Preleukemic clonal evolution revealed using single-cell DNA sequencing and computational modelling

Gladys Poon[1]✉, **Aditi Vedi**[2,3], **Mathijs Sanders**[4], **Elisa Laurenti**[2], **Peter Valk**[4], **and Jamie R. Blundell**[1]✉

[1]Early Cancer Institute, University of Cambridge, UK
[2]Wellcome - MRC Cambridge Stem Cell Institute, University of Cambridge, UK
[3]Cambridge University Hospital NHS Foundation Trust, UK
[4]Department of Hematology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, the Netherlands

The acquisition of cancer driver mutations can predate diagnosis by decades. However, our understanding of the evolution that occurs in these early stages of cancer remains limited because most studies focus only on sequencing the malignant clone that "wins" the evolutionary race. The representation of driver mutations in the stem cells which do not transform into the malignant clone provides a window into this precancerous evolution. Here, we isolate preleukemic haematopoietic stem cells (pHSCs) from the bone marrow of 16 patients with Acute Myeloid Leukaemia (AML) and perform single-cell DNA sequencing on thousands of pHSCs to reconstruct phylogenetic trees for the preleukemic evolution in each patient. In the majority of cases AML required the acquisition of 3 or 4 driver mutations in the same cell lineage. However the reconstructed phylogenetic trees are highly variable across the 16 patients. The structure of the trees and the sizes of intermediate preleukemic clones encode important information about the roles of selection, mutation and genetic drift operating during precancerous evolution. Using simulations and machine learning approaches, we show that the highly variable patterns observed in pHSC trees emerge naturally from a model of somatic evolution in which there is strong positive selection acting throughout the entire multi-hit trajectory towards AML. We infer that the acquisition of each driver mutation provides an additional selective advantage of $\sim 15\%$ per year to pHSC clones relative to their immediate ancestors. Using these inferred levels of positive selection we show that individuals destined to develop AML late in life are likely to harbour multiple-mutant clones that have arisen anomalously early in life. Our findings suggest that the identification of a single HSC carrying 2 driver mutations in early adulthood may provide a rational strategy to identify individuals at high risk of AML progression.

Introduction

Large cancer genome sequencing efforts over the last 20 years have revealed that most cancers harbour 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 only mutation acquisition as the key factor seemingly capture the age-incidence relationships across a range of cancers[6]. However, using estimates of somatic mutation rates[7] and stem cell numbers these theories struggle to provide plausible estimates for overall cancer incidence[8].

A key element missing from these early multi-hit theories of cancer is the concept of positive selection acting on precancerous mutations. 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[9–20]. 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[21,22]. The haematopoietic system provides an ideal model for using single-cell approaches as haematopoietic stem cells (HSCs) can be easily isolated and sequenced. These approaches have revealed that large numbers of HSCs maintain the blood in healthy individuals[23,24], that the driver events of certain blood cancers are acquired in utero[5] and that there is extensive clonal heterogeneity in acute myeloid leukaemia (AML)[25–28]. An innovative approach to studying precancerous evolution is the isolation of ostensibly healthy preleukemic HSCs (pHSCs) from AML patients at diagnosis. Coupling this idea with single-cell sequencing has previously been used to demonstrate the step-wise nature of preleukemic evolution and how these clones can seed relapse post therapy[29]. However, the quantitative information encoded in the phylogenetic trees and intermediates clone sizes constructed from pHSCs have yet to be used to understand how positive selection operates during the entire multi-hit trajectory of preleukemic evolution.

Here we considered pHSCs from AML patients carrying *DNMT3A* and *NPM1* hotspot mutations and used single-cell sequencing to reconstruct the key evolutionary events that preceded leukemia onset. We then built a computational model of AML evolution that includes the key elements of mutation, genetic drift, positive selection on clones and the multi-hit nature of AML transformation. Using this com-

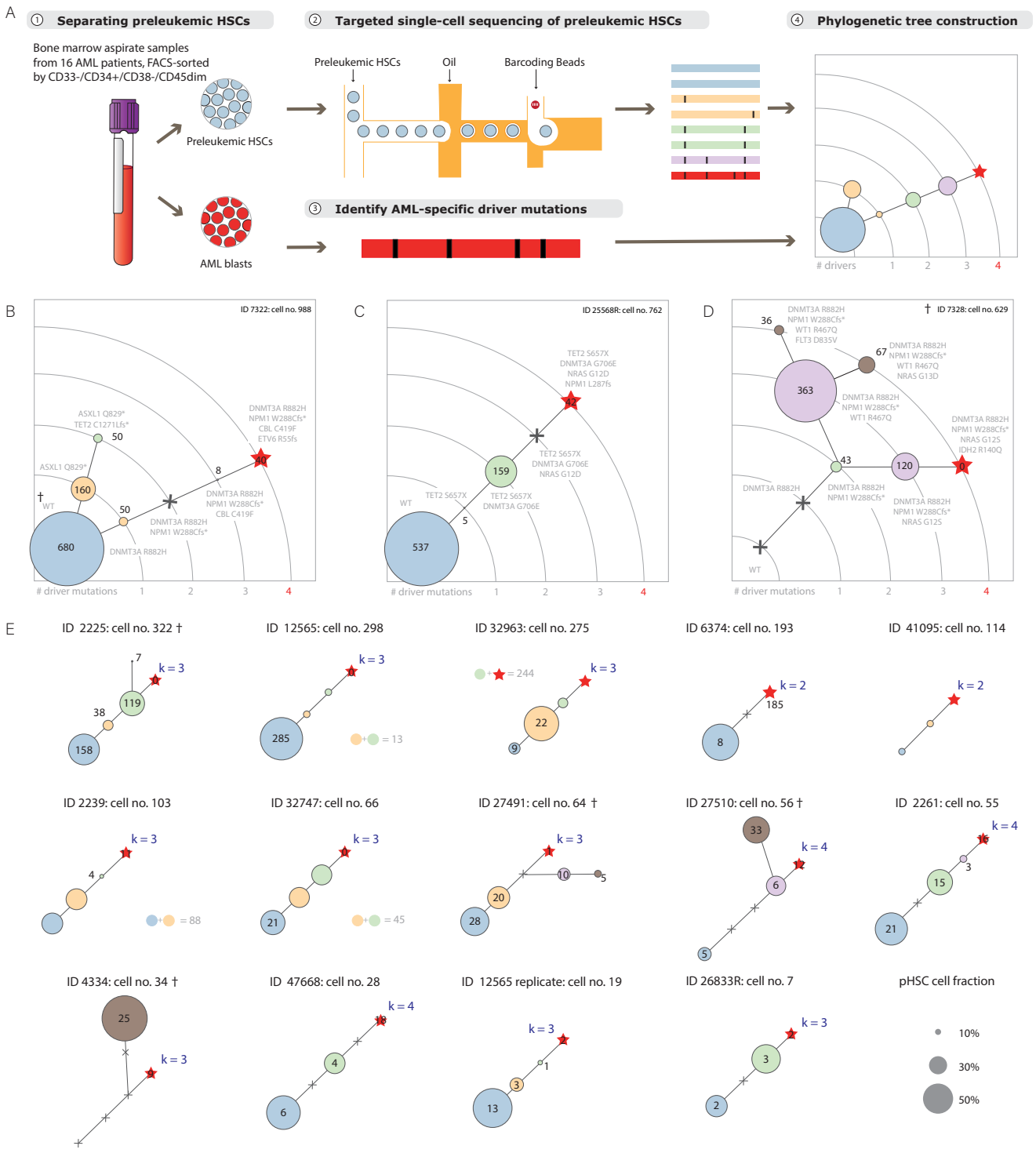

Fig. 1. Construction of phylogenetic trees from preleukemic haematopoietic stem cells from AML patients at diagnosis (A) Preleukemic haematopoietic stem cells were sorted from bone marrow aspirate samples acquired from AML patients at diagnosis. The sorted samples were then passed onto the single-cell DNA sequencing platform for genotype information on a targeted panel of 45 commonly mutated myeloid genes. Variant information at 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. **(B-D)** Three 4-hit phylogenetic trees from three patients are shown. Clones are denoted by circles and ordered on ascending arcs based on number of drivers detected. The number of cells assigned to each clone is labelled on the circle and replaced by a grey cross if zero. The genotype of a clone is also labelled and represented by a red star if it belongs to the blast genotype. **(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 replicate from patient ID12565, also see Supplementary Figure S13) and 2 trees were identified as 2-hit events. Samples that contain multiple k-th mutants, i.e. multiple possible AML blast genotypes, are marked with a †.

putational framework we performed billions of simulations to generate hundreds-of-thousands of virtual AMLs. Neural networks trained on this large number of virtual cancers 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. We demonstrate that the patterns observed in the experimental trees are consistent with a preleukemic process where positive selection acts on HSCs along the entire multi-hit trajectory with selection strengths which are similar to those observed in high-risk clonal haematopoiesis. The emergence and expansion of anomalously early multiple-mutant clones are a highly specific feature of future of AMLs and suggest more rational approaches to predicting individual cancer risk.

Results

Single-cell phylogenies from preleukemic HSCs. To construct phylogenetic trees from preleukemic HSCs (pHSCs) we obtained bone marrow aspirate samples from 16 AML patients collected at diagnosis from the multicenter HOVON-SAKK clinical trials (see 'Sample selection' in Supplementary note 1). To minimize inter-patient variation we selected samples with a genetically restricted AML genotype (*DNMT3A*^{mut}/*NPM1*^{mut}). Preleukemic HSCs were isolated by FACS sorting for the lin-CD33-/CD34+/CD38-/CD45dim compartment (Figure 1A, see 'Isolation of pHSCs' in Supplementary note 1). We then performed single-cell DNA sequencing of these pHSCs targeting 45 myeloid genes using the Tapestri platform (Mission Bio). Combining this 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 4013 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 Supplementary note 1).

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 harbour 3 ($n = 8$) or 4 ($n = 6$) driver mutations (Figure 1B-E). 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* ($n = 4$) (Supplementary Figure S1-4). 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 the 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 observe evidence of branched evolution indicating multiple evolutionary paths to higher fitness in HSCs. Our data suggests that these parallel competing lineages can coexist for a considerable time as the coalescence events can occur early in a tree (Figure 1B&D). In the remaining 10 samples, we observed patterns of linear evolution which may indicate strong preferences for particular evolutionary paths in the fitness landscape. However these patterns can also result from finite sampling effects. In addition to the qualitative differences among the trees we also observe clear quantitative diversity at the level of clone sizes. In approximately a third of cases (6/16) we observe that wild-type HSCs – cells carrying no detectable driver mutations – remain the dominant clone in the HSC pool at the time of diagnosis (Figure 1B, C and e.g. ID12565 and ID2261 in E). On the other hand, in 5 cases we observe that the wild-type HSCs have largely been outcompeted by later occurring clones carrying multiple driver mutations (Figure 1D, e.g. ID4334 and ID32963 in E). In the remaining samples there appears to be coexistence of intermediate clones – harbouring different number of driver mutations – at similar clone sizes (ID47668 and ID26833R in Figure 1E).

A stochastic model of preleukemic evolution. 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 (Figure 1) 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 (Figure 2) which is informed by experimental observations. Data from single-cell phylogenies of healthy HSCs[23,24], and from clone size statistics[11] shows that HSCs can self renew and terminally differentiate at balanced rates during adulthood. To account for this in our model, each HSC undergoes stochastic cell fate decisions at rate λ (Figure 2A). Large-scale genomic studies of AML have shown that leukemic blasts harbour between $2-5$ clonal driver mutations at diagnosis[2,33] and studies of clonal haematopoiesis have demonstrated that these mutations can confer selective advantages to the preleukemic clones[11,24,34–36]. To model these effects, each HSC can acquire driver mutations (stars) at a constant rate (μ). Each driver mutation acquired increases the fitness of the cell by a fitness effect s. To avoid over-fitting to the data we impose the simplifying assumption that each driver mutation is acquired at the same rate and confers the same advantage to cells regardless of mutation identity and the effects of driver mutations combine additively. Transformation to AML occurs once k driver mutations occur in the same cell (red clone, $k = 4$ case in Figure 2A&B). This framework

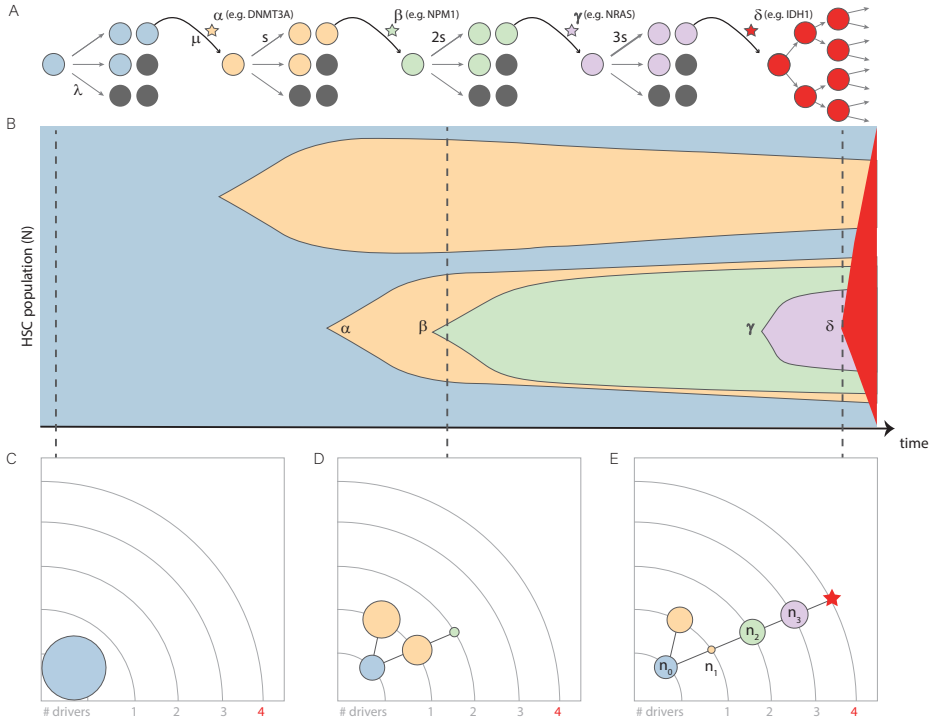

Fig. 2. An evolutionary model of cancer development. **(A)** A model of HSC dynamics as a multi-class branching process where stem cells divide symmetrically at rate λ and acquire driver mutations stochastically at a rate of μ per year. Driver mutations confer a selective advantage s per year to the cell. Under the staircase model, every additional mutation confers an additional selective advantage to the lineage, i.e. lineages harbouring two drivers are twice as fit as lineages that harbour only one driver mutation, and so on. Cancer onset is defined when any lineage acquires k driver mutations with $k = 4$ in this example. **(B)** A constant population of HSCs evolving according to this multistage branching process. Due to the constraint of a constant HSC population, clones acquiring different numbers of driver mutations compete with one another over time. Clone sizes reflect the interaction of mutation, genetic drift and positive selection. **(C)** At the start of preleukemic evolution none of the HSCs harbour any driver mutations and a phylogenetic tree constructed at this time will consists of one single wild-type clone sitting at the smallest arc corresponding to 0 drivers. **(D)** Driver mutations are acquired in the HSC pool as it evolves in time. In this example, two mutations have arisen in separate clones, one of which was followed by a second mutation acquired in the same clone. This results in a phylogenetic tree with branching at the wild-type clone level, where two branches extend up to the arc corresponding to 2 and 1 driver mutations respectively. **(E)** The phylogenetic tree detected at the time of cancer onset preserves evolutionary patterns resulting from preleukemic evolution. The sizes of clones belonging to the ancestral branch of AML i.e. the clone size vector $\boldsymbol{n} = (n_0, n_1, n_2, n_3)$ encodes information about the effects of positive selection, mutation and genetic drift.

can be used to simulate phylogenies of HSCs at the time of transformation (Figure 2C-E). 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, $\boldsymbol{n}$, which quantifies the size of the wild-type, single mutant, double mutant etc. clones along the main evolutionary branch ancestral to the AML clone (Figure 2E).

Inferring evolutionary parameters from phylogenies. The statistical properties of the clone size vector, $\boldsymbol{n}$, may reflect different levels of mutation (μ) and selection (s) operating during preleukemic evolution. However the function that links $\boldsymbol{n}$ to these parameters is analytically and numerically challenging to calculate due to stochasticity in mutation occurrence times, non-linearities induced by clonal competition and fluctuations introduced by finite HSC sample sizes. We reasoned that a neural network could be trained to learn this function using large numbers of simulated examples (Figure 3A, Supplementary note 2). To test this, 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[11,12,23,24,35,37](see 'ML model for parameter inference' in Supplementary note 2). 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. For each of these virtual AMLs we considered the tree of preleukemic evolution and simulated the subsampling process to generate clone size vectors to train a neural network model that was given the task of classifying which parameter regime each vector was generated from. After training on tens of thousands of examples, the model is able to distinguish between trees generated by neutral evolution ($s = 0$), moderate positive selection ($s = 0.15$) and very strong positive selection ($s = 0.30$) with a high degree of accuracy (Figure 3B). The model is able to distinguish between these evolutionary scenarios by learning the marked differences in tree patterns observed across the three different regimes (Figure 3B). Neutral evolution results in 'stochastic tunnelling' where the virtual AML clone typically emerges before any of the intermediate clones are detectable. In contrast, moderate to very strong positive selection results in the intermediate clones expanding to dominate the HSC pool. The clone size patterns aris-

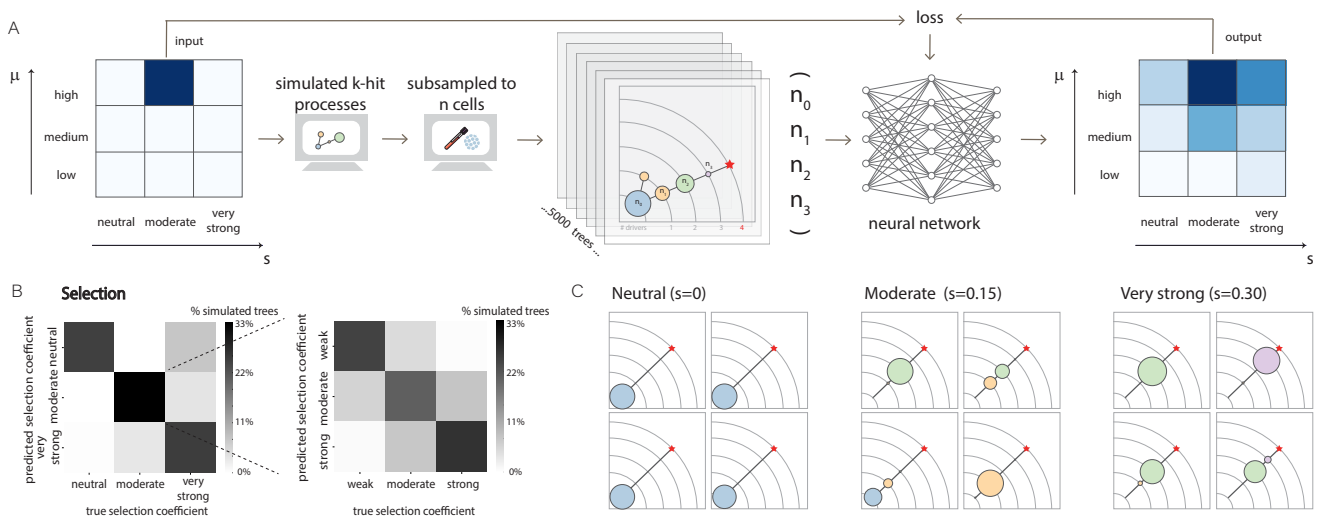

Fig. 3. A statistical learning framework infers levels of selection and mutation encoded in preleukemic phylogenetic trees. (**A**) Stochastic simulations of the evolutionary dynamics were performed with known parameters for mutation rate μ and fitness effect s in a constant population, N of HSCs, dividing symmetrically every $1/\lambda$ years. The process is stopped once the $k-$th driver mutation is acquired. Five thousand k-hit events with diagnoses before age 80 were recorded for every set of parameters. For each simulation the phylogenetic tree was recorded and sub-sampled to n cells ($n = 30, 100, 300, 1000$) to model the effects of subsampling during the process of single-cell HSC sequencing. The resulting clone size vectors were then used to train a neural network classifier to infer the correct levels of selection and mutation. The classifier was trained on two-thirds of all simulated events and tested on the one-third of events that was held out. (**B**) We first trained the models to classify across a wider range of selection strengths : neutral ($s = 0\%$ per year), moderate ($s = 15\%$ per year), and very strong ($s = 30\%$ per year) before subsequently training models over a narrower range of selection strengths: weak ($s = 9\%$), moderate ($s = 15\%$), strong ($s = 21\%$) per year. The performance of the classifier model in recognising selection strengths trained for three-hit processes after subsampling to $n = 1000$ is illustrated as an example. This particular classifier model achieves $> 50\%$ accuracy in assigning events to the correct class among 9 classes spanning across different selection strengths and mutation rates (see 'ML model for parameter inference' in Supplementary note 2). (**C**) Representative clone size vectors simulated under the operation of different selection strengths.

ing from more subtle differences in selection strengths are less visually clear (Supplementary Figure S12). However, the neural network is also able to distinguish these more subtle differences in selection strengths (Figure 3B). The accuracy of these inferences depends on the HSC sample size with $n \geq 30$ HSCs typically being required for high performance i.e. test accuracy $\geq 50\%$ (Supplementary Figure S8). In summary, a neural network trained on tens of thousands of virtual AML trees can accurately infer selection strengths and mutation rates from a single example tree.

To infer what mutation rates and selection strengths could have given rise to our experimentally observed trees, we used the observed clone size vectors as input to our neural network that had been fully trained on 'virtual AMLs'. Of the sixteen reconstructed trees, eleven were unambiguously genotyped for all driver mutations and thus could be used as input into our machine learning framework for parameter inference. The majority of the observed trees are consistent with preleukemic evolution whereby intermediate clones are under substantial strengths of positive selection ($s \approx 0.15$, Figure 4A). Moderate selection strengths cause intermediate clones to expand to high cell fractions, often leading to the co-existence of wildtype and multiple mutant clones in the HSC pool, consistent with observations (Figure 4B-L). Positive selection on intermediate clones also increases the probability of acquiring three or four driver mutations within a single cell over a lifetime to within a range that is consistent with the observed lifetime risk of AML (Supplementary Figure S6). Our data suggests, however, AML may arise through diverse levels of positive selection on intermediate clones. One of our trees (ID7322) was predicted to be consistent with neutral evolution. The patterns of preleukemic evolution in this tree were more consistent with 'stochastic tunnelling' whereby the AML clone emerges from intermediate clones with little to no clonal expansion. This finding is notable because a model of neutral evolution is harder to reconcile with the lifetime risk of AML (Supplementary Figure S6).

In silico modelling 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[11,23] 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, Figure 5B and C). This is consistent with what is known about

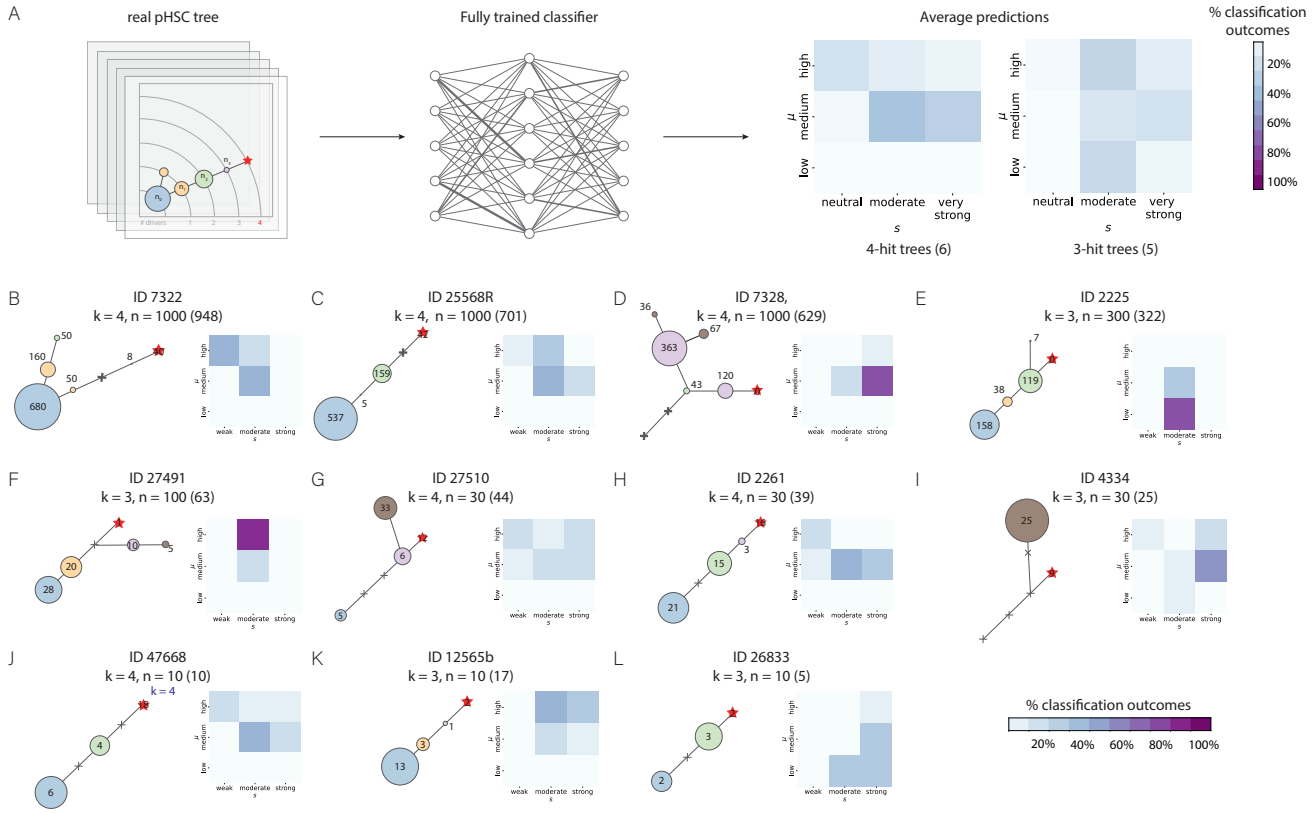

Fig. 4. Phylogenetic trees from preleukemic HSCs suggest moderate levels of positive selection drives AML. (**A**) The average classification outcomes are shown for all 6 four-hit trees and for all 5 three-hit trees which were unambiguously genotyped. (**B-L**) Individual classification outcomes of the eleven experimentally observed phylogenetic trees (6 four-hit trees and 5 three-hit trees) are shown. For each tree, ten classifier models of the relevant n and k (labelled on top of tree, Supplementary figure S9) were used to generate the final classification outcome, displayed on the right of the tree. This classification outcome is the sum of the ten models' classification results spanning selection strength: $s = 0\%, 15\%$ or 30% per year and mutation rate: $\mu = 10^{-6}, 10^{-5}$ or 10^{-4} per year in the parameter regime (see 'ML model for parameter inference' in Supplementary note 2).

the prevalence of clonal haematopoiesis in healthy individuals[13,14] 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 harbouring two driver mutations) is predicted to be rare in healthy individuals under the age of 30. In contrast, more than half of future 'virtual AML' cases have already acquired a double mutant by the same age (green shading, Figure 5B and C). While the relative risk conferred by the emergence of a double mutant is considerable, its absolute risk remains modest (Figure 5D) (see Supplementary Figure S11 for $k = 3$). The emergence of a triple mutant clones before the age of 50 confers both a large relative and absolute risk of developing a four-hit 'virtual AML' before age 70. 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 which 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 harbour subsets of the driver mutations present in the AML clone, coexist in the HSC/multipotent progenitor (MPP) pool at the time 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. Despite focusing on a genetically restricted set of AML (i.e. *DNMT3A*^{mut}/*NPM1*^{mut}), we found that clone size patterns revealed by the phylogenetic trees were highly variable across the 16 individuals studied. Using evolutionary modelling we show that these variable patterns emerge naturally from a toy model of AML evolution where there is a vital role for positive selection on preleukemic clones. Specifically, each driver mutation occurs stochastically and confers an additional selective advantage to the HSCs. Stem cells that acquire the correct complement of $3-4$ driver mutations then define transformation to AML.

Whilst this model is clearly an oversimplification of the underlying biology (neglecting spatial effects, epistasis between driver mutations, extrinsic perturbations altering the fitness of clones, explicit effects of ageing and many other factors) it nevertheless highlights that the patterns observed in the phylogenetic trees can be explained by a far simpler model

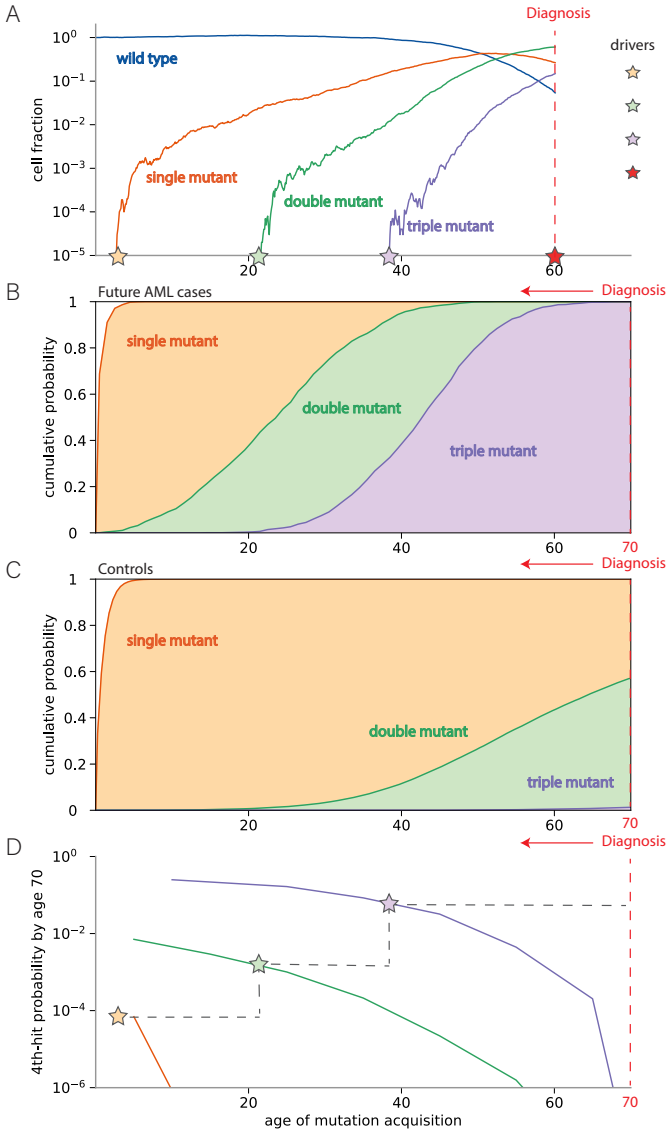

Fig. 5. Early double-mutant and triple-mutant clones identify a high-risk group for AML transformation. **(A)** An example of the preleukemic clonal history of a simulated individual who acquires an AML-defining 4th-hit at age 60 where AML drivers confer moderate levels of positive selection. **(B)** Distribution of mutation acquisition times in simulated individuals (under moderate selection and at medium mutation level) who acquired 4th-hit before age 70. **(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 transformation before age 70 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 upwards (grey dashed line) whenever the individual acquired a further mutation (star).

which has as its defining feature the requirement for positive selection on intermediate clones. 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) to highlight that the emergence of multiple mutant clones (those harbouring 2 or 3 drivers in a single cell) is a key determinant of AML risk. More specifically, our framework highlights that the divergence between healthy evolutionary dynamics and premalignant evolutionary dynamics in blood occurs at the double mutant stage. We propose that the detection of rare clones harbouring 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.

There are a number of technical limitations in our work that also need to be considered when interpreting the results. Chief amongst 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. Whilst it is well established that these factors do indeed shape clone sizes, many other effects, both technical and biological, can contribute to clone size differences. 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, this could cause systematic biases in our inference of selection strengths. While it is hard to rule out this effect entirely, the observation that clones harbouring the same set of drivers exhibit highly variable cell fractions across samples suggests that these systematic effects are not the dominant contribution to observed clone sizes. To further rule out the role of purely technical effects we performed a technical replicate of our experimental procedure on the bone marrow aspirates from one patient (ID12565 in Supplementary Figure S13) and confirmed that our model inferences were unaffected.

Large cellular populations, such as the pool of HSCs that maintain human blood, can acquire large numbers of mutations in parallel. In these populations mutations conferring a fitness advantage to cells are common and drive evolutionary dynamics characterized by clonal interference and multiple mutants[38]. Even in the simplest theoretical models of these systems the evolutionary dynamics are surprisingly rich[38]. The key challenge in using these models to understand the evolution of AML in our data was to learn how parameters in this evolutionary model dictate the statistical properties of the phylogenetic trees in the anomalous cases where an AML clone emerges. Because the emergence of an AML clone is a rare event, calculating the statistical properties of these phylogenies is analytically challenging. However, we have shown that neural networks trained on large-scale simulated data are sufficiently flexible function approximators that they can learn how parameters (such as mutation rates, selections strengths, population sizes and numbers of driver mutations needed for cancer) map on to the statistical properties of the phylogenetic trees. Here we have shown that this data-inspired framework of preleukemic evolution can provide insights about the factors shaping future cancer risk, highlighting the key roles of positive selection acting on preleukemic clones and the emergence cells harbouring multiple driver mutations. Future work incorporating more complex effects including differences between driver mutations, epitasis and effects of external perturbations could be used to guide more rational approaches to early detection, risk stratification and chemo-prevention strategies.

Acknowledgements

We thank Daniel Fisher for helpful discussions and the team at Mission Bio for their technical support throughout this project. We thanks all members of the Blundell lab for input. Y.P.G.P and J.R.B are funded by a UKRI Future Leaders Fellowship and by the CRUK Cambridge Cancer Centre. We thank 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 Gates Cambridge Trust PhD Scholarship (10350885) until 2022. Work in E.L. group is funded by BIRAX (47BX16ELLS) and core support grants by 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). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Data and code availability

All code used in this study will be available on the Blundell lab Github page.

References

1. Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020). Number: 7793 Publisher: Nature Publishing Group.
2. Cancer Genome Atlas Research Network, Ley, T.J., Miller, C., Ding, L., Raphael, B.J. et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England Journal of Medicine* **368**, 2059–2074 (2013).
3. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013). Number: 7471 Publisher: Nature Publishing Group.
4. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020). Number: 7793 Publisher: Nature Publishing Group.
5. Williams, N., Lee, J., Mitchell, E., Moore, L., Baxter, E.J. et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).
6. Armitage, P. & Doll, R. The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. *British Journal of Cancer* **8**, 1–12 (1954). Number: 1 Publisher: Nature Publishing Group.
7. Araten, D.J., Golde, D.W., Zhang, R.H., Thaler, H.T., Gargiulo, L. et al. A quantitative measurement of the human somatic mutation rate. *Cancer Research* **65**, 8111–8117 (2005).
8. Durrett, R. & Moseley, S. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology* **77**, 42–48 (2010).
9. Williams, M.J., Sottoriva, A. & Graham, T.A. Measuring Clonal Evolution in Cancer with Genomics. *Annual Review of Genomics and Human Genetics* **20**, 309–329 (2019). _eprint: https://doi.org/10.1146/annurev-genom-083117-021712.
10. Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K. et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
11. Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020). Publisher: American Association for the Advancement of Science Section: Research Article.
12. Poon, G.Y.P., Watson, C.J., Fisher, D.S. & Blundell, J.R. Synonymous mutations reveal genome-wide levels of positive selection in healthy tissues. *Nature Genetics* **53**, 1597–1605 (2021). Number: 11 Publisher: Nature Publishing Group.
13. Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A. et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *New England Journal of Medicine* **371**, 2477–2487 (2014).
14. Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *The New England Journal of Medicine* **371**, 2488–2498 (2014).
15. Loh, P.R., Genovese, G., Handsaker, R.E., Finucane, H.K., Reshef, Y.A. et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350 (2018).
16. Loh, P.R., Genovese, G. & McCarroll, S.A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* 1–6 (2020). Publisher: Nature Publishing Group.
17. Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (New York, N.Y.)* **348**, 880–886 (2015).
18. Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
19. Lawson, A.R.J., Abascal, F., Coorens, T.H.H., Hooks, Y., O'Neill, L. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science (New York, N.Y.)* **370**, 75–82 (2020).
20. Moore, L., Leongamornlert, D., Coorens, T.H.H., Sanders, M.A., Ellis, P. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020). Number: 7805 Publisher: Nature Publishing Group.
21. Coorens, T.H.H., Moore, L., Robinson, P.S., Sanghvi, R., Christopher, J. et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
22. Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019). Number: 7779 Publisher: Nature Publishing Group.
23. Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473 (2018).
24. Mitchell, E., Spencer Chapman, M., Williams, N., Dawson, K.J., Mende, N. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022). Number: 7913 Publisher: Nature Publishing Group.
25. Miles, L.A., Bowman, R.L., Merlinsky, T.R., Csete, I.S., Ooi, A.T. et al. Single cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).
26. Corces-Zimmerman, M.R. & Majeti, R. Pre-leukemic evolution of hematopoietic stem cells: the importance of early mutations in leukemogenesis. *Leukemia* **28**, 2276–2282 (2014).
27. Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T. et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature Communications* **11**, 5327 (2020). Number: 1 Publisher: Nature Publishing Group.
28. Ediriwickrema, A., Aleshin, A., Reiter, J.G., Corces, M.R., Köhnke, T. et al. Single-cell mutational profiling enhances the clinical evaluation of AML MRD. *Blood Advances* **4**, 943–952 (2020).
29. Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L. et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Science Translational Medicine* **4**, 149ra118 (2012).
30. Majeti, R. Mutation order in acute myeloid leukemia identifies uncommon patterns of evolution and illuminates phenotypic heterogeneity, PREPRINT (Version 1) (2023).
31. Figueroa, M.E., Wahab, O.A., Lu, C., Ward, P.S., Patel, J. et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer cell* **18**, 553–567 (2010).
32. Gaidzik, V.I., Paschka, P., Späth, D., Habdank, M., Köhne, C.H. et al. 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 (2012).
33. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P. et al. Genomic Classification and Prognosis in Acute

Myeloid Leukemia. *New England Journal of Medicine* **374**, 2209–2221 (2016).
34. Abelson, S., Collord, G., Ng, S.W.K., Weissbrod, O., Cohen, N.M. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400 (2018).
35. Fabre, M.A., de Almeida, J.G., Fiorillo, E., Mitchell, E., Damaskou, A. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022). Number: 7913 Publisher: Nature Publishing Group.
36. Weinstock, J.S., Gopakumar, J., Burugula, B.B., Uddin, M.M., Jahn, N. et al. Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. *Nature* 1–11 (2023). Publisher: Nature Publishing Group.
37. Bolton, K.L., Ptashkin, R.N., Gao, T., Braunstein, L., Devlin, S.M. et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nature Genetics* 1–8 (2020). Publisher: Nature Publishing Group.
38. Desai, M.M. & Fisher, D.S. Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics* **176**, 1759–1798 (2007).
39. Jongen-Lavrencic, M., Grob, T., Hanekamp, D., Kavelaars, F.G., Hinai, A.a. et al. Molecular Minimal Residual Disease in Acute Myeloid Leukemia. *New England Journal of Medicine* (2018).
40. mission bio. 2020, Tapestri Single-Cell DNA Sequencing Requirements Guide, accessed June 2020. Library Catalog: support.missionbio.com.
41. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
42. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114–2120 (2014).
43. Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013).
44. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
45. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
46. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491–498 (2011).
47. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
48. Acute myeloid leukaemia (AML) statistics (2015).

Supplementary Note 1: Experimental Methods

Sample selection.

Bone marrow aspirate samples containing mononuclear cells were collected between 1998-2017 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 harbouring 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 (labelled on Supplementary Figure S1, S2, S3, S4).

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[39]. 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 Supplementary Table 1 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 minutes 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 10min 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 2. 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 harbour 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 (Supplementary Figure S5). 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 catalogue myeloid panel which consists of 312 amplicons (65 kb) that target 45 commonly mutated myeloid genes[40]. 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 $\approx 10-1000$ single cells were well genotyped at all driver mutation positions (Supplementary Figure S1, S2, S3, S4).

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[41,42] and reads were mapped to the hg19 reference genome using the BWA-MEM algorithm[43,44]. Barcodes within a small Hamming or Levenshtein distance from corrupted barcodes were corrected. Genome Analysis Toolkit[45] with a joint calling approach that follows GATK Best Practices recommendations[46,47] 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

Sample number	Thawed count	CD34+ selection	Isolation strategy	Sorted count
7322	1.10E+08	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim	25221
7328	4.88E+07	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim	38995
12565a	1.42E+08	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim	5406
32963	1.65E+08	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim/CD38-/CD45RA-	25049
32747	9.50E+07	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim/CD38-/CD45RA-	2500
6374	5.75E+07	Miltenyi	Live/CD19-/CD33-/CD34+	5600
41095	1.50E+07	EasySep	Live/CD19-/CD33-/CD34+	3000
2225	3.25E+07	EasySep	Live/CD19-/CD33-/CD34+/CD45dim	6700
47668	1.16E+08	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim	19836
2261	2.63E+07	EasySep	Live/CD19-/CD33-/CD34+/CD45dim	8202
4334	5.20E+07	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim	7427
27491	1.08E+08	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim	42810
12565b (replicate)	9.30E+07	Miltenyi	Live/CD19-/CD33-/CD34+/CD45dim/CD38-/CD45RA-	5713
27510	1.15E+07	Miltenyi	Live/CD19-/CD33-/CD34+	68292
26833R	5.70E+07	Miltenyi	Live/CD19-/CD33-/CD34+	7902
25568R	2.00E+07	EasySep	Live/CD19-/CD33-/CD34+	60305
2239	2.50E+07	EasySep	Live/CD19-/CD33-/CD34+	9721

Table 1. PHSC isolation strategies.

Antibody	Dilution	Manufacturer	Catalog #
Mouse monoclonal anti-CD45RA, Alexa 700	1:300	BioLegend	304120
Mouse monoclonal anti-CD33, APC	1:200	BD Biosciences	345800
Mouse monoclonal anti-CD34, APC-Cy7	1:100	BioLegend	343514
Mouse monoclonal anti-Zombie, Aqua (BV510)	1:800	BioLegend	423101
Mouse monoclonal anti-CD45, BV785	1:100	BioLegend	304048
Mouse monoclonal anti-CD19, FITC	1:100	BioLegend	302206
Mouse monoclonal anti-CD90, PE	1:50	BioLegend	328110
Mouse monoclonal anti-CD49f, PE-Cy5	1:100	BD Biosciences	551129
Mouse monoclonal anti-CD38, PE-Cy7	1:100	BioLegend	303516

Table 2. List of all antibodies used for FACS in this study.

our samples due to their low variant allele frequencies even when present (ID 6374 and 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$\approx 14\%$) 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 labelled as either harbouring 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 (Supplementary Figure S1, S2, S3, S4) 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 harbour. 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 harbouring high-confidence driver mutations whose presence cannot be ascertained are labelled 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. The assumption is that these mutations caused the AML transformation event and belong to the blast genotype (e.g. in sample ID12565a, ID2225, 32747, 7328, 7322). 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 blast cells) were considered.

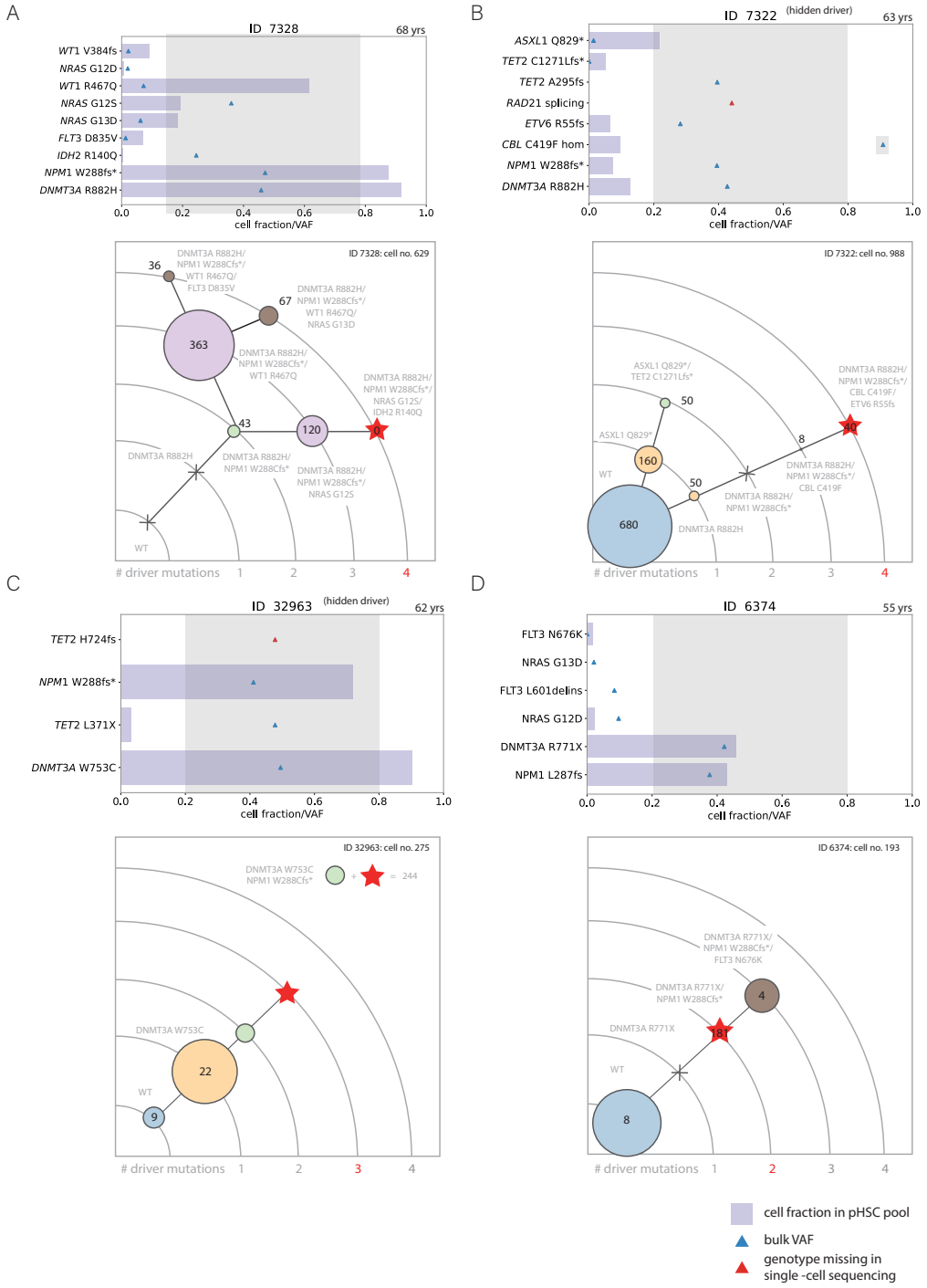

Fig. S1. Single-cell DNA sequencing was used to construct pHSC phylogenetic trees for individual patients (first batch - 1). (A-D) Variant information from bulk sample (triangle) and cell fractions detected in single cell sequencing (purple horizontal bar) were used to construct phylogenetic trees for the preleukemic HSC population in each patient. Variants falling within VAF $0.2-0.8$ (VAF≥ 0.9 for homozygous) were considered as part of the high-confidence driver mutation set and the number of these drivers harboured by a clone was used to assign the clone to corresponding arc levels in the phylogenetic tree. If variants were not covered by the single-cell sequencing panel or not genotyped, the order of acquisition for those mutations cannot be ascertained. These samples are labelled with 'hidden driver'. Only the total clonal sizes summed across the ambiguous clones were reported (upper right corner equations in tree plots). Clonal genotypes and sizes were labelled on the constructed phylogenetic trees of each sample. Sizes of circles are proportional to clonal sizes and extinct clones are indicated with grey crosses.

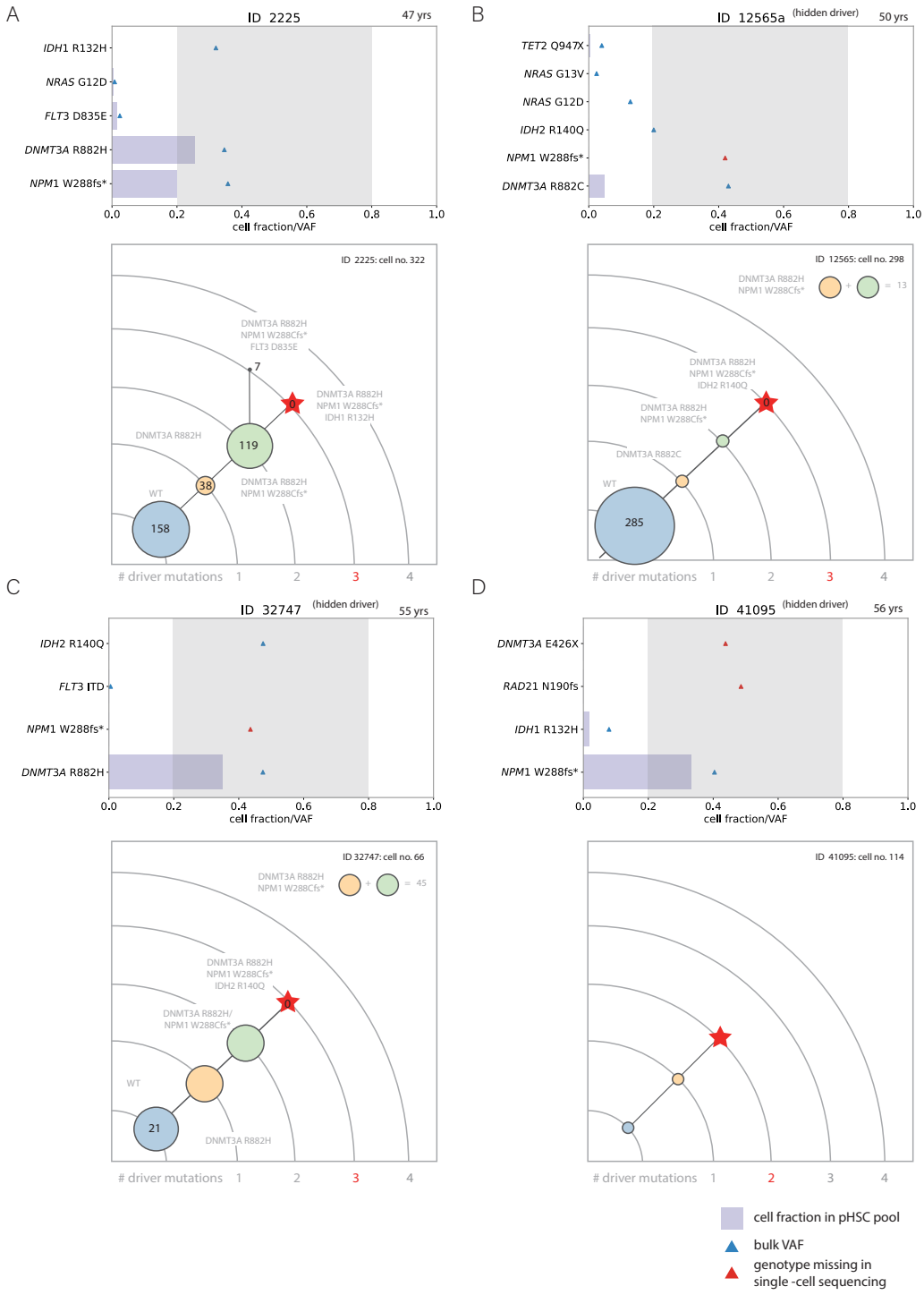

Fig. S2. Single-cell DNA sequencing was used to construct pHSC phylogenetic trees for individual patients (first batch - 2). (A-D) Variant information from bulk sample (triangle) and cell fractions detected in single cell sequencing (purple horizontal bar) were used to construct phylogenetic trees for the preleukemic HSC population in each patient. Variants falling within VAF $0.2-0.8$ (VAF≥ 0.9 for homozygous) were considered as part of the high-confidence driver mutation set and the number of these drivers harboured by a clone was used to assign the clone to corresponding arc levels in the phylogenetic tree. If variants were not covered by the single-cell sequencing panel or not genotyped, the order of acquisition for those mutations cannot be ascertained. These samples are labelled with 'hidden driver'. Only the total clonal sizes summed across the ambiguous clones were reported (upper right corner equations in tree plots). Clonal genotypes and sizes were labelled on the constructed phylogenetic trees of each sample. Sizes of circles are proportional to clonal sizes and extinct clones are indicated with black crosses.

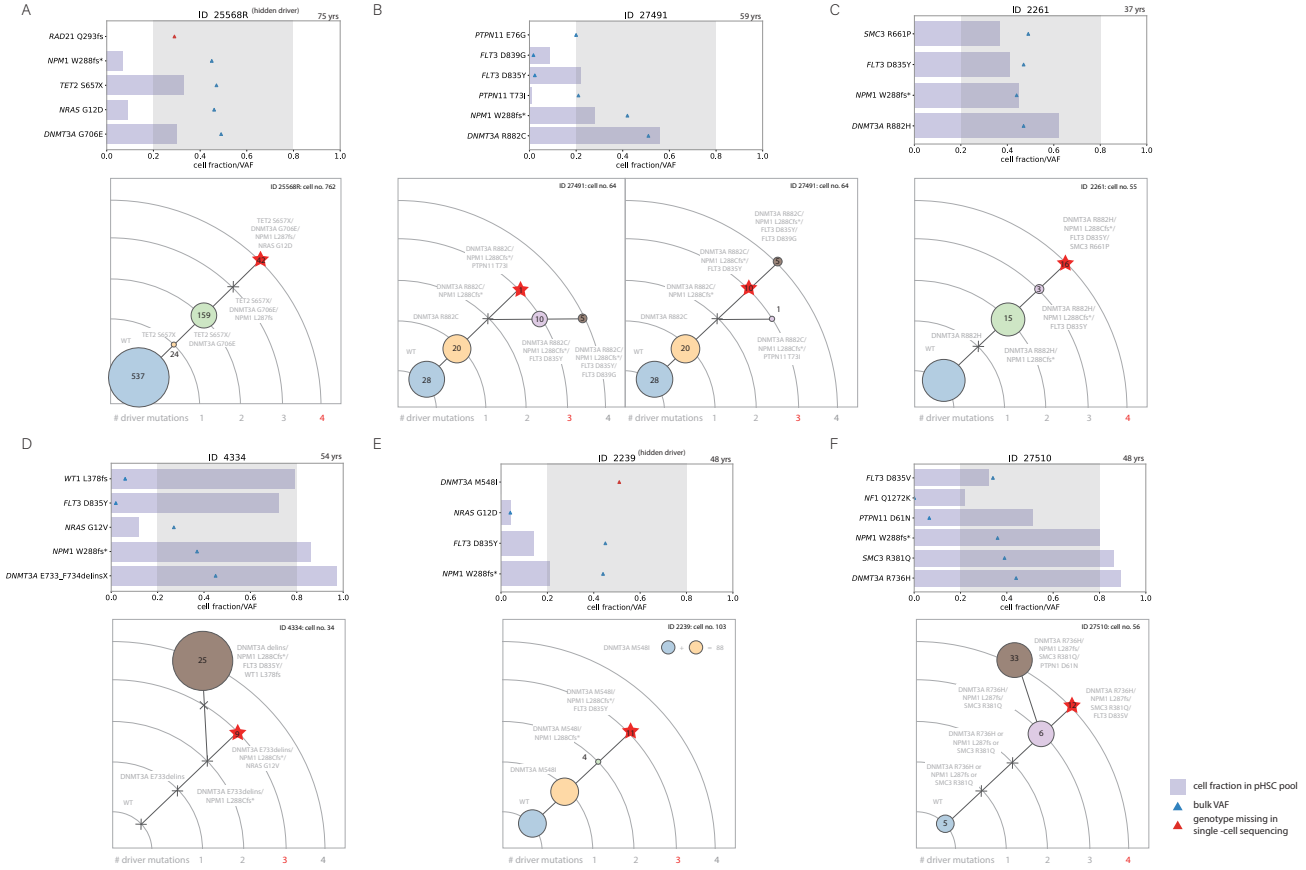

Fig. S3. Single-cell DNA sequencing was used to construct pHSC phylogenetic trees for individual patients (second batch). (A-F) Variant information and constructed phylogenies are shown for six additional samples. For ID27491 there are two possible phylogenetic trees both of which produce similar clone size vectors.

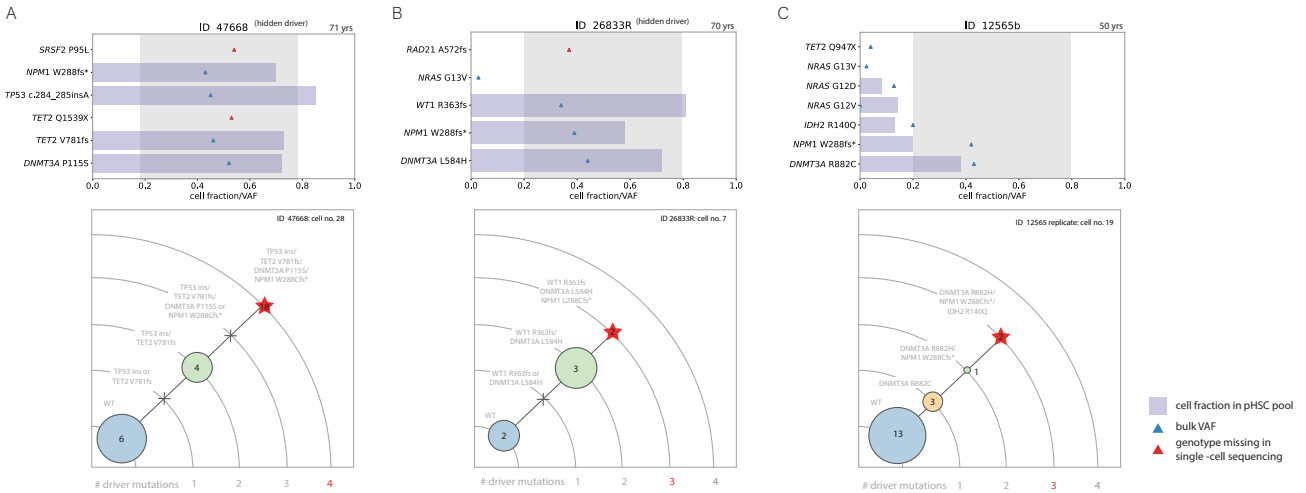

Fig. S4. Single-cell DNA sequencing was used to construct pHSC phylogenetic trees for individual patients (second batch). (A-C) Variant information and constructed phylogenies are shown for three additional samples.

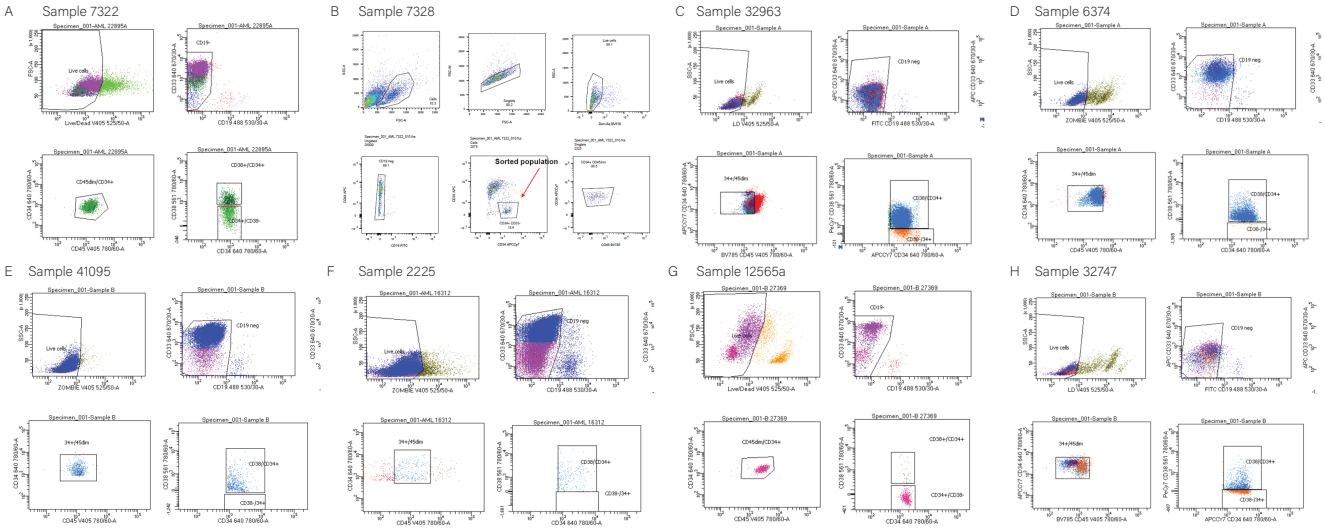

Fig. S5. Fluorescence-activated cell sorting gates (first batch).

Supplementary Note 2: Inferring evolutionary parameters from clone size vectors

Simulating preleukemic evolution using the k-hit staircase model

We simulated the process of mutation acquisition and clonal expansion before AML onset using a multi-type branching process. Stem cells divide symmetrically at rate $\lambda = 1$ per year and acquire driver mutations stochastically at a rate of μ per cell per year. Every driver mutation confers the same selective advantage s, depending on the selection level within the stem cell population, to the lineage and their effects combine additively, i.e. lineages harbouring two driver mutations are twice as fit as lineages harbouring only one driver mutation. This biases cell fate towards symmetric renewal and such cells produce $1 + s$ offspring on average in each generation. Clonal competition was modelled by allowing lineages to grow stochastically at a fitness relative to a growing mean fitness as the number of fit lineages increases. This maintains the total HSC population at a relatively constant size of $N = 10^5$ [11,23]. Stochasticity of mutations and clonal expansions was modelled using Poisson processes.

Cancer onset is defined as the event where any lineage acquires k driver mutations. We recorded 5000 3rd-hit and 1000 4th-hit events for each pair of (μ, s) values in the parameter space that produces cancer onset incidence 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 $k = 3$. Cancer onset rates vary depending on the pair of (μ, s) values used (Supplementary Figure S6). Using higher μ and s values generates a higher number of 'virtual AMLs' for the same number of simulated processes (Supplementary Figure S6). As our analysis of clone size patterns concerns the preleukemic evolutionary process in patients with the *DNMT3A*^{mut}/*NPM1*^{mut} AML subtype, selection of which is influenced by additional factors such as number of driver mutations, we do not expect the predicted cancer event rates to fully corroborate with observed epidemiological rates of AML. We focus on the question: can clone size patterns alone reveal the underlying (μ, s) that had driven the preleukemic process in the patient cohort? To do this we generated equal numbers of 'virtual AMLs' (5000 events) from each pairs (μ, s) so that all pairs of (μ, s) were fairly represented.

The 'virtual AMLs' were then split into training (3335 ($k = 4$) events / 26680 ($k = 3$) events) and testing (1665 ($k = 4$) events/ 13320 ($k = 3$) events) datasets before any subsampling. Subsampling procedures were performed to extract clone size vectors $\boldsymbol{n}$ from the final clonal composition using multinomial subsampling where the probability of a cell being from a certain clone is proportional to its clonal size relative to the total HSC population. Subsampling was performed on the training dataset to create ten different sets of clone size vectors for each subsampling size n (Supplementary Figure S7).

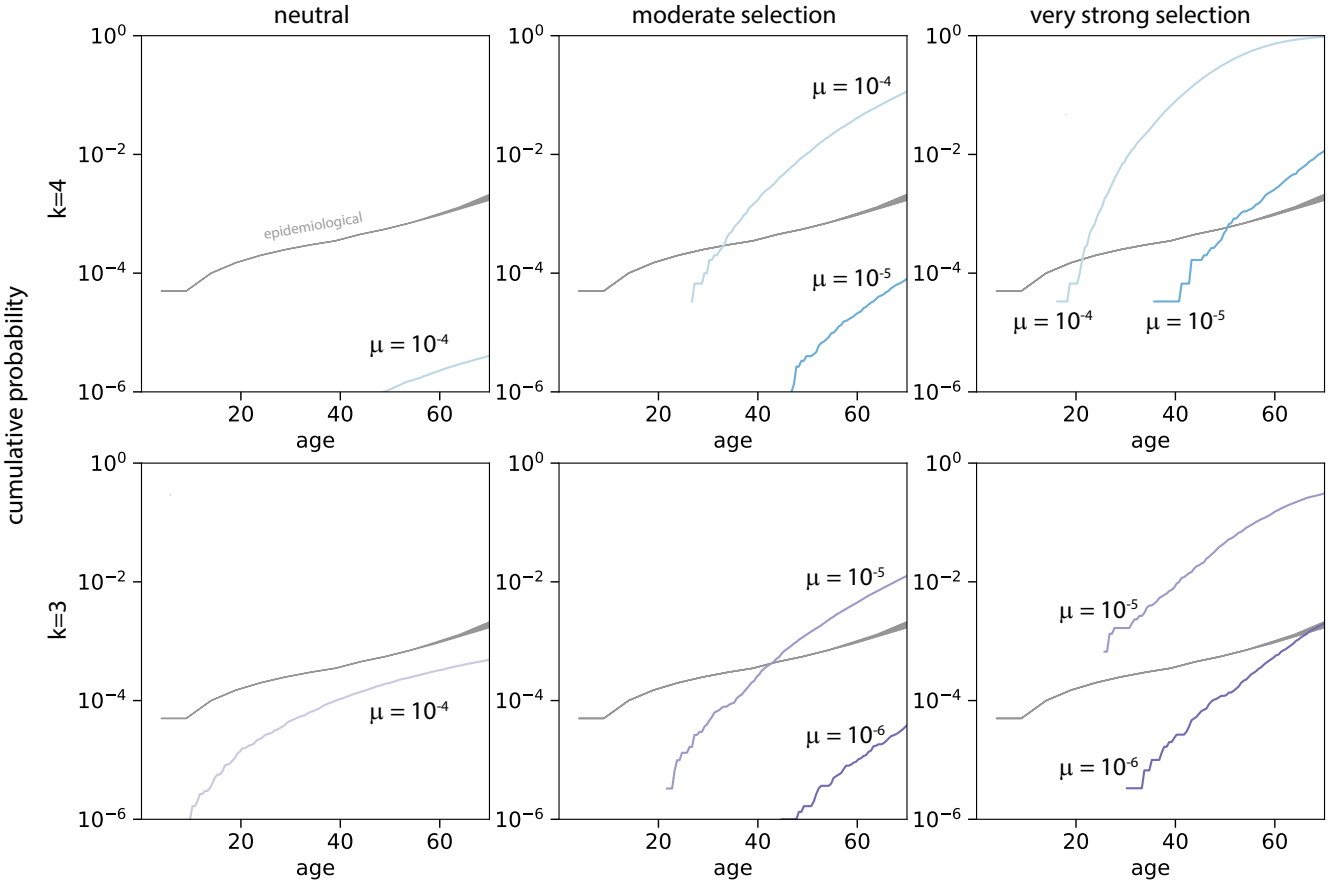

Fig. S6. Cancer onset rates based on k-th staircase model compared to epidemiological rates. The k-th hit staircase model predicts cumulative rates of 'virtual AMLs' across time. For decreasing selection strengths, the corresponding mutation rates needed to produce reasonable cumulative incidence rates increases. Predicted cancer onset rates for $k = 3$ and $k = 4$ are shown alongside observed epidemiological data[48].

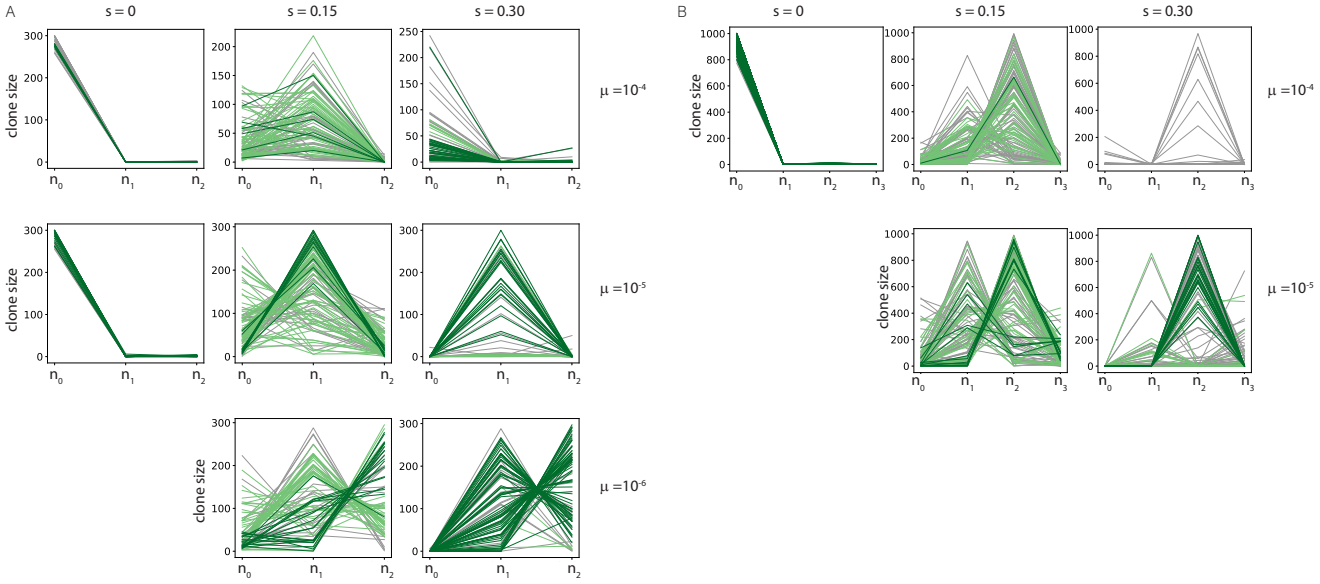

Fig. S7. Visualization of clone size vectors belonging to the training set. (A) One example set of clone size vectors from subsampling 'virtual AMLs' ($k = 3$) down to $n = 300$. **(B)** One example set of clone size vectors from subsampling 'virtual AMLs' ($k = 4$) down to $n = 1000$. Models after training correctly recover the ground-truth i.e. correct (μ, s) with full consensus (10 out of 10) for clone size vectors highlighted in dark green and 80% consensus for those highlighted in light green.

ML model for parameter inference

We trained a fully connected neural network with 4 hidden layers for each subsampled dataset and k to classify a clone size statistics vector $\boldsymbol{n}$ by the parameter regime it was generated from. For each pair of (k, n) values, ten models were trained on ten different sets of clone size vectors extracted from the training set of 'virtual AMLs'. 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 $k = 4$ (Supplementary Figure S8). 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%. For each real patient pHSC tree, we applied 10 models trained on independently sampled datasets to infer its parameter regime based on the closest n to its sample size (Supplementary Figure S9). 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 (Supplementary Figure S10, Figure 4B-L). This ensemble approach provides an estimate of variance for the classifier in its predictions. We considered model predictions for all three-hit trees to estimate the parameter space for three-hit trees and similarly for four-hit trees (Figure 4A).

All eleven unambiguous trees were fed into ensemble classifiers based on the corresponding number of hits k and n closest to sample size (Supplementary Figure S9). The following real trees were classified with the ($n = 1000$, $k = 4$) ensemble classifier: ID7322, ID7328, and ID25568R. Tree ID2225 was 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. Most trees were classified as under moderate selection (9 out of 11 unambiguous trees) where only one tree (ID7322) was classified as neutral evolution and two trees (ID7328 and ID4334) were classified as under very strong positive selection (Supplementary Figure S10).

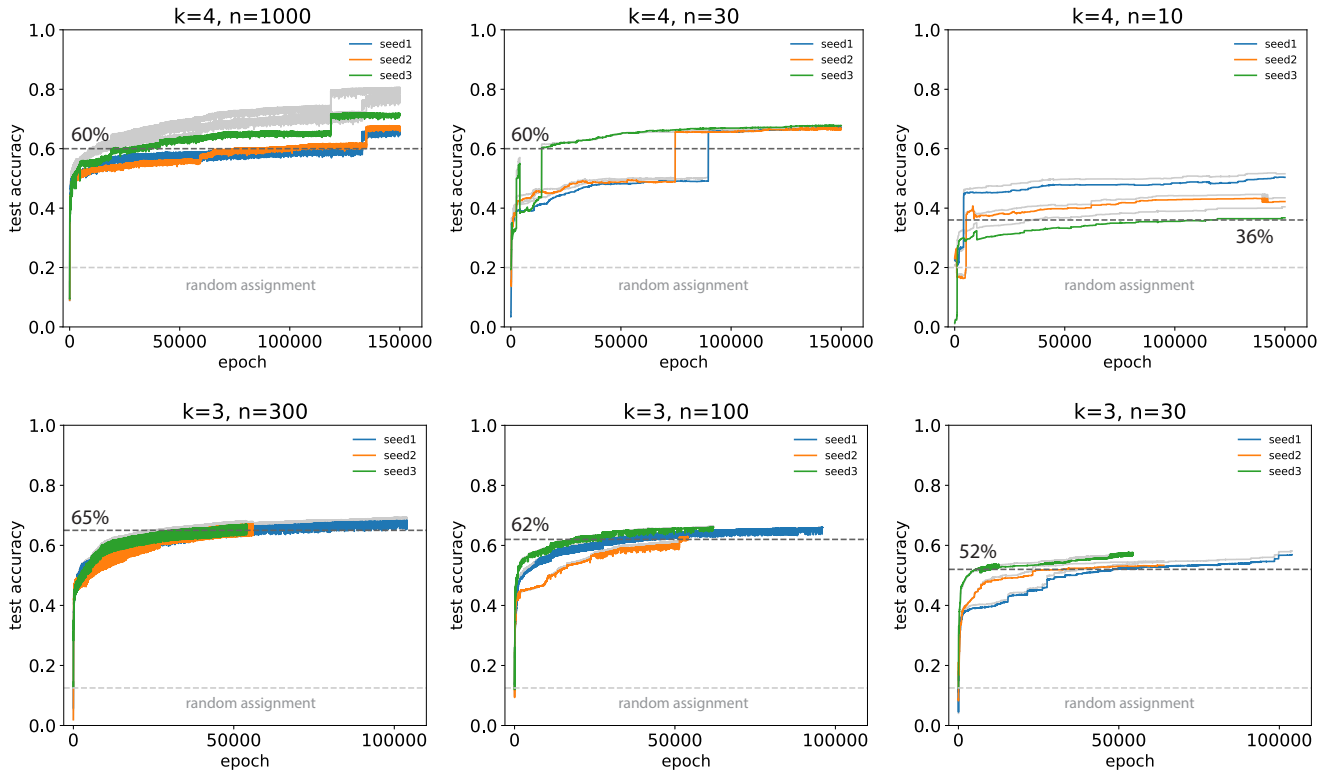

Fig. S8. Typical test accuracies after training. Using one single training/testing set of clone size vectors for each (k, n), we trained models using different random seeds to show that the models typically converge to the labelled test accuracies. All classifier models used to classify real patients pHSC trees had been subsequently trained until they reach test accuracies above or similar to these labelled accuracies. The expected test accuracies based on random assignment are shown for comparison. Training accuracies are also plotted as grey lines for each seed which shows that there is no over-fitting.

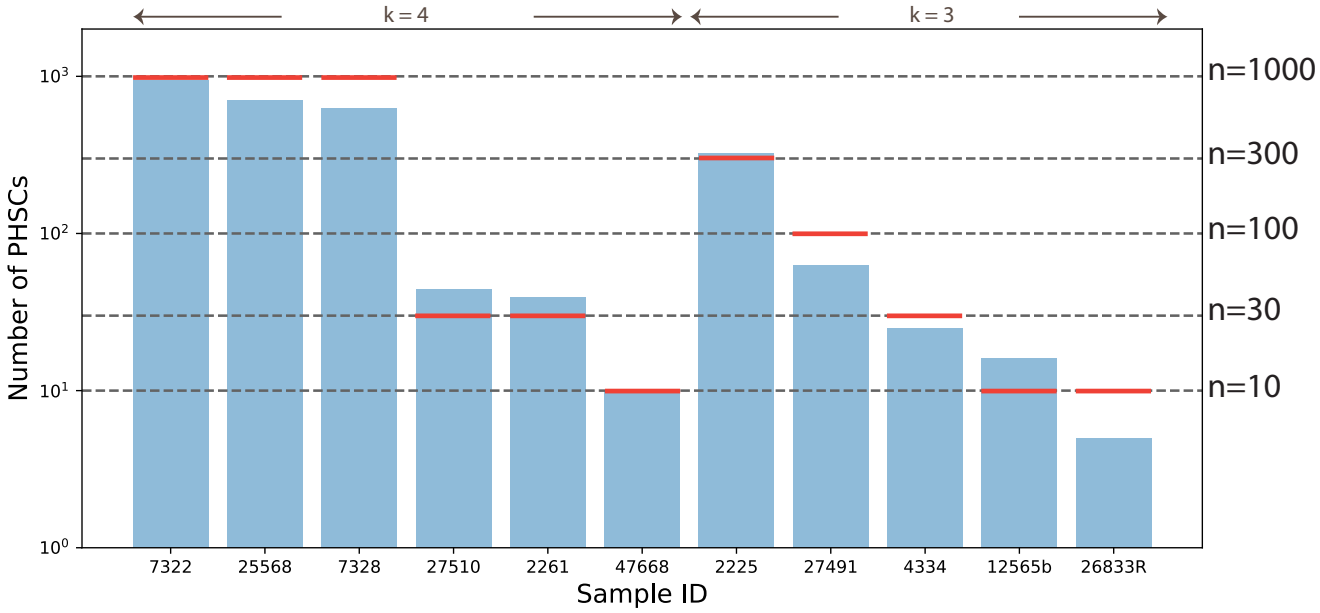

Fig. S9. Choice of n. Each sample was matched to the closest n in logarithmic scale (red line) based on the number of preleukemic HSCs yielded after single-cell DNA sequencing. The number of driver mutations determines k. All eleven samples with unambiguous pHSC trees fed into the machine learning framework are shown.

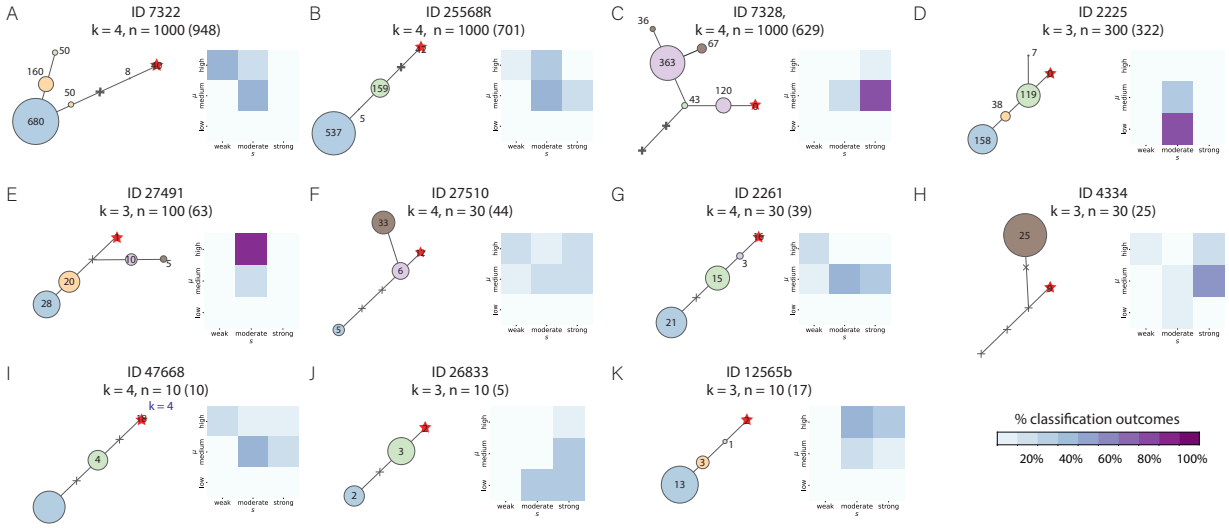

Fig. S10. Classifier predictions for patient preleukemic HSC trees. Preleukemic HSC trees of the 11 *DNMT3A*^{mut}/*NPM1*^{mut} patients are shown alongside classifier predictions if exact genotypes can be delineated for every single cell in the sample i.e. they are unambiguous trees. Clone size vectors extracted from these trees were fed to ensemble classifiers based on number of hits k and n closest to true pHSC number (bracketed number). Most unambiguous trees were classified as under moderate selection (9 out of 11 unambiguous trees) where only one tree (ID7322) was classified as neutral evolution and ID7328, ID4334 were classified as under very strong positive selection.

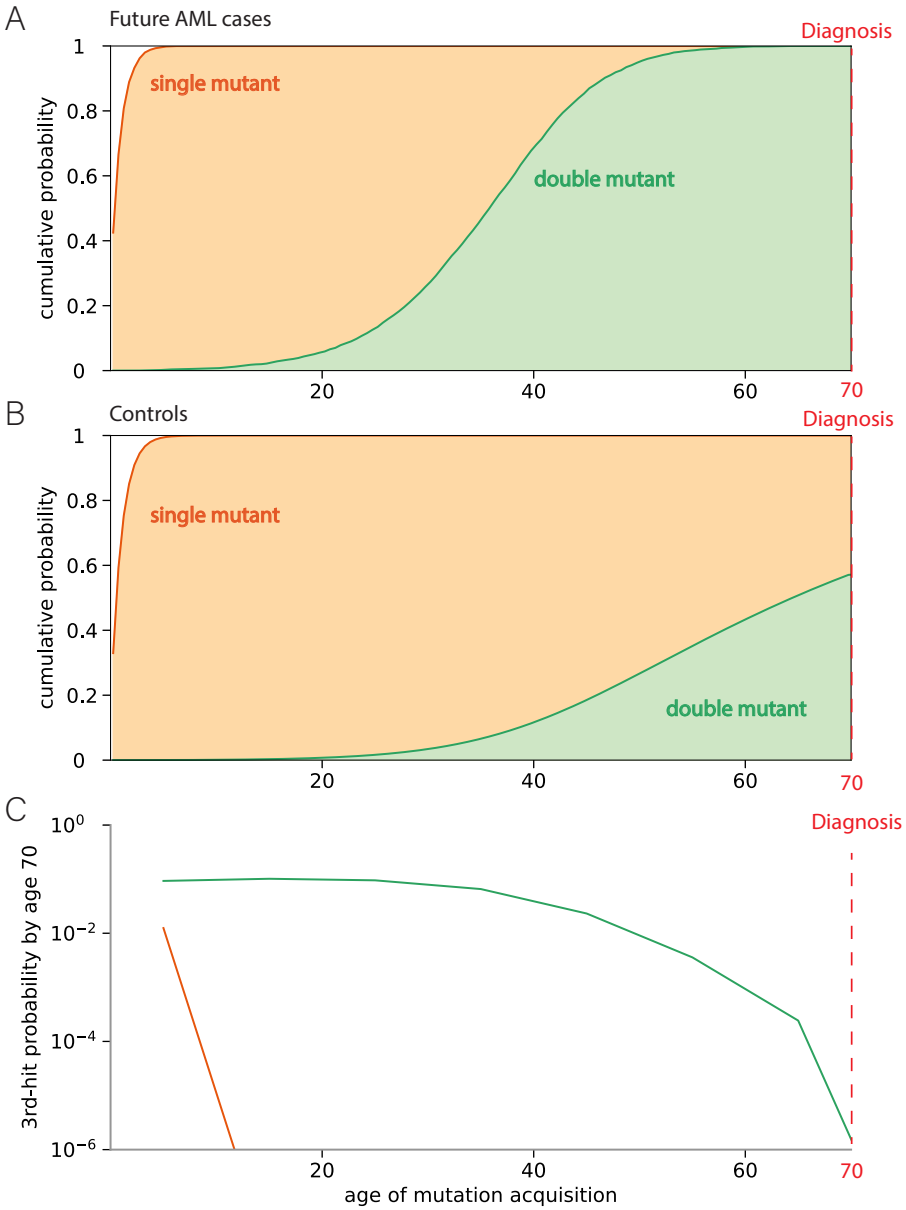

Fig. S11. Early double-mutant clones identify a high-risk group for AML transformation. **(A)** Distribution of mutation acquisition times in simulated individuals (under moderate selection and at medium mutation level) who acquired 3rd-hit by age 70. **(B)** Distribution of mutation acquisition times in unselected simulated individuals, showing marked differences compared to simulated individuals destined to develop AML by age 70. **(C)** The absolute risk of 'virtual AML' (three-hits) transformation before age 70 is dependent on when single (orange) and double (green) mutant clones are acquired.

ML model for parameter inference with finer resolution. To further test the performance of our machine learning framework, we trained models based on 'virtual AMLs' that result from a narrower range of selection strengths. As we found that most real trees were under moderate selection, we chose the range of selection strengths to be $s = 0.09$, $s = 0.15$ and $s = 0.21$ across mutation rate levels: 10^{-4}, 10^{-5},10^{-6}. We performed again in excess of a billion of simulations to generate 5000 'virtual AMLs' for every pair of (μ, s). Clone size vectors were extracted the same way by subsampling to small n and we achieve model accuracies $\approx 50\%$ for $n \geq 1000$ (Figure 3B).

We carried out a replicate experiment using two ampoules from the same patient (ID12565) and found that although exact cell numbers differ the classification outcomes are highly consistent (Supplementary Figure S13).

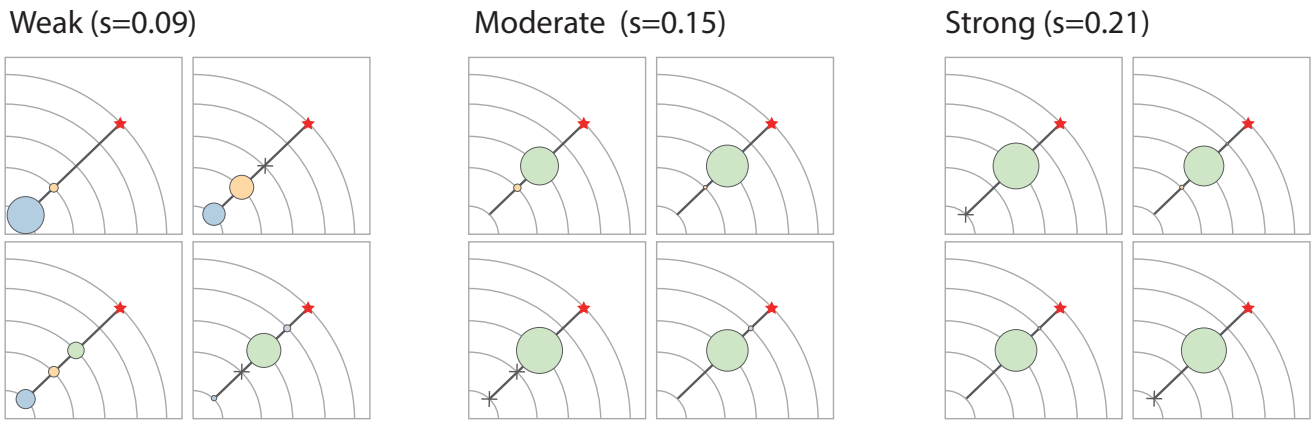

Fig. S12. Tree pattern differences are more subtle across a narrow range of selection strengths. Classifier models were trained on 'virtual AMLs' simulated over a narrower range of selection strengths: $s = 0.09$, $s = 0.15$, $s = 0.21$ where the preleukemic HSC trees show less markedly visible differences.

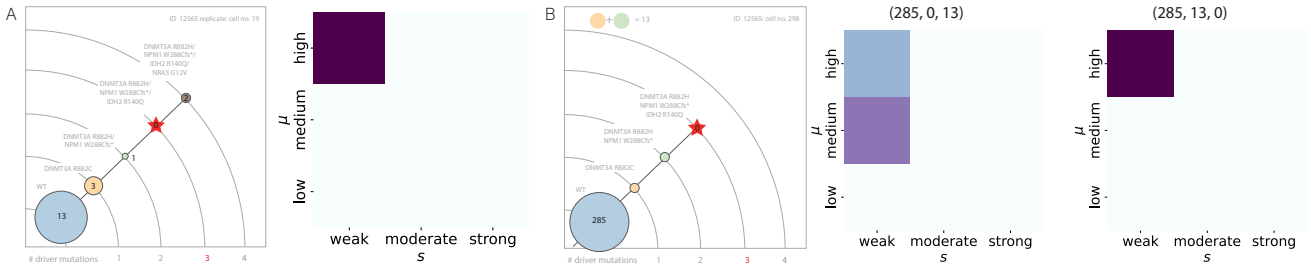

Fig. S13. Sample 12565 and its replicate. Although the exact clonal sizes are different in the pair of technical replicates (ID12565a and ID12565b), the statistical learning model predicts similar outcomes for the level of selection and mutation. **(A)** The ($k = 3$, $n = 30$) classifier predicts weak selection and high mutation rate for the first replicate shown ($\boldsymbol{n} = (13, 3, 1)$). **(B)** The ($k = 3$, $n = 300$) classifier predicts weak selection for the second replicate. As there are uncertainties in the genotype of the single and double mutant clones, the classifier model was fed with two clone size vectors ($\boldsymbol{n} = (285, 0, 13)$ and $\boldsymbol{n} = (285, 13, 0)$) which are at the extreme bounds of possibility. Both clone size vectors were inferred to be under weak selection ($s = 9\%$) albeit at varying levels of mutation rates.