

# Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics

Lars Velten <sup>1,2,17</sup>✉, Benjamin A. Story <sup>3,4,17</sup>, Pablo Hernández-Malmierca <sup>5,6,17</sup>, Simon Raffel<sup>5,6,7</sup>, Daniel R. Leone<sup>3</sup>, Jennifer Milbank<sup>3</sup>, Malte Paulsen<sup>8</sup>, Aykut Demir <sup>7</sup>, Chelsea Szu-Tu<sup>1,2</sup>, Robert Frömel<sup>1,2</sup>, Christoph Lutz<sup>7</sup>, Daniel Nowak <sup>9</sup>, Johann-Christoph Jann <sup>9</sup>, Caroline Pabst<sup>7,10</sup>, Tobias Boch<sup>4,6,9</sup>, Wolf-Karsten Hofmann<sup>9</sup>, Carsten Müller-Tidow <sup>7</sup>, Andreas Trumpp <sup>5,6,11</sup>, Simon Haas <sup>5,6,11,12,13,14</sup> & Lars M. Steinmetz<sup>3,15,16</sup>✉

Cancer stem cells drive disease progression and relapse in many types of cancer. Despite this, a thorough characterization of these cells remains elusive and with it the ability to eradicate cancer at its source. In acute myeloid leukemia (AML), leukemic stem cells (LSCs) underlie mortality but are difficult to isolate due to their low abundance and high similarity to healthy hematopoietic stem cells (HSCs). Here, we demonstrate that LSCs, HSCs, and pre-leukemic stem cells can be identified and molecularly profiled by combining single-cell transcriptomics with lineage tracing using both nuclear and mitochondrial somatic variants. While mutational status discriminates between healthy and cancerous cells, gene expression distinguishes stem cells and progenitor cell populations. Our approach enables the identification of LSC-specific gene expression programs and the characterization of differentiation blocks induced by leukemic mutations. Taken together, we demonstrate the power of single-cell multi-omic approaches in characterizing cancer stem cells.

<sup>1</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. <sup>2</sup>Universitat Pompeu Fabra (UPF), Barcelona, Spain. <sup>3</sup>European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. <sup>4</sup>Swiss Federal Institute of Technology (ETH) Zurich, Department of Biosystems Science and Engineering, Basel, Switzerland. <sup>5</sup>Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), Heidelberg, Germany. <sup>6</sup>Division of Stem Cells and Cancer, Deutsches Krebsforschungszentrum (DKFZ) and DKFZ-ZMBH Alliance, Heidelberg, Germany. <sup>7</sup>Department of Internal Medicine V, Hematology, Oncology and Rheumatology, University of Heidelberg, Heidelberg, Germany. <sup>8</sup>European Molecular Biology Laboratory (EMBL), Flow Cytometry Core Facility, Heidelberg, Germany. <sup>9</sup>Department of Hematology and Oncology, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany. <sup>10</sup>Molecular Medicine Partnership Unit (MMPU), University of Heidelberg and European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. <sup>11</sup>German Cancer Consortium (DKTK), Heidelberg, Germany. <sup>12</sup>Berlin Institute of Health (BIH), Berlin, Germany. <sup>13</sup>Charité-Universitätsmedizin, Berlin, Germany. <sup>14</sup>Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany. <sup>15</sup>Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. <sup>16</sup>Stanford Genome Technology Center, Palo Alto, CA, USA. <sup>17</sup>These authors contributed equally: Lars Velten, Benjamin A. Story, Pablo Hernández-Malmierca. ✉email: [lars.velten@crg.eu](mailto:lars.velten@crg.eu); [larsms@embl.de](mailto:larsms@embl.de)

Issues with high cellular turnover, such as the hematopoietic system or the intestine, depend on “professional” adult stem cells for their continuous regeneration<sup>1</sup>. Oncogenic mutations in these cells can cause cancers that maintain a hierarchical organization reminiscent of the tissue of origin. Only cancer stem cells (CSCs), residing at the top of the hierarchy, are able to fuel long-term cancer growth and drive relapse, whereas the bulk of the cancer consists of rapidly dividing cells with limited capacity for self-renewal, i.e., cells that exhaust their replicative potential after a finite number of divisions<sup>2–4</sup>. Owing to their stem cell-like properties, CSCs constitute an important driver of relapse, but their low division rates make them difficult to target therapeutically. Tools that permit the confident identification and characterization of CSCs are therefore urgently needed.

Acute myeloid leukemia (AML) serves as a paradigm for the study of cancer stem cells<sup>5</sup>. In 10–20% of healthy individuals over age 70, the acquisition of pre-leukemic mutations in hematopoietic stem cells (HSCs) results in the dominance of a small number of HSC-derived clones, a process termed Clonal Hematopoiesis of Indeterminate Potential (CHIP)<sup>6,7</sup>. While such pre-leukemic stem cells (pre-LSCs) are capable of giving rise to healthy blood and immune cells, additional mutations can cause a complete block in differentiation and thereby result in the malignant expansion of aberrant progenitor cells<sup>8</sup>. The accumulation of these so-called “blast” cells is ultimately fueled by the presence of leukemic stem cells (LSCs). Classic chemotherapy regimens primarily target actively cycling “blast” cells and initially lead to remission. Since quiescent or protected LSCs often avoid eradication, relapse rates are high with 5-year survival rates below 15% for patients over the age of 60. A key goal is, therefore, to identify therapeutic strategies for targeting LSCs, while sparing healthy HSCs<sup>9–11</sup>. Characterizing gene expression differences between HSCs, pre-LSCs and LSCs would be a valuable step towards that goal.

Previously, LSC-specific gene expression patterns were characterized by isolating cells positive for stem cell-specific surface markers used in the healthy hematopoietic system, such as CD34 (refs. 12–15). More recently, leukemic engraftment rates in xenotransplant models were correlated with gene expression<sup>16–19</sup>. However, these approaches all measure impure populations of cells. Here, we propose that by measuring mutational status and gene expression in single cells simultaneously, cancer stem cells can be uniquely distinguished from both mature cancer cells (based on gene expression) and healthy stem cells (based on mutational status) (Fig. 1a). Finally, pre-LSCs are thought to typically carry mutations associated with CHIP (e.g., in *DNMT3A*) but not mutations associated with leukemia (e.g., in *NPM1*)<sup>7,20,21</sup>, potentially enabling their identification by profiling both known leukemic and known preleukemic mutations.

While we and others have demonstrated the utility of single-cell genomics for mapping hematopoietic differentiation hierarchies<sup>22–24</sup>, tracking mutations or clones in single-cell gene expression data remains difficult. Previous work has amplified somatic mutations from complementary DNA (cDNA)<sup>25–28</sup>, extracted mutational information from single-cell RNA-seq reads<sup>29</sup>, or processed both genomic DNA and RNA from single cells<sup>30–32</sup>. However, these protocols suffer from a lack of confidence in assigning cells to clones and/or require prior knowledge of genomic mutation sites. As an alternative tool for clonal tracking, the use of endogenous mitochondrial mutations as clonal markers has been proposed, obviating the need for prior knowledge of genomic mutations<sup>33,34</sup>. However, the application of these methods to characterize LSCs has not been demonstrated, and in particular requires the ability to reliably detect clonal expansion events, associate clinically relevant coding

mutations to clones with high confidence, and draw statements on gene expression changes between clones.

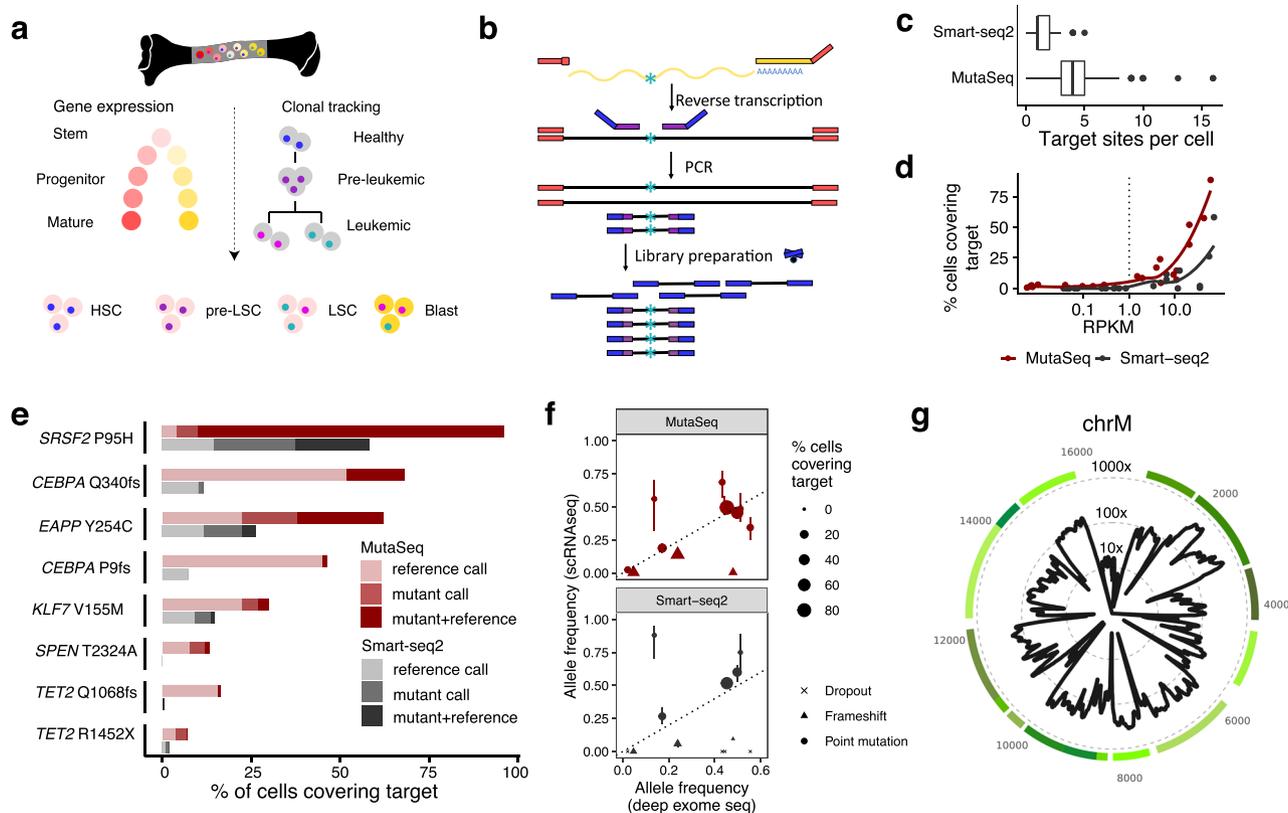
Here, we introduce MutaSeq, a workflow that amplifies nuclear mutations from cDNA, and mitoClone, a computational tool that achieves high-confidence clonal assignments and de novo discovery of clones using mitochondrial marker mutations when available. MutaSeq data from four AML patients allows us to distinguish HSCs, pre-LSCs, LSCs, and progenitor/blast populations. Thereby, we identify transcriptomic consequences of leukemic and pre-leukemic mutations relevant to stem cells. Additionally, we characterize the contribution of different leukemic and pre-leukemic clones to healthy and disease-specific bone marrow populations with unprecedented detail. Altogether, our results demonstrate cancer stem cell identification and characterization by simultaneous mapping of genomic and mitochondrial mutations in single-cell transcriptomes.

## Results

**MutaSeq provides high coverage of genomic and mitochondrial mutations.** To establish a robust experimental setup for the clonal tracking of human cells in single-cell transcriptomic data, we evaluated various modifications of the Smart-seq2 protocol aimed at increasing coverage at polymorphic genomic sites of interest (Supplementary Fig. 1a). We found that inclusion of targeting primers during reverse transcription frequently resulted in the formation of undesired byproducts, especially when targeting a higher number of sites (Supplementary Fig. 1a–d). By contrast, when sites of interest were targeted during cDNA amplification, we obtained high-quality transcriptome data while increasing the average number of target sites captured per cell by 2–4-fold compared to a non-targeted approach (Fig. 1b, c and Supplementary Fig. 1a, b). An automated pipeline for primer design that minimizes off-target sites and potential primer-dimer formation is available at <https://github.com/veltenlab/PrimerDesign> (see also the Methods section). MutaSeq stably works with up to 30–40 primer pairs (targets); when higher numbers of primers are included, library quality progressively decreases (Supplementary Fig. 1e). In a test with only highly expressed target genes, target amplicons are created from all primer pairs in virtually all single cells (Supplementary Fig. 1f, g).

To evaluate MutaSeq, we performed deep exome sequencing of an AML patient (P1) and designed primers targeting 14 nuclear mutations (Supplementary Data 1 and 4). We then systematically compared the performance of MutaSeq and non-targeted Smart-seq2 on CD34+ cells from this patient. MutaSeq increased the number of target sites covered per single cell from a median of 1 to a median of 4 (Fig. 1c–e) and recapitulated the variant allele frequencies estimated by exome sequencing with higher accuracy than Smart-seq2 (Fig. 1f), while maintaining comparable transcriptome data quality (Supplementary Fig. 1h, i). Both methods underestimated the abundance of frameshift mutations, possibly as a consequence of nonsense mediated decay<sup>35</sup> (Fig. 1e, f). While our data do not provide statistical evidence for an effect of target gene length or sequence complexity on dropout, we cannot exclude such effects.

Importantly, both methods provide an excellent coverage of the mitochondrial genome (mean ~100X mitochondrial coverage given a mean sequencing depth of ~788,000 total reads per cell, Fig. 1g), unlike most other single-cell RNA-seq protocols applied in the context of clonal tracking<sup>27–29,31</sup> (Supplementary Fig. 1j). Altogether, these results demonstrate that MutaSeq efficiently covers the mitochondrial genome in single-cell RNA-sequencing experiments and provides improved coverage of genomic target sites compared to Smart-seq2. Importantly, it requires no changes to existing Smart-seq2 pipelines, except for the addition of targeting primers during cDNA amplification.



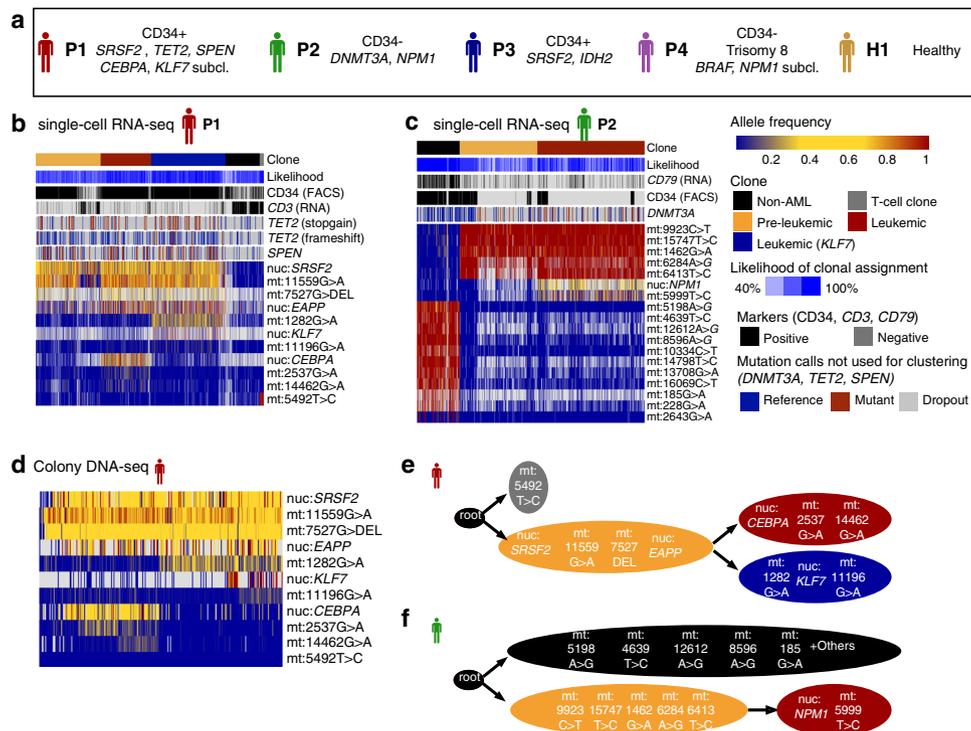
**Fig. 1 MutaSeq for high-quality single-cell RNA-seq data with clonal information.** See also Supplementary Fig. 1. **a** Overview of the study. HSC: hematopoietic stem cell, (pre-)LSC: (pre-)leukemic stem cell, Blast: mature leukemic blast. **b** Overview of the MutaSeq method. Targeting primers (purple) are included during the cDNA amplification step of the Smart-seq2 protocol. Targeting primers are directly fused to illumina library adapters (blue) and therefore get amplified efficiently during library preparation. Tagmentation introduces the same adapters to the full-length cDNA product. **c** Number of target sites covered per cell, across  $n = 206$  (Smart-seq2) or  $n = 658$  CD34+ (MutaSeq) bone marrow cells from patient P1 (see the Methods section Data visualization, for a definition of boxplot elements). **d** Mean gene expression of genes containing mutations of interest is plotted against the fraction of cells in which the mutation is covered. **e** Fractions of cells covering key non-synonymous mutations observed in the patient. Reference call: The reference allele was observed. Mutant call: the mutant allele, as defined by bulk exome sequencing (Supplementary Data 1) was observed. Mutant+Reference: both alleles were observed. **f** Allele frequency estimates derived from deep exome sequencing compared to allele frequency estimates derived from MutaSeq (red dots,  $n = 2208$  single cells) or Smart-seq2 (gray dots,  $n = 206$  single cells). Dot size indicates coverage at target site. Point shape indicates the type of mutation. Error bars indicates interquartile range. See figure source data for complete specification of sample size used to derive statistics ( $n$ ). **g** Coverage of the mitochondrial genome in MutaSeq data. See Supplementary Fig. 1j for a comparison across methods. Green line segments correspond to genes in the mitochondrial genome.

**Simultaneous mapping of mitochondrial and genomic mutations permits high-confidence tracking of leukemic, pre-leukemic, and healthy clones.** To investigate if MutaSeq can distinguish leukemic, pre-leukemic and residual healthy clones, we generated data from four AML patients with heterogeneous genotypes and phenotypes (Figs. 2a, 3b and Supplementary Fig. 2). To allow for a better characterization of stem cells in each patient, cells were sorted such that putative stem and progenitor cells (CD34+) and putatively more mature cells (CD34-) were approximately covered at equal portions (Supplementary Fig. 2). Of note, two of the patients (P2, P4), exhibited bone marrow consisting of >99% of CD34- cells (Supplementary Fig. 2a). Owing to different capture rates across individuals and gates, the final data set consisted of between 618 to 1430 cells per patient, of which between 190 and 968 were CD34+ (Supplementary Fig. 2b, c). We therefore avoid statements on quantitative shifts in population size between patients throughout this manuscript.

We then called nuclear genomic mutations, as well as mitochondrial mutations, at the single-cell level in order to cluster cells into clonal hierarchies. Bulk exome sequencing of the patients had identified known pre-leukemic mutations present at high allele frequency and known leukemic mutations present at a

somewhat lower allele frequency (Fig. 2a and Supplementary Data 1. Patient 1: mutations in *SRSF2*, *TET2*/*CEBPA* and *SRSF2*, *TET2*/*KLF7*; Patient 2: mutations in *DNMT3A*/*NPM1*; Patient 3: mutations in *SRSF2*/*IDH2*; Patient 4: leukemic Trisomy 8 and *BRAF* mutations). While some statements on clonal hierarchies could be drawn solely based on calls of these nuclear somatic mutations (Supplementary Fig. 3a), the relatively high dropout of these sites impeded robust assignments of cells to clones (Supplementary Fig. 3b–d). Moreover, the result was biased by the expression levels of the mutated genes of interest: in cells with low expression, dropout was higher, leading to a higher fraction of false-negative calls, i.e., false classifications of mutant cells as reference (Fig. 1e and Supplementary Fig. 3c, d). Similar issues were faced by other methods using related approaches of mutation amplification from cDNA<sup>27,28</sup>.

To overcome these limitations, we next determined if mitochondrial mutations can be used to refine clonal hierarchies jointly with the nuclear mutations. Since mitochondrial RNA is extensively edited, it is important to cluster cells based solely on mutations, and not on seemingly polymorphic sites that are the result of post-transcriptional events with no relationship to clonal structure. Here, we developed and tested a filtering strategy that



**Fig. 2 Mitochondrial mutations serve as high-confidence clonal markers in AML.** See also Supplementary Figs. 2–6. **a** Overview of the patients. CD34+ and CD34- indicate the dominant surface phenotype of the leukemic blasts (see Fig. 3b and Supplementary Fig. 2 for quantification). For each patient, genes containing mutations are printed in italic (see Supplementary Data 1 for a complete list). Subcl., sub-clonal mutation. **b** Heatmap depicting variant allele frequencies (color coded, see right of Fig. 2c for a key) observed in single-cell RNA-seq data of  $n = 1430$  cells from P1. Gr indicates missing values. Cells and mutations are arranged according to the clustering result obtained by PHISCS<sup>36</sup> as described in the “Methods” section, *Analysis of mitochondrial mutations and reconstruction of clonal hierarchies*. Calculation of the likelihood is described in the same section. Mutations with low coverage (in *TET2*, and *SPEN*) were not included in the clustering and are depicted in the heatmaps as metadata, however, in all cases except *TET2* frameshift there is quantitative evidence for their association with specific clones (Supplementary Fig. 5b). For reproducing the computations, see the vignettes accompanying the mitoClone package. For mutations, nuc is nuclear genome; mt is mitochondrial genome. **c** Like panel **b**, but using  $n = 1066$  cells from P2. *DNMT3A* is included as a low coverage mutation with significant association to the pre-leukemic and leukemic clone (Supplementary Fig. 5b). **d** Heatmap depicting variant allele frequencies (color coded, see legend to the right of Fig. 2c) observed in targeted DNA-seq data from  $n = 288$  single-cell derived colonies from P1. **e** PHISCS<sup>36</sup> was run on the mutational data from P1 to reconstruct a clonal hierarchy, see the Methods section *Analysis of mitochondrial mutations and reconstruction of clonal hierarchies*. Take note that while the order of mutations is based on the PHISCS model, the grouping of mutations into clones is based on an arbitrary cutoff to provide a useful clustering for further analyses. See Clone in Fig. 2c legend for color codes. **f** Like **e**, but for P2.

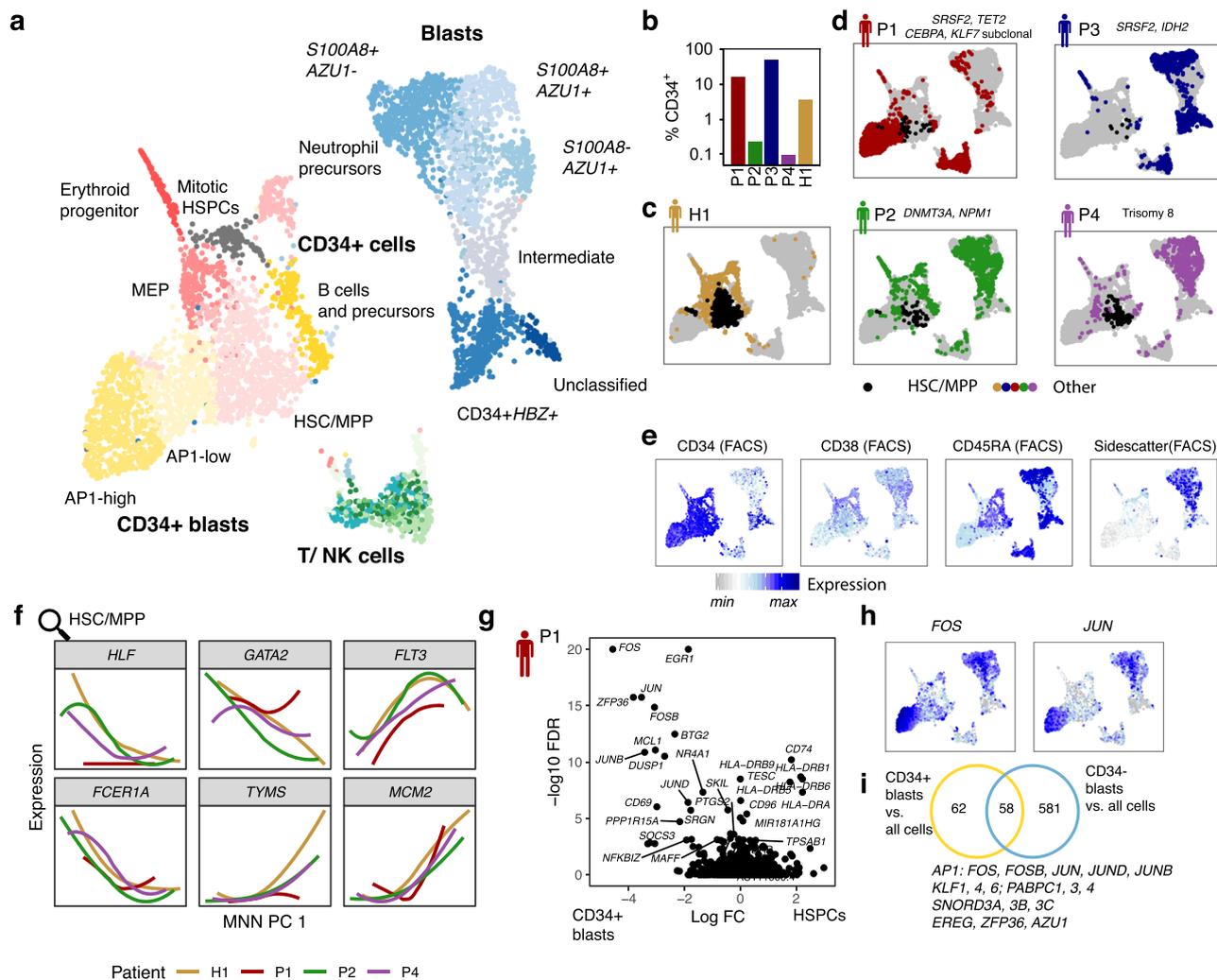
only makes use of mitochondrial mutations for clonal tracking if they uniquely occur in individual patients (Supplementary Fig. 4a, see also Methods). This idea assumes that RNA editing events are typically shared between individuals, whereas somatic mutations are not. Using the whole-exome sequencing data, we validated that this approach, at the level of genomic sites, correctly distinguishes mutations and RNA editing events with a precision of 97% (Supplementary Fig. 4b, c). We further analyzed various control data sets with known associations between mitochondrial mutations and clones<sup>33</sup> to validate that this approach enables the detection of relevant mitochondrial mutations without a need for a DNA-based reference, and further enables the unsupervised identification of clones (Supplementary Fig. 4d–f). We have implemented all the required filtering and blacklisting routines required for the identification of high-confidence somatic mitochondrial variants in the mitoClone R package (<https://github.com/veltenlab/mitoClone>).

We then computed clonal hierarchies from both nuclear and mitochondrial somatic variants using a mathematical model that accounts for allelic dropout<sup>36</sup> (see Methods for detail, all required tools are contained in the mitoClone package). In patients P1 and P2, pre-leukemic as well as sub-clonal leukemic mutations were significantly associated with distinct sets of well-covered mitochondrial variants (Supplementary Fig. 5a, b), such that clonal

hierarchies could be delineated, and a confident assignment of cells to clones became possible (Fig. 2b, c and see Supplementary Fig. 3b for a quantitative analysis). Unlike genomic mutation calling from cDNA, identification of clonal identities from mitochondrial mutations is mostly not, and in one case weakly, affected by gene expression levels or library quality (Supplementary Figs. 3c, d, 5c), and is possible at lower sequencing depths (Supplementary Fig. 5d–f), since mitochondrial genes are consistently highly expressed.

Importantly, the clonal structure was validated by targeted DNA-sequencing from single-cell derived colonies (Patient 1, Fig. 2d). Across the patients, we identified clones carrying known pre-leukemic mutations (e.g., in *SRSF2*, *DNMT3A*) and sub-clones carrying known leukemic mutations (e.g., *CEBPA*, *NPM1*) (Fig. 2e, f). Below we functionally characterize these clones as leukemic or pre-leukemic based on their contribution to healthy blood production (Fig. 4).

In patient P3, allele frequencies of leukemic and pre-leukemic mutations were near 50%, indicating the presence of a single leukemic clone (Supplementary Data 1). In patient P4, an 18-year old individual with a leukemia driven by triplication of chromosome 8, no mitochondrial markers were identified, even though exome sequencing suggested the presence of sub-clonal variants (Supplementary Data 1). In this case, MutaSeq still



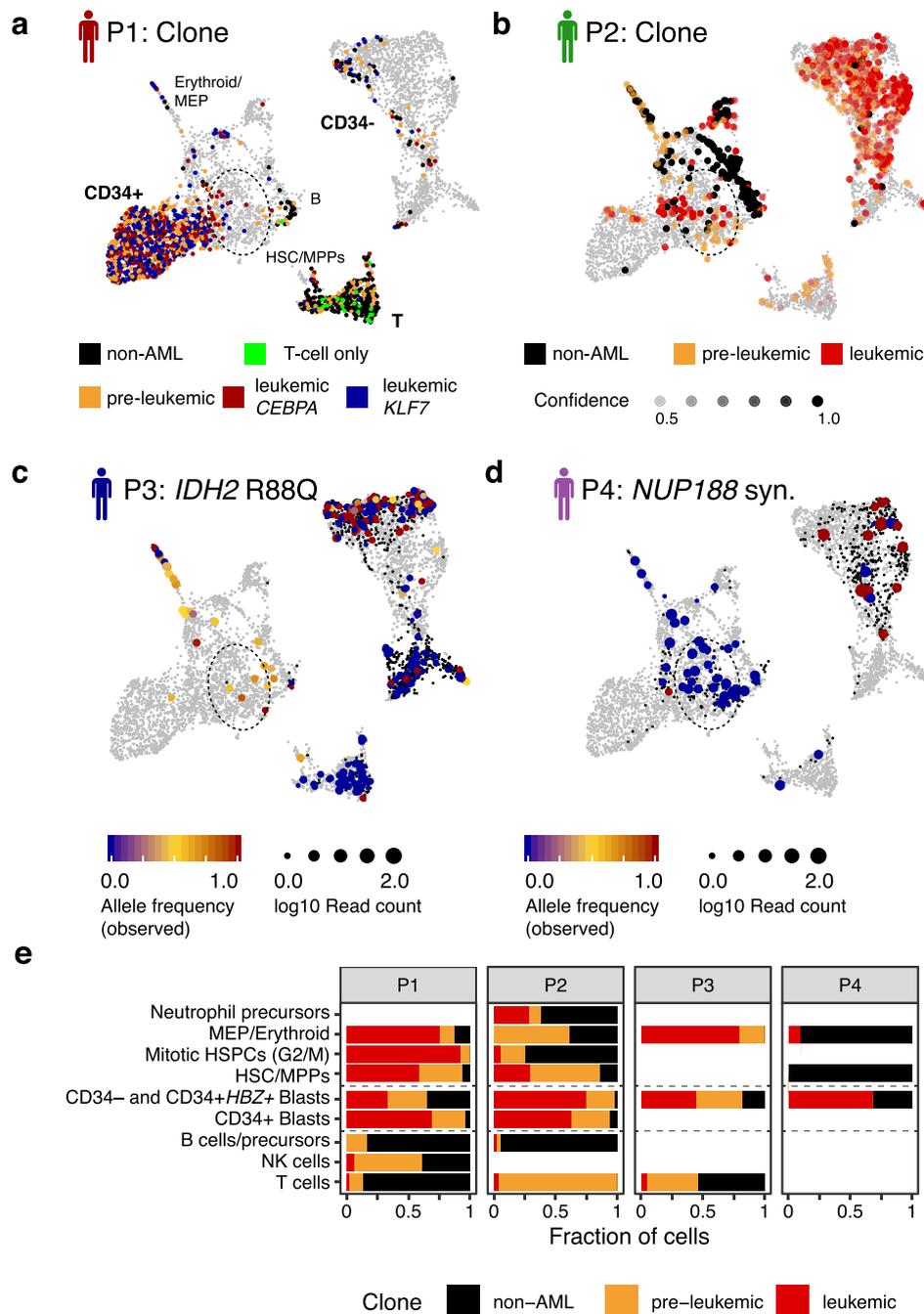
**Fig. 3 An overview of cell types observed in bone marrow of healthy and leukemic individuals.** See also Supplementary Figs. 2 and 7. **a** Data from the five individuals (Fig. 1a) were integrated using scanorama<sup>60</sup> and visualized in two dimensions using uMAP<sup>62,63</sup>. Clusters are color-coded. *N* = 5228 cells. **b** Fraction of CD34+ cells in total bone marrow from the five individuals. CD34+ cells were enriched during FACS sorting for a higher representation in the single-cell RNA-seq data (Supplementary Fig. 2). **c** CD34+ cells from a healthy individual<sup>22</sup> highlighted on the uMAP. Black dots correspond to cells from the HSC/MPP cluster. **d** Expression of key FACS markers highlighted on the uMAP (Supplementary Fig. 7a). **e** Logicle-transformed expression of several marker genes of healthy HSC/MPP subsets<sup>22</sup> is plotted over the first dimension of variability identified by MNN (Supplementary Fig. 7f). **f** Data from *n* = 667 cells from the HSC/MPP cluster were integrated across individuals using MNN<sup>64</sup>. The smoothed expression of several marker genes of healthy HSC/MPP subsets<sup>22</sup> is plotted over the first dimension of variability identified by MNN (Supplementary Fig. 7f). **g** Volcano plot of the log<sub>10</sub> expression change (Log FC) in *n* = 569 AP1-high CD34+ blasts vs. *n* = 667 HSC/MPP-like cells, plotted against corrected *p*-values from MAST, using a model that accounts for differences in library quality and patient identity/batch (see the Methods section Single-cell gene expression data analysis). AP1-high CD34+ blasts were chosen for this comparison since AP1-low blasts, in terms of all marker genes, appear to constitute an intermediate state between Healthy-like HSC/MPPs and AP1-high blasts. **h** Log-normalized expression of *FOS* and *JUN* on the uMAP from panel **a**. See panel **e** for a color scale. **i** Venn diagram displaying genes with significant overexpression in AP1-high CD34+ blasts and CD34- blasts, compared to all other cells from the data set.

permits a qualitative analysis using genomic mutations alone (see below). Taken together, our approach allows for the identification of putatively leukemic, pre-leukemic, and healthy clones and can assign cells to clones with high confidence if mitochondrial somatic variation is present.

**Identification and characterization of clones de novo.** We next investigated if the use of mitochondrial somatic variants enables the identification and characterization of clones without prior knowledge of nuclear mutations. To that end, we made use of a data set from patient P1 generated without amplification of nuclear sites (i.e., standard Smart-seq2). A clear clonal structure was identified in an unsupervised manner based solely on

mitochondrial variants (Supplementary Fig. 6a). In order to examine whether the presence of somatic genetic variability is associated with the different clones, we then queried the mutational status of 13,797 genomic sites annotated as mutated in AML in the COSMIC database<sup>37</sup> using a beta-binomial model (see Methods). This unsupervised analysis revealed a highly significant association of the *SRSF2* P95H mutation with the leukemic and pre-leukemic sub-clones (Supplementary Fig. 6b, c). Their malignant nature was further evidenced by a markedly reduced ability to contribute to the T cell lineage (Supplementary Fig. 6d and see also below).

To further demonstrate our ability to identify clones de novo, we highlight a clonal expansion of non-leukemic cells in P2 (Fig. 2f).



**Fig. 4 Effects of (pre-)leukemic mutations on cellular differentiation.** See also Supplementary Fig. 8. **a** Clonal identity of the cells from P1 highlighted on the uMAP (Figs. 2b, e, 3a). Gray dots correspond to cells from other patients. The dotted ellipse serves as a guide to identify the location of the HSC/MPP population. **b** Clonal identity of the cells from P2 highlighted on the uMAP (Figs. 2c, f, 3a). **c** Observed variant allele frequencies for the *IDH2* R88Q mutation from P3 highlighted on the uMAP. Small black dots correspond to cells with no coverage of the mutation (see Supplementary Fig. 8a for an estimate of target capture rates). **d** Observed variant allele frequencies for the synonymous *NUP188* mutation from P4 highlighted on the uMAP. This mutation was observed at an allele frequency of 50% in exome data of total bone marrow. Note that CD34+ cells were enriched more than 100-fold during sorting for single-cell RNA-seq (Supplementary Fig. 2 and see Supplementary Fig. 8a for an estimate of target capture rates). **e** Estimate of the contribution of different clones to the cell types. For P1 and P2, clonal identities from Fig. 2 were used. For P3, cells were classified as leukemic if the *IDH2* mutation was observed, as pre-leukemic if the *SRSF2* mutation was observed, or as non-leukemic if the reference allele was observed for both mutations. Cells without coverage were excluded from the analysis. For P4, cells were classified as leukemic if the *NUP188* mutation was observed, or as non-leukemic if the reference allele was observed. Bars are only shown for populations covered with at least 10 cells (see Supplementary Fig. 8b for absolute numbers and Supplementary Fig. 8c, d for a quantitative analysis).

This clone would have been missed by approaches relying on genomic mutations alone<sup>27,28,31</sup>. Interestingly, these cells were not associated with the pre-leukemic *DNMT3A* mutation. By again querying sites from the COSMIC database<sup>37</sup> using a beta-binomial

model, we identified that they had uniquely acquired a mutation in the *RPL3* gene (Supplementary Fig. 6b, e). These results suggest that this clonal expansion event is independent of the leukemia and associated with the acquisition of unrelated nuclear mutations.

We also take note of a putative non-leukemic clone in P1 marked by a single mitochondrial variant (5492T>C). With one exception, all cells carrying this variant are positive for the T-cell marker CD3 (Figs. 2b, 4a). Hence, this variant was likely acquired in a T-cell precursor or T-cell clone, although we cannot formally exclude that it corresponds to a T-cell-specific RNA editing event.

Taken together, these results demonstrate that our approach can identify and characterize clones de novo without prior knowledge of nuclear genomic mutations. The mitoClone package implements all routines for clonal clustering and mutation calling.

**A map of cell states in leukemic bone marrow identifies HSC-like cells.** We next used single-cell transcriptome data to distinguish stem cells, progenitor cells and leukemic blasts. To define cell populations in our samples, we integrated the gene expression data from all four patients along with data from CD34+ bone marrow cells of a healthy individual<sup>22</sup> into a two-dimensional representation, and characterized cell types based on marker gene expression (Fig. 3a, Supplementary Fig. 7a–c, f, Supplementary Data 2).

All four bone marrow samples contained cells that clustered with HSCs and multipotent progenitor cells (HSC/MPP) from the healthy individual (Fig. 3b–d) and displayed a CD34<sup>+</sup>CD38<sup>low</sup>, FSC/SSC<sup>low</sup> phenotype (Fig. 3e). Unsupervised analysis separated these cells into quiescent immature HSC-like cells (*HLF*+), more proliferative erythromyeloid primed progenitors (*GATA2*+), and lymphomyeloid primed progenitors (*FLT3*+ ) (Fig. 3f and Supplementary Fig. 7f). Cells resembling healthy erythroid progenitors and MEPs were also identified, alongside various types of B-cell precursors and T/NK cells (Supplementary Fig. 7b, c).

In contrast to non-leukemic populations, the majority of cells from leukemic bone marrow samples (“blasts”) were very different between the patients, in line with previous observations<sup>27</sup> (Fig. 3d): we observed (a) differentiated blasts (CD34<sup>+</sup>SSC<sup>hi</sup>/FSC<sup>hi</sup>) that expressed neutrophil genes such as calprotectin (*S100A8*, *S100A9*) and *AZU1* (Fig. 3e and Supplementary Fig. 7a); (b) CD34+ blasts that were highly mitotic and expressed the fetal hemoglobin *HBZ*; and c) CD34<sup>+</sup>CD38<sup>low</sup> blasts expressing markers typical of hematopoietic stem and progenitor cells (HSPCs), such as *PROM1* (CD133) and *MEIS1* (Supplementary Fig. 7a, d, e). The latter population appeared to be connected to the HSC/MPP population across a continuum of states that gradually upregulate AP1 transcription factor expression (*FOX*, *JUN*, *FOXB*, *JUNB*, *JUND*) while down-regulating MHC class II (Fig. 3g, h and Supplementary Data 3). A global analysis of highly expressed genes across all populations revealed that high expression levels of AP1, as well as several Krüppel-like factors and poly-A binding proteins, were common to all different blast populations from the patients and distinguished them from healthy progenitors (Fig. 3h, i). Of note, high AP1 and KLF activity have recently been identified as a hallmark conserved across genetically distinct types of AML by bulk-sequencing studies that only investigated blast populations<sup>38</sup>. Our results demonstrate that indeed these transcription factors appear to be relevant in all, phenotypically very different, blast populations.

Taken together, all four leukemia samples could be stratified into stem cells, progenitors, and blasts. Furthermore, all patients retain cells highly similar to healthy HSCs, although in variable abundance.

**Clonal tracking identifies cellular differentiation states and gene expression patterns associated with pre-leukemic and leukemic mutations.** Gene expression information alone was insufficient to distinguish cancerous from non-cancerous HSPCs

(Supplementary Fig. 7f). To definitively characterize cells as (pre-) LSCs or residual healthy cells, we therefore integrated single-cell gene expression data and clonal tracking results (Fig. 4a–d). If mitochondrial somatic variability was present, we were able to assign clonal identities with high confidence, allowing us to draw quantitative statements (Patients P1 and P2). In the absence of mitochondrial somatic variability, we used nuclear mutation calls in *SRSF2*, *IDH2* and *NUP188* for purely qualitative statements (Patients P3 and P4). The capture rates of these marker sites ranged from 70% (*SRSF2*) to 11% (*NUP188*) (Supplementary Fig. 8a).

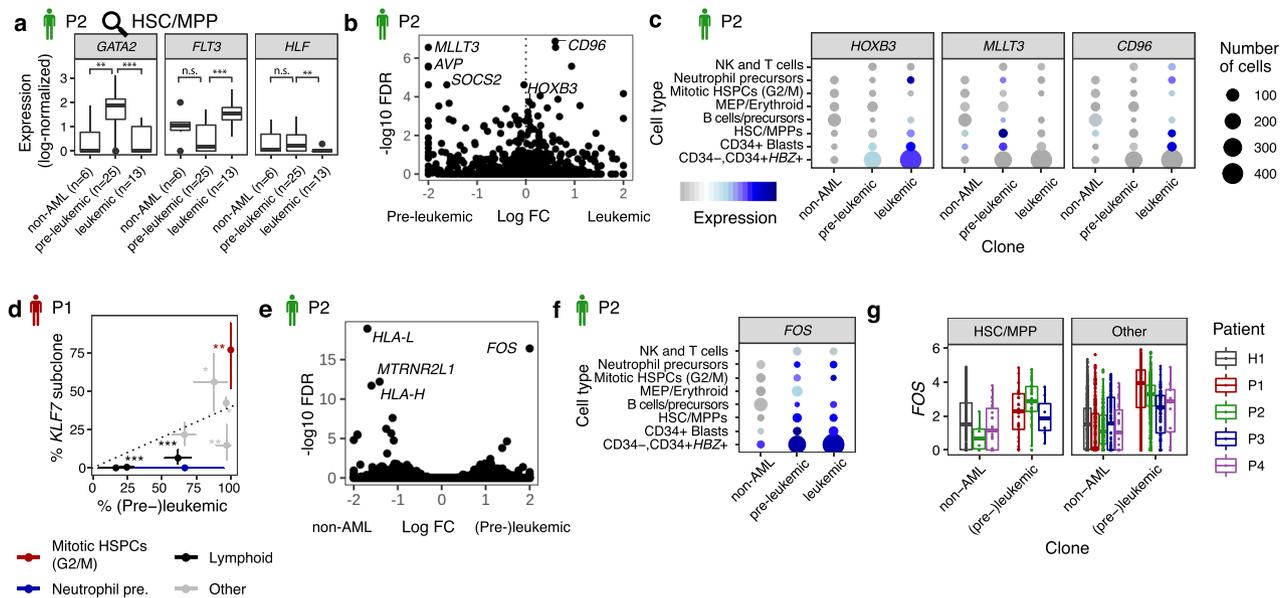
Clones associated with leukemic mutations were most prevalent in the blast compartments and were also detected in the HSPC compartment, but were almost absent in lymphoid (B, NK, T) lineages. By contrast, clones associated with pre-leukemic mutations were found in all lineages, but mostly displayed a decreased prevalence in lymphoid lineages (Fig. 4e and Supplementary Fig. 8b–d). These observations confirm the designation of these clones as “leukemic” and “pre-leukemic”. Furthermore, these results highlight that the leukemic mutations may initiate differentiation blocks at various levels, as previously reported<sup>4,27,39</sup>. For example, in patients P1 and P3, leukemic cells had retained the ability to contribute to the erythroid lineage, while in patient P2, this activity was restricted to the pre-leukemic and non-leukemic clones. Importantly, the leukemic cells in P1, P2, and P3 were observed in a cell state that is highly reminiscent of healthy HSCs/MPPs on a molecular level, and retains the ability to contribute to various lineages, i.e., is functionally multipotent.

Next, we investigated the molecular effects of distinct mutations in detail. Previously, the consequences of specific leukemic mutations were commonly studied in mouse models. MutaSeq data allows us to compare gene expression between clones differing only in a single mutation, thereby elucidating the specific effects of that mutation on human hematopoiesis.

Mutations in the de novo DNA methyl transferase *DNMT3A* are the most common cause of benign clonal expansions of HSCs in individuals of advanced age<sup>6,7</sup>. In patient P2, *DNMT3A*-mutated pre-leukemic HSPCs were rather stem-like (with relatively high expression of *HLF*) or primed into the erythromyeloid direction (with relatively high expression of *GATA2*) (Figs. 4b and 5a). These results are in line with recent findings from *DNMT3A* knock-out mouse models<sup>40</sup>. Interestingly, independent of cell state, this coincided with an upregulation of *MLL3*, a gene whose enforced expression promotes erythroid-megakaryocytic output from HSCs<sup>41</sup> (Fig. 5b, c and Supplementary Data 3).

Mutations in the multifunctional ribonucleoprotein *NPM1* are identified as drivers for acute myeloid leukemia in 30% of patients and frequently co-occur with pre-leukemic *DNMT3A* mutations<sup>21</sup>. In patient P2, the erythromyeloid bias of the *DNMT3A* clone was lost upon acquiring the leukemic *NPM1* mutation. *NPM1* mutated cells upregulated *HOXB3*, again in line with data from mouse models<sup>42–44</sup> (Fig. 5b, c). Independent of cell state, these cells further exhibited upregulation of *CD96* RNA expression, which has previously been identified as a leukemia stem cell-specific marker<sup>12</sup>. *CD96* was also highly expressed on leukemic HSC/MPP-like cells of patient P3, but not in patient P1, further illustrating the patient-specific nature of LSC markers (Supplementary Fig. 8e).

Mutations in *KLF7* are not commonly observed in leukemia; however, krüppel-like factors are highly expressed by genetically and phenotypically different blasts<sup>38</sup> (see above). In patient P1, the *KLF7* mutated clone displayed a higher proportion of cells in G2/M phase (Fig. 5d). Based on a reanalysis of ATAC-seq data from human CD34+ cells<sup>45</sup> we found that enhancers containing *KLF7*-binding sites were enriched near tumor suppressor genes<sup>46</sup>,



**Fig. 5 Molecular consequences of leukemic and pre-leukemic mutations.** See also Supplementary Fig. 8. **a** Log-normalized expression of *GATA2*, *FLT3* and *HLF* in HSC/MPPs from P2, stratified by clonal identity (see the Methods section Single-cell gene expression data analysis, for detail on data normalization). Asterisk indicate significance from a two-sided Wilcoxon test, \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , n.s.: not significant (see the Methods section Data Visualization, for a definition of boxplot elements). **b** Volcano plot of the  $\log_{10}$  expression change (log FC) in pre-leukemic ( $n = 55$ ) vs. leukemic ( $n = 50$ ) CD34+ cells of P2, plotted against corrected  $p$ -values (FDR) from MAST<sup>65</sup>, using a model that accounts for differences in cell type and library quality (see Methods). Only the following CD34+ cell types from Fig. 3a were included in the test: HSC/MPP, CD34+ Blasts (both subsets), Neutrophil precursors, and MEPs. **c** Dot plot comparing the expression of relevant genes across non-leukemic, pre-leukemic, and leukemic cells in the different cell types. Symbol size scales with the number of cells per cell type. **d** Scatter plot depicting the fraction of (pre-)leukemic cells in relation to the fraction of cells from the *KLF7*-mutated clone in various cell types in P1. Dotted line indicates the mean ratio across all cells, error bars denote 95% confidence intervals from a beta distribution, and asterisk indicate significant deviation from the mean ratio, as follows: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ .  $p$ -values are from a two-sided binomial test and were not adjusted for multiple testing (see Supplementary Fig. 8c and figure source data for number of single cells underlying each group). **e** Volcano plot of the  $\log_{10}$  expression change (log FC) in (pre-)leukemic ( $n = 105$ ) vs. non-leukemic ( $n = 41$ ) CD34+ cells of P3, plotted against corrected  $p$ -values (FDR) from MAST<sup>65</sup>, using a model that accounts for differences in cell type and library quality (see Methods). Only the following cell types from Fig. 3a were included in the test: HSC/MPP-like, CD34+ Blasts (AP1 high and low), Neutrophil precursors, and MEPs. **f** Dot plot comparing the expression of *FOS* across non-leukemic, pre-leukemic and leukemic cells in the different cell types. Dot size and color represent the quantity of cells and gene expression level, respectively (see legend in Fig. 5c). **g** Boxplots comparing the log-normalized expression levels of *FOS* between cells with evidence of originating from the non-leukemic clone(s), and cells with evidence of originating from the leukemic or pre-leukemic clones. Cells were assigned to clones as in Fig. 4e (see the Methods section Single-cell gene expression data analysis, for detail on data normalization, see figure source data for number of single cells underlying each group and see the Methods section Data Visualization, for a definition of boxplot elements).

including *CDK6*, *PTEN*, *RUNX1*, and *FLT3*, compared to active enhancers not containing *KLF7* binding sites ( $p = 0.002$ , Supplementary Fig. 8f).

In sum, we have used a small heterogeneous patient cohort to demonstrate, as a proof-of-concept, that the acquisition of specific mutations is frequently linked to an altered gene expression program, which is consistent with data obtained from mouse models<sup>40,42–44</sup>.

Finally, we compared gene expression between all (pre-)leukemic and non-leukemic cells, with the goal of identifying potential markers or drug targets present in all (pre-)leukemic cells, but not in residual healthy cells. In patient P2, our analysis identified a small number of hits (Fig. 5e and Supplementary Data 3), most notably *FOS*, which was expressed in all (pre-)leukemic cells across cell types, but not by non-leukemic HSC/MPPs (Fig. 5f), and the MHC class I genes *HLA-L* and *HLA-H*, which were expressed by all healthy cells, but not by (pre-)leukemic cells. Across all patients, *FOS* was consistently over-expressed in cells carrying (pre-)leukemic lesions, both in HSCs/MPPs, and in other cell types (Fig. 5g).

Taken together, these results demonstrate the ability of MutaSeq and mitoClone to delineate developmental and molecular effects of clonal evolution caused by leukemic and pre-leukemic mutations.

## Discussion

Herein, we have described a joint single-cell transcriptomics and clonal tracking approach (MutaSeq and mitoClone) for characterizing LSCs, charting their differentiation capabilities, and mapping the molecular consequences of oncogenic mutations. While single-cell gene expression profiling permits the identification of cells with a stem-cell signature, clonal tracking using genomic and mitochondrial mutations allows for a clean separation between healthy and cancerous clones. Thereby we distinguish LSCs, HSCs, pre-LSCs, healthy progenitors, and blasts. We have demonstrated this approach in the context of acute myeloid leukemia, and we propose that similar approaches may be applied to other types of cancers.

By applying our approach to bone marrow samples from four AML patients, we have demonstrated its capabilities:

**Charting the differentiation capacities of hematopoietic clones.** By quantifying the contribution of (pre-)leukemic cells to blood lineages, we have shown that in patients P1–P3, leukemic clones not only form blasts but also exist in an HSC-like state. Pre-leukemic clones additionally contribute to erythroid and lymphoid lineages. Multipotent HSCs are, therefore, the likely cell of origin in these patients<sup>4</sup>. In patient P4, with one exception, only cells with a mature phenotype displayed leukemic mutations, illustrating that the disease can also be fueled by cells with a

committed phenotype<sup>39</sup>. Alternatively, the LSC population in this patient might be rare among CD34+ cells.

**Identification of clones de novo.** In the presence of mitochondrial somatic variability, our approach does not rely on previously known nuclear mutations to detect clones. In the example of patient 2, we have thereby identified an expanded clone present at very low frequency in total bone marrow, but highly abundant in the CD34+ fraction. We have demonstrated that this clone is associated with a nuclear mutation in the *RPL3* gene, which we discovered de novo. This result illustrates the low clonal complexity of hematopoiesis in individuals of advanced age, which is often not associated with candidate driver mutations<sup>20</sup>.

**Identification and characterization of LSCs.** Importantly, our approach was capable of identifying leukemic cells highly reminiscent of healthy HSCs. These cells will be important to target therapeutically without ablating their healthy counterparts<sup>9–11</sup>. While stemming from a study cohort of limited size, our data suggest that *FOS* might constitute a potential target for further investigation, as it is expressed throughout the disparate leukemic populations including the most HSC-like cells, pre-LSCs, and blasts, but not in healthy HSCs. Further, we have confirmed that CD96 is a specific marker for LSCs in some patients<sup>12</sup>. Studies in larger cohorts will be required to assess how generally applicable these findings are.

Taken together, our study expands upon earlier work on the molecular phenotypes of AML blasts<sup>27,38</sup> and constitutes the first detailed characterization of LSCs by single-cell transcriptomics. This advance is owed to three crucial aspects of experimental design.

**Enrichment of relevant starting populations.** LSCs are excessively rare and generally present at below 0.1% of total bone marrow<sup>16</sup>. In order to characterize these cells by single-cell genomics, a prior enrichment, e.g., by sorting for CD34 expression, is essential. In future studies, this approach can be complemented with markers that also label CD34- LSCs, such as GPR56 (ref. 18).

**Deep transcriptome sequencing.** In some cases, the bulk of leukemic cells displays gene expression signatures highly similar to stem cells, as observed here for the CD34+ blasts of patient P1. The differences between LSCs and residual healthy HSCs are even more subtle. Previous work using shallow, microwell-based sequencing of AML cells<sup>27</sup> has not identified differences between LSCs, CD34+ blasts and residual healthy HSCs.

**High-confidence clonal tracking.** When available, the use of mitochondrial variants enables the confident assignment of cells to clones, and thereby, a quantitative analysis of gene expression. In two out of four AML patients, we identified high-confidence, specific mitochondrial genetic markers for pre-leukemic and leukemic sub-clones. In the third patient, allele frequencies of leukemic and pre-leukemic mutations were around 50%, indicating dominance of a single leukemic clone. In the final patient, sub-clonal genomic mutations were observed, but not associated with mitochondrial variability. Interestingly, this patient was only 18 years old and exhibited a leukemia possibly driven by a “catastrophic” triplication of chromosome 8. The length of the pre-leukemic phase, the buildup of mitochondrial variants accumulated during normal ageing, as well as unknown factors affecting mitochondrial mutation rates might all contribute to the presence of mitochondrial marker mutations.

In the context of recent methods for clonal tracking within single-cell transcriptomics<sup>27–29,31,33</sup>, the use of the MutaSeq protocol and mitoClone computational pipeline combines the strengths of previous approaches relying on either nuclear or mitochondrial variants, but also has limitations. Specifically, droplet- or microwell-based protocols for single-cell RNA-seq and clonal tracking<sup>27–29</sup> provide low coverage of mitochondrial

genomes and high dropout of nuclear genomic sites, and therefore do not enable a confident assignment of cells to clones. Single-cell ATAC-seq<sup>33,34,47</sup> and Smart-seq2 efficiently cover mitochondrial genomes, but their coverage of nuclear genomic sites is absent or low, hardening interpretability of the data. Finally, while TARGET-seq addresses the limitation of dropout of nuclear sites, depending on the implementation it does not offer sufficient mitochondrial coverage, or is of very limited throughput. Of all methods available to date, the approach presented here converges on crucial aspects, specifically: (a) high mitochondrial coverage, allowing us to identify benign expanded clones de novo, even in the absence of known genetic markers, (b) decreased dropout of relevant genomic mutations, permitting the association of clones with genomic mutations, if present, and (c) a highly confident assignment of cells to clones, enabling quantitative analyses of clone-specific gene expression (Supplementary Fig. 9). These capabilities expand the potential applications of our approach to the study of clonal dynamics during ageing and oncogenesis beyond the hematopoietic system.

**Limitations.** The major limitation of our pipeline is that it requires natural somatic variability to resolve clones at high confidence. In the absence of mitochondrial somatic variation, the MutaSeq protocol can be used to draw qualitative statements on clonal differentiation capacities (similar to ref. 28, and with an improvement over Smart-seq2), but due to dropout neither enables the statistically confident assignment of cells to clones, nor differential expression testing between clones. In the presence of mitochondrial somatic variation, an association between clones and nuclear mutations is only possible for mutations in highly expressed genes, and can be limited by the nature of the mutation (e.g., frameshift mutations) and possibly other factors such as sequence complexity. The third limitation of MutaSeq is its relatively low throughput. This limitation is currently shared with Smart-seq2 and alternative single-cell RNA-seq methods allowing high-confident assignment of cells to clones<sup>31,33</sup>. Future work will focus on the inclusion of full-length coverage of the mitochondrial genome in droplet-based single-cell RNA-seq platforms.

## Methods

**Patient and sample collection.** The AML samples were collected from diagnostic bone marrow aspirations at the University hospitals in Heidelberg, Germany, and Mannheim, Germany after obtaining informed written consent. Bone marrow mononuclear cells were isolated by density gradient centrifugation and stored in liquid nitrogen until further use. All experiments involving human samples were conducted in compliance with the Declaration of Helsinki and all relevant ethical regulations and were approved by the ethics committees of the medical faculties Heidelberg and Heidelberg-Mannheim of the University of Heidelberg, the Bioethics Internal Advisory Committee (BIAC) at EMBL and the CRG bioethics committee (CEIC-Parc de Salut Mar).

**Deep exome sequencing and target selection.** For exome sequencing, DNA was extracted from  $9 \times 10^3$  flow sorted CD34+ cells (for CD34+ leukemias: P1, P3) or total bone marrow. As healthy controls, we used a buccal swap (P1), FACS-sorted CD45-CD105+MSCs (P3) or in vitro expanded MSCs (P2 and P4)<sup>48</sup>. Sequencing libraries were constructed using the SureSelect HS XT Target Enrichment System v6 (Agilent), and a mean on-exon sequencing coverage of at least >70X was obtained for each patient. Genomic alignments were performed using BWA MEM v0.7.15 (ref. 49) and cancer variants were identified using Mutect2 v3.8 (P1 and P3) and v4.0.9 (P2 and P4)<sup>50</sup>, following the GATK best practice recommendations. Variants were annotated using ANNOVAR<sup>51</sup>. Output from Mutect2 was filtered to remove variants that did not overlap with known genes. The final list of candidate variants included only those with allele frequencies (AF) >4% in the cancer exome sample and with an AF fourfold larger than in the healthy exome sample (Supplementary Data 1). Finally, the candidates for targeting were hand-selected from this list with a focus on cancer relevant genes, highly expressed genes, and potential sub-clonal markers.

**FACS sorting.** Bone marrow mononuclear cells were stained according to standard protocols. In brief, cells were thawed, washed once in medium (IMDM, 10% FCS,

20 U/mL DNase I) and 3 million cells were resuspended in 100  $\mu$ L medium containing antibodies diluted as described in Supplementary Table 1. Following incubation for 30 min on ice, cells were resuspended in phosphate-buffered saline with 2% FCS. For single-cell liquid cultures and MutaSeq, cells were stained with fluorescent-labeled antibodies against lineage markers (CD4, CD8, CD19, CD20, CD41a, CD235a) and additional markers (CD45RA, CD135, GPR56, CD34, CD38, CD90, CD33, Tim3), and sorted according to the gating scheme illustrated in Supplementary Fig. 2. BD FACS Fusion (BD Biosciences) equipped with 405, 488, 561, and 640 nm lasers were used. Of note, in P1 and P4, Lin<sup>+</sup> cells could not be efficiently processed into libraries for unknown reasons and are, therefore, excluded. A list of all antibodies used can be found in Supplementary Table 1.

**Cell culture.** K562 cells were purchased from ATCC (catalog number CCL-243) and cultivated in RPMI-1640 (Thermo 21875034) supplemented with 10% FBS and P/S.

**Primer design.** Primers for MutaSeq, for other single-cell targeting protocols tested (Supplementary Fig. 1), as well as for targeted DNA sequencing were designed using the computational pipeline available at <http://git.embl.de/velten/PrimerDesign>. For MutaSeq, the refgene transcripts spanning each genomic site of interest were selected as template; if multiple refgene transcripts were found for one site, a consensus transcript containing only exonic sequences present in all variants was created. We then used primer3 (ref. 52) to design five possible pairs of primers for each intended target, with an amplicon length of 90–145 bp and a melting temperature of (nominally) 60 °C. BLAST was used to remove primer pairs, which potentially form off-target amplicons. Then, the pair complementarity (i.e., potential to form dimers) was computed for each possible combination of primers across all target sites (forward-reverse, forward-forward and reverse-reverse). In order to identify a set of primers that covered the maximal number of genes while strictly forbidding primers with high complementarity scores, a graph was constructed that connected all primers with different targets to each other if their complementarity score was lower than 15. A maximum clique-finding algorithm<sup>53</sup> was then used to identify the largest mutually connected component in the graph. Thereby, the largest number of targets that efficiently avoids dimer formation was selected.

Nextera adapters were added to all primers designed accordingly (fwd: GTCGT CGGCAGCGTCAGATGTGTATAAGAGACAG, rev: GTCTCGTGGGCTCGGA GATGTGTATAAGAGACAG).

For targeted DNA-seq experiments, the genomic sequence surrounding the target was used as template and nested PCR primers were designed. Inner primers were designed as in the case of MutaSeq, and outer primers surrounding the inner PCR product with an amplicon length of 200–350 bp and a nominal annealing temperature of 58 °C were added. A list of all primers used for this study is included in Supplementary Data 4.

### Single-cell RNA sequencing with targeting of genomic sites of interest

**(MutaSeq).** MutaSeq is based on the Smart-seq2 protocol<sup>54,55</sup> with the modifications introduced by ref. 22. For lysis, we used 5  $\mu$ L of a buffer containing 0.1  $\mu$ L RNAsin + (Promega), 0.04  $\mu$ L 10% Triton X-100 (SigmaAldrich), 0.1  $\mu$ L of 100  $\mu$ M Smart-seq2 Oligo-dT primer (SigmaAldrich) and 1  $\mu$ L dNTP mix (10 mM each, NEB). In P1, we had additionally included 0.075  $\mu$ L of a 1:1,000,000 dilution of ERCC spike-in mix 1 (Ambion), as well as a control spike-in to quantify the false-positive detection rate of mutations (Supplementary Fig. 10). Plates were snap frozen directly after sorting and later thawed at 10 °C in a PCR machine for 5 min and denatured at 72 °C for 3 min. 5  $\mu$ L of a buffer containing 0.25  $\mu$ L RNAsin +, 2  $\mu$ L 5x SMART FS buffer, 0.5  $\mu$ L DTT 20 mM, 1  $\mu$ L SmartScribe enzyme (all TaKaRa) and 0.2  $\mu$ L 50  $\mu$ M Smart-seq2 TSO (Exiqon) were then added and RT was performed for 90 min at 42 °C, 10 cycles of [50 °C, 2 min and 42 °C, 2 min], and enzyme inactivation at 70 °C for 15 min. Then, we added 15  $\mu$ L PCR mix containing 12.5  $\mu$ L KAPA HiFi HS mastermix (Merck), 0.25  $\mu$ L 10  $\mu$ M Smart-seq2 ISPCR primer (SigmaAldrich) and 0.5  $\mu$ L of a pool of all targeting primers, present at 1  $\mu$ M each. cDNA amplification was performed by 98 °C 3 min, 21 cycles of [98 °C, 20 sec, 67 °C, 60 sec, 72 °C 6 min], and 72 °C, 5 min. cDNA was cleaned up using an equal volume (25  $\mu$ L) of CleanPCR beads (CleanNA) and tagged using homemade Tn5 (ref. 56). In brief, cDNA was diluted to ~150 pg/ $\mu$ L, Tn5 was diluted 1:10–1:100 and combined with (20 mM Tris-HCl pH7.5, 20 mM MgCl<sub>2</sub>, 50% DMF) and diluted DNA in a 1:2:1 volume ratio at total volume of 5  $\mu$ L. The mix was incubated at 55 °C for 3 min and afterwards shifted to ice and inactivated by the addition of 1.25  $\mu$ L 0.2% SDS. PCR was performed by adding 6.75  $\mu$ L of KAPA HiFi HS mastermix, 0.75  $\mu$ L of DMSO and 1.25  $\mu$ L each of the forward i5 and reverse i7 library primer (Supplementary Data 4) at 10  $\mu$ M. PCR program was 72 °C 3 min, 95 °C 30 sec, 12 cycles of [98 °C, 20 sec, 58 °C, 15 sec, 72 °C 30 sec], and 72 °C, 3 min.

**Single-cell cultures.** Bone Marrow mononuclear cells from patient P1 were stained and Lin<sup>-</sup> or Lin<sup>-</sup>CD34<sup>+</sup> single cells were index-sorted into ultra-low attachment 96-well plates (Corning) containing 100  $\mu$ L StemSpan SFEM media (Stem Cell Technologies). Media was supplemented with penicillin/streptomycin (100 ng/mL), l-glutamine (100 ng/mL) 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), G-CSF (20 ng/mL), EPO (40 ng/mL), IL-5 (20 ng/mL),

M-CSF (20 ng/mL), GM-CSF (50 ng/mL). After 21 days at 5% CO<sub>2</sub> and 37 °C, colonies were imaged by microscopy, and processed as detailed in the following.

**Targeted DNA sequencing by nested PCR amplification.** Single-cell derived colonies were transferred into 50  $\mu$ L buffer RLT (Qiagen). Cleanup was performed using CleanPCR beads (CleanNA) at a 1.8x volume ratio and eluted in 20  $\mu$ L 10 mM Tris-HCl pH 7.8. 4.5  $\mu$ L were transferred to a PCR plate containing 7.5  $\mu$ L Kapa HiFi HS mastermix and 3  $\mu$ L of a pool of all outer primers (Supplementary Data 4, each primer at 0.5  $\mu$ M) were added, followed by a PCR program of 98 °C 3 min, 30 cycles of [98 °C, 20 sec, 63 °C, 60 sec, 72 °C 10 sec] and 72 °C, 5 min and subsequent enzymatic cleanup with 2.5  $\mu$ L 10x ExoI buffer, 0.4  $\mu$ L ExoI (NEB) and 0.4  $\mu$ L FastAP (ThermoFisher), 30 min incubation at 37 °C and 5 min inactivation at 95 °C. Afterwards, 1  $\mu$ L was transferred to a PCR tube containing 5.9  $\mu$ L water, 7.5  $\mu$ L Kapa HiFi HS mastermix and 0.6  $\mu$ L of a pool of all inner primers (Supplementary Data 4, each primer at 0.5  $\mu$ M), followed by a PCR program of 98 °C 3 min, 15 cycles of [98 °C, 20 sec, 65 °C, 15 sec, 72 °C 30 sec], 72 °C, 5 min and enzymatic cleanup as above. One microliter was then transferred into a PCR with Nextera indexing primers (Supplementary Data 4) and amplified with 98 °C 3 min, 10 cycles of [98 °C, 20 sec, 60 °C, 15 sec, 72 °C 30 sec] and 72 °C, 5 min.

**Processing of next generation sequencing data.** Raw sequencing reads from MutaSeq and Smart-seq2 experiments were processed using the BBDuk software to trim both the standard Illumina Nextera adapters and the ISPCR adapter. Reads were then mapped to the hg38 human genome (Ensembl release 95) using STAR v2.6 (ref. 57), with the outFilterMismatchNmax parameter set to 5. Exonic gene counts were tabulated, keeping only reads that did not overlap with targeted regions, overlapped with only one annotated gene, and with lengths greater than 30 nt. For the colony DNA sequencing experiment, reads were mapped to the hg38 human genome using bwa mem (v0.7.17)<sup>49</sup>.

**Analysis of mitochondrial mutations and reconstruction of clonal hierarchies (mitoClone package).** The following set of routines are implemented in the mitoClone package available from <https://github.com/veltenlab/mitoClone>, and documented further in the package vignettes.

**Construction of allele count tables.** For each cell and position of mitochondrial genome as well as other genomic sites of interest, count tables for all nucleotides (A/C/G/T), and deletions were created from the BAM files. To this end, the mitoClone package implements the baseCountsFromBamList function, which essentially serves as a wrapper to the bam2R function from the deepSNV R package<sup>58</sup>.

**Filtering of mitochondrial variants.** To identify relevant somatic variants, we implemented the *mutationCallsFromCohort* function. In short, we select coordinates in the mitochondrial genome containing at least five reads each in at least 20 cells. To distinguish RNA editing events and true mitochondrial mutations at the level of genomic sites, we then identify mitochondrial variants that occur in several individuals. For this purpose only, individual cells are called as “mutant” in a given site of the mitochondrial genome if at least 10% of the reads from that cell were from a minor allele (i.e., distinct from the reference). Mutations present in at least 1% of cells in a given patient, but no more than 10 cells in any other individual, are then included into the final data set and counts supporting the reference and mutant alleles are computed as for sites of interest in the nuclear genome. Mutations present in several individuals are stored as a blacklist and were used further for filtering some of the data analyzed in Supplementary Fig. 4. Importantly, the result from this step is simply a list of genomic sites that are likely to display genetic variability across single cells. The vignette “Variant calling and blacklist creation of the mitoClone package” provides further recommendation for the choice of filtering parameters.

**Construction of clonal hierarchies.** To construct a basic clonal hierarchy, we implemented the *muta\_cluster* function. In short, all nuclear and mitochondrial variants with coverage in at least 20% of cells are selected. Observed variant allele frequencies (VAF) are then computed for each cell. From these values, we compute a ternary matrix of observed variant calls  $N_{c,g}$ ; for each cell  $c$  and genomic site  $g$ , cells were classified as mutant (1) if the VAF was above 5%, reference (0) if it was below 5%, or dropout (NA) if the site was not covered. We thereby assign a genotype “mutant” or “non-mutant” to single cells at all genomic sites selected in the filtering step, using a less stringent cutoff for calling the site as mutant. Then, PhISCS<sup>36</sup> is used to reconstruct a clonal tree. Unlike conventional algorithms for the reconstruction of phylogenetic hierarchies, PhISCS is very robust with regard to noise. For the figures presented in the main text, we ran PhISCS assuming a false-positive rate (FPR) of 3% and an allelic dropout rate (AD) of 10% across all genes. We additionally estimated the dropout rate on a per-gene level from the number of complete dropouts, and varied the resulting parameter vector using Latin hypercube sampling around the means using a beta distribution with concentration parameter of 10. Across 80 sampling runs, the same tree was consistently obtained except that the order of nodes within clones was swapped (Supplementary Data 5). PhISCS results are, therefore, robust to variations in the parameters of the

statistical technique used. The false-positive rate of MutaSeq had been empirically estimated using spike-in controls (Supplementary Fig. 10).

**Clustering of mutations into clones and assignment of cells to clones.** PhISCS provides a maximum likelihood phylogenetic tree, enforcing an ordering of mutations. In reality, however, not all intermediate evolutionary steps are represented by cells present in the biological sample (for example, in P1, there are few or no cells displaying the *SRSF2* mutation, but not the mt:11559 G > A mutation). Hence, the order of the nodes in the maximum likelihood tree is to some extent arbitrary and driven by noise; even if there is some statistical support for a specific order, it may be attractive in practice to merge mutations into clones so as to obtain a biologically meaningful, interpretable analysis. We therefore implemented the clusterMetaclones function, which employs a likelihood-based approach. The maximum likelihood tree is split into contiguous linear branches (e.g., for P1, nuc: *SRSF2*, mt:11559G > A, mt:7527DEL and nuc:*EAPP* constitute one such branch). Within each branch, all nodes are then swapped with each other and the likelihood of the data given the altered structure is calculated using the PhISCS model:

$$\mathcal{L} = \prod_c \prod_g \begin{cases} \alpha \text{ if } M_{cg} = 1 \ \& \ N_{cg} = 0 \\ \beta \text{ if } M_{cg} = 0 \ \& \ N_{cg} = 1 \\ 1 \text{ if } N_{cg} = \text{NA} \\ (1 - \alpha) * (1 - \beta) \text{ else} \end{cases} \quad (1)$$

Here,  $M_{cg}$  indicates whether according to the model, cell  $c$  is mutant at genomic site  $g$ .  $\alpha$  is the allelic dropout rate and  $\beta$  is the false-positive rate.

The branch is then split into clones such that within each clone, the average difference in log-likelihood incurred by swapping nodes was smaller than 1 per cell. This threshold is set arbitrarily to obtain a practically useful grouping of mutations into clones. The vignette “Computation of clonal hierarchies and clustering of mutations” of the mitoClone package provides further practical recommendations.

The same model was used to compute the likelihood of clonal assignments for each cell. For the analysis in Supplementary Fig. 3b, this estimate was then transformed to bits of information using the Kullback–Leibler distance:

$$D_{\text{KL}} = \sum_c \mathcal{L}_c \log_2(\mathcal{L}_c * |C|) \quad (2)$$

where  $|C|$  is the total number of clones.

**Quantitative analyses.** For all quantitative analyses of clones (differential expression testing, contribution of clones to cell types) cells with a likelihood of clonal assignment of <0.8 were removed.

**De novo mutation calling.** To identify nuclear mutations associated with the clones, we performed variant calling at a list of candidate sites from COSMIC<sup>37</sup>. We therefore focused on a subset of COSMIC including putatively pathogenic SNVs and small InDels, variants in expressed genes (mean > 20 reads per cell), and variants associated with a primary site “haematopoietic\_and\_lymphoid\_tissue”, obtaining a list of 13,797 sites of interest. SNVs commonly observed as germline variants in the 1000 genomes project data set were removed<sup>59</sup>. Allele count tables were created for each site as described above. Finally, a beta-binomial model with the same probability for mutant in all cells was compared to a beta-binomial model with a different probabilities for mutant in each clone using Akaike’s Information Criterion.

**Single-cell gene expression data analysis.** Cells with <500 distinct genes observed and genes that appeared in <5 cells were removed. Additionally, data from a healthy individual (“H1”) was downloaded from the NCBI Gene Expression Omnibus (GSE75478). Data from all individuals was then integrated using scanorama according to the workflow described by the authors<sup>60</sup>. The low-dimensional data representation obtained by scanorama was then loaded into Seurat<sup>61</sup>, and default Seurat implementations of UMAP<sup>62,63</sup> and graph-based clustering were used for data visualization and clustering, respectively, using the first 15 scanorama components.

For more detailed analyses of the T/NK-cell and HSC/MPP populations, cells with these identities were selected and the MNN<sup>64</sup> data integration workflow was repeated using raw expression counts of these specific cell populations as input.

Differential expression testing was performed using MAST<sup>65</sup>, using a linear model containing the variable of interest (e.g., clonal identity), a library quality covariate (i.e., the number of genes observed per cell), and, when applicable, additional covariates accounting for patient and cell type.

For the display of gene expression values only, data were normalized according to the Seurat defaults (i.e., divided by the total count of RNA in the cell, multiplied by a scale factor of 10,000 and log-transformation).

**Data visualization.** All plots were generated using the ggplot2 (v. 3.2.1) and pheatmap (v. 1.0.12) packages in R 3.6.2. Boxplots are defined as follows: the middle line corresponds to the median; lower and upper hinges correspond to first and third quartiles. The upper whisker extends from the hinge to the largest value no further than 1.5 \* IQR from the hinge (where IQR is the inter-quartile range, or

distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 \* IQR of the hinge. Data beyond the end of the whiskers are called “outlying” points and are plotted individually<sup>66</sup>.

**Statistics and reproducibility.** Statistical analyses were performed using R 3.6.2. Statistical details for each experiment are provided in the figure legends. FlowJo v10 TreeStar was used for the analysis of flow cytometry data.

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

## Data availability

Count tables and other processed data necessary to reproduce all analysis from the manuscript are deposited in figshare with <https://doi.org/10.6084/m9.figshare.12382685.v1> (ref. 67). Raw sequencing data are deposited under a Data Access Agreement to protect patient privacy in the European Genome-Phenome Archive with the accession id EGAS00001003414. Requests for data access shall be addressed to LMS (larsms@embl.de). The following restrictions apply: Research with the goal of identifying characteristics of the patient not related to the leukemia (such as surname inference and ancestry research) are excluded. The use of the data for projects not related to cancer research is excluded, exceptions may apply in the context of research aiming to develop new bioinformatics methods. Sequencing data from the healthy individuals are deposited in GEO with the accession id GSE75478. The source data underlying Figs. 1f, 5d, 5h, and Supplementary Fig. 8c–e are provided as a Source Data file. All the other data supporting the findings of this study are available within the article and its supplementary information files and from the corresponding author upon reasonable request. Source data are provided with this paper.

## Code availability

Code for MutaSeq primer design is available at <https://github.com/veltenlab/PrimerDesign><sup>68</sup>. A package for data analysis is available at <https://github.com/veltenlab/mitoClone><sup>69</sup>.

Received: 2 June 2020; Accepted: 5 February 2021;

Published online: 01 March 2021

## References

- Clevers, H. What is an adult stem cell? *Science* **350**, 1319–1320 (2015).
- Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–337 (2013).
- Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
- Shlush, L. I. et al. Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).
- Greaves, M. Leukaemia ‘firsts’ in cancer research and treatment. *Nat. Rev. Cancer* **16**, 163–172 (2016).
- Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
- Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
- Bowman, R. L., Busque, L. & Levine, R. L. Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell Stem Cell* **22**, 157–170 (2018).
- Trumpp, A. & Wiestler, O. D. Mechanisms of disease: cancer stem cells—targeting the evil twin. *Nat. Clin. Pract. Oncol.* **5**, 337–347 (2008).
- Essers, M. A. G. & Trumpp, A. Targeting leukemic stem cells by breaking their dormancy. *Mol. Oncol.* **4**, 443–450 (2010).
- Pollyea, D. A. & Jordan, C. T. Therapeutic targeting of acute myeloid leukemia stem cells. *Blood* **129**, 1627–1635 (2017).
- Hosen, N. et al. CD96 is a leukemic stem cell-specific marker in human acute myeloid leukemia. *Proc. Natl Acad. Sci. USA* **104**, 11008–11013 (2007).
- Majeti, R. et al. Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc. Natl Acad. Sci.* **106**, 3396–3401 (2009).
- Majeti, R. et al. CD47 is an adverse prognostic factor and therapeutic antibody target on human acute myeloid leukemia stem cells. *Cell* **138**, 286–299 (2009).
- Jan, M. et al. Prospective separation of normal and leukemic stem cells based on differential expression of TIM3, a human acute myeloid leukemia stem cell marker. *Proc. Natl Acad. Sci. USA* **108**, 5009–5014 (2011).
- Eppert, K. et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* **17**, 1086–1093 (2011).
- Ng, S. W. K. et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**, 433–437 (2016).

18. Pabst, C. et al. GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood* **127**, 2018–2027 (2016).
19. Raffel, S. et al. BCAT1 restricts aKG levels in AML stem cells leading to IDHmut-like DNA hypermethylation. *Nature* **551**, 384–388 (2017).
20. Zink, F. et al. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
21. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
22. Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
23. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
24. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
25. Giustacchini, A. et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
26. Filbin, M. G. et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331–335 (2018).
27. van Galen, P. et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281.e24 (2019).
28. Nam, A. S. et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature* **571**, 355–360 (2019).
29. Petti, A. A. et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat. Commun.* **10**, 3660 (2019).
30. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
31. Rodriguez-Meira, A. et al. Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. *Mol. Cell* **73**, 1292–1305.e8 (2019).
32. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
33. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
34. Xu, J. et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife* **8**, 1–14 (2019).
35. Coban-Akdemir, Z. et al. Identifying genes whose mutant transcripts cause dominant disease traits by potential gain-of-function alleles. *Am. J. Hum. Genet.* **103**, 171–187 (2018).
36. Malikic, S. et al. PHISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Res.* **29**, 1860–1877 (2019).
37. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
38. Assi, S. A. et al. Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat. Genet.* **51**, 151–162 (2019).
39. Shlush, L. I. et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
40. Izzo, F. et al. DNA methylation disruption reshapes the hematopoietic differentiation landscape. *Nat. Genet.* **52**, 378–387 (2020).
41. Pina, C., May, G., Soneji, S., Hong, D. & Enver, T. MLLT3 regulates early human erythroid and megakaryocytic cell fate. *Cell Stem Cell* **2**, 264–273 (2008).
42. Uckelmann, H. J. et al. Therapeutic targeting of preleukemia cells in a mouse model of NPM1 mutant acute myeloid leukemia. *Science* **367**, 586–590 (2020).
43. Brunetti, L. et al. Mutant NPM1 maintains the leukemic state through HOX expression. *Cancer Cell* **34**, 499–512.e9 (2018).
44. Vassiliou, G. S. et al. Mutant nucleophosmin and cooperating pathways drive leukemia initiation and progression in mice. *Nat. Genet.* **43**, 470–475 (2011).
45. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
46. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* **44**, D1023–D1031 (2016).
47. Lareau, C. A. et al. Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-020-0645-6> (2020).
48. Schallmoser, K. et al. Rapid large-scale expansion of functional mesenchymal stem cells from unmanipulated bone marrow without animal serum. *Tissue Eng. Part C. Methods* **14**, 185–196 (2008).
49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at *arXiv* <https://arxiv.org/abs/1303.3997> (2013).
50. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
52. Untergasser, A. et al. Primer3 — new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
53. Östergård, P. R. J. A fast algorithm for the maximum clique problem. *Discret. Appl. Math.* **120**, 197–207 (2002).
54. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
55. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
56. Hennig, B. P. et al. Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3* **8**, 79–89 (2018).
57. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Gerstung, M. et al. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* **3**, 811 (2012).
59. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
60. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
61. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
62. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–47 (2019).
63. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* <https://arxiv.org/abs/1802.03426> (2018).
64. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
65. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
66. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
67. Velten, L., Story, B. A. & Steinmetz, L. M. Dataset for the manuscript “Identification of leukemic and pre-leukemic stem cells by clonal tracking from single-cell transcriptomics”. *figshare* <https://doi.org/10.6084/m9.figshare.12382685.v1> (2021).
68. Velten, L., Story, B. A. & Steinmetz, L. M. PrimerDesign package v1.0. <https://doi.org/10.5281/zenodo.4443088> (2021).
69. Velten, L., Story, B. A. & Steinmetz, L. M. mitoClone package v1.0. <https://doi.org/10.5281/zenodo.4443074> (2021).

## Acknowledgements

We thank Niko Beerenwinkel for discussions, and Andreas Gschwind for contributing code for primer design. We thank the members of the Steinmetz and Haas labs for discussions and support, and we thank the EMBL and DKFZ Genomics core facilities, and the EMBL flow cytometry core facility for technical support. This project was financially supported by the Deutsche José Carreras Leukämie Stiftung grant DJCLS 20R/2017 (to L.V., S.H., L.M.S., and A.T.), the Emerson foundation grant 643577 (to L.V. and L.M.S.) and the German Bundesministerium für Bildung und Forschung (BMBF) through the Juniorverbund in der Systemmedizin “LeukoSysTem” (FKZ 01ZX1911D to L.V., S.H., and S.R.). Contributions by S.R. were further supported by Emmy Noether Fellowship RA 3166/1-1 (DFG). Contributions by C.P. were supported by a Max-Eder Grant (German Cancer Aid 70111531). Contributions by D.N., J.C.J., W.K.H., and T.B. were supported by the Gutermuth Foundation, the H.W. & J. Hector fund, Baden-Württemberg, and the Dr. Rolf M. Schwiete Fund, Mannheim. D.N. is an endowed professor of the Deutsche José Carreras Leukämie Stiftung (DJCLS H 03/01).

## Author contributions

L.V. developed MutaSeq and performed single-cell RNA-seq experiments with assistance by J.M., D.R.L., and C.S.T. P.H.M. performed sample characterizations, single-cell culture experiments, and FACS sorting with support by M.P., A.D., S.R., and supervision by S.H. B.A.S. and L.V. developed mitoClone and analyzed the data with contributions from R.F. L.V., S.H., A.T., and L.M.S. conceived the study. L.V. and B.A.S. wrote the manuscript with contributions from P.H.M., S.H., S.R., and L.M.S. C.L., D.N., J.C.J., C.P., T.B., W.K.H., and C.M.T. collected samples and performed initial sample characterization. All authors have read and commented on the manuscript.

## Competing interests

L.M.S. is co-founder of Sophia Genetics and Levitas Bio and consultant for several companies on genetic analysis. All other authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-21650-1>.

**Correspondence** and requests for materials should be addressed to L.V. or L.M.S.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021