two driver mutations are twice as fit as lineages harboring only one driver mutation. This biases cell fate toward symmetric renewal and such cells produce $1+s$ offspring on average in each generation. We modeled lineage growth according to the summed fitness across its mutations minus the mean fitness of the population, which maintains the total HSC population at a relatively constant size of $N = 10^5$[9,10] and models clonal competition. All stochastic processes -the mutational process and lineage growth process-were modeled as Poisson. 'Cancer diagnosis' is defined as the event where any lineage acquires k driver mutations. This does not necessarily correspond to transformation into AML blasts. Matched to the time when experimental samples were collected from AML patients (which is upon 'cancer diagnosis'), this timepoint serves as the single-time snapshot that encodes the evolutionary history toward AML.

Using this computational model we performed simulations to generate 4000 3 rd-hit events per s across $s = 0\%, s = 3\%, s = 6\%, s = 9\%, s = 12\%, s = 15\%, s = 18\%, s = 21\%, s = 24\%, s = 27\%, s = 30\%$ per year and 4000 4 th-hit events per s across $s = 9\%, s = 12\%, s = 15\%, s = 18\%, s = 21\%, s = 24\%, s = 27\%, s = 30\%$ per year, all at $\mu = 10^{-5}$ per year. Values of s below 9% per year produce implausibly low AML occurrence rates for $k = 4$ so were not considered except for the analysis on individual variation where we examined 1000 'virtual AMLs' per s down to $s = 6\%$ per year. The clonal frequencies of clones ancestral to the $k-th$ mutant at the time of 'cancer diagnosis' were recorded for each event and stored as the corresponding clone size vector $*n = (n_0, n_1, \ldots, n_{k-1})$.

Tree similarity and set similarity

We extracted the clone size vectors for all the unambiguous preleukemic phylogenies (6 three-hit and 6 four-hit observed preleukemic trees) constructed from our single-cell experiments using the clonal frequencies of the main branches. We measured the set similarity of the 12 observed clone size vectors by summing the tree similarity of each observed clone size vector with its best-matched virtual AML tree. The set simlarities displayed in Figures 2 and 3 were normalized to show similarity calibrated to 1 for a perfect match i.e., all pairs of clone size vectors are completely identical. The tree similarity between any two clone size vector $*n$ and $*n'$ is calculated as $\sum_i^{k-1} e^{-|n_i - n_i'|}$. This metric is less sensitive to outliers than using the exponentiation of euclidean distance. We compared real, observed trees ($*n$, purple lines) from our single-cell experiments with simulated 'virtual AMLs' ($*n'$, pink lines) and found that the regime at $s = 12\%$ and $\mu = 10^{-5}$ shows the highest total similarity to the set of 6 three-hit and 6 four-hit trees from our single-cell experiments.

Benchmarking

We benchmarked our inference method using 100 virtual AMLs from each s and compared inference results with their true values. We benchmarked our set comparison method by bootstrapping the 100 virtual AMLs into 5 sets of 20 and identified the parameter that produces the best-matched 20 virtual AMLs for each set. Best-matched-ness is defined as the parameter with the highest set similarity based on clone size vector comparison ('all'). We also benchmarked our inference method using the distribution of 100 (single trees) best-matched virtual AMLs across the parameter space. We show that both set comparison and single comparison can infer fitness effect s where set comparison is less influenced by stochasticity.

To show that the wild type mutant size is the most influential statistic, we carried out the inference using a similarity metric (termed 'preleukemic') based solely on the relative sizes of wild-type and first mutants: $\frac{n_0}{n_0+n_1}$. Similarity across any two trees is measured as the difference of this metric exponentiated with a negative sign. We show that this 'preleukemic' metric can infer the true underlying s both by set comparison and single tree comparison, with slightly worse performance at higher s's than our method based on all clonal sizes. We also show that the inferred s's based on these two different similarity metrics are positively correlated and highly consistent for Poon et al. and Morita et al. AMLs.

We benchmarked four-hit virtual AMLs using the same procedure and found that set comparison is again more robust than single-tree comparison. Measuring tree similarity by taking into account all clones in the clone size vector ('all') is more powerful at the larger s's than using only the relative sizes of wild-type and first mutants: $\frac{n_0}{n_0+n_1}$ ('preleukemic'). Compared to three-hit AMLs, these positively

correlated inferred s's show lower consistency due to the loss of clonal information from n_2 and n_3. However, almost all our experimental trees have highly consistent inferred s values across the 'all' and 'preleukemic' metrics, owing to the preleukemic sorting procedure applied. Inferences for sample 27510 are inconsistent due to the specific form of the clone size vector (both n_1 and n_2 being zero). Overall, this shows that our inferences for almost all samples (except potentially 27510, 7328 and 4334) are robust.

Validation using published single-cell data

Morita et al.[27] reported the clonal architecture and mutational histories of 123 acute myeloid leukemia (AML) patients reconstructed from single-cell data. Of the 123 patients, 98 patients were analyzed for the single-timepoint sample collected at pre-treatment ($N = 98$). There were 93 (out of 123) *de novo* AML samples, among which 88 were not treated. These samples were sequenced on a narrower panel focusing on 19 AML genes. Based on additional published information on tree topologies from Schwede et al.,[60] we extracted 44 three-hits (27 linear) and 23 four-hits (15 linear) clone size vectors.

We compared the sets of $k = 3$ and $k = 4$ cases separately with 'virtual AMLs' simulated from the broader classes of parameters and found that $s = 12\%$ shows the highest similarity to the observed sets. This results holds and is not affected by how we divide the observed tree set into subsets (e.g., based on branching patterns, k…etc.). It remains highly consistent even when we reduce the number of simulated AMLs considered from each parameter regime. Trees from Morita et al. were directly inferred from bone marrow and peripheral blood mononuclear cells and in principle suffers less from subsampling biases. The highly consistent results with our inference despite low cell numbers in our experimental pipeline serves as a strong validation for our inference method.

Extracting clone size vectors from stochastic simulations to account for subsampling

To evaluate the impact of subsampling effects we performed the same simulations as described in STAR Methods to generate hundreds-of-thousands of virtual AMLs. However, to construct the clone size vectors this time we consider the absolute cell number that belongs to each clonal genotype after multinomial subsampling, instead of using clonal frequencies. Each virtual AML was subsampled n times and the probability of subsampling from a certain clonal genotype is equal to the cell fraction that clone constitutes in the total stem cell pool. The entries of a clone size vector therefore sum to n minus the number of subsampled side branch cells. We recorded 5000 3 rd-hit and 1000 4 th-hit events for each pair of (μ, s) values in the parameter space that produces cancer diagnosis rates within reasonable bounds: (10^{-4}, 0%), (10^{-4}, 15%), (10^{-4}, 30%), (10^{-5}, 15%), (10^{-5}, 30%) for $k = 4$ and (10^{-4}, 0%), (10^{-4}, 15%), (10^{-4}, 30%), (10^{-5}, 0%),(10^{-5}, 15%), (10^{-5}, 30%),(10^{-6}, 15%), (10^{-6}, 30%) for $k = 3$.

ML model for parameter inference

The 'virtual AMLs' were split into training (3335 ($k = 4$) events/26680 ($k = 4$) events) and testing (1665 ($k = 4$) events/13320 ($k = 3$) events) datasets. Each dataset was subsampled ten times to create ten subsampled datasets. We trained a fully connected neural network with 4 hidden layers for each subsampled dataset to classify the clone size statistics vector $*n$ (with corresponding k and n) by the parameter regime it was generated from. The models were trained using stochastic gradient descent and cross-entropy loss at a learning rate of 10^{-8} in batch sizes of 32 over a number of epochs. The number of epochs was decided based on the typical convergence in test accuracy across models trained using different seeds. Model accuracies converge to $\approx$65% ($n = 1000$), $\approx$ 62% ($n = 30$), $\approx$52% ($n = 10$) for $k = 3$ and $\approx$60% ($n = 300$), $\approx$60% ($n = 30$), $\approx$36% ($n = 10$) for $k = 4$. Down-sampling reduces the signal-to-noise ratios and typically results in lower classification accuracies. However models for $n \geq 30$ can still achieve test accuracy beyond 50%. By training on this large number of virtual cancers the models were able to learn how patterns in the phylogenetic trees encoded the selection strengths and mutation rates that were operating during preleukamic evolution.

We applied the fully trained neutral network to each of the 16 experimental trees to infer the patient-specific levels of selection and mutation which were operating during preleukemic evolution. For each real patient pHSC tree, we applied 10 models trained on independently sampled datasets to infer its parameter regime based on the corresponding number of hits k and the closest n to its sample size. The number of models producing a certain inference on the parameter space was counted and the sum across the ten models represents the classification outcome of the sample. This ensemble approach provides an estimate of variance for the classifier in its predictions.

The following real trees were classified with the ($n = 1000$, $k = 4$) ensemble classifier: ID7322, ID7328, and ID25568R. Tree ID2225 and ID32963 were classified with the ($n = 300$, $k = 3$) ensemble classifier and tree ID27491 was classified using the ($n = 100, k = 3$) ensemble classifier. For trees with lowest cell output, tree ID4334 was classified with the ($n = 30, k = 4$) ensemble classifier and trees ID27510 and ID2261 were classified using ($n = 30$, $k = 4$) ensemble classifier. The smallest ambiguous trees ID12565b, ID26833R and ID47668 was fed to the ($n = 10, k = 3$) and ($n = 10, k = 4$) ensemble classifier respectively. The remaining patient trees were not applied to our models as clonal genotype information was insufficient.

The classification outcomes applied to observed preleukemic trees indicate that classifiers trained on independently subsampled datasets typically converge on the inferred selection levels. The inferred selection levels are largely consistent with our previous inferences which indicates that biases introduced by subsampling only influence a small subset of the observed trees: 27510 and 26833 (and possibly ID25568R, 2261, 47668 to a lesser extent).