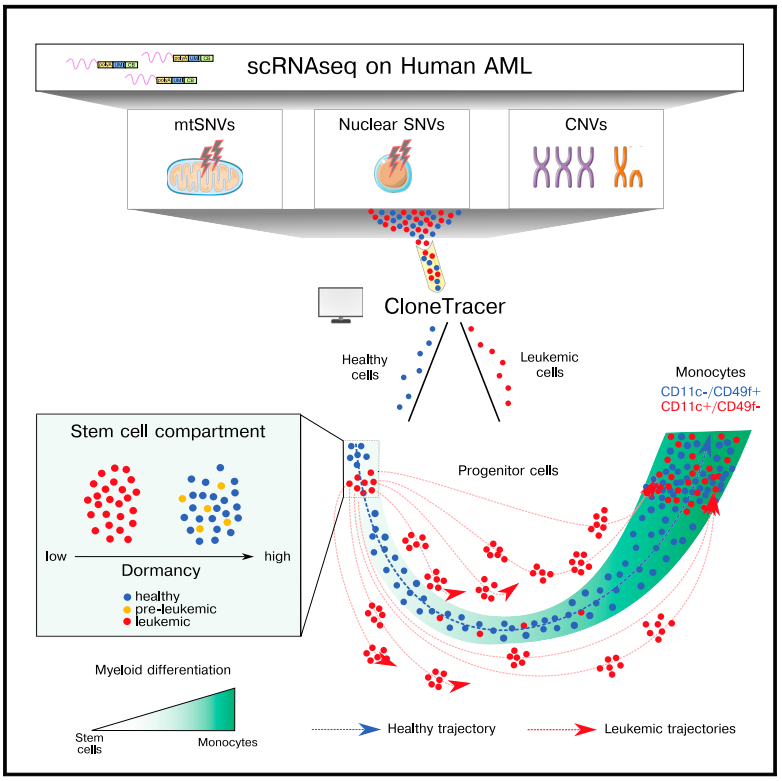
Cell Stem Cell

Resource

Clonally resolved single-cell multi-omics identifies routes of cellular differentiation in acute myeloid leukemia

Graphical abstract



Highlights

- CloneTracer extracts clonal information from single-cell RNA-seq data
- Data resource of healthy and leukemic stem and progenitor cells from 19 AML cases
- Dormant hematopoietic stem cells in AML patients are healthy or preleukemic
- Leukemic stem cells resemble active HSCs but form aberrant myeloid progenitors

Authors

Sergi Beneyto-Calabuig, Anne Kathrin Merbach, Jonas-Alexander Kniffka, ..., Simon Raffel, Carsten Müller-Tidow, Lars Velten

Correspondence

carsten.mueller-tidow@med.uni-heidelberg.de (C.M.-T.), lars.velten@crg.eu (L.V.)

In brief

Velten and colleagues develop CloneTracer, a computational method to identify clones in single-cell RNA-seq data. Applied to immature cells from 19 acute myeloid leukemia patients, CloneTracer shows that dormant hematopoietic stem cells (HSCs) are healthy or preleukemic. Leukemic stem cells resemble healthy active HSCs but give rise to aberrant progenitors.

Beneyto-Calabuig et al., 2023, Cell Stem Cell *30*, 706–721
May 4, 2023
https://doi.org/10.1016/j.stem.2023.04.001

CellPress

CellPress
OPEN ACCESS

Cell Stem Cell

Resource

Clonally resolved single-cell multi-omics identifies routes of cellular differentiation in acute myeloid leukemia

Sergi Beneyto-Calabuig,[1,3,15] Anne Kathrin Merbach,[2,4,15] Jonas-Alexander Kniffka,[2,16] Magdalena Antes,[2,5,6,16] Chelsea Szu-Tu,[1,16] Christian Rohde,[2,4] Alexander Waclawiczek,[5,6] Patrick Stelmach,[2,6] Sarah Gräßle,[7,8,9] Philip Pervan,[1] Maike Janssen,[2,4] Jonathan J.M. Landry,[10] Vladimir Benes,[10] Anna Jauch,[11] Michaela Brough,[11] Marcus Bauer,[12] Birgit Besenbeck,[2] Julia Felden,[2] Sebastian Bäumer,[13] Michael Hundemer,[2] Tim Sauer,[2] Caroline Pabst,[2,4] Claudia Wickenhauser,[12] Linus Angenendt,[13,14] Christoph Schliemann,[13] Andreas Trumpp,[5,6] Simon Haas,[5,6,7,8,9] Michael Scherer,[1] Simon Raffel,[2] Carsten Müller-Tidow,[2,4,*] and Lars Velten[1,3,17,*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
[2]Department of Medicine, Hematology, Oncology and Rheumatology, University Hospital Heidelberg, 69120 Heidelberg, Germany
[3]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[4]Molecular Medicine Partnership Unit, European Molecular Biology Laboratory (EMBL), University of Heidelberg, 69117 Heidelberg, Germany
[5]Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), 69120 Heidelberg, Germany
[6]Division of Stem Cells and Cancer, Deutsches Krebsforschungszentrum (DKFZ) and DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany
[7]Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, 10117 Berlin, Germany
[8]Charité-Universitätsmedizin, 10117 Berlin, Germany
[9]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 10115 Berlin, Germany
[10]Genomics Core Facility, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
[11]Institute of Human Genetics, University of Heidelberg, 69120 Heidelberg, Germany
[12]Institute of Pathology, University Hospital Halle (Saale), Martin-Luther-University Halle-Wittenberg, 06112 Halle, Germany
[13]Department of Medicine A, Hematology and Oncology, University Hospital, Muenster, Germany
[14]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[15]These authors contributed equally
[16]These authors contributed equally
[17]Lead contact
*Correspondence: carsten.mueller-tidow@med.uni-heidelberg.de (C.M.-T.), lars.velten@crg.eu (L.V.)
https://doi.org/10.1016/j.stem.2023.04.001

SUMMARY

Inter-patient variability and the similarity of healthy and leukemic stem cells (LSCs) have impeded the characterization of LSCs in acute myeloid leukemia (AML) and their differentiation landscape. Here, we introduce CloneTracer, a novel method that adds clonal resolution to single-cell RNA-seq datasets. Applied to samples from 19 AML patients, CloneTracer revealed routes of leukemic differentiation. Although residual healthy and preleukemic cells dominated the dormant stem cell compartment, active LSCs resembled their healthy counterpart and retained erythroid capacity. By contrast, downstream myeloid progenitors constituted a highly aberrant, disease-defining compartment: their gene expression and differentiation state affected both the chemotherapy response and leukemia's ability to differentiate into transcriptomically normal monocytes. Finally, we demonstrated the potential of CloneTracer to identify surface markers misregulated specifically in leukemic cells. Taken together, CloneTracer reveals a differentiation landscape that mimics its healthy counterpart and may determine biology and therapy response in AML.

INTRODUCTION

Our understanding of blood formation has fundamentally changed in the last decade. Single-cell RNA sequencing (scRNA-seq)-based studies have demonstrated that hematopoietic stem cells acquire priming early, at phenotypically immature stages.[1] The oligopotent progenitor types that were previously thought to drive hematopoiesis, such as common myeloid progenitors (CMPs), consist of mixtures of fully lineage committed cells.[2] Rather than passing through a CMP stage, lineage differentiation occurs along two major branches, a lymphomyeloid and an erythromyeloid branch.[3,4] These results are supported by various functional assays.[5–7] By contrast, the aberrations that characterize the differentiation landscape in myeloid

CellPress
OPEN ACCESS

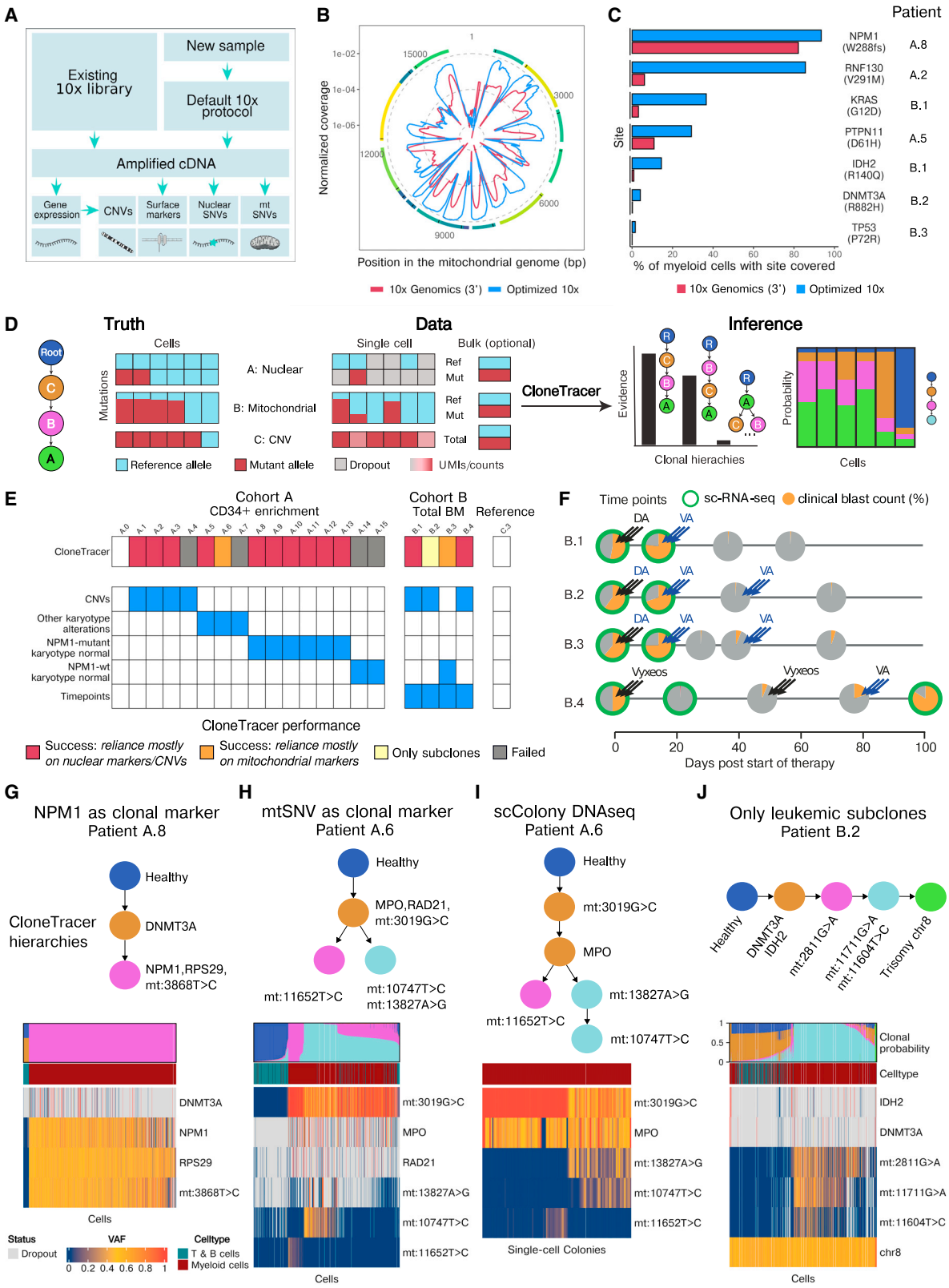


Figure 1. CloneTracer and Optimized 10x enable clonal tracking in droplet-based scRNA-seq

(A) Scheme of Optimized 10x.

(B) Normalized coverage across the mitochondrial genome obtained by default and Optimized 10x.

(legend continued on next page)

malignancies remain unknown. In particular, many diseases were thought to affect, or originate from, CMPs, which do not represent a defined cell type.

Since the healthy and diseased hematopoietic systems co-exist in myeloid malignancies, investigating malignant differentiation landscapes requires clonally resolved scRNA-seq methods. Recent studies have profiled *CALR*-mutant[8] and *JAK2*-mutant[9] myeloproliferative neoplasm, as well as *DNMT3A*-mutant clonal hematopoiesis[10,11] and revealed the expansion of particular differentiation states at the expense of others.[12] However, the shape of the cellular differentiation landscape in full-blown acute myeloid leukemia (AML) remains unknown: Are healthy routes of lineage differentiation co-opted in this disease, or are novel, aberrant cellular identities created? In addition, more specifically, are leukemic stem cells (LSCs) a consistent cell type resembling healthy HSCs, or are they heterogeneous groups of leukemic cells that possess stemness properties? Answering these questions is of key importance to prioritize cellular targets for therapies and identify novel prognostic factors.

Here, we have developed a new strategy to add clonal resolution to high-throughput (droplet-based) scRNA-seq data that robustly work across many of the heterogeneous AML genotypes. Existing approaches use single nucleotide variants (SNVs) or mitochondrial SNVs (mtSNVs) as qualitative markers to identify healthy and malignant cells from scRNA-seq data.[8–10,13,14] However, these measurements are noisy, and methods for quantitative analyses are lacking. Copy-number variants (CNVs) can be inferred from scRNA-seq data[15–17] but are not always present. Our new computational method, CloneTracer, integrates information from SNVs, mtSNVs, and infers CNVs (when present) through a statistical model appropriate for noisy single-cell data. CloneTracer thereby identifies clonal hierarchies and probabilistically assigns single cells to clones.

We applied CloneTracer to bone marrow samples from 19 AML patients. We showed that CloneTracer could unanimously identify most healthy and leukemic cells in 14 of these patients. By integrating data across all patients, we identified a population of HSCs expressing a dormancy gene signature that was dominated by residual healthy and preleukemic cells, as well as leukemic cells resembling active HSCs (active leukemic stem cells [aLSCs]) that often retained erythroid capacity. Downstream of aLSCs, differentiation-blocked, aberrant myeloid progenitors affected chemotherapy responses and fed into qualitatively normal myeloid differentiation. Together, our data established a healthy-like differentiation landscape that may determine biology and therapy response in leukemia.

RESULTS

Improving coverage of nuclear and mitochondrial SNVs in droplet-based scRNA-seq

Conventional droplet-based scRNA-seq protocols exhibit low coverage of nuclear and mitochondrial mutations. We mitigated this issue by splitting the cDNA pool after amplification and constructing sequencing libraries specifically covering RNA expression, surface antigen expression, nuclear SNVs, and mitochondrial genomes (Figure 1A; STAR Methods). In particular, we constructed libraries that unlike default 10x Genomics cover the mitochondrial genome full length, similar to a recent report[14] (Figures 1B and S1A–S1C). In addition, 55%–85% of these libraries mapped to the mitochondrial genome, allowing for a cost-effective deep sequencing of mitochondrial genomes. False positive observations of mitochondrial genetic variants occurred at negligible background rates (Figures S1D–S1G).

For genotyping of nuclear SNVs, we utilized a modified version of TAP-seq (targeted Perturb-seq)[18] with nested, mutation-specific primers (Figures S1H–S1J and https://github.com/veltenlab/CloneTracer/tree/master/primer_design for primer design software that also assists with identifying mutations suitable for amplification). Thereby, coverage on relevant mutations was substantially improved, compared with default 10x Genomics 3′ (Figures 1C and S1K) and similar to results from related approaches.[8,9,19] Mitochondrial and nuclear SNV-targeted sequencing libraries can be constructed from existing (e.g., already sequenced) full-length cDNA libraries from 10x Genomics 3′, making this method ("Optimized 10x") applicable to characterize existing samples in more depth. Sequencing depth requirements, as well as a comparison of long-read[20] and short-read sequencing, are presented in the Methods S1. Libraries included in the final dataset were analyzed with short-read sequencing.

We compared the performance of Optimized 10x with a plate-based RNA-seq protocol (MutaSeq, a modified Smart-Seq2 protocol[21]) and a droplet-based assay for transposase-accessible chromatin with sequencing (ATAC-seq) method focused on tracking mitochondrial mutations (sc-mito-ATAC-seq[22]) (Figures S2A–S2C). Optimized 10x maximized the mutational information available from default 10x Genomics libraries, and, unlike low-throughput high-confidence plate-based methods,[21,23] it displayed the throughput required for ambitious scRNA-seq oncology projects (Figure S2C).

CloneTracer, a statistical model to infer clonal hierarchies and identities from scRNA-seq data

Despite the improved coverage of leukemic point mutations and mtSNVs in Optimized 10x, data from scRNA-seq-based

(C) Coverage of nuclear mutations from various AML patients. Only immature and early myeloid cells are included. See also Figure S1K.
(D) Illustration of the statistical challenge addressed by CloneTracer.
(E) Overview of two AML cohorts, see also Methods S1.
(F) Overview of longitudinal sampling in cohort B. Pie charts indicate clinical blast counts. DA/VA indicate treatment with Daunorubicin/Ara-C and Venetoclax/Azacitidine, respectively.
(G) Top row: inferred clonal hierarchy for patient A.8. Middle row: stacked bar chart illustrating each cell's probability to derive from the different clones shown in top panel. Bottom rows: heatmap depicting the variant allele frequency of all clonal markers in all cells.
(H) Like (G), except that data from patient A.6 is shown.
(I) Clonal hierarchy of patient A.6 identified from sequencing of single-cell-derived colonies, see STAR Methods.
(J) Like (G), except that data from patient B.2 is shown. For CNVs, the scaled number of counts on the specified chromosome are shown.

protocols are inherently noisy, as illustrated by frequent allelic dropout even of highly expressed genes such as *NPM1* (see Figures 1C and S1K). For a confident interpretation of the data and quantitative analyses, statistical methods are needed that identify the most likely hierarchy among the mutations and thereby, for example, clarify whether a mitochondrial mutation or CNV is present in all cancer cells or demarcates a sub-clone. Furthermore, dropout and false positive rates (FPRs) need to be systematically accounted for when assigning cells to (sub-) clones.

We therefore developed CloneTracer, a Bayesian model that identifies the hierarchical relationship between mutations and assigns the cells to the clones. Our model considers previous information, such as allele frequencies from exome sequencing (exome-seq), and most importantly, it accounts for the technical noise associated with single-cell measurements of CNVs, SNVs, and mtSNVs (Figure 1D; see Methods S1 for detail). Thereby, it first compares possible clonal hierarchies. Second, for the mutational hierarchy with the highest evidence, it computes the posterior probability of each cell to belong to any particular clone.

We applied CloneTracer to 19 AML patients from two cohorts (Figures 1E and 1F; Table S3). Cohort A consisted of diagnostic bone marrow samples that were subjected to a CD34 enrichment before single-cell CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing); a median of 2,232 single cells per patient passed stringent quality control filters. Somatic variants were identified *a priori* by ATAC-seq (for mitochondrial variants) and exome-seq (for nuclear variants) of myeloid and T cells (Figure S1G; Tables S3 and S5). Cohort B included paired longitudinal samples from four individual patients at the time of diagnosis, after therapy, and (in one case) at the time of relapse. A median of 12,034 single cells per patient passed quality filters. Somatic variants were identified *a priori* by panel sequencing. In both cohorts, cells were stained with CITE-seq surface antibodies (see Table S1). Overall, we analyzed 88,602 single cells from 25 specimens.

To demonstrate the performance of CloneTracer, we chose to highlight three representative patients (A.8, A.6, and B.2). Detailed analyses of all patients are described in Methods S1.

Patient A.8 represents the performance of CloneTracer in cases with well-covered leukemic mutations:

Here, a mutation in *NPM1* was covered in 94% of the cells. Mutations in RPS29 and a mitochondrial gene co-occurred with this mutation. A preleukemic *DNMT3A* mutation occurred upstream of *NPM1* but displayed a higher dropout rate. CloneTracer confidently assigned cells as part of the leukemic clone if any of the downstream (*NPM1*, *RPS29*, or the mtSNV) mutations were observed (Figure 1G). In their absence, there was often no conclusive evidence if the cell was healthy or preleukemic, due to the dropout of *DNMT3A*. We grouped these cells as 'healthy' in downstream analyses and followed up on preleukemia for selected patients (below, see Figures 4 and S5). Similar results were obtained for 11 further patients with a well-covered mutation (mostly *NPM1* or a CNV) on top of the clonal hierarchy.

Patient A.6 represents the behavior of CloneTracer in cases with moderately covered leukemic mutations and co-occurring mitochondrial markers:

Here, a nuclear SNV with a high allele frequency in bulk exome-seq was located in *MPO*, which was covered in 22.8% of cells. The mutant *MPO* allele was only observed in cells carrying a mitochondrial mutation (3019G>C) (Figure 1H). The mitochondrial mutation was a suitable clonal marker, as it had likely occurred before the nuclear mutation or there were only cells carrying both mutations. To verify these results, we grew single-cell-derived colonies and genotyped *MPO* and the mitochondrial mutations, confirming that the mitochondrial mutation is a high-confidence clonal leukemia marker (Figure 1I). Similar to patient A.6, mitochondrial mutations drove the assignment of leukemic and healthy cells in patient B.3.

In the case of patient B.2, we observed several sub-clonal mitochondrial mutations downstream of the co-occurring *IDH2* and *DNMT3A* mutations that, when occurring together, likely constitute a leukemic, and not preleukemic, event[24] (Figure 1J). Since these genes displayed high dropout and no reliable clonal marker was identified, there was often considerable uncertainty regarding the assignment of healthy vs. leukemic cells. Hence, this patient was excluded from the clonal analysis. In the remaining 4 patients, there were no well-covered clonal markers (i.e., mtSNVs, CNVs, or well-covered SNVs).

Together, these examples illustrate the importance of using statistical models when interpreting single-cell genotyping data. All subsequent analyses involving clonal identities pertain to the 14 patients with high-confidence healthy/leukemic assignments. Importantly, "leukemic" is here defined purely by the presence of a mutation not observed in the T cell lineage (usually *NPM1* or a CNV, and in some cases, a mitochondrial marker mutation) and not by a functional ability to induce leukemia. Recent work has used clonal tracking to demonstrate that not all stem cells carrying a leukemic driver are functionally leukemogenic.[25]

Validation of CloneTracer

Overall, 91% of the cells from the 14 patients could be assigned as healthy or leukemic with high confidence (Figure 2A; STAR Methods). We validated these assignments using established parameters for AML diagnosis: In AML, most myeloid cells are leukemic, whereas lymphoid cells are usually healthy.[26,27] We identify lymphoid and myeloid cells from scRNA-seq data and used this assignment as an indicator for healthy and leukemic, respectively. Under this assumption, the median area under the receiver-operating characteristics curve (AUROC) of CloneTracer was 0.96 (range 0.88–1). The discretized CloneTracer assignments had a median FPR across patients of 9% (range 0.1%–20%) and a median false negative rate (FNR) of 1% (range 0%–20%, Figures 2B and 2C). For patients with SNVs as clonal markers, statistically naive assignments[8,28] that classify a cell as leukemic if at least one mutant allele is observed and otherwise as healthy if at least one healthy allele is observed, sub-optimally balanced between FPR and FNR (Figure 2C).

Notably, not all myeloid cells in AML patients are leukemic. We therefore considered leukemia-associated immunophenotypes (LAIPs) to distinguish leukemic vs. healthy myeloid cells. LAIPs are cell state-specific, aberrantly expressed markers identified during the routine clinical flow cytometric analysis of AML diagnostic samples.[29,30] Three patients carried a significant number of residual healthy cells along the full myeloid differentiation spectrum, as well as a clinically described LAIP. In these

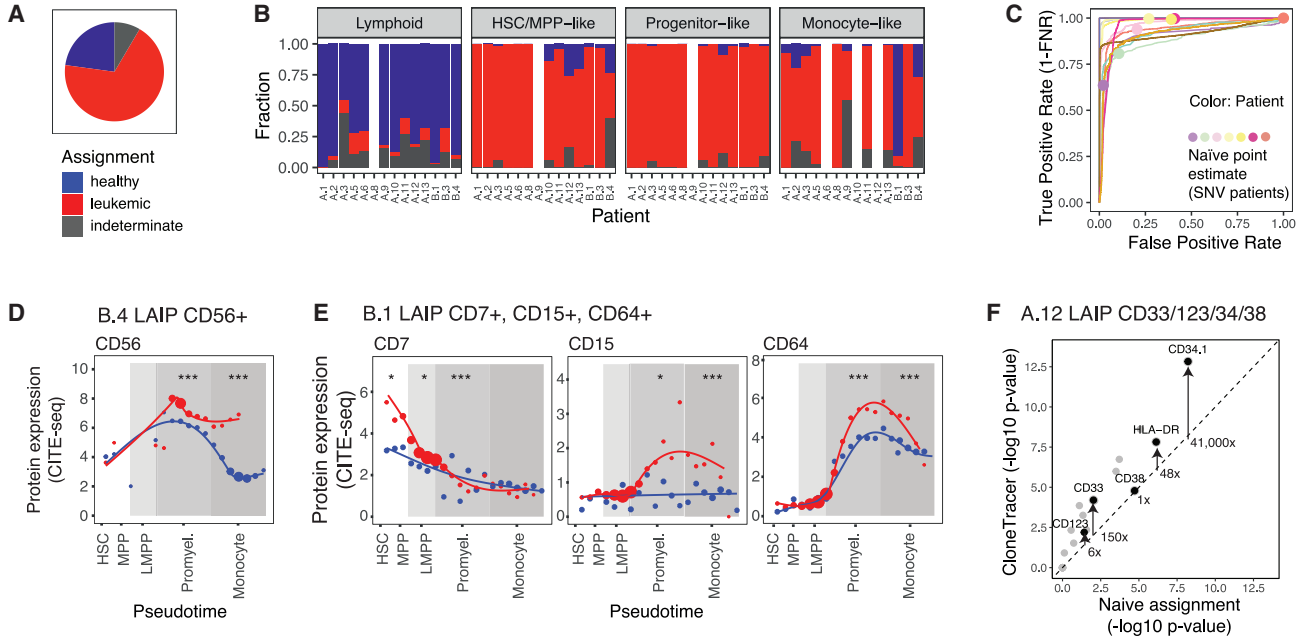


Figure 2. Validation of CloneTracer clonal assignments

(A) Pie chart summarizing the assignment of cells as healthy or leukemic.

(B) Barchart depicting the fraction of cells assigned as healthy (blue), leukemic (red), or indeterminate (gray), stratified by cell type. Cell types with less than 50 cells covered are excluded.

(C) ROC curves computed from CloneTracer leukemia probabilities, assuming that lymphoid cells are healthy and myeloid cells are leukemic. Dots depict statistically naive point estimates for the patients without CNVs.

(D) Gene expression data were projected to a healthy reference,[4] and for myeloid progenitors, a pseudotime was computed. Plots depict smoothened expression of LAIP markers over pseudotime, stratified by the clone. Points indicate mean expression within 20 equally sized bins, and point size indicates number of cells per bin. Asterisks indicate significance of differential expression. *** FDR < 0.001, ** FDR < 0.01, * FDR < 0.1. p values are from a Wilcoxon test of library-size normalized ADT counts.

(E) As in (D), but for patient B.1.

(F) Scatterplot depicting p values from statistical tests comparing surface marker expression between healthy and leukemic immature cells. x axis shows estimates obtained using a statistically naive assignment, and y axis shows estimates obtained using CloneTracer assignments.

patients, we projected each cell to a pseudotime of myeloid differentiation (see also below) and plotted the protein expression of clinically identified LAIP markers over differentiation pseudotime separately for leukemic and healthy cells. As expected, leukemic, but not residual healthy cells, expressed LAIP in a differentiation state-dependent manner (Figures 2D and 2E). For example, CD7 was only expressed by leukemic stem-like cells in patient B.1. The statistical power for identifying LAIP markers was increased by using CloneTracer, compared with the statistically naive assignment (Figure 2F).

Together, these analyses demonstrate that CloneTracer correctly assigned cells as healthy and leukemic and outperformed statistically naive assignments.

Differentiation hierarchies in leukemia

To identify differentiation landscapes, we integrated gene expression data from all 19 patients with data from two healthy individuals (A.0 from this study and C.3 from Triana et al.[4]). The integration strategy was selected to preserve real biological differences between samples (Figure S2D) while accounting for technical batch effects (Figure S2E; see STAR Methods). This analysis showed that cells from the healthy individuals arrange according to the known differentiation trajectories (Figures 3A and S3A, points in color). By contrast, cells from leukemic patients abundantly existed in cell states not observed in healthy patients (Figure 3A, gray points). Many of these cell states were observed in single or few patients, highlighting inter-patient heterogeneity (Figures 3B, 3C, and S3B).

Highlighting CloneTracer assignments on the uMAP showed that most lymphoid cells from AML patients were healthy and myeloid cells were leukemic (Figure 3D; see also Figure 2B). Few lymphoid cells assigned as leukemic likely constituted false positive calls. By contrast, significant numbers of healthy monocytes and healthy HSC/MPP (multipotent progenitor)-like cells occurred in 6 and 5 patients, respectively (Figures 3D and S3B).

We next aimed to identify leukemic and healthy stem cells. We observed a cluster, C6 (see Figure S4A for exact cluster labels), that contained HSCs from the healthy reference individuals (Figure 3A), as well as both healthy and leukemic cells from the different AML patients (Figures 3B and 3D). Interestingly, C6 contained cells from 16 of 19 AML patients, whereas most other progenitor clusters were dominated by cells from only one patient each (Figures 3B and 3C). Other progenitor populations

CellPress
OPEN ACCESS

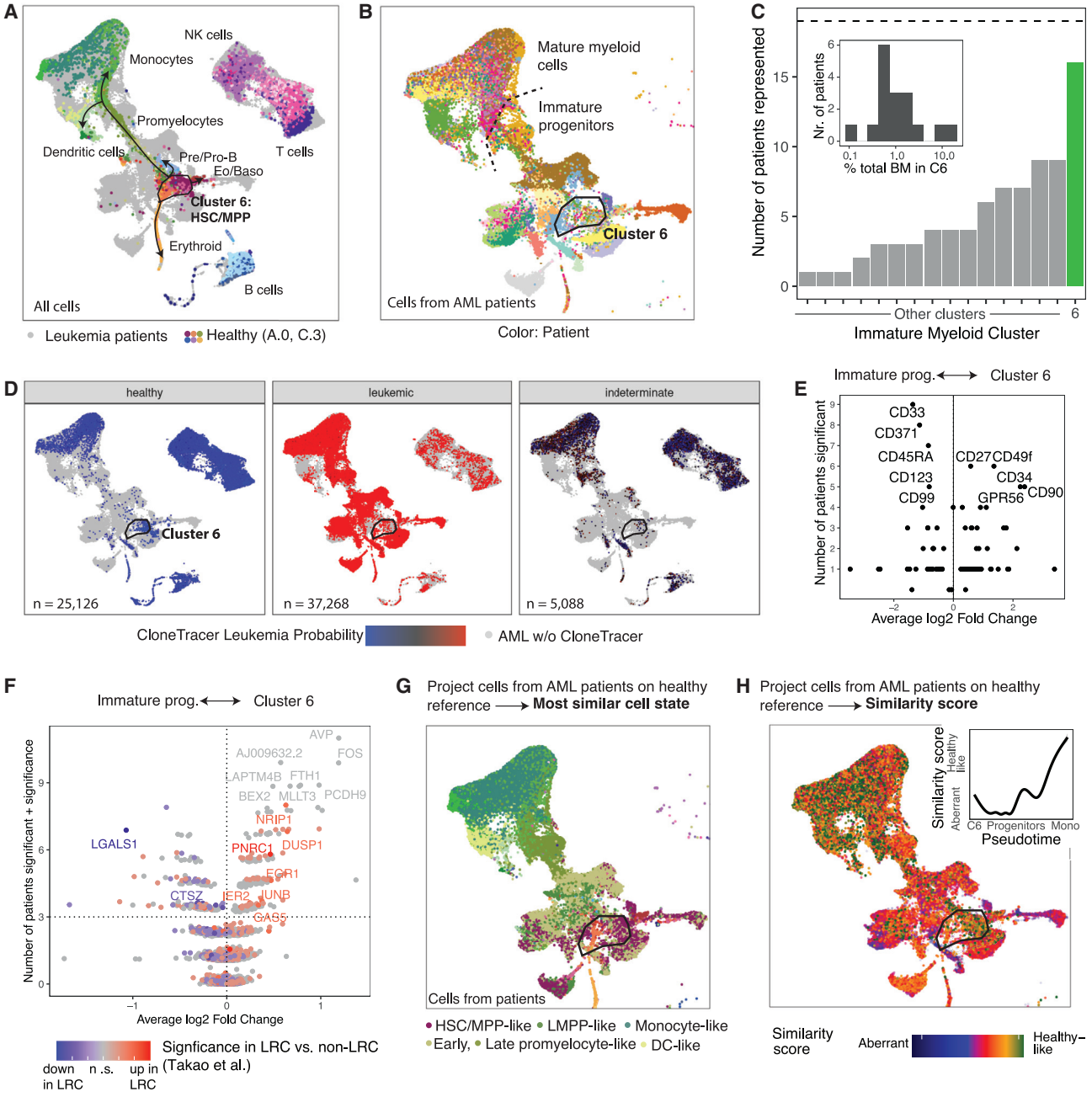


Figure 3. Differentiation landscapes in AML

(A) Uniform manifold approximation and projection (uMAP) depicting integrated data from both cohorts. Color: cell type for cells from healthy individuals (A.0 and C.3[4]), see also Figure S3A. Gray: cells from leukemia patients.

(B) Same uMAP highlighting patient identity.

(C) Bar chart summarizing the number of patients represented in each cluster with at least 5 cells. Inset: histogram depicting the size of C6 as a fraction of total bone marrow. n = 16 patients with cells in C6 are included.

(D) uMAP highlighting CloneTracer leukemia probabilities for 14/19 AML patients. Gray: cells from remaining individuals.

(E) Volcano plot highlighting the number of patients where a given surface marker was significantly ($p < 0.05$) differentially expressed between cells from C6 and other immature myeloid cells from the same patient, vs. the average $\log_2$-fold change across patients. n = 16 patients with C6 represented and n = 2 healthy individuals were analyzed.

(F) Volcano plot as in (E), but for RNA expression. To avoid overplotting, a score from 0 to 1 that depended on the log sum of p values was added to the patient number on the y axis. [30]Color: number of patients where the gene appeared as significant in label-retaining (LRC) vs. non-LRC AML cells.[31]

(legend continued on next page)

appeared to "emerge" from cluster C6, which was more evident in a 3D uMAP (see https://veltenlab.crg.eu/clonetracer/).

Compared with other progenitor cells, cells from cluster C6 tended to express stem cell surface markers (Figure 3E: expressions of CD34, CD90, and CD49f and lower expressions of CD38 and CD45RA). Interestingly, C6 overexpressed genes that were identified as upregulated in label-retaining AML cells (LRCs) during xenotransplant assays, compared with the non-label-retaining fraction[31] (Figure 3F; Table S2). LRCs were characterized as the population responsible for causing disease in xenotransplants and for drug resistance.[31] Across the different AML patients, cluster C6 contained a median of 1% (range 0.1%–17%) of the total bone marrow (Figure 3C, inset), which is higher than the LSC number estimated from xenotransplants.[35]

To evaluate the similarity of cells from C6 to healthy stem cells, we projected all cells to a healthy reference[4] and assigned each cell to the most similar healthy cell state (see STAR Methods; Figure 3G) and a score that quantifies the similarity (Figure 3H). Leukemic cells from cluster C6 were very similar to healthy HSCs, whereas leukemic cells outside of C6 mapped to HSCs or downstream myeloid progenitor states but displayed lower similarity. At the level of monocytes and dendritic cells, the transcriptomic similarity between leukemic cells and healthy cells increased (Figure 3H, inset). We therefore conceptually structure leukemic differentiation in three stages, putative healthy-like stem cells ("C6"), highly heterogeneous and aberrant progenitors, and mature, healthy-like monocytes/dendritic cells. Importantly, most leukemic cells in cluster C6 were marked by mutations in *NPM1* or CNVs, which are typically associated with leukemia and not preleukemia[27] (Figures S3C and S3D).

The dormant stem cell compartment is healthy or preleukemic

To further characterize the putative stem cell cluster C6, we focused on the six patients for whom both healthy and leukemic cells occurred within this cluster. Healthy and leukemic cells in C6 were generally separated by the principal component analysis based on gene expression (Figure 4A). Healthy cells expressed genes characterized as "dormant HSC" genes in (1) a recent scRNA-seq study of highly purified human HSCs,[32] (2) label retention assays in mice,[33] and (3) "low-output HSC" genes identified using clonal tracking[36] (Figures 4B and S4B; Table S2). A dormant HSC gene expression signature robustly separated healthy from the large majority of leukemic C6 cells across patients (Figure 4C); cells expressing the dormant signature were consistently CD34+CD38− (Figure S4C). These results suggest that in AML, the dormant stem cell compartment, where present or observed, contains healthy stem cells (dHSCs). By contrast, the active stem cell compartment was predominantly leukemic: these cells were predominantly healthy only in 2 of the 14 investigated patients (A.9, A.13) (Figure 4C). Of note, five of the seven patients where we observe dormant healthy HSCs belong to the karyotypically normal, *NPM1* mutant subtype (Figure 4C).

To evaluate the potential leukemic and preleukemic content of the dormant stem cell population, we increased the numbers of analyzed cells.

First, we focused on 3 patients with *NPM1* mutations where CD34 expression was rare and nearly exclusive to the dormant stem cell population (Figure 4C: patients A.10, A.11, and A.12). We sorted CD34+ cells and performed MutaSeq,[21] a well-based single-cell method that allows us to efficiently capture mutations in lowly expressed genes such as *DNMT3A*. The sorting strategy resulted in a significant number of cells expressing the dormancy gene signature (Figure 4D). For all of these dormant stem cells, genotype data indicated that they were healthy or preleukemic (Figure 4E).

To follow up on dormant stem cells in two patients where CD34 expression was not exclusive to the dormant stem cell compartment, we sequenced 23,110 additional CD34+ and total BM cells from two patients, A.2 and A.9 (Figure S5A). In the case of A.9, we thereby identified 267 C6 cells, a subset of which expressed the dormancy signature (Figures S5B and S5D). All but two C6 cells here were healthy or preleukemic (i.e., *DNMT3A*-mutant). In the case of A.2, we identified 1,109 cluster C6 cells that were mostly leukemic and lacked the expression of the dormancy score (Figures S5B and S5C). These data further allowed us to demonstrate that the main dataset was sufficiently powered to detect all cell states (Figures S5C and S5D).

Taken together, our data suggest that the dormant stem cell compartment is predominantly healthy or preleukemic. By contrast, the active stem cell compartment was leukemic in 12 of the 14 patients. Our results cannot rule out the existence of rare leukemic dormant stem cells that might be relevant for relapse.

LSCs retain erythroid capacity

We next investigated the leukemic fraction of C6 and its routes of differentiation. In some patients, leukemic cells from C6 expressed "active HSC" genes[32,33] or "high-output HSC" genes[36] relative to healthy cells from C6 (Figures 4B and S4B; Table S2). To identify whether these cells are truly distinct from other leukemic progenitors, we performed differential expression testing, contrasting leukemic C6 cells to other leukemic myeloid progenitor cells from the same patient. Although most leukemic cells expressed genes associated with lymphomyeloid priming, leukemic C6 cells highly expressed genes associated with early erythromyeloid (erythroid, megakaryocytic, and eosinophilic/basophilic) priming,[1] AP-1 transcription factors, and genes associated with LRCs[31] (Figures 5A, S4D, and S4E; Table S2). In line with this observation, in 10 of the 14 patients with confident CloneTracer assignments, we observed erythroid progenitors carrying leukemic mutations (Figure 5B). Typically, these cells carried *NPM1* mutations and/or CNVs and were hence derived from the leukemic clone, and not from a preleukemic clone (Figures S3C and S3D). The abundance of mutation-carrying erythroid progenitor cells correlated with the abundance of C6

(G) Cells from leukemia patients were projected to the healthy reference[4]. Color: most similar healthy cell type for each cell (STAR Methods). See Figure S3A for color legend.

(H) uMAP highlighting the similarity score, i.e., similarity to the 5 most similar healthy reference cells (STAR Methods). Inset: smoothened average of the similarity score over pseudotime.

CellPress
OPEN ACCESS

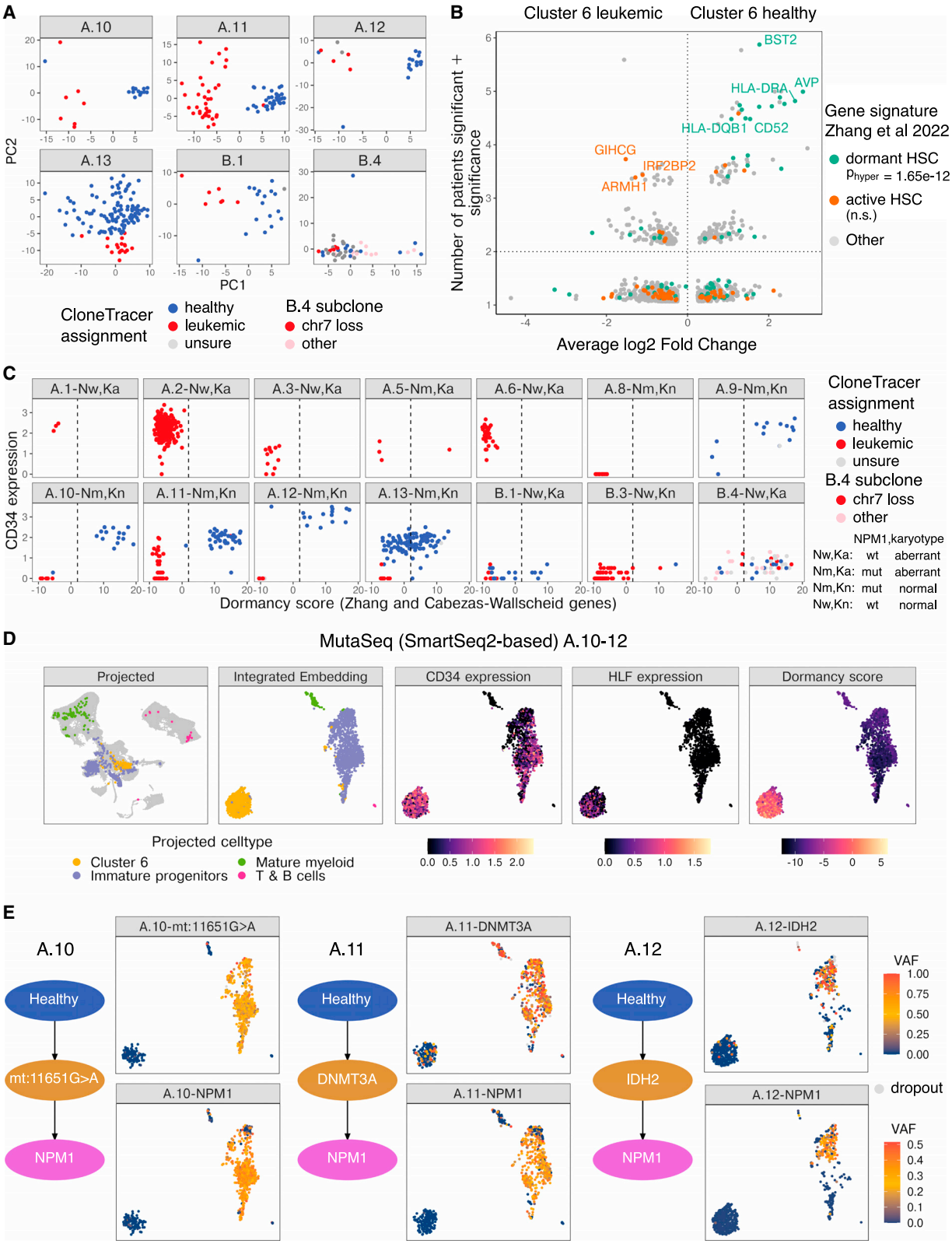


Figure 4. Characterization of LSCs

(A) PCA of RNA expression data of cells from C6 was performed separately for each patient. Score plots from n = 6 patients with both healthy and leukemic cells represented in C6 are shown.

(legend continued on next page)

(Figure 5B). These results indicate that most leukemias, specifically C6, can differentiate into the erythroid lineage at low rates.

Together, these results allowed us to designate leukemic cells from C6 as aLSCs, a rare population of stem-like cells that exists in most AML patients and that often retain erythroid capacity.

The state and extent of the differentiation block determine the phenotypic manifestation of AML

We next focused on the leukemic progenitors downstream of LSCs. To determine the healthy cell state most strongly resembling these cells, their transcriptome was projected onto healthy progenitor cells, ranging from early MPPs to promyelocytes[4] (see also Figures 3G and 3H). We thereby obtained an average pseudotime (i.e., a value describing each cell's progression along the stem cell to monocyte trajectory). The average projected pseudotime was associated with therapy response (Figure 5C; only n = 14 patients treated with anthracycline and cytarabine induction therapies were included here). We found that patients with the most immature leukemic progenitors had blast persistence or died during the first induction therapy, whereas patients with LMPP (lymphomyeloid primed progenitors)-like leukemic progenitors went into complete remission ($p = 0.04$, Wilcoxon test). Although the cohort size underlying these analyses was small, the results are consistent with a recent report studying the relationship between differentiation arrest and survival in a large bulk RNA-seq cohort.[37] The presence or size of the C6 stem cell population was not correlated with chemotherapy response. Taken together, these results suggest that the stage of the differentiation block may play a role in determining chemotherapy response.

We next investigated the ability of the leukemic progenitors to give rise to mature monocytes, which was highly variable between patients. As expected, the genotype (e.g., *NPM1* mutant) only partly explained the degree of monocytic differentiation. We hypothesized that leukemic progenitors with a larger resemblance to their healthy equivalent might display a weaker differentiation block. We computed for each progenitor cell a similarity score, describing how close it resembled the most similar healthy cell (see Figure 3H). We found that this score, after accounting for the genotype, correlated closely with the fraction of mature monocytes or dendritic cells in the bone marrow (Figure 5D). Patients with aberrant progenitor cells had a few mature myeloid cells. By contrast, patients with progenitors more closely resembling their healthy counterparts had large numbers of mature cells. Taken together, these results suggest that the "degree" of the differentiation block, together with leukemia's genotype, determines the fraction of monocytes in the bone marrow.

Of note, we observed patients with mostly healthy monocytes and other patients with mostly leukemic monocytes (see Figure S3B). The fraction of healthy monocytes correlated inversely with the overall number of monocytes (Figure 5E): in leukemia cells with a few monocytes (e.g., FAB [French-American-British AML classification] M0,M1), these monocytes were derived from residual healthy stem cells. An unsupervised analysis of gene expressions revealed that monocytes exist in two cell states. Leukemia-derived monocytes were enriched in a cell state with higher expressions of MHC-II and interferon response genes (*IFITM1* and *IFITM3*) (Figure 5F; Table S2). A similar signature was recently described for monocytes in clonal hematopoiesis.[38]

Together, these results suggest that leukemia-derived monocytes originated from incomplete differentiation blocks at the progenitor level and matured along normal differentiation pathways. The stage of the differentiation block at MPP, LMPP, or promyelocyte stages was linked to the first-line chemotherapy response. Furthermore, the strength of the differentiation block was also encoded at the progenitor level and determined the degree of monocytic differentiation. The stage and the degree of the differentiation block are independent properties.

Our results suggested that the stage of the differentiation block was an important feature of the AML and thus may be subject to clonal selection. To investigate this hypothesis, we focused on three patients with co-existing sub-clones marked by relevant driver mutations (Figure S6). In these patients, sub-clones shifted toward more immature differentiation blocks, compared with the parental clones, possibly because evolutionary pressures may favor differentiation blocks at more immature states. Given the small patient number available for the analyses of sub-clones, we cannot rule out that other properties or genetic drift led to the expansion of the sub-clones.

CloneTracer enables the discovery of leukemia and healthy specific markers

Our results indicated that LSCs, as well as leukemia-derived monocytes, are rather healthy like and difficult to distinguish from their healthy counterparts. This raised the question of whether specific markers can be used to identify healthy vs. leukemic cells of various differentiation stages, including stem cells and monocytes. Such markers can possibly be identified by comparing healthy and leukemic cells from the same patient, thereby avoiding batch effects, genetic background, and other variables typically confounding healthy-cancer comparisons. To investigate this idea, we first used the CITE-seq data of cohort B, since a larger number of surface markers and cells per patient were covered. We asked whether there are markers that are overexpressed or depleted in leukemic cells of various differentiation stages, compared with healthy cells of the same stage. These comparisons identified CD11c

(B) Volcano plot as in Figure 3F, comparing healthy and leukemic cells from C6. n = 6 patients were analyzed, as in (A). Genes are colored by human HSC gene signatures.[32] p_{hyper}: hypergeometric test for enrichment.

(C) PCA of cells from cluster 6 performed jointly across all patients but using exclusively genes from the dHSC signature.[32,33] Cells to the right of the dotted line are putative dormant stem cells.

(D) Rare putatively healthy dormant CD34+ stem cells from three patients were sorted and subjected to a well-based scRNA-seq protocol.[21] uMAP plots, from left to right, show: (i) Projection on original uMAP from Figure 3, (ii) uMAP of Smart-Seq2 data, (iii) CD34 expression, (iv) *HLF*[34] expression, and (v) a dormancy score computed from the gene list in.[32,33]

(E) uMAPs as in (D), highlighting the variant allele frequency of relevant preleukemic and leukemic mutations.

CellPress
OPEN ACCESS

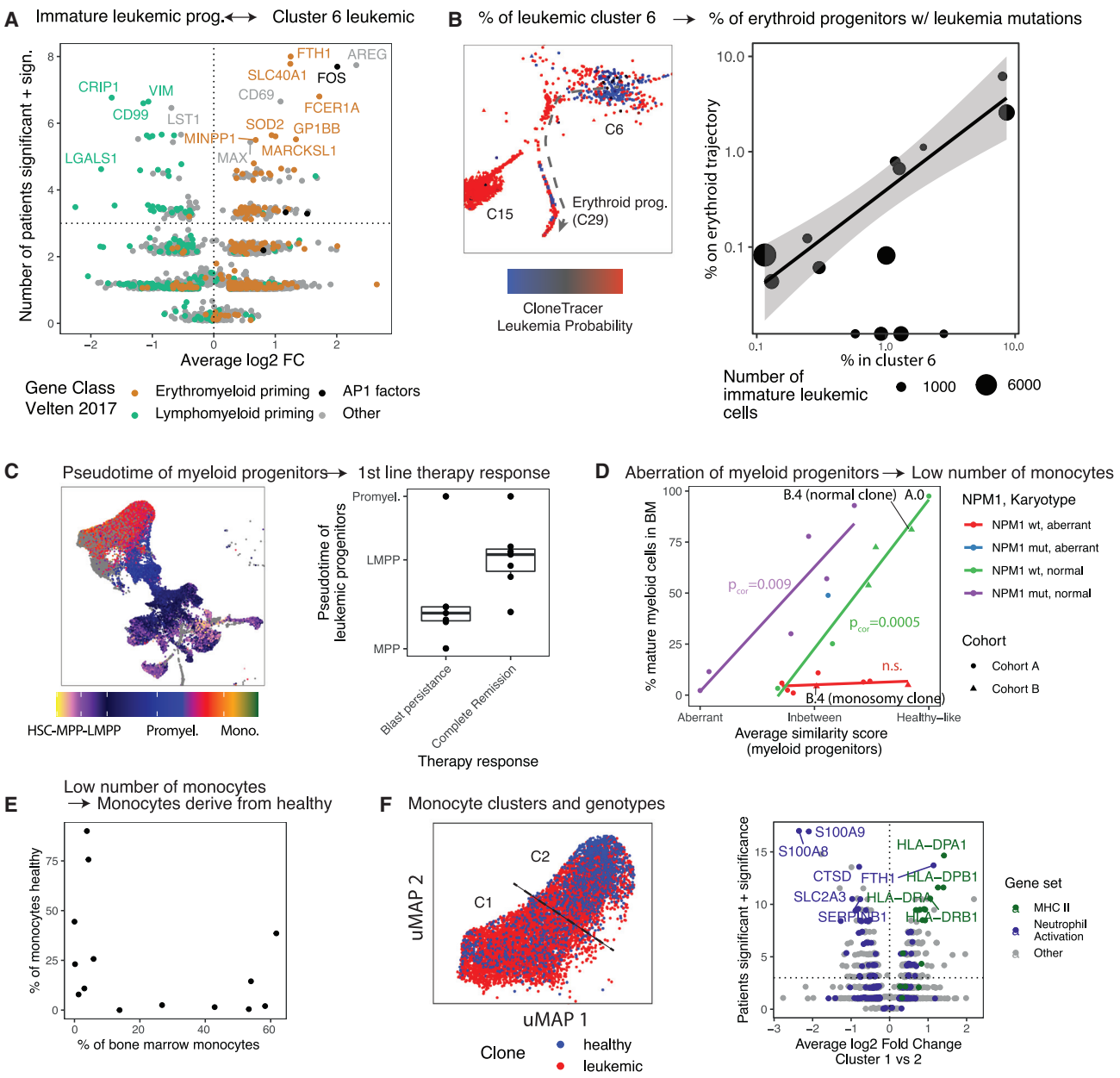


Figure 5. Differentiation pathways downstream of LSCs

(A) Volcano plot as in Figure 3F, comparing leukemic cells from C6 to leukemic immature myeloid cells from the same patient. n = 14 patients were analyzed. Color: priming gene signatures.[1]

(B) Left panel: zoom-in on Figure 3D, displaying only clusters 6, 15, and 29. C15 is a patient-specific cluster displaying aberrant expression of hemoglobins. Right panel: scatterplot depicting the size of cluster 6 and cluster 29 as a fraction of immature leukemic cells.

(C) Left panel: uMAP highlighting myeloid pseudotime obtained from projection to a healthy reference[4]. Right panel: boxplot contrasting the therapy response of each patient with the average pseudotime of the patients' immature leukemic cells.

(D) Scatterplot depicting the fraction of mature myeloid cells in diagnostic bone marrow samples across n=19 leukemic and one healthy individual (A.0) as a function of the average similarity score of the immature myeloid progenitors, see also Figure 3H. Genotype is color coded. For patient B.4, the fraction of mature myeloid cells was computed separately for the two sub-clones.

(E) Scatterplot depicting the fraction of monocytes in total bone marrow (x axis) and the fraction of monocytes that are healthy (y axis) in n = 14 patients.

(F) Left panel: CloneTracer assignments on the uMAP of cluster 1 and 2 (monocytes). Right panel: volcano plot as in Figure 3F, comparing cluster C1 to cluster C2 from the same patient across n = 19 patients and n = 2 healthy individuals.

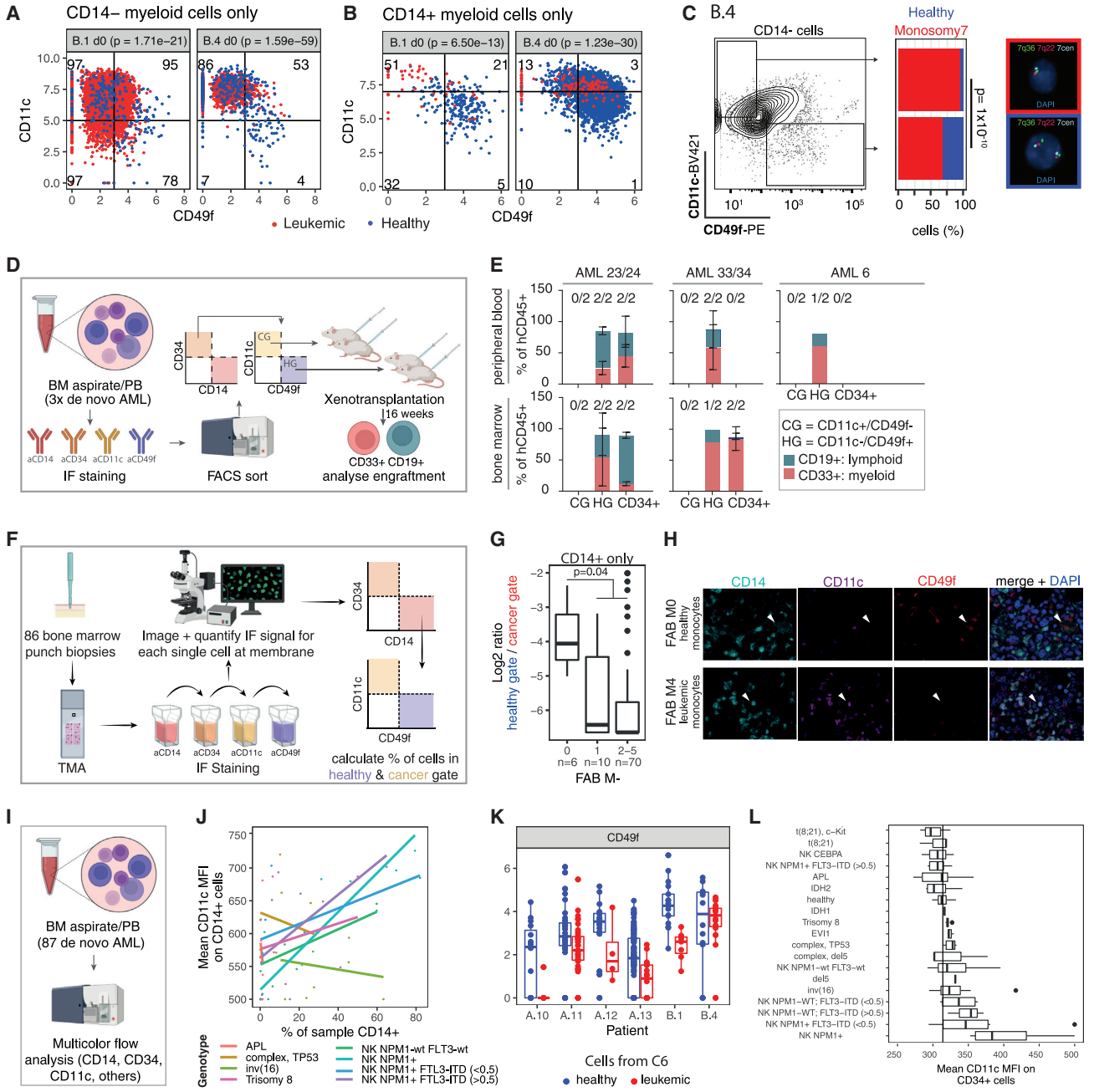


Figure 6. Discovery of leukemia and healthy specific markers

(A) Scatterplots of the expression of CD49f and CD11c highlighting the clonal identity for patients B.1 and B.4. Only CD14- cells are shown. Numbers indicate the percentage of leukemic cells in each of the four quadrants. p values are from a Fisher test for the association of quadrant with clonal identity.

(B) Like (A), except that only CD14+ cells are shown.

(C) Validation of CD11c and CD49f as leukemia/healthy markers by FACS sorting followed by FISH analysis. See STAR Methods. Gates were arbitrarily set to achieve sufficient cell numbers. p values were calculated using a Fisher test. Right panel: representative images show hybridization of FISH probes.

(D) Schematic overview of the xenotransplantation experiments.

(E) Engraftment and lineage potential of CD34+CD14−CD11c+CD49f− (CG), CD34+CD14−CD11c−CD49f+ (HG), and CD34+CD14− peripheral blood and bone marrow cells isolated from three *de novo* AML patients. Numbers indicate the quantity of mice with engraftment versus the total number of mice. Scale bars indicate the standard deviation.

(F) Scheme illustrating the use of TMAs.

(G) TMA data from n = 86 patients. Ratio between the healthy and cancer gate was calculated as a function of FAB classification. p value was calculated using a Wilcoxon test.

(legend continued on next page)

as overexpressed by leukemic cells and CD49f as enriched in healthy cells (Figure S7A). Accordingly, we observed an enrichment of leukemic and healthy cells in the CD11c+CD49f− and CD11c−CD49f+ fraction, respectively (Figures 6A and 6B). Since CD11c expression changed as a function of differentiation (Figure S7B), enrichment analyses of leukemic and healthy cells were performed for the immature (CD14−) and mature (CD14+) compartments separately (Figures 6A and 6B).

We next confirmed the specificity of the CD11c/CD49f marker combination by FACS sorting followed by fluorescent *in situ* hybridization (FISH). In patient B.4, leukemic cells carrying the monosomy 7 were enriched in the CD11c+CD49f− fraction, whereas healthy cells diploid for chromosome 7 were enriched in the CD11c−CD49f+ fraction (Figures 6C and S7C) ($p = 10^{-10}$). Similar enrichments were demonstrated in an independent patient with trisomy 8 (Figure S7D).

To further demonstrate that CD11c and CD49f can be used to enrich for functionally healthy cells, we performed xenotransplantation assays. We sorted CD34+CD14− peripheral blood and bone marrow cells from three *de novo* AML patients into CD11c+CD49f− and CD11c−CD49f+ fractions and transplanted each fraction into two immunocompromised mice. We observed that the putatively healthy CD11c−CD49f+ fractions gave rise to both myeloid and lymphoid engraftment in 8/10 NSG mice from 3 of the 3 patients, indicating healthy hematopoiesis (Figures 6D and 6E). The putatively leukemic CD11c+CD49f− fractions did not engraft (Figures 6D and 6E), as is frequently observed for *de novo* AML samples.[39]

Thus, the marker combination CD11c and CD49f identified CD34+ progenitor populations in AML specimens, which repopulate NSG mice with healthy cells. In the samples studied, the marker combination could not be used to enrich rare LSCs.

Finally, we evaluated these markers in two larger cohorts with different techniques:

We used immunohistochemistry for tissue microarrays (TMAs) from 86 AML patients and analyzed expression of CD14, CD34, CD11c, and CD49f (Figures 6F–6H and S7E). A distinct cohort of 87 AML patients was analyzed by flow cytometry for expression of CD14, CD34, and CD11c (Figures 6I and 6J). Together, these results allowed us to provide a perspective on the specificity of CD11c and CD49f and their potential relevance as markers for healthy vs. leukemic cells. In particular, we observed that in more differentiated leukemias (FAB M2–M5), the putatively leukemia-derived CD14+ cells were predominantly CD11c+CD49f−. In undifferentiated leukemias (M0), which contain only a small number of monocytes, however, residual, putatively healthy CD14+ cells showed a CD11c−CD49f+ phenotype (TMA data, Figures 6F–6H and S7E) and decreased CD11c expression (Figures 6I and 6J). Thus, at the level of monocytes, CD11c and CD49f constituted a robust combination of markers to quantitate the fraction of leukemia content. This might be particularly helpful in AML diagnosis if large numbers of monocytes are present.

In stem cells, the CD11c+/CD49f− combination was informative in a subset of patients. In 5 of the 6 patients analyzed by CloneTracer who contained healthy and leukemic cells in cluster C6, CD49f was more highly expressed by healthy (i.e., dormant) stem cells[40] (Figure 6K). The expression of CD11c on CD34+ cells was variable across and within genotypes (Figures 6L and S7F). Accordingly, data integration by single-cell transcriptomics was overall superior in identifying the multipotent leukemia stem cell cluster, as well as stem-like progenitor cells, compared with flow cytometry.

DISCUSSION

To investigate routes of cellular differentiation in AML, we have introduced CloneTracer, a computational method for adding clonal resolution and identifying leukemic and healthy cells in scRNA-seq data. Tailored to scRNA-seq, CloneTracer extends on DNA-seq-specific error models,[40–42] as well as models that require previous knowledge of the clonal hierarchy.[43] In the AML context, CloneTracer confidently identified healthy and leukemic cells in 14/19 patients. CloneTracer assignments relied on the presence of a clonal CNV (observed in 5 patients), a clonal mutation in a highly expressed nuclear gene (observed in 7 patients), and/or a clonal mitochondrial mutation (observed in 6 patients; full detail for all patients is provided in the Methods S1). By combining the three layers of information, CloneTracer outperformed methods that look at individual layers only.[8,14,17] Previous knowledge of the mutations is required to run CloneTracer, and we recommend calling these mutations from bulk data (e.g., exome sequencing, bulk ATAC-seq for mitochondrial variants, and karyotyping), although mitochondrial variants and CNVs can also be called *de novo* from single-cell data.[14,17]

The availability of clonal information enabled us to clarify routes of leukemic differentiation. Of note, through the integration of data from all patients, we identified a cluster of stem cells that consisted of dormant, healthy, or preleukemic stem cells and predominantly leukemic, active SCs with retained erythroid potential and a gene expression signature resembling label-retaining cells in xenotransplants.[31] Since both dHSCs/dpre (dormant preleukemic cells)-LSCs and aLSCs exhibited relatively consistent gene expression signatures across patients and healthy individuals, data integration of scRNA-seq datasets represented a robust strategy for their identification. By contrast, these cells are difficult to enrich through flow sorting strategies: the large degree of inter-patient heterogeneity at the level of progenitors renders it difficult to develop universal purification schemes.

Downstream of LSCs, we observed a highly heterogeneous and aberrant compartment of immature myeloid cells. The

(H) Representative mid-optical sections of a CD14, CD11c, and CD49f stained tissue micro array used for quantification in (G) (see also Figure S7E for scale bar). Arrows, upper row: CD14+/CD11c−/CD49f+ cell. Arrow, lower row: CD14+/CD11c+/CD49f− cell.
(I) Scheme illustrating the flow cytometry experiment.
(J) Scatterplot relating the CD11c mean fluorescent intensity on CD14+ cells to the fraction of bone marrow that is CD14+, and the genotype. Flow cytometry data from n = 59 individuals of selected genotypes is shown.
(K) Boxplot comparing the expression of CD49f in healthy and leukemic cells from cluster C6.
(L) Bar chart relating the expression of CD11c on CD34+ cells to genotype across n = 87 patients profiled by flow cytometry.

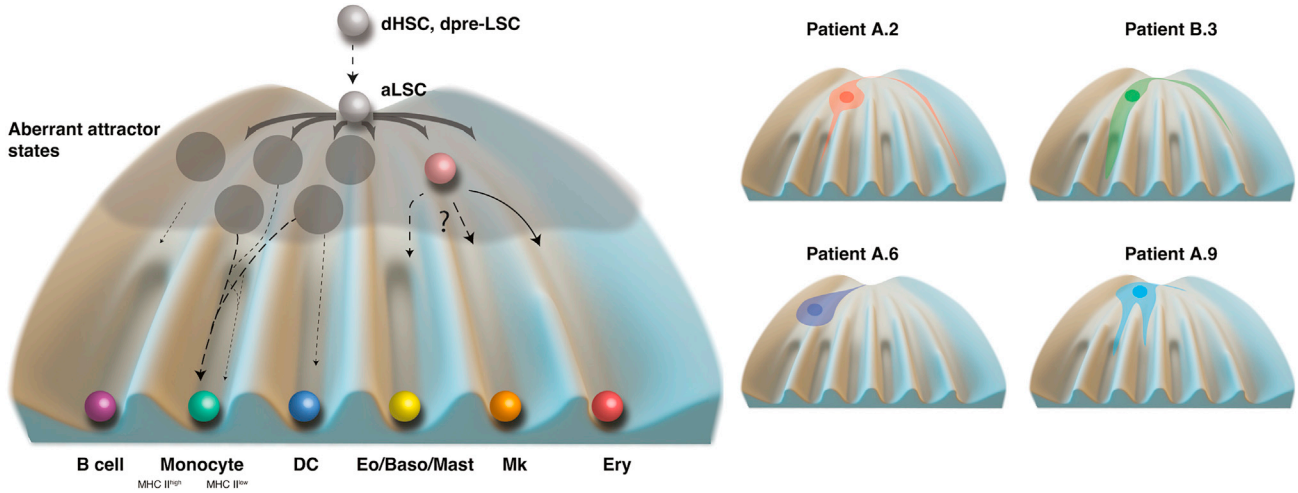


Figure 7. Model of leukemia differentiation
Adapted from a model of healthy hematopoietic differentiation.[1]

patient-specific stage of the differentiation arrest observed in this compartment determined the initial chemotherapy response of the patient. The strength of the block independently determined the overall degree of monocytic differentiation. Unlike cellular hierarchies identified from bulk data,[37] the availability of single-cell resolution allowed us to distinguish stem cells (C6) from various immature, "stem-like" progenitors and pinpoint a poor first-line chemotherapy response to the latter.

Figure 7 summarizes our model of leukemic differentiation pathways and suggests overall similarities to healthy hematopoietic differentiation, but with an aberrant myeloid progenitor compartment. Our model suggests that leukemic mutations, although present in stem cells and monocytes, may prominently exert their effect in the cellular context of progenitors: Seven of the ten most commonly mutated AML driver genes[44] were more strongly expressed in progenitors, compared with stem cells and monocytes (Figure S7G). Differences in expression levels might lead to specific effects of the mutated gene in each cellular compartment.

This implies that AML evolution requires mutations in slowly dividing stem cells, although selection occurs at the level of progenitors. Such a model would be in line with the low number of genetic aberrations observed in most AMLs. Of note, our data are static and cannot exclude the possibility that progenitor cells in AML might also de-differentiate to give rise to stem cells.

In sum, these data may carry implications for the future development of therapeutic strategies: our results indicate that in most if not all AML patients, there is a stem cell compartment distinct from the most immature progenitor cells. Hence, targeted therapies aimed at immature progenitors may increase the initial therapeutic response, but unless these therapies also target the actual stem cell compartment, the effect on relapse and long-term survival might be limited.

Limitations of the study

At the level of the specific single-cell methodology employed for clonal tracking, a limitation is that in droplet-based scRNA-seq protocols, SNVs in lowly expressed genes such as *TET2* are difficult to amplify and large chromosomal inversions or translocations in non-coding regions cannot be mapped. For *DNMT3A*, coverage was obtained in approximately 20% of the CD34+ cells; hence, preleukemic cells were difficult to distinguish from healthy cells with high confidence. Preleukemic cells were therefore followed up on with well-based protocols (Figures 4D and 4E) or simply by sequencing larger numbers of cells (Figure S5). In the future, methods that combine DNA-based genotyping[45] with RNA-seq in droplets might overcome the limitations of CloneTracer but might initially suffer from worse quality of the RNA-seq data.

At the level of the cohort analyzed, a limitation of the study is that statements are drawn from only 19 patients (14 of whom have clonal tracking information). Hence, the results may not be entirely representative of the large heterogeneity of AML genotypes and phenotypes observed, and studies with larger cohorts are going to systematically link genotypes and scRNA-seq phenotypes.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects
  - Animals
- METHOD DETAILS
  - Collection of bone marrow
  - Panel, exome, and bulk ATAC sequencing
  - Antibody-oligo conjugation
  - CITEseq surface labeling and FACS sorting
  - Single-cell RNA sequencing

- Optimized 10x: Mitochondrial libraries
- Optimized 10x: Targeted genotyping libraries
- Plate-based single-cell RNA-seq (MutaSeq)
- Genotyping of single cell derived cultures
- Raw 10x Genomics data processing
- Analysis of single cell gene expression data
- Dormancy score calculation
- Analysis of DNAseq from single-cell derived colonies
- Processing of MutaSeq scRNAseq data
- Raw data processing of MAESTER data
- Fluorescent *In Situ* Hybridization
- Tissue microarrays
- Large cohort flow cytometry analysis
- Xenotransplantations

● QUANTIFICATION AND STATISTICAL ANALYSIS
- CloneTracer model
- Differential expression testing
- Data visualization

● ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.stem.2023.04.001.

ACKNOWLEDGMENTS

We thank Fengbiao Zhou, Anna Mathioudaki, and Judith Zaugg for discussions and Laleh Haghverdi and Valerie Marot-Laussazaie for providing feedback on the mathematical model. We thank all members of GeneCore (EMBL) for assistance with the CITE-seq experiments, the DKFZ Single-Cell Open Lab (scOpenLab) for assistance with the MutaSeq/SmartSeq2 experiment, the FACS facilities of DKFZ, Clinics HD, and CRG/UPF, and the genomics units at CRG and CNAG. We thank the NCT CLB, Sektion Cell Biobanking, for processing and providing bone marrow samples. Figure 1A was created with BioRender.com. Icons used in the graphical abstract are from Servier Medical Art and licensed under CC-BY 3.0. This work was financially supported by the German Bundesministerium für Bildung und Forschung (BMBF) through the Juniorverbund in der Systemmedizin "LeukoSyStem" (FKZ 01ZX1911D to L.V. and S.R.) as well as the Verbundprojekt SMART-CARE (031L0212A to C.M.-T.), the Emerson foundation grant 643577 (to L.V.), grant PID2019-108082GA-I00 and PRE2020-093229 by the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI/FEDER, UE), the German Research Foundation (DFG; projects MU1328/18-1 and MU1328/21-1 and MU1328/23-1 to C.M.-T.), and the German Cancer Aid (DKH; project 70113908 to C.M.-T.). L.V. acknowledges support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme/Generalitat de Catalunya. C.M.-T., A.K.M., and J.-A.K. gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1314-1 FUGG and INST 35/1503-1 FUGG. J.-A.K. acknowledges support of the Deutsche Gesellschaft für Hämatologie und Medizinische Onkologie e.V. (DGHO) and Deutsche José Carreras Leukämie-Stiftung e.V. through the José Carreras-DGHO-Promotionsstipendium.

AUTHOR CONTRIBUTIONS

A.K.M., S.R., C.M.-T., and L.V. conceived the project. C.S.-T., A.K.M., M.A., J.-A.K., M.J., A.W., P.S., and M.B. generated the data and developed the laboratory protocols. S.B.-C., J.-A.K., A.K.M., M.B., P.P., and L.V. analyzed the data with support from M.S. and C.R. S.B.-C. and L.V. developed the statistical model. J.J.-M.L. and V.B. generated and processed raw sequencing data. S.B. advised on antibody-oligo conjugation. A.W. performed xenotransplant experiments. A.J. and M.B. performed FISH. L.V. and C.M.-T. supervised research. All other authors were involved in the acquisition and characterization of clinical specimen. L.V., A.K.M., S.B.-C., and C.M.-T. wrote the manuscript and generated the figures. All authors have read and commented on the manuscript.

DECLARATION OF INTERESTS

The Department of Medicine V (Director C.M.-T.) receives research funding from multiple pharmaceutical and biotech companies especially for clinical trials but also for translational research.

Received: July 29, 2022
Revised: February 5, 2023
Accepted: March 30, 2023
Published: April 24, 2023

REFERENCES

1. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281. https://doi.org/10.1038/ncb3493.
2. Paul, F., Arkin, Y., Giladi, A., Jaitin, D., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell *163*, 1663–1677. https://doi.org/10.1016/j.cell.2015.11.013.
3. Tusi, B.K., Wolock, S.L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J.R., Klein, A.M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. Nature *555*, 54–60. https://doi.org/10.1038/nature25741.
4. Triana, S., Vonficht, D., Jopp-Saile, L., Raffel, S., Lutz, R., Leonce, D., Antes, M., Hernández-Malmierca, P., Ordoñez-Rueda, D., Ramasz, B., et al. (2021). Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. Nat. Immunol. *22*, 1577–1589. https://doi.org/10.1038/s41590-021-01059-0.
5. Perié, L., Duffy, K.R., Kok, L., de Boer, R.J., and Schumacher, T.N. (2015). The branching point in erythro-myeloid differentiation. Cell *163*, 1655–1662. https://doi.org/10.1016/j.cell.2015.11.059.
6. Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212–216. https://doi.org/10.1038/nature25168.
7. Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science *351*, aab2116. https://doi.org/10.1126/science.aab2116.
8. Nam, A.S., Kim, K.T., Chaligne, R., Izzo, F., Ang, C., Taylor, J., Myers, R.M., Abu-Zeinah, G., Brand, R., Omans, N.D., et al. (2019). Somatic mutations and cell identity linked by Genotyping of transcriptomes. Nature *571*, 355–360. https://doi.org/10.1038/s41586-019-1367-0.
9. van Egeren, D., Escabi, J., Nguyen, M., Liu, S., Reilly, C.R., Patel, S., Kamaz, B., Kalyva, M., DeAngelo, D.J., Galinsky, I., et al. (2021). Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. Cell Stem Cell *28*, 514–523.e9. https://doi.org/10.1016/j.stem.2021.02.001.
10. Nam, A.S., Dusaj, N., Izzo, F., Murali, R., Mouhieddine, T.H., Myers, R.M., Sotelo, J., Benbarche, S., Gaiti, F., Tahri, S., et al. (2022). Single-cell multi-omics in human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. https://doi.org/10.1101/2022.01.14.476225.
11. Izzo, F., Lee, S.C., Poran, A., Chaligne, R., Gaiti, F., Gross, B., Murali, R.R., Deochand, S.D., Ang, C., Jones, P.W., et al. (2020). DNA methylation disruption reshapes the hematopoietic differentiation landscape. Nat. Genet. *52*, 378–387. https://doi.org/10.1038/s41588-020-0595-4.

12. Nam, A.S., Chaligne, R., and Landau, D.A. (2021). Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat. Rev. Genet. *22*, 3–18. https://doi.org/10.1038/s41576-020-0265-5.

13. Petti, A.A., Williams, S.R., Miller, C.A., Fiddes, I.T., Srivatsan, S.N., Chen, D.Y., Fronick, C.C., Fulton, R.S., Church, D.M., and Ley, T.J. (2019). A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. Nat. Commun. *10*, 3660. https://doi.org/10.1038/s41467-019-11591-1.

14. Miller, T.E., Lareau, C.A., Verga, J.A., DePasquale, E.A.K., Liu, V., Ssozi, D., Sandor, K., Yin, Y., Ludwig, L.S., el Farran, C.A., et al. (2022). Mitochondrial variant enrichment from high-throughput single-cell RNA sequencing resolves clonal populations. Nat. Biotechnol. *40*, 1030–1034. https://doi.org/10.1038/s41587-022-01210-8.

15. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401. https://doi.org/10.1126/science.1254257.

16. Gao, R., Bai, S., Henderson, Y.C., Lin, Y., Schalck, A., Yan, Y., Kumar, T., Hu, M., Sei, E., Davis, A., et al. (2021). Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat. Biotechnol. *39*, 599–608. https://doi.org/10.1038/s41587-020-00795-2.

17. Gao, T., Soldatov, R., Sarkar, H., Kurkiewicz, A., Biederstedt, E., Loh, P.R., and Kharchenko, P.V. (2023). Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. Nat. Biotechnol. *41*, 417–426. https://doi.org/10.1038/s41587-022-01468-y.

18. Schraivogel, D., Gschwind, A.R., Milbank, J.H., Leonce, D.R., Jakob, P., Mathur, L., Korbel, J.O., Merten, C.A., Velten, L., and Steinmetz, L.M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat. Methods *17*, 629–635. https://doi.org/10.1038/s41592-020-0837-5.

19. Gohl, D.M., Magli, A., Garbe, J., Becker, A., Johnson, D.M., Anderson, S., Auch, B., Billstein, B., Froehling, E., McDevitt, S.L., et al. (2019). Measuring sequencer size bias using REcount: a novel method for highly accurate Illumina sequencing-based quantification. Genome Biol. *20*, 85. https://doi.org/10.1186/s13059-019-1691-6.

20. Lebrigand, K., Magnone, V., Barbry, P., and Waldmann, R. (2020). High throughput error corrected nanopore single cell transcriptome sequencing. Nat. Commun. *11*, 4025. https://doi.org/10.1038/s41467-020-17800-6.

21. Velten, L., Story, B.A., Hernández-Malmierca, P., Raffel, S., Leonce, D.R., Milbank, J., Paulsen, M., Demir, A., Szu-Tu, C., Frömel, R., et al. (2021). Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics. Nat. Commun. *12*, 1366. https://doi.org/10.1038/s41467-021-21650-1.

22. Lareau, C.A., Ludwig, L.S., Muus, C., Gohil, S.H., Zhao, T., Chiang, Z., Pelka, K., Verboon, J.M., Luo, W., Christian, E., et al. (2021). Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. Nat. Biotechnol. *39*, 451–461. https://doi.org/10.1038/s41587-020-0645-6.

23. Rodriguez-Meira, A., Buck, G., Clark, S.A., Povinelli, B.J., Alcolea, V., Louka, E., McGowan, S., Hamblin, A., Sousos, N., Barkas, N., et al. (2019). Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol. Cell *73*, 1292–1305.e8. https://doi.org/10.1016/j.molcel.2019.01.009.

24. Zhang, X., Wang, X., Wang, X.Q.D., Su, J., Putluri, N., Zhou, T., Qu, Y., Jeong, M., Guzman, A., Rosas, C., et al. (2020). Dnmt3a loss and Idh2 neomorphic mutations mutually potentiate malignant hematopoiesis. Blood *135*, 845–856. https://doi.org/10.1182/blood.2019003330.

25. Fennell, K.A., Vassiliadis, D., Lam, E.Y.N., Martelotto, L.G., Balic, J.J., Hollizeck, S., Weber, T.S., Semple, T., Wang, Q., Miles, D.C., et al. (2022). Non-genetic determinants of malignant clonal fitness at single-cell resolution. Nature *601*, 125–131. https://doi.org/10.1038/s41586-021-04206-7.

26. Pelcovits, A., and Niroula, R. (2020). Acute myeloid leukemia: a review. R. I. Med. J. (2013) *103*, 38–40.

27. Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., Kennedy, J.A., Schimmer, A.D., Schuh, A.C., Yee, K.W., et al. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. Nature *506*, 328–333. https://doi.org/10.1038/nature13038.

28. van Galen, P., Hovestadt, V., Wadsworth, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. Cell *176*, 1265–1281.e24. https://doi.org/10.1016/j.cell.2019.01.031.

29. Feller, N., van der Velden, V.H.J., Brooimans, R.A., Boeckx, N., Preijers, F., Kelder, A., de Greef, I., Westra, G., te Marvelde, J.G., Aerts, P., et al. (2013). Defining consensus leukemia-associated immunophenotypes for detection of minimal residual disease in acute myeloid leukemia in a multicenter setting. Blood Cancer J. *3*, e129. https://doi.org/10.1038/bcj.2013.27.

30. Al-Mawali, A., Gillis, D., Hissaria, P., and Lewis, I. (2008). Incidence, sensitivity, and specificity of leukemia-associated phenotypes in acute myeloid leukemia using specific five-color multiparameter flow cytometry. Am. J. Clin. Pathol. *129*, 934–945. https://doi.org/10.1309/FY0UMAMM91VPMR2W.

31. Takao, S., Morell, V., Brown, F.C., Koche, R., and Kentsis, A. (2022). Epigenetic mechanisms controlling human leukemia stem cells and therapy resistance. https://doi.org/10.1101/2022.09.22.509005.

32. Zhang, Y.W., Mess, J., Aizarani, N., Mishra, P., Johnson, C., Romero-Mulero, M.C., Rettkowski, J., Schönberger, K., Obier, N., Jäcklein, K., et al. (2022). Hyaluronic Acid–GPRC5C Signalling Promotes Dormancy in Haematopoietic Stem Cells (Springer). https://doi.org/10.1038/s41556-022-00931-x.

33. Cabezas-Wallscheid, N., Buettner, F., Sommerkamp, P., Klimmeck, D., Ladel, L., Thalheimer, F.B., Pastor-Flores, D., Roma, L.P., Renders, S., Zeisberger, P., et al. (2017). Vitamin A-retinoic acid signaling regulates hematopoietic stem cell dormancy. Cell *169*, 807–823.e19. https://doi.org/10.1016/j.cell.2017.04.018.

34. Lehnertz, B., Chagraoui, J., MacRae, T., Tomellini, E., Corneau, S., Mayotte, N., Boivin, I., Durand, A., Gracias, D., and Sauvageau, G. (2021). HLF expression defines the human hematopoietic stem cell state. Blood *138*, 2642–2654. https://doi.org/10.1182/blood.2021010745.

35. Bonnet, D., and Dick, J.E. (1997). Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. Nat. Med. *3*, 730–737. https://doi.org/10.1038/nm0797-730.

36. Rodriguez-Fraticelli, A.E., Weinreb, C., Wang, S.W., Migueles, R.P., Jankovic, M., Usart, M., Klein, A.M., Lowell, S., and Camargo, F.D. (2020). Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. Nature *583*, 585–589. https://doi.org/10.1038/s41586-020-2503-6.

37. Zeng, A.G.X., Bansal, S., Jin, L., Mitchell, A., Chen, W.C., Abbas, H.A., Chan-Seng-Yue, M., Voisin, V., van Galen, P., Tierens, A., et al. (2022). A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. Nat. Med. *28*, 1212–1223. https://doi.org/10.1038/s41591-022-01819-x.

38. Brett Heimlich, J., Bhat, P., Parker, A.C., Jenkins, M.T., Vlasschaert, C., Ulloa, J., Potts, C.R., Olson, S., Silver, A.J., Ahmad, A., et al. (2022). Mutated cells mediate distinct inflammatory responses in clonal hematopoiesis. https://doi.org/10.1101/2022.12.01.518580.

39. Krevvata, M., Shan, X., Zhou, C., dos Santos, C., Habineza Ndikuyeze, G., Secreto, A., Glover, J., Trotman, W., Brake-Silla, G., Nunez-Cruz, S., et al. (2018). Cytokines increase engraftment of human acute myeloid leukemia cells in immunocompromised mice but not engraftment of human myelodysplastic syndrome cells. Haematologica *103*, 959–971. https://doi.org/10.3324/haematol.2017.183202.

40. Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. Genome Biol. *17*, 86. https://doi.org/10.1186/s13059-016-0936-x.

41. Malikic, S., Mehrabadi, F.R., Ciccolella, S., Rahman, M.K., Ricketts, C., Haghshenas, E., Seidman, D., Hach, F., Hajirasouliha, I., and Sahinalp, S.C. (2019). PhISCS: a combinatorial approach for subperfect tumor

phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. Genome Res. *29*, 1860–1877. https://doi.org/10.1101/gr.234435.118.

42. Zafar, H., Navin, N., Chen, K., and Nakhleh, L. (2019). SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. Genome Res. *29*, 1847–1859. https://doi.org/10.1101/gr.243121.118.

43. McCarthy, D.J., Rostom, R., Huang, Y., Kunz, D.J., Danecek, P., Bonder, M.J., Hagai, T., Lyu, R.; HipSci Consortium, and Wang, W., et al. (2020). Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. Nat. Methods *17*, 414–421. https://doi.org/10.1038/s41592-020-0766-3.

44. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. N. Engl. J. Med. *374*, 2209–2221. https://doi.org/10.1056/NEJMoa1516192.

45. 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.

46. Hennig, B.P., Velten, L., Racke, I., Tu, C.S., Thoms, M., Rybin, V., Besir, H., Remans, K., and Steinmetz, L.M. (2018). Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. G3 (Bethesda) *8*, 79–89. https://doi.org/10.1534/g3.117.300257.

47. Gong, H., Holcomb, I., Ooi, A., Wang, X., Majonis, D., Unger, M.A., and Ramakrishnan, R. (2016). Simple method to prepare oligonucleotide-conjugated antibodies and its application in multiplex protein detection in single cells. Bioconjug. Chem. *27*, 217–225. https://doi.org/10.1021/acs.bioconjchem.5b00613.

48. Paczulla, A.M., Rothfelder, K., Raffel, S., Konantz, M., Steinbacher, J., Wang, H., Tandler, C., Mbarga, M., Schaefer, T., Falcone, M., et al. (2019). Absence of NKG2D ligands defines leukaemia stem cells and mediates their immune evasion. Nature *572*, 254–259. https://doi.org/10.1038/s41586-019-1410-1.

49. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Cell Syst. *8*, 281–291.e9. https://doi.org/10.1016/j.cels.2018.11.005.

50. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002.

51. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. F1000Res *10*, 33. https://doi.org/10.12688/f1000research.29032.2.

52. Kiselev, V.Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. Nat. Methods *15*, 359–362. https://doi.org/10.1038/nmeth.4644.

53. Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat. Biotechnol. *37*, 685–691. https://doi.org/10.1038/s41587-019-0113-3.

54. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

55. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50. https://doi.org/10.1038/s41592-021-01336-8.

56. Bauer, M., Vaxevanis, C., Bethmann, D., Massa, C., Pazaitis, N., Wickenhauser, C., and Seliger, B. (2020). Multiplex immunohistochemistry as a novel tool for the topographic assessment of the bone marrow stem cell niche. Methods Enzymol. *635*, 67–79. https://doi.org/10.1016/bs.mie.2019.05.055.

57. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N.D. (2018). Pyro: deep universal probabilistic programming. https://doi.org/10.48550/arXiv.1810.09538.

58. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
For a complete list of the antibodies used in this study, see Table S1.	N/A	N/A
Biological samples		
For a list of the biological samples used in this study, see Table S3.	N/A	N/A
Chemicals, peptides, and recombinant proteins		
DBCO-PEG5-NHS Ester	Jena Bioscience	CLK-CSTM
Cell Staining buffer	Biolegend	420201
Human TrueStain FcX	Biolegend	422302
TrueStain Monocyte Blocker	Biolegend	426102
UltraPure BSA	Thermo Fisher	AM2616
DRAQ7	Biolegend	424001
Incucyte Caspase3/7 Red	VWR International	MSPP-4704
Homemade Tn5[46]	CRG Protein Technologies Unit	N/A
N,N Dimethylformamide	Sigma Aldrich	D4551-250ML
Betain Lösung	Sigma Aldrich	B0300-5VL
Recombinant RNase Inhibitor	TaKaRa	2313B
Maxima H Minus Reverse Transcriptase	ThermoFisher	EP0752
KAPA HiFi HotStart ReadyMix	Roche	KK2602
SCF	Peprotech	300-07
Flt3-L	Peprotech	300-19
TPO	Peprotech	300-18
IL-3	Peprotech	200-03
IL-6	Peprotech	200-06
UM729	Stem Cell Technologies	72332
Critical commercial assays		
Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1	10x genomics	PN-1000121
Chromium Next GEM Single Cell 3' Kit v3.1	10x genomics	PN-1000268
Chromium Next GEM Chip G Single Cell Kit	10x genomics	PN-1000120
Single Index Kit T Set A	10x genomics	PN-1000213
Dual Index Kit TT Set A	10x genomics	PN-1000215
CleanPCR beads	CleanNA	CPCR-0050
AmpureXP Beads	Beckman Coulter	A63881
SPRIselect beads	Beckman Coulter	B23318
Qubit High Sensitivity dsDNA Assay	ThermoFisher	Q32851
Agilent Bioanalyzer High Sensitivity	Agilent	5067-4626
Opal seven-color IHC kit	Akoya Biosciences	NEL811001KT
Deposited data		
Count tables and metadata	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.20291628
Full input and output of CloneTracer model	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.21982496
Additional data: Patient A.9 follow up (Figure S5)	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.21982490
Additional data: Patient A.2 follow up (Figure S5)	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.21982454

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Additional data: MutaSeq data (Figures 4D and 4E)	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.21982424
All single cell RNA-seq datasets, raw data	EGA	EGA: EGAS00001007078
Reference data[4]	Figshare	Figshare: https://doi.org/10.6084/m9.figshare.13397987.v3
Reference data[31]	Zenodo	Zenodo: https://doi.org/10.5281/zenodo.6496279
Experimental models: Organisms/strains		
NOD.*Prkdc*^{scid}.*Il2rg*^{null} (NSG) mice	Jackson Laboratory	005557
Oligonucleotides		
For a complete list of oligonucleotides used in this study, see Table S4.	N/A	N/A
FISH probes: 6q21/8q24	MetaSystems	D-5802-100-OG
FISH probes: 7cen/7q22/7q36	MetaSystems	D-5043-100-TC
Software and algorithms		
CloneTracer and Primer Design code	Zenodo	Zenodo:
ComplexHeatmap	CRAN	v. 2.6.2
FlowJo	FlowJo, LLC	v. 10.8.1, 10.6.1
ggplot2	CRAN	v. 3.3.5
Htseq (https://pypi.org/project/HTSeq/)	PyPI	v. 2.02
mitoClone	Zenodo	Zenodo: https://doi.org/10.5281/zenodo.4443074
Mgatk (https://github.com/caleblareau/mgatk)	Github	v. 0.1.1
Pheatmap	CRAN	v. 1.0.12
PhISCS (https://github.com/sfu-compbio/PhISCS)	Github	v. 1.0.0
R	CRAN	v. 4.0.2
Seurat	CRAN	v. 4.3.0
Scrublet (https://github.com/swolock/scrublet)	Github	v. 0.2.3
Scmap	Bioconductor	https://doi.org/10.18129/B9.bioc.scmap
Scanorama (https://github.com/brianhie/scanorama)	Github	v. 1.7.3
Spectre (https://github.com/ImmuneDynamics/Spectre)	Github	v. 1.0.0
STAR (https://github.com/alexdobin/STAR)	Github	v. 2.5.4
Other		
StemSpan SFEM media	Stem Cell Technologies	09650

RESOURCE AVAILABILITY

Lead contact

Requests for further information, resources and reagents should be directed to and will be fulfilled by the lead contact, Lars Velten (lars.velten@crg.eu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Datasets including processed and integrated gene expression data, cell type annotation, clonal assignments, metadata and dimensionality reduction are publicly available as Seurat v3 objects through figshare. The DOI is listed in the key resources table. To protect patient privacy and as requested by the relevant ethics boards, raw sequencing data is available from the European Genome-Phenome Archive upon submitting a data access agreement. To obtain these data, contact the lead contact. All accession numbers of data analyzed in this manuscript are listed in the key resources table.
- The implementation of the model and code for primer design and data processing of Optimized 10x libraries is deposited at Zenodo and publicly accessible. The DOI is listed in the key resources table. Code can also be found at https://github.com/veltenlab/CloneTracer.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Bone marrow samples from AML patients were obtained at the Heidelberg University Hospital after informed written consent using ethic application number S-169/2017. For demographic characteristics of sample donors, see Table S3. All experiments involving human samples were approved by the ethics committee of the University Hospital Heidelberg and were in accordance with the Declaration of Helsinki.

Animals

NOD.*Prkdcscid.Il2rgnull* (NSG) mice were bred and housed under specific pathogen-free conditions in the central animal facility of the German Cancer Research Center (DKFZ). Animal experiments were approved and performed in accordance with all regulatory guidelines of the official committee (Regierungspräsidium Karlsruhe). Immune compromised, healthy, female NSG mice 8-12 weeks of age and an average weight of 18-25 g were sublethally irradiated (175 cGy) 24 h before xenotransplantation assays.

METHOD DETAILS

Collection of bone marrow

Bone marrow aspirates were collected from iliac crest. Mononuclear cells were isolated by Ficoll (GE Healthcare, Chicago, Illinois, USA) density gradient centrifugation and stored in liquid nitrogen until further use.

Panel, exome, and bulk ATAC sequencing

For bone marrow samples form cohort A, CD3- and CD3+ cells were sorted by FACS and subjected to exome sequencing as described before.[21] GATK best practices were followed. Mutect2 with Tumor with match normal option was used for the identification nuclear mutations specific for each patient. We considered CD3- cells as tumor and CD3+ as normal. Results are summarized in Table S3.

Additionally, samples A.1, A.3, A.5, A.6, A.7, A.11, A.12, A.13, A.15 were subjected to bulk ATAC sequencing to identify mitochondrial mutations. Again, Mutect2 with Tumor with match normal option was used to identify variants in the mitochondrial genome. Results are summarized in Table S5.

Bone marrow samples from cohort B were sequenced at diagnosis time point with the Illumina TruSight Myeloid Sequencing Panel (Illumina, San Diego, USA) to determine the mutation status of leukemia driver mutations.

Antibody-oligo conjugation

For markers where no commercial conjugates were available, azide-modified oligonucleotides were conjugated to purified antibodies (anti-human CD166, Clone 3A6 (Biolegend, 343902); anti-human GPR56, Clone 4C3 (Biolegend, 391902)) by the use of a DBCO-PEG5-NHS Ester (Santa Cruz Biotechnology, Dallas, USA) in a copper-free click reaction.[47]

In brief, azide-containing storage buffer of purified antibodies was exchanged to PBS (pH 8.5) using the Amicon Ultra-0.5 NMWL 30 kDa Centrifugal Filter (EMD Millipore, Billerica, USA).

100 μg of PBS-buffered antibody was incubated with 2mM DBCO-PEG5-NHS in a final reaction volume of 100μL for 30 minutes at room temperature. The reaction was stopped by the addition of 100mM Tris HCl (pH 8) for 5 minutes at room temperature and non-reactive DBCO-PEG5-NHS was removed using the Amicon Ultra-0.5 NMWL 30 kDa Centrifugal Filter.

Azide-modified oligonucleotides were reconstituted in PBS before adding 30pmol per 1μg DBCO-functionalized antibody. The click reaction was conducted at 4°C for 18 hours. Unreacted oligonucleotides were removed using the Amicon Ultra-0.5 NMWL 50 kDa Centrifugal Filter and the final volume was adjusted to 100μL using PBS (pH 8.5).

Conjugation products were confirmed on Ethidiumbromide (EtBr) stained 2% agarose gels, Coomassie brilliant blue (CBB) stained 4-12% polyacrylamide gels and by absorbance spectroscopy.

Azide-modified oligonucleotides were purchased from Biolegio (Biolegio, Nijmegen, Netherlands) and contained an antibody-specific barcode (bold), a PCR handle (italic) and a capture sequence (underlined). * indicates a phosphorothioated bond to prevent nuclease degradation:

CD166: 5'/Azide/CCTTGGCACCCGAGAATTCCA**CATTAACAGCGCCAA**CAAAAAAAAAAAAAAAAAAAAAAAAAAAAA*A*A

GPR56: 5'/Azide/CCTTGGCACCCGAGAATTCCA**TCATATCCGTTGTCC**CAAAAAAAAAAAAAAAAAAAAAAAAAAAAA*A*A

CITEseq surface labeling and FACS sorting

Human bone marrow samples were thawed and stained using the CITEseq antibody pool (Table S1), as well as sorting antibodies according to the BioLegend protocol https://www.biolegend.com/en-us/protocols/totalseq-a-antibodies-and-cell-hashing-with-10x-single-cell-3-reagent-kit-v3-3-1-protocol

In cohort A, sorting was done using fluorophore-tagged antibodies from BioLegend (San Diego, USA) against human CD3 (clone UCHT; 1:30), CD34 (clone 581; 1:100), and GPR56 (clone CG4; 1:20). FACS sorting of live bone marrow cells was performed on a BD FACSAria equipped with a 70 μm nozzle to enrich for the following populations: CD3+, CD3-CD34+, CD3-CD34-, while aiming for a representation of 25%, 50%, 25%. When insufficient CD34+ were available, a maximum of CD34+ cells were sorted and GPR56 was

used as a second sorting marker to enrich stem cells from the CD34- fraction. Population frequencies were recorded and accounted for in quantitative analysis of the single-cell RNA-seq data set. Sorted cells were loaded onto the Next GEM Chip G for a targeted cell recovery of 5000 cells following the manufacturer's instructions (10x Genomics, CG000206 Rev D).

In cohort B, fluorophore-tagged antibodies against human CD34 (clone 581) and CD45 (clone HI30) (patients B.1, B.2, B.3), or CD56 (clone QA17A16) and CD45 (patient B.4) were used to enrich for following populations:

	B.1		B.2		B.3		B.4		
	d0	d15	d0	d15	d0	d15	d0	d21	d105
total BM/CD45+	85%	99.6%	100%	100%	92%	100%	100%	100%	98%
CD45dim/CD34+	15%	0.4%	0%	0%	8%	0%	-	-	-
CD45dim/CD56+	-	-	-	-	-	-	0%	0%	2%

Population frequencies were recorded and accounted for in quantitative analysis of the single-cell RNA-seq data set. In particular, in analyses that investigate the absolute frequencies of cell types in bone marrow (Figure 5D) the frequency of the cell type was computed per sorted population, and multiplied with the frequency of the sorting gates in the bone marrow sample.

In cases where different biological samples were combined in the same GEM generation run, cells were labeled additionally with oligonucleotide coupled cell hashing antibodies (Biolegend, San Diego, USA). FACS sorting of live bone marrow cells was performed using DRAQ7 (1:1000; Biolegend, San Diego, USA) and Incucyte Caspase3/7 Red (1:5000; VWR International, Radnor, Pennsylvania, USA) on a BD FACSAria™ Fusion equipped with a 100μm nozzle. Sorted cells were loaded onto the Next GEM Chip G for a targeted cell recovery of 8000 cells following the manufacturer's instruction (10x Genomics, CG000206 Rev D).

Single-cell RNA sequencing

cDNA libraries were generated using the 10x Genomics 3' gene expression kit version 3.1 according to the manufacturer's instructions (10x Genomics, CG000206 Rev D). At the cDNA amplification step (step 2.2 of the 10x Genomics protocol), additive primers for amplification of the ADT and HTO libraries were added according to the manufacturer's instructions (Biolegend protocol: TotalSeq™-A Antibodies and Cell hashing with 10x Single Cell 3' Reagent Kit v3 or v3.1 (Single Index) Protocol, Step II). Following cDNA amplification (10x Genomics protocol: step 2.3A), cDNA was split: 10 μL were used for generating Gene Expression (GEX) libraries and 5μL were used for generating antibody-derived tags (ADT) and hashtag oligo (HTO) libraries, respectively, according to manufacturer's instructions (GEX: 10x Genomics protocol: CG000206 Rev D, Step 3; ADT and HTO: Biolegend protocol: TotalSeq™-A Antibodies and Cell hashing with 10X Single Cell 3' Reagent Kit v3 or v3.1 (Single Index) Protocol, Step III). The remaining material was used to construct mitochondrial and targeted mutation libraries.

Final GEX, ADT and HTO libraries were quantified by Qubit and QC'ed on the Bioanalyzer.

Final GEX and ADT libraries were sequenced on separate lanes on a NovaSeq (Cohort A) or HiSeq4000 (Cohort B) with a targeted sequencing depth of 50,000-100,000 reads/cell (GEX) and 300 reads/antibody/cell (ADT), respectively. HTO libraries were sequenced with a targeted sequencing depth of 4000 reads/cell on a NextSeq500.

Optimized 10x: Mitochondrial libraries

For the full-length amplification of mitochondrial cDNA, mitochondrial primers were pooled so that each mitochondrial primer is present at a final concentration of 0.9μM (mito primer mix). See Table S4 for all primer sequences used in this protocol. 10 ng of amplified cDNA was added to a PCR master mix containing 50 μl 2X KAPA HiFi HotStart ReadyMix (Roche), 4 μl 10 uM PartialRead1 primer, 2.5 μl mito primer mix, in a total volume of 100μL. PCR was run as follows: 1 cycle of 95C for 3 mins, 11 cycles of [98C for 20 secs, 67C for 1 min, 72C for 1 min], and 1 cycle of 72C for 5 mins followed by a 4C hold. PCR product was then cleaned with 1.5X (v/v) CleanPCR beads (CleanNA), followed by two washes of 80% ethanol, and eluted in 30 μl EB (Qiagen), after which it was quantified by Qubit and QC'ed by running 1-2 ng of DNA on an Agilent Bioanalyzer High Sensitivity chip. Sample bioanalyzer traces after this step are shown in Figure S1A.

Mitochondrial mutation libraries were then generated by tagmentation with an in-house produced wild-type transposase (Tn5).[46] Briefly, transposome assembly and linker loading was carried out by adding 1 μl of 2 mg/ml Tn5 and 1 μl of annealed linker Tn5ME-B/Tn5MErev to 9 μl of water followed by incubation at 23C whilst shaking at 300 RPM for 30 mins. Assembled transposome was then diluted 1:100 with water. In our experience, it typically required four parallel tagmentation reactions to generate adequate yield for sequencing for a given sample. In a single tagmentation reaction, 1.5 ng of cDNA were added to 10 μl of diluted Tn5 and 10 μl of 4X tagmentation buffer (40 mM Tris-HCl, pH 7.4; 40 mM $MgCl_2$), 10μL DMF for a total of 40 μL. Tagmentation reaction in the PCR was run as follows: 1 cycle of 55C for 3 mins, then a 10C hold. It is important that the PCR is already at 55C when the PCR tubes are placed in the instrument. After tagmentation, 10 μl of 0.2% SDS was added to the tagmented mixture and incubated at room temperature for 5 mins to neutralize the reaction. Once the transposase has been neutralized, the tagmented sample was added to a PCR master mix of 54 μl 2X KAPA HiFi HotStart ReadyMix, 6 μl of 100% DMSO (Thermo Scientific), 10 μl of 10 μM Targeted 10X primer, and 10 μl of 10 μM N7XX primer. This reaction mix was split into two PCR tubes and PCR was run as follows: 1 cycle of 72C for 3 mins, 1 cycle of 95C for 30 secs, 12 cycles of [98C for 20 secs, 60C for 15 secs, 72C for 30 secs], and 1 cycle of 72C

for 3 mins followed by a 10C hold. After the PCR, all reactions were pooled and underwent two rounds of successive bead cleanup. In the first bead cleanup, 0.6X (v/v) CleanPCR beads were used, followed by two 80% ethanol washes and eluted in 50 μl EB. In the second cleanup, 0.6X (v/v) CleanPCR beads were again used, followed by two 80% ethanol washes and eluted in 15 μl EB. The final library was then quantified by Qubit and QC'ed on the Bioanalyzer. Representative bioanalyzer traces are shown in Figure S1B.

Optimized 10x: Targeted genotyping libraries

Nuclear mutations were selected from panel or exome sequencing data by choosing non-synonymous variants in expressed genes, located < 1.5kb away from the end of the gene. Primers targeting mutations of interest were designed using a customized version of the TAPseq Bioconductor package,[18] https://github.com/veltenlab/CloneTracer/tree/master/primer_design

Four rounds of PCRs were then used to generate nuclear mutation libraries. The first three rounds of PCRs are gene-specific nested PCRs and sequencing adaptors and indices were added in the last PCR.

In the first round of PCR (PCR1), 10 ng of amplified cDNA from 10x was added to a PCR master mix containing 2.5 μl of pooled outer gene-specific primers (final concentration of each individual primer in the final pool 10 μM-100 μM), 5 μl of 1 μM Partial_Read1 primer, 20 μl of 5 M Betaine, 50 μl of 2X KAPA HiFi HotStart ReadyMix, and topped up to 100 μl with nuclease-free water. PCR with a heated lid was run as follows: 1 cycle of 95C for 3 mins, 11 cycles of [98C for 20 secs, 67C for 1 min, 72C for 1 min], and 1 cycle of 72C for 5 mins followed by a 4C hold. PCR product was then cleaned with 1.5X (v/v) CleanPCR beads (CleanNA), followed by two washes of 80% ethanol, and eluted in 15 μl EB (Qiagen). After each round of post-PCR cleanups, PCR products were quantified by Qubit and QC'ed by running 1-2 ng of DNA on an Agilent Bioanalyzer High Sensitivity chip. Example bioanalyzer traces are shown in Figure S1H.

In the second round of PCR (PCR2), 10 ng from PCR1 was added to a PCR master mix containing 2.5 μl of pooled middle gene-specific primers, 5 μl of 1 μM Partial_Read1 primer, 20 μl of 5 M Betaine, 50 μl of 2X KAPA HiFi HotStart ReadyMix (Roche), and topped up to 100 μl with nuclease-free water. PCR was run as follows: 1 cycle of 95C for 3 mins, 10 cycles of [98C for 20 secs, 67C for 1 min, 72C for 1 min], and 1 cycle of 72C for 5 mins followed by a 4C hold. Again, PCR product was cleaned with 1.5X (v/v) CleanPCR beads (CleanNA), followed by two washes of 80% ethanol, and eluted in 30 μl EB (Qiagen). Example bioanalyzer traces are shown in Figure S1I.

The third round of PCR (PCR3) was run separately for each target gene. 10 ng from PCR2 was added to a PCR master mix containing 2.5 μl of pooled staggered gene-specific primers (concentration of each primer in the final pool: 25 μM), 5 μl of 1 μM Partial_Read1 primer, 20 μl of 5 M Betaine, 50 μl of 2X KAPA HiFi HotStart ReadyMix (Roche), and topped up to 100 μl with nuclease-free water. PCR was run as follows: 1 cycle of 95C for 3 mins, 10 cycles of [98C for 20 secs, 67C for 1 min, 72C for 1 min], and 1 cycle of 72C for 5 mins followed by a 4C hold. PCR product was cleaned with 1.5X (v/v) CleanPCR beads (CleanNA), followed by two washes of 80% ethanol, and eluted in 30 μl EB (Qiagen).

Finally, libraries were uniquely indexed for each sample. To this end, 10 ng from PCR3 was added to a PCR master mix containing 2.5 μl of 10 μM SI primer, 2.5 μl of 10 μM RPI-N7XX primer (see Table S1), 50 μl of 2X KAPA HiFi HotStart ReadyMix (Roche), and topped up to 100 μl with nuclease-free water. PCR was run as follows: 1 cycle of 95C for 3 mins, 8 cycles of [98C for 20 secs, 52C for 15 sec, 72C for 45 sec], and 1 cycle of 72C for 5 mins followed by a 4C hold. PCR product was cleaned with 1.5X (v/v) CleanPCR beads (CleanNA), followed by two washes of 80% ethanol, and eluted in 15 μl EB (Qiagen). Example bioanalyzer traces are shown in Figure S1J.

Plate-based single-cell RNA-seq (MutaSeq)

Defrosted bone marrow mononuclear cells were stained with following antibodies: Lineage antibodies (CD3, CD19, CD20, CD235a) and additional antibodies (CD34, CD38, CD36, CD45RA, CD90, CD49f, NKG2DL[48]). The single cell index sort was performed on a BD FACSAria Fusion (BD Biosciences) equipped with 355, 405, 488, 561, and 640 nm lasers and Lin- CD34+ cells were sorted into single wells containing lysis buffer to enrich the stem cell compartment except for 3 rows per 384-well plate, in which CD34- cell populations were sorted. After the sort plates were flash frozen and stored at -80°C until library preparation.

Primer design and single-cell RNA-seq was performed as described.[21] See Table S4 for all primers used. Cells were sorted into 384-well instead of 96-well plates and therefore the reaction volumes were downscaled by a factor 2.5-5 depending on the reaction. Lysis volume was 1.2 μl per well. For reverse transcription 2 μl of a buffer containing 0.1 μl Maxima H Minus Reverse Transcriptase (200 U/μl), 0.6 μl 5x RT buffer (both Thermo Scientific), 0.07 μl Recombinant RNase Inhibitor (TaKaRa), 0.45 μl PEG 50%, 0.09 μl 100 μM Smart-seq2 TSO (Eurogentec) and 0.69 μl nuclease-free H_2O were added and RT was performed for 90 min at 42 °C followed by enzyme inactivation at 70 °C for 15 min. The PCR reaction was downscaled to 3 μl. The cDNA was cleaned up using a 0.9x volume (5 μl) of AmpureXP Beads (Beckman Coulter) and tagmented using homemade Tn5[46] at a dilution of 1:50. cDNA was used at a concentration of 1-3 ng/μl and 0.4 μl was tagmented by addition of 1.2 μl of Tn5 dilution mixed 1:1 with 2x tagmentation buffer (20 mM Tris-HCl pH7.5, 20 mM MgCl2, 50% DMF) at 55 °C for 10 min and afterwards shifted to ice. 0.4 μl 0.1% SDS was added for inactivation and incubated for 5 min on ice. PCR was performed by adding 2.7 μl of KAPA HiFi HS mastermix, 0.3 μL of DMSO and 0.4 μL each of the forward i5 and reverse i7 library primer at 3 μM. The PCR program was 72 °C 3 min, 98°C 30 sec, 12 cycles of [98°C, 20s ec, 63°C, 15 sec, 72°C 30 sec], and 72° C, 3 min. Libraries were pooled and purified with 0.9x AmpureXP Beads.

Genotyping of single cell derived cultures

Single cell cultures and genotyping were performed as described [21] with the following modifications: Bone Marrow mononuclear cells from patient A.6 were stained with following antibodies (CD3, CD45RA, CD33, CD98, CD49f, CD38, CD11c, CD371 and HLA-DR).

Lin- or Lin-CD34 + single cells were index-sorted into U-bottom 96-well plates (Sarstedt) containing 100 μl StemSpan SFEM media (Stem Cell Technologies). Media was supplemented with penicillin/streptomycin (100 ng/mL), UM729 (1 μM, Stem Cell Technologies) and the following human cytokines (all from Peprotech): SCF (20 ng/mL), Flt3-L (20 ng/mL), TPO (50 ng/mL), IL-3 (20 ng/mL), IL-6 (20 ng/mL). After two weeks at 5% CO_2 and 37 °C, colonies were imaged by microscopy, and harvested in 12 μl buffer RLT (Qiagen) for subsequent DNA isolation.

Raw 10x Genomics data processing

Gene expression data was processed using cellranger version 4.0.0 with default parameters for feature barcoding. Doublets were removed using scrublet (v. 0.2.3).[49] For cohort A, cells with <1800 genes detected or >10% mitochondrial reads were removed. For cohort B, cells with <1000 genes detected or >40 % mitochondrial reads were removed. The data from the healthy reference individual (C.3) was downloaded from https://doi.org/10.6084/m9.figshare.13397987.v3 and not subjected to further quality filters.

Mitochondrial libraries were processed following the DropSeq standard workflow[50] except that reads were aligned to the mitochondrial genome (GRCh38). Consensus mitochondrial reads were called using the fgbio tool CallMolecularConsensusReads (v. 1.3.0). Only reads from cell barcodes which were detected in the gene expression dataset were used for the downstream analysis. Nucleotide counts were extracted for each single cell using pysam (v. 0.15.3). The final output of the workflow is a list of single-cell matrices in which for each position of the mitochondrial genome the number of A,T,C and Gs UMIs are stored. Mitochondrial variants were identified as previously described,[21] and using bulk ATAC sequencing, where available (most of cohort A). The workflow was implemented in snakemake[51] and can be found in https://github.com/veltenlab/CloneTracer/tree/master/library_processing/mitochondria

Nuclear SNV libraries were processed similarly to the mitochondrial libraries with the difference that reads were aligned to the complete human genome (GRCh38). Only reference and alternative alleles (identified by exome or panel sequencing) were considered for the final count table. Due to the high number of PCR amplification steps, only UMIs supported by at least two reads were included in the analysis. The final count table contains the number of reference and alternative UMIs for each single cell and targeted mutation. The workflow to process nuclear SNVs libraries was written in snakemake and can be found in https://github.com/veltenlab/CloneTracer/tree/master/library_processing/nuclear-snv

Analysis of single cell gene expression data

For projecting single cell data onto a reference atlas of healthy bone marrow, we used a workflow based on scmap[52] as described.[4] Sample code for reference atlas projection is available at https://git.embl.de/triana/nrn/-/tree/master/Projection_Vignette. Thereby, we obtained uMAP coordinates, cell type labels, and myelocyte pseudotime, where applicable.

For unsupervised integration of all data sets, scanorama[53] was used with default parameters to integrate across the three cohorts A, B and C, using the cohort as the batch. Scanorama components were then imported into Seurat and uMAP, nearest neighbor graphs, and clustering were computed using the default Seurat pipeline with default parameters.[54] scanorama was selected based on a systematic comparison study,[55] where it was described as a method that maintains biologically true difference between samples, which we considered relevant in the context of a highly heterogeneous disease such as AML.

To illustrate this, in Figure S2D, cluster C15, C16 and C32 all resemble HSCs/MPPs when projected to a healthy reference. Compared to other immature myeloid clusters all cells from these clusters express stem cell markers (CD34, MHC class II) and lack markers of myeloid commitment (MPO, AZU1); but they also differ in the expression of genes that are usually co-expressed in HSCs[1] such as MECOM (exclusive to cluster C16), HOX genes (strongly overexpressed in cluster C32), or, in the case of cluster C15, displayed a strong interferon response signature. Since these are real biological, and not technical differences, we think it is important that they are represented in the uMAP and unsupervised clustering. Thereby data integration by scanorama is highly complementary to the projection to a healthy reference, that we employ e.g. in the context of Figures 3A, 3G, and 3H: This two-tiered approach (projection and weak batch integration) allows us to identify biologically different cell states that all resemble e.g. MPPs, but vastly differ in the (aberrant) expression of key genes.

As a further validation of the integration strategy employed, we included a healthy control individual in each cohort. We then integrated all three healthy individuals (A.0, B.0 and the Reference individual from[4]) using the same unsupervised data integration steps used to generate the main figures of the manuscript, with identical parameters. The resulting uMAP (Figure S2E) demonstrates that our data integration strategy effectively accounts for technical differences between cohorts.

Dormancy score calculation

To compute a dormancy score for each individual cell, we first selected genes differentially expressed in Zhang et al.[32] (adjusted p-value <0.01) and the dHSC gene list from Cabezas-Wallscheid et al.[33] (adjusted p-value <0.05). We then normalized the gene expression count matrix by library size and centered and scaled the data. We computed principal components for cluster C6 cells from all patients with <100 of cells in this population using the prcomp() function in R. The dormancy score corresponded to the first principal component. To compute the score in other cells we used the function predict() with the principal component loadings computed as described above as a model and the scaled gene expression data as new input.

Analysis of DNAseq from single-cell derived colonies

Raw sequencing reads were aligned to the human genome using STAR (v. 2.5.4). Nucleotide count tables were generated from single-colony BAM files using the function baseCountsFromBamList from the package mitoClone (v.1.0). For each mutation, colonies were labelled as mutant when > 5% of the reads were mutant and healthy otherwise. If the mutation was not covered it was labelled as dropout. The binarized table (colonies x mutations) was the input to PhISCS[41] which was ran with default parameters to infer the clonal hierarchy.

Processing of MutaSeq scRNAseq data

Raw gene expression data was aligned using STAR (v. 2.5.4) and count matrices were generated using htseq (v. 2.02) with default parameters. Only cells with > 2000 genes detected and <10% mitochondrial reads were kept for downstream analysis. Gene expression data for all patients was integrated using scanorama.[53] The 5000 most variable genes were included. Genes which were amplified with mutation-specific primers were also excluded from the integration process. Each patient was used as batch key. The first 100 scanorama components were used to compute the integrated uMAP following the default Seurat pipeline. Reference and mutant counts for SNVs and mtSNVs were obtained using the function baseCountFromBamList from the mitoClone package (v.1.0)[21]

Raw data processing of MAESTER data

Raw fastq files from the MAESTER library of a human clonal hematopoiesis sample were downloaded from SRA (SRR15598777) and processed as described.[14] In brief, 24bp primer sequences were trimmed from the 5' of read2 fastq file using homertools (v. 4.11). Read 2 was tagged with cell and umi barcodes and aligned to the complete human genome using STAR (v.2.5.4). Only reads from cells present in the final Seurat object (downloaded from https://www.dropbox.com/s/vna1k3k7khazd7j/BPDCN712_Seurat_Final.rds?dl=0) were kept for downstream analysis. mgatk (v. 0.1.1) was used to obtain single-cell nucleotide count matrices with default parameters and –mr = 3.

Fluorescent *In Situ* Hybridization

Human bone marrow samples were thawed and stained with fluorophore-tagged antibodies against CD45, CD3, CD49f, CD11c, CD14 and CD34 as described above (CITEseq surface labeling, FACS sorting and GEM generation). For antibody clones and titrated amounts, see Table S1. Cells were collected by FACS sorting on a BD FACSAria™,or BD FACSAria™ Fusion, respectively, each equipped with a 100μm nozzle. Sorted cells were fixed on glass slides in methanol/acetic acid. Hybridization was performed according to the manufacturer's instructions by using FISH probes for chromosome regions 6q21/8q24 and 7cen/7q22/7q36 (MetaSystems, Altlussheim, Germany), respectively. Interphase nuclei were validated using an automated scanning system (Applied Spectral Imaging, Edingen/Neckarhausen, Germany).

Tissue microarrays

The frequency of different cell subsets in the bone marrow microenvironment in AML patients was analyzed by multispectral imaging (MSI). Formalin-fixed and paraffin embedded (FFPE), decalcified bone marrow samples were stained as described elsewhere.[56] The marker panel used for staining included antibodies directed against CD34, CD14, CD11c, CD49f. For antibody clones and dilutions, see Table S1. All primary antibodies were incubated for 30 min. Tyramide signal amplification (TSA) visualization was performed using the Opal seven-color IHC kit containing fluorophores Opal 520, Opal 540, Opal 570, Opal 690 (Akoya Biosciences., Marlborough, MA, USA), and DAPI. Stained slides were imaged employing the PerkinElmer Vectra Polaris platform. To unify the spatial distribution analysis, × 20 MSI fields (1872 × 1404 pixel, 0.5 μm/pixel) were analyzed. Cell segmentation and phenotyping of the cell subpopulations were performed using the inForm software (PerkinElmer Inc., USA). The frequency of all immune cell populations analyzed and the cartographic coordinates of each stained cell type were obtained.

Large cohort flow cytometry analysis

Human BM samples were stained as described above and analyzed using the BD Symphony (see Table S1 for a list of antibodies used). Cells were pre-analyzed using FlowJo v10.8.1. Doublets and dead cells were excluded, as well as artefacts using a time gate. Remaining cells were exported using the channel values and imported into R. Further, the package Spectre v1.0.0 was used for batch correction, clustering, dimension reduction and visualization following the "discovery workflow with batch alignment using CytoNorm". Using the summary table of the package, CD11c expression was investigated on CD34+ and CD14+ cell clusters.

Xenotransplantations

Female NSG mice 8-12 weeks of age were sublethally irradiated (175 cGy) 24 h before xenotransplantation assays. FACS sorted primary AML samples were injected into the femoral BM cavity of sublethally irradiated mice. Mice were daily monitored, and femur bone marrow aspirates were taken at 16 weeks to determine engraftment and lineage potential. Human leukemic engraftment in mouse BM was evaluated by flow cytometry using anti-human-CD45-FITC (clone HI30), anti-human-CD34-BUV395 (clone 581), anti-human-CD19-FITC (clone HIB19), anti-human-CD33-PE-Cy7 (clone WM53), CD3-BV510 (clone OKT3), and anti-mouse-CD45-Alexa700 (clone 30-F11).

QUANTIFICATION AND STATISTICAL ANALYSIS

CloneTracer model

See Methods S1 for a full description of the CloneTracer model. Posterior predictive checks[57] were used to determine if the data met the assumptions of the statistical model, as detailed in the Methods S1.

Differential expression testing

For differential expression testing of surface antigens, we used Wilcoxon tests following library size normalization. For differential expression testing of RNA, we used MAST.[58] In all cases, comparisons were performed separately by patients, and the number of patients where the change was significant was used as an overall measure of significance and consistency.

Data visualization

All plots were generated using the ggplot2 (v. 3.3.5), ComplexHeatmap (v. 2.6.2) and pheatmap (v. 1.0.12) packages in R 4.0.2 or FlowJo (v. 10.6.1, BD). Boxplots are defined as follows: the middle line corresponds to the median; the lower and upper hinges correspond to first and third quartiles, respectively; the upper whisker extends from the hinge to the largest value no further than 1.5X the inter-quartile range (or the distance between the first and third quartiles) from the hinge and the lower whisker extends from the hinge to the smallest value at most 1.5X the inter-quartile range of the hinge. Data beyond the end of the whiskers are called 'outlying' points and are plotted individually.

Detail on statistical tests used in the different figures and definition of relevant summary statistics are included in the figure legends.

ADDITIONAL RESOURCES

Interactive 2D and 3D versions of most uMAPs from this paper are available at https://veltenlab.crg.eu/clonetracer/.