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

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

1. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
2. Universitat Pompeu Fabra (UPF), Barcelona, Spain
3. European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany
4. Swiss Federal Institute of Technology (ETH) Zurich, Department of Biosystems Science and Engineering, 4058 Basel, Switzerland
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. Department of Internal Medicine V, Hematology, Oncology and Rheumatology, University of Heidelberg, 69120 Heidelberg, Germany
8. European Molecular Biology Laboratory (EMBL), Flow Cytometry Core Facility, 69117 Heidelberg, Germany
9. Department of Hematology and Oncology, Medical Faculty Mannheim, Heidelberg University, 68167 Mannheim, Germany.
10. Molecular Medicine Partnership Unit (MMPU), University of Heidelberg and European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
11. German Cancer Consortium (DKTK), 69120 Heidelberg, Germany
12. Berlin Institute of Health (BIH), 10178 Berlin, Germany
13. Charité-Universitätsmedizin, 10117 Berlin, Germany
14. Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 10115 Berlin, Germany
15. Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA
16. Stanford Genome Technology Center, Palo Alto, California 94304, USA

* Correspondence should be addressed to lars.velten@crg.eu or larsms@embl.de

§ These authors contributed equally: Lars Velten, Benjamin A. Story and Pablo Hernández-Malmierca

Abstract:

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.

Keywords: Acute Myeloid Leukemia / Cancer evolution / Cancer stem cells / Single-cell genomics

Introduction

Tissues with high cellular turnover, such as the hematopoietic system or the intestine, depend on 'professional' adult stem cells for their continuous regeneration[1]. 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[2–4]. Due 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[5]. 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)[6,7]. 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[8]. 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[9–11]. 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[12–15]. More recently, leukemic engraftment rates in xenotransplant models were correlated with gene expression[16–19]. 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) (Figure 1a). Finally, pre-LSCs are thought to typically carry mutations associated with CHIP (e.g. *Dnmt3a*) but not mutations associated with leukemia (e.g. *Npm1*)[7,20,21], 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[22–24], tracking mutations or clones in single-cell gene expression data remains difficult. Previous work has amplified somatic mutations from cDNA[25–28], extracted mutational information from single-cell RNA-seq reads[29], or processed both genomic DNA and RNA from single cells[30–32]. 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[33,34]. 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 allowed us to distinguish HSCs, pre-LSCs, LSCs, and progenitor/blast populations. Thereby, we identified transcriptomic consequences of leukemic and pre-leukemic mutations relevant to stem cells. Additionally, we characterized the contribution of different leukemic and pre-leukemic clones to healthy and disease-specific bone marrow populations with unprecedented detail. Together, 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 (Figure S1a). 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 (Figure S1a-d). In 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 (Figures 1b,c, and S1a,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 Methods section). MutaSeq stably works with up to 30-40 primer pairs (targets); when higher numbers of primers are included, library quality progressively decreases (Figure S1e). In a test with only highly expressed target genes, target amplicons are created from all primer pairs in virtually all single cells (Figure S1f,g).

121 To evaluate MutaSeq, we performed deep exome sequencing of an AML patient (P1) and designed
122 primers targeting 14 nuclear mutations (Table S1, S5). We then systematically compared the
123 performance of MutaSeq and non-targeted Smart-seq2 on CD34+ cells from this patient. MutaSeq
124 increased the number of target sites covered per single cell from a median of 1 to a median of 4 (Figure
125 1c-e) and recapitulated the variant allele frequencies estimated by exome sequencing with higher
126 accuracy than Smart-seq2 (Figure 1f), while maintaining comparable transcriptome data quality
127 (Figure S1h,i). Both methods underestimated the abundance of frameshift mutations, possibly as a
128 consequence of nonsense mediated decay[35] (Figure 1e,f). While our data do not provide statistical
129 evidence for an effect of target gene length or sequence complexity on dropout, we cannot exclude
130 such effects.
131 Importantly, both methods provide an excellent coverage of the mitochondrial genome (mean ~100X
132 mitochondrial coverage given a mean sequencing depth of ~788,000 total reads per cell, Figure 1g),
133 unlike most other single-cell RNA-seq protocols applied in the context of clonal tracking[27–29,31] (Figure
134 S1j). Together, these results demonstrate that MutaSeq efficiently covers the mitochondrial genome in
135 single-cell RNA-sequencing experiments and provides improved coverage of genomic target sites
136 compared to Smart-seq2. Importantly, it requires no changes to existing Smart-seq2 pipelines, except
137 for the addition of targeting primers during cDNA amplification.
138
139 *Simultaneous mapping of mitochondrial and genomic mutations permits high-confidence tracking of*
140 *leukemic, pre-leukemic and healthy clones*
141 To investigate if MutaSeq can distinguish leukemic, pre-leukemic and residual healthy clones, we
142 generated data from four AML patients with heterogeneous genotypes and phenotypes (Figure 2a, and
143 see also Figures 3b, S2). To allow for a better characterization of stem cells in each patient, cells were
144 sorted such that putative stem and progenitor cells (CD34+) and putatively more mature cells (CD34-
145) were approximately covered at equal portions (Figure S2). Of note, two of the patients (P2, P4),
146 exhibited bone marrow consisting of >99% of CD34- cells (Figure S2a). Due to different capture rates
147 across individuals and gates, the final data set consisted of between 618 to 1430 cells per patient, of
148 which between 190 and 968 were CD34+ (Figure S2b,c). We therefore avoid statements on
149 quantitative shifts in population size between patients throughout this manuscript.
150
151 We then called nuclear genomic mutations, as well as mitochondrial mutations, at the single-cell level
152 in order to cluster cells into clonal hierarchies. Bulk exome sequencing of the patients had identified
153 known pre-leukemic mutations present at high allele frequency and known leukemic mutations
154 present at a somewhat lower allele frequency (Figure 2a, Table S1. Patient 1: SRSF2,TET2/CEBPA and
155 SRSF2,TET2/KLF7; Patient 2: DNMT3A/NPM1; Patient 3: SRSF2,IDH2; Patient 4: leukemic Trisomy 8
156 and BRAF mutations). While some statements on clonal hierarchies could be drawn solely based on
157 calls of these nuclear somatic mutation (Figure S3a), the relatively high dropout of these sites impeded
158 robust assignments of cells to clones (Figure S3b-d). Moreover, the result was biased by the expression
159 levels of the mutated genes of interest: in cells with low expression, dropout was higher, leading to a

160 higher fraction of false negative calls, i.e. false classifications of mutant cells as reference (Figure S3c-
161 d and see also Figure 1e). Similar issues were faced by other methods using related approaches of
162 mutation amplification from cDNA[27,28].
163
164 To overcome these limitations, we next determined if mitochondrial mutations can be used to refine
165 clonal hierarchies jointly with the nuclear mutations. Since mitochondrial RNA is extensively edited, it
166 is important to cluster cells based solely on mutations, and not on seemingly polymorphic sites that
167 are the result of post-transcriptional events with no relationship to clonal structure. Here we
168 developed and tested a filtering strategy that only makes use of mitochondrial mutations for clonal
169 tracking if they uniquely occur in individual patients (Figure S4a, see also Methods). This idea assumes
170 that RNA editing events are typically shared between individuals whereas somatic mutations are not.
171 Using the whole exome sequencing data, we validated that this approach, at the level of genomic sites,
172 correctly distinguishes mutations and RNA editing events with a precision of 97% (Figure S4b,c). We
173 further analyzed various control datasets with known associations between mitochondrial mutations
174 and clones[33] to validate that this approach enables the detection of relevant mitochondrial mutations
175 without a need for a DNA-based reference, and further enables the unsupervised identification of
176 clones (Figure S4d-f). We have implemented all the required filtering and blacklisting routines
177 required for the identification of high-confidence somatic mitochondrial variants in the *mitoClone* R
178 package (https://github.com/veltenlab/mitoClone).
179
180 We then computed clonal hierarchies from both nuclear and mitochondrial somatic variants using a
181 mathematical model that accounts for allelic dropout[36] (see Methods for detail, all required tools are
182 contained in the *mitoClone* package). In patients P1 and P2, pre-leukemic as well as sub-clonal
183 leukemic mutations were significantly associated with distinct sets of well-covered mitochondrial
184 variants (Figure S5a,b), such that clonal hierarchies could be delineated, and a confident assignment
185 of cells to clones became possible (Figure 2b,c and see Figure S3b for a quantitative analysis). Unlike
186 genomic mutation calling from cDNA, identification of clonal identities from mitochondrial mutations
187 is mostly not, and in one case weakly, affected by gene expression levels or library quality (Figure S5c
188 and see also S3c,d), and is possible at lower sequencing depths (Figure S5d-f), since mitochondrial
189 genes are consistently highly expressed.
190 Importantly, the clonal structure was validated by targeted DNA-sequencing from single-cell derived
191 colonies (Patient 1, Figure 2d). Across the patients, we identified clones carrying known pre-leukemic
192 mutations (e.g. SRSF2, DNMT3A) and sub-clones carrying known leukemic mutations (e.g. CEBPA,
193 NPM1) (Figure 2e,f). Below we functionally characterize these clones as leukemic or pre-leukemic
194 based on their contribution to healthy blood production (see Figure 4).
195
196 In patient P3, allele frequencies of leukemic and pre-leukemic mutations were near 50%, indicating
197 the presence of a single leukemic clone (Table S1). In patient P4, an 18-year old individual with a
198 leukemia driven by triplication of chromosome 8, no mitochondrial markers were identified, even

199 though exome sequencing suggested the presence of sub-clonal variants (Table S1). In this case,
200 MutaSeq still permits a qualitative analysis using genomic mutations alone (see below). Taken
201 together, our approach allows for the identification of putatively leukemic, pre-leukemic, and healthy
202 clones and can assign cells to clones with high confidence if mitochondrial somatic variation is present.
203
204 *Identification and characterization of clones* de novo
205 We next investigated if the use of mitochondrial somatic variants enables the identification and
206 characterization of clones without prior knowledge of nuclear mutations. To that end, we made use of
207 a dataset from patient P1 generated without amplification of nuclear sites (i.e. standard Smart-seq2).
208 A clear clonal structure was identified in an unsupervised manner based solely on mitochondrial
209 variants (Figure S6a). In order to examine whether the presence of somatic genetic variability is
210 associated with the different clones, we then queried the mutational status of 13,797 genomic sites
211 annotated as mutated in AML in the COSMIC database[37] using a beta-binomial model (see Methods).
212 This unsupervised analysis revealed a highly significant association of the SRSF2 P95H mutation with
213 the leukemic and pre-leukemic sub-clones (Figure S6b,c). Their malignant nature was further
214 evidenced by a markedly reduced ability to contribute to the T cell lineage (Figure S6d and see also
215 below).
216
217 To further demonstrate our ability to identify clones *de novo*, we highlight a clonal expansion of non-
218 leukemic cells in P2 (Figure 2f). This clones would have been missed by approaches relying on genomic
219 mutations alone[27,28,31]. Interestingly, these cells were not associated with the pre-leukemic *DNMT3A*
220 mutation. By again querying sites from the COSMIC database[37] using a beta-binomial model, we
221 identified that they had uniquely acquired a mutation in the *RPL3* gene (Figure S6b,e). These results
222 suggest that this clonal expansion event is independent of the leukemia and associated with the
223 acquisition of unrelated nuclear mutations.
224 We also take note of a putative non-leukemic clone in P1 marked by a single mitochondrial variant
225 (5492T>C). With one exception, all cells carrying this variant are positive for the T cell marker CD3
226 (Figure 2b, see also Figure 4a), but they display diverse TCR alpha and beta chain sequences (not
227 shown). Hence, this variant was likely acquired in a T cell precursor, although we cannot formally
228 exclude that it corresponds to a T cell-specific RNA editing event,
229
230 Taken together, these results demonstrate that our approach can identify and characterize clones *de*
231 *novo* without prior knowledge of nuclear genomic mutations. The *mitoClone* package implements all
232 routines for clonal clustering and mutation calling.
233
234 *A map of cell states in leukemic bone marrow identifies HSC-like cells*
235 We next used single-cell transcriptome data to distinguish stem cells, progenitor cells and leukemic
236 blasts. To define cell populations in our samples, we integrated the gene expression data from all four
237 patients along with data from CD34+ bone marrow cells of a healthy individual[22] into a two-

238 dimensional representation, and characterized cell types based on marker gene expression (Figure 3a,
239 S7a,b,c,f, Table S2).
240
241 All four bone marrow samples contained cells that clustered with HSCs and multipotent progenitor
242 cells (HSC/MPP) from the healthy individual (Figure 3b-d) and displayed a $CD34^{+}CD38^{low}$, FSC/SSC^{low}
243 phenotype (Figure 3e). Unsupervised analysis separated these cells into quiescent immature HSC-like
244 cells (*HLF*+), more proliferative erythromyeloid primed progenitors (*GATA2*+), and lymphomyeloid
245 primed progenitors (*FLT3*+) (Figure 3f, S7f). Cells resembling healthy erythroid progenitors and MEPs
246 were also identified, alongside various types of B-cell precursors and T/NK cells (Figure S7b,c).
247
248 In contrast to non-leukemic populations, the majority of cells from leukemic bone marrow samples
249 ('blasts') were very different between the patients, in line with previous observations[27] (Figure 3d):
250 we observed a) differentiated blasts ($CD34\text{-}SSC/FSC^{hi}$) that expressed neutrophil genes such as
251 calprotectin (S100A8, S100A9) and *AZU1 (*Figures 3e, S7a); b) CD34+ blasts that were highly mitotic
252 and expressed the fetal hemoglobin *HBZ*; and c) $CD34^{+}CD38^{low}$ blasts expressing markers typical of
253 hematopoietic stem and progenitor cells (HSPCs) such as *PROM1* (CD133) and *MEIS1* (Figure S7a,d,e).
254 The latter population appeared to be connected to the HSC/MPP population across a continuum of
255 states that gradually upregulate AP1 transcription factors (FOS, JUN, FOSB, JUNB, JUND) while down-
256 regulating MHC class II (Figure 3g,h, Table S3). A global analysis of highly expressed genes across all
257 populations revealed that high expression levels of AP1, as well as several Krüppel-like factors and
258 poly-A binding proteins, were common to all different blast populations from the patients and
259 distinguished them from healthy progenitors (Figure 3h,i). Of note, high AP1 and KLF activity have
260 recently been identified as a hallmark conserved across genetically distinct types of AML by bulk-
261 sequencing studies that only investigated blast populations[38]. Our results demonstrate that indeed
262 these transcription factors appear to be relevant in all, phenotypically very different, blast populations.
263
264 Taken together, all four leukemia samples could be stratified into stem cells, progenitors, and blasts.
265 Furthermore, all patients retain cells highly similar to healthy HSCs, although in variable abundance.
266
267 *Clonal tracking identifies cellular differentiation states and gene expression patterns associated with pre-*
268 *leukemic and leukemic mutations*
269 Gene expression information alone was insufficient to distinguish cancerous from non-cancerous
270 HSPCs (Figure S7f). To definitively characterize cells as (pre-)LSCs or residual healthy cells, we
271 therefore integrated single-cell gene expression data and clonal tracking results (Figure 4a-d). If
272 mitochondrial somatic variability was present, we were able to assign clonal identities with high
273 confidence, allowing us to draw quantitative statements (Patients P1+P2). In the absence of
274 mitochondrial somatic variability, we used nuclear mutation calls in *SRSF2*, *IDH2* and *NUP188* for
275 purely qualitative statements (Patients P3+P4). The capture rates of these marker sites ranged from
276 70% (SRSF2) to 11% (NUP188) (Figure S8a).

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. In contrast, clones associated with pre-leukemic mutations were found in all lineages, but mostly displayed a decreased prevalence in lymphoid lineages (Figure 4e, S8b-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 [4,27,39]. 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[6,7]. 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*) (Figure 4b, 5a). These results are in line with recent findings from DNMT3A knock out mouse models[40]. Interestingly, independent of cell state, this coincided with an upregulation of *MLLT3*, a gene whose enforced expression promotes erythroid-megakaryocytic output from HSCs [41] (Figure 5b,c, Table S3).

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[21]. 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[42–44] (Figure 5b,c). Independent of cell state, these cells further exhibited upregulation of *CD96*, which has previously been identified as a leukemia stem cell specific marker[12]. *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 (Figure S8e).

Mutations in *KLF7* are not commonly observed in leukemia; however, krüppel-like factors are highly expressed by genetically and phenotypically different blasts (see above and ref. [38]). In patient P1, the KLF7 mutated clone displayed a higher proportion of cells in G2/M phase (Figure 5d). Based on a reanalysis of ATAC-seq data from human CD34+ cells[45] we found that enhancers containing KLF7

316 binding sites were enriched near tumor suppressor genes[46] including *CDK6*, *PTEN*, *RUNX1* and *FLT3*,
317 compared to active enhancers not containing KLF7 binding sites (p = 0.002, Figure S8f).
318
319 In sum, we have used a small heterogeneous patient cohort to demonstrate, as a proof-of-concept, that
320 the acquisition of specific mutations is frequently linked to an altered gene expression program, which
321 is consistent with data obtained from mouse models[40,42–44].
322
323 Finally, we compared gene expression between all (pre-)leukemic and non-leukemic cells, with the
324 goal of identifying potential markers or drug targets present in all (pre-)leukemic cells, but not in
325 residual healthy cells. In patient P2, our analysis identified a small number of hits (Figure 5e, Table
326 S3), most notably *FOS*, which was expressed in all (pre-)leukemic cells across cell types, but not by
327 non-leukemic HSC/MPPs (Figure 5f), and the MHC class I genes *HLA-L and HLA-H*, which were
328 expressed by all healthy cells, but not by (pre-)leukemic cells. Across all patients, *FOS* was consistently
329 overexpressed in cells carrying (pre-)leukemic lesions, both in HSCs/MPPs, and in other cell types
330 (Figure 5g).
331
332 Taken together, these results demonstrate the ability of MutaSeq and *mitoClone* to delineate
333 developmental and molecular effects of clonal evolution caused by leukemic and pre-leukemic
334 mutations.
335
336 **Discussion**
337 Herein, we have described a joint single-cell transcriptomics and clonal tracking approach (MutaSeq
338 and *mitoClone*) for characterizing LSCs, charting their differentiation capabilities, and mapping the
339 molecular consequences of oncogenic mutations. While single-cell gene expression profiling permits
340 the identification of cells with a stem-cell signature, clonal tracking using genomic and mitochondrial
341 mutations allows for a clean separation between healthy and cancerous clones. Thereby we distinguish
342 LSCs, HSCs, pre-LSCs, healthy progenitors, and blasts. We have demonstrated this approach in the
343 context of acute myeloid leukemia, and we propose that similar approaches may be applied to other
344 types of cancers.
345
346 By applying our approach to bone marrow samples from four AML patients, we have demonstrated its
347 capabilities:
348 *Charting the differentiation capacities of hematopoietic clones.* By quantifying the contribution of (pre-
349)leukemic cells to blood lineages, we have shown that in patients P1-P3, leukemic clones not only form
350 blasts but also exist in an HSC-like state. Pre-leukemic clones additionally contribute to erythroid and
351 lymphoid lineages. Multipotent HSCs are therefore the likely cell of origin in these patients[4]. In patient
352 P4, with one exception, only cells with a mature phenotype displayed leukemic mutations, illustrating
353 that the disease can also be fueled by cells with a committed phenotype[39]. Alternatively, the LSC
354 population in this patient might be rare among CD34+ cells.

355 *Identification of clones de novo.* In the presence of mitochondrial somatic variability, our approach does
356 not rely on previously known nuclear mutations to detect clones. In the example of patient 2, we have
357 thereby identified an expanded clone present at very low frequency in total bone marrow, but highly
358 abundant in the CD34+ fraction. We have demonstrated that this clone is associated with a nuclear
359 mutation in the *RPL3* gene, which we discovered *de novo*. This result illustrates the low clonal
360 complexity of hematopoiesis in individuals of advanced age, which often is not associated with
361 candidate driver mutations[20].
362 *Identification and characterization of LSCs.* Importantly, our approach was capable of identifying
363 leukemic cells highly reminiscent of healthy HSCs. These cells will be important to target
364 therapeutically without ablating their healthy counterparts[9–11]. While stemming from a study cohort
365 of limited size, our data suggest that *FOS* might constitute a potential target for further investigation,
366 as it is expressed throughout the disparate leukemic populations including the most HSC-like cells,
367 pre-LSCs, and blasts, but not in healthy HSCs. Further, we have confirmed that CD96 is a specific
368 marker for LSCs in some patients[12]. Studies in larger cohorts will be required to assess how generally
369 applicable these findings are.
370
371 Taken together, our study expands upon earlier work on the molecular phenotypes of AML blasts[27,38]
372 and constitutes the first detailed characterization of LSCs by single-cell transcriptomics. This advance
373 is owed to three crucial aspects of experimental design.
374 *Enrichment of relevant starting populations.* LSCs are excessively rare and generally present at below
375 0.1% of total bone marrow[16]. In order to characterize these cells by single-cell genomics, a prior
376 enrichment, e.g. by sorting for CD34 expression, is essential. In future studies, this approach can be
377 complemented with markers that also label CD34- LSCs, such as GRP56[18].
378 *Deep transcriptome sequencing.* In some cases, the bulk of leukemic cells displays gene expression
379 signatures highly similar to stem cells, as observed here for the CD34+ blasts of patient P1. The
380 differences between LSCs and residual healthy HSCs are even more subtle. Previous work using
381 shallow, microwell based sequencing of AML cells[27] has not identified differences between LSCs,
382 CD34+ blasts and residual healthy HSCs.
383 *High-confidence clonal tracking.* When available, the use of mitochondrial variants enables the
384 confident assignment of cells to clones, and, thereby, a quantitative analysis of gene expression. In two
385 out of four AML patients, we identified high-confidence, specific mitochondrial genetic markers for
386 pre-leukemic and leukemic sub-clones. In the third patient, allele frequencies of leukemic and pre-
387 leukemic mutations were around 50%, indicating dominance of a single leukemic clone. In the final
388 patient, sub-clonal genomic mutations were observed, but not associated with mitochondrial
389 variability. Interestingly, this patient was only 18 years old and exhibited a leukemia possibly driven
390 by a 'catastrophic' triplication of chromosome 8. The length of the pre-leukemic phase, the buildup of
391 mitochondrial variants accumulated during normal ageing, as well as unknown factors affecting
392 mitochondrial mutation rates might all contribute to the presence of mitochondrial marker mutations.

393 In the context of recent methods for clonal tracking within single cell transcriptomics[27–29,31,33], the use
394 of the MutaSeq protocol and *mitoClone* computational pipeline combines the strengths of previous
395 approaches relying on either nuclear or mitochondrial variants, but also has limitations. Specifically,
396 droplet- or microwell-based protocols for single-cell RNA-seq and clonal tracking[27–29] provide low
397 coverage of mitochondrial genomes and high dropout of nuclear genomic sites, and therefore do not
398 enable a confident assignment of cells to clones. Single-cell ATAC-sequencing[33,34,47] and Smart-seq2
399 efficiently cover mitochondrial genomes, but their coverage of nuclear genomic sites is absent or low,
400 hardening interpretability of the data. Finally, while TARGET-seq addresses the limitation of dropout
401 of nuclear sites, depending on the implementation it does not offer sufficient mitochondrial coverage,
402 or is of very limited throughput. Of all methods available to date, the approach presented here
403 converges on crucial aspects, specifically: a) high mitochondrial coverage, allowing us to identify
404 benign expanded clones *de novo*, even in the absence of known genetic markers, b) decreased dropout
405 of relevant genomic mutations, permitting the association of clones with genomic mutations, if present
406 and c) a highly confident assignment of cells to clones, enabling quantitative analyses of clone-specific
407 gene expression (Figure S9). These capabilities expand the potential applications of our approach to
408 the study of clonal dynamics during ageing and oncogenesis beyond the hematopoietic system.
409
410 *Limitations*. The major limitation of our pipeline is that it requires natural somatic variability to resolve
411 clones at high confidence. In the absence of mitochondrial somatic variation, the MutaSeq protocol can
412 be used to draw qualitative statements on clonal differentiation capacities (similar to ref. [28], and with
413 an improvement over Smart-seq2), but due to dropout neither enables the statistically confident
414 assignment of cells to clones, nor differential expression testing between clones. In the presence of
415 mitochondrial somatic variation, an association between clones and nuclear mutations is only possible
416 for mutations in highly expressed genes, and can be limited by the nature of the mutation (e.g.
417 frameshift mutations) and possibly other factors such as sequence complexity. The third limitation of
418 MutaSeq is its relatively low throughput. This limitation is currently shared with Smart-seq2 and
419 alternative single-cell RNA-seq methods allowing high-confident assignment of cells to clones[31,33].
420 Future work will focus on the inclusion of full-length coverage of the mitochondrial genome in droplet-
421 based single-cell RNA-seq platforms.
422
423
424 **Material and Methods**
425 *Patient and sample collection*
426 The AML samples were collected from diagnostic bone marrow aspirations at the University hospitals
427 in Heidelberg, Germany, and Mannheim, Germany after obtaining informed written consent. Bone
428 marrow mononuclear cells were isolated by density gradient centrifugation and stored in liquid
429 nitrogen until further use. All experiments involving human samples were conducted in compliance
430 with the Declaration of Helsinki and all relevant ethical regulations and were approved by the ethics

431 committees of the medical faculties Heidelberg and Heidelberg-Mannheim of the University of
432 Heidelberg.
433
434 *Deep exome sequencing and target selection*
435 For exome sequencing, DNA was extracted from $9x10^3$ flow sorted CD34+ cells (for CD34+ leukemias:
436 P1, P3) or total bone marrow. As healthy controls, we used a buccal swap (P1), FACS-sorted CD45-
437 CD105+ MSCs (P3) or *in vitro* expanded MSCs (P2 and P4)[5]. Sequencing libraries were constructed
438 using the SureSelect HS XT Target Enrichment System v6 (Agilent), and a mean on-exon sequencing
439 coverage of at least >70X was obtained for each patient. Genomic alignments were performed using
440 BWA MEM v0.7.15[49] and cancer variants were identified using Mutect2 v3.8 (P1 and P3) and v4.0.9
441 (P2 and P4)[50], following the GATK best practice recommendations. Variants were annotated using
442 ANNOVAR[51]. Output from Mutect2 was filtered to remove variants that did not overlap with known
443 genes. The final list of candidate variants included only those with allele frequencies (AF) greater than
444 4% in the cancer exome sample and with an AF four-fold larger than in the healthy exome sample
445 (Table S1). Finally, the candidates for targeting were hand-selected from this list with a focus on cancer
446 relevant genes, highly expressed genes, and potential sub-clonal markers.
447
448
449 *FACS sorting*
450 Bone marrow mononuclear cells were stained for 30 minutes on ice according to standard protocols.
451 For single-cell liquid cultures and MutaSeq, cells were stained with fluorescent-labelled antibodies
452 against lineage markers (CD4, CD8, CD19, CD20, CD41a, CD235a) and additional markers (CD45RA,
453 CD135, GPR56, CD34, CD38, CD90, CD33, Tim3), and sorted according to the gating scheme illustrated
454 in Figure S2. BD FACS Fusion (BD Biosciences) equipped with 405nm, 488nm, 561nm and 640nm
455 lasers were used. Of note, in P1 and P4, Lin+ cells could not be efficiently processed into libraries for
456 unknown reasons and are therefore excluded. A list of all antibodies used can be found in Table S4.
457
458 *Primer Design*
459 Primers for MutaSeq, for other single-cell targeting protocols tested (Figure S1), as well as for targeted
460 DNA sequencing were designed using the computational pipeline available at
461 http://git.embl.de/velten/PrimerDesign . For MutaSeq, the refgene transcripts spanning each genomic
462 site of interest were selected as template; if multiple refgene transcripts were found for one site, a
463 consensus transcript containing only exonic sequences present in all variants was created. We then
464 used primer3[52] to design five possible pairs of primers for each intended target, with an amplicon
465 length of 90-145bp and a melting temperature of (nominally) 60°C. BLAST was used to remove primer
466 pairs which potentially form off-target amplicons. Then, the pair complementarity (i.e. potential to
467 form dimers) was computed for each possible combination of primers across all target sites (forward-
468 reverse, forward-forward and reverse-reverse). In initial experiments, primers with high
469 complementarity scores were found to abundantly form primer dimers (not shown). In order to

470 identify a set of primers that covered the maximal number of genes while strictly forbidding primers
471 with high complementarity scores, a graph was constructed that connected all primers with different
472 targets to each other if their complementarity score was lower than 15. A maximum clique-finding
473 algorithm[53] was then used to identify the largest mutually connected component in the graph. Thereby,
474 the largest number of targets that efficiently avoids dimer formation was selected.
475 Nextera adapters were added to all primers designed accordingly (fwd:
476 GTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG, rev:
477 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG).
478 For targeted DNA-seq experiments, the genomic sequence surrounding the target was used as
479 template and nested PCR primers were designed. Inner primers were designed as in the case of
480 MutaSeq, and outer primers surrounding the inner PCR product with an amplicon length of 200-350bp
481 and a nominal annealing temperature of 58°C were added. A list of all primers used for this study is
482 included in Table S5.
483
484 *Single-cell RNA sequencing with targeting of genomic sites of interest (MutaSeq)*
485 MutaSeq is based on the Smart-seq2 protocol[54,55] with the modifications introduced by[22]. For lysis, we
486 used 5µL of a buffer containing 0.1µL RNAsin+ (Promega), 0.04µL 10% Triton X-100 (SigmaAldrich),
487 0.1µL of 100µM Smart-seq2 Oligo-dT primer (SigmaAldrich) and 1µL dNTP mix (10mM each, NEB). In
488 P1, we had additionally included 0.075µL of a 1:1,000,000 dilution of ERCC spike-in mix 1 (Ambion),
489 as well as a control spike-in to quantify the false positive detection rate of mutations (see Figure S10).
490 Plates were snap frozen directly after sorting and later thawed at 10°C in a PCR machine for 5' and
491 denatured at 72°C for 3'. 5µL of a buffer containing 0.25µL RNAsin+, 2µL 5x SMART FS buffer, 0.5µL
492 DTT 20mM, 1µL SmartScribe enzyme (all TaKaRa) and 0.2µL 50µM Smart-seq2 TSO (Exiqon) were
493 then added and RT was performed for 90' at 42°C, 10 cycles of [50°C, 2' and 42°C, 2'], and enzyme
494 inactivation at 70°C for 15'. Then, we added 15µL PCR mix containing 12.5µL KAPA HiFi HS mastermix
495 (Merck), 0.25µL 10µM Smart-seq2 ISPCR primer (SigmaAldrich) and 0.5µL of a pool of all targeting
496 primers, present at 1µM each. cDNA amplification was performed by 98°C 3', 21 cycles of [98°C, 20'',
497 67°C, **60''**, 72°C 6'] and 72°C, 5'. cDNA was the cleaned up using an equal volume (25µL) of CleanPCR
498 beads (CleanNA) and tagmented using homemade Tn5[56].
499
500 *Single-cell cultures*
501 Bone Marrow mononuclear cells from patient P1 were stained and Lin- or Lin-CD34+ single cells were
502 index-sorted into ultra-low attachment 96-well plates (Corning) containing 100µL StemSpan SFEM
503 media (Stem Cell Technologies). Media was supplemented with penicillin/streptomycin (100ng/mL),
504 L-glutamine (100ng/mL) and the following human cytokines (all from Peprotech): SCF (20ng/mL),
505 Flt3-L (20ng/mL), TPO (50ng/mL), IL-3 (20ng/mL), IL-6 (20ng/mL), G-CSF (20ng/mL), EPO
506 (40ng/mL), IL-5 (20ng/mL), M-CSF (20ng/mL), GM-CSF (50ng/mL). After 21 days at 5% CO_2 and 37°C,
507 colonies were imaged by microscopy, and processed as detailed in the following.
508

509 *Targeted DNA sequencing by nested PCR amplification*
510 Single-cell derived colonies were transferred into 50µL buffer RLT (Qiagen). Cleanup was performed
511 using CleanPCR beads (CleanNA) at a 1.8x volume ratio and eluted in 20µL 10mM Tris-HCl pH 7.8.
512 4.5µL were transferred to a PCR plate containing 7.5µL Kapa HiFi HS mastermix and 3 µL of a pool of
513 all outer primers (Table S5, each primer at 0.5µM) were added, followed by a PCR program of 98°C 3',
514 30 cycles of [98°C, 20'', 63°C, 60'', 72°C 10''] and 72°C, 5' and subsequent enzymatic cleanup with 2.5µL
515 10x ExoI buffer, 0.4µL ExoI (NEB) and 0.4µL FastAP (ThermoFisher), 30' incubation at 37°C and 5'
516 inactivation at 95°C. Afterwards, 1µL was transferred to a PCR tube containing 5.9µL water, 7.5µL Kapa
517 HiFi HS mastermix and 0.6µL of a pool of all inner primers (Table S5, each primer at 0.5µM), followed
518 by a PCR program of 98°C 3', 15 cycles of [98°C, 20'', 65°C, 15'', 72°C 30''], 72°C, 5' and enzymatic
519 cleanup as above. 1µL was then transferred into a PCR with Nextera indexing primers as described
520 in[56] and amplified with 98°C 3', 10 cycles of [98°C, 20'', 60°C, 15'', 72°C 30''] and 72°C, 5'.
521
522 *Processing of Next Generation Sequencing data*
523 Raw sequencing reads from MutaSeq and Smart-seq2 experiments were processed using the BBDuk
524 software to trim both the standard Illumina Nextera adapters and the ISPCR adapter. Reads were then
525 mapped to the hg38 human genome (Ensembl release 95) using STAR v2.6 (Dobin et al, 2013), with
526 the outFilterMismatchNmax parameter set to 5. Exonic gene counts were tabulated, keeping only reads
527 that did not overlap with targeted regions, overlapped with only one annotated gene, and with lengths
528 greater than 30 nt. For the colony DNA sequencing experiment, reads were mapped to the hg38 human
529 genome using bwa mem (v0.7.17)[49].
530
531 *Analysis of mitochondrial mutations and reconstruction of clonal hierarchies (mitoClone package)*
532 The following set of routines are implemented in the mitoClone package available form
533 https://github.com/veltenlab/mitoClone , and documented further in the package vignettes.
534 *Construction of allele count tables.* For each cell and position of mitochondrial genome as well as other
535 genomic sites of interest, count tables for all nucleotides (A/C/G/T), and deletions were created from
536 the BAM files. To this end, the *mitoClone* package implements the *baseCountsFromBamList* function,
537 which essentially serves as a wrapper to the bam2R function from the deepSNV R package[57].
538 *Filtering of mitochondrial variants.* To identify relevant somatic variants, we implemented the
539 *mutationCallsFromCohort* function. In short, we select coordinates in the mitochondrial genome
540 containing at least 5 reads each in at least 20 cells. To distinguish RNA editing events and true
541 mitochondrial mutations at the level of genomic sites, we then identify mitochondrial variants that
542 occur in several individuals. For this purpose only, individual cells are therefore called as "mutant" in
543 a given site of the mitochondrial genome if at least 10% of the reads from that cell were from a minor
544 allele (i.e. distinct from the reference). Mutations present in at least 1% of cells in a given patient, but
545 no more than 10 cells in any other individual, are then included into the final dataset and counts
546 supporting the reference and mutant alleles are computed as for sites of interest in the nuclear
547 genome. Mutations present in several individuals are stored as a blacklist and were used further for

filtering some of the data analyzed in Figure S4. 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, PhISICS[36] 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 (not shown). 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 (Figure S10).

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:11559G>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:

$$\mathcal{L} = \prod_{c}\prod_{g}\begin{cases} AD \text{ if } M_{cg} = 1 \;\&\; N_{cg} = 0 \\ FPR \text{ if } M_{cg} = 0 \;\&\; N_{cg} = 1 \\ 1 \text{ if } N_{cg} = NA \\ (1 - AD) * (1 - FPR) \text{ else} \end{cases}$$

Here, M_{cg} indicates whether according to the model, cell c is mutant at genomic site g.

584 The branch is then split into clones such that within each clone, the average difference in log-likelihood
585 incurred by swapping nodes was smaller than 1 per cell. This threshold is set arbitrarily to obtain a
586 practically useful grouping of mutations into clones. The vignette 'Computation of clonal hierarchies
587 and clustering of mutations' of the mitoClone package provides further practical recommendations.
588 The same model was used to compute the likelihood of clonal assignments for each cell. For the
589 analysis in Figure S3b, this estimate was then transformed to bits of information using the Kullback-
590 Leibler distance:

591 $$D_{KL} = \sum_{c} \mathcal{L}_c \ \log_2(\mathcal{L}_c * \ |C|)$$

592 where $|C|$ is the total number of clones.
593 *Quantitative analyses.* For all quantitative analyses of clones (differential expression testing,
594 constribution of clones to cell types) cells with a likelihood of clonal assignment of <0.8 were removed.
595 *De novo mutation calling.* To identify nuclear mutations associated with the clones, we performed
596 variant calling at a list of candidate sites from COSMIC[37]. We therefore focused on a subset of COSMIC
597 including putatively pathogenic SNVs and small InDels, variants in expressed genes (mean >20 reads
598 per cell), and variants associated with a primary site "haematopoietic_and_lymphoid_tissue", obtaining
599 a list of 13,797 sites of interest. SNVs commonly observed as germline variants in the 1000 genomes
600 project dataset were removed[58]. Allele count tables were created for each site as described above.
601 Finally, a beta-binomial model with the same probability for mutant in all cells was compared to a beta-
602 binomial model with a different probabilities for mutant in each clone using Akaike's Information
603 Criterion.
604
605 *Single-cell gene expression data analysis*
606 Cells with less than 500 distinct genes observed and genes that appeared in less than 5 cells were
607 removed. Additionally, data from a healthy individual ("H1") was downloaded from the NCBI Gene
608 Expression Omnibus (GSE75478). Data from all individuals was then integrated using scanorama
609 according to the workflow described by the authors[59]. The low dimensional data representation
610 obtained by scanorama was then loaded into Seurat[60], and default Seurat implementations of UMAP[61]
611 and graph-based clustering were used for data visualization and clustering, respectively, using the first
612 15 scanorama components.
613 For more detailed analyses of the T/NK-cell and HSC/MPP populations, cells with these identities were
614 selected and the MNN[62] data integration workflow was repeated using raw expression counts of these
615 specific cell populations as input.
616 Differential expression testing was performed using MAST[63], using a linear model containing the
617 variable of interest (e.g. clonal identity), a library quality covariate (i.e. the number of genes observed
618 per cell), and, when applicable, additional covariates accounting for patient and cell type.
619 For the display of gene expression values only, data were normalized according to the Seurat defaults
620 (i.e. divided by the total count of RNA in the cell, multiplied by a scale factor of 10,000 and log-
621 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[64].

Statistics and Reproducibility

Statistical analyses were performed using R. Statistical details for each experiment are provided in the figure legends.

Data availability

Count tables and other processed data necessary to reproduce all analysis from the manuscript are deposited in figshare with DOI 10.6084/m9.figshare.12382685.v1[65]. 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. Sequencing data from the healthy individuals are deposited in GEO with the accession id GSE75478.

Code availability

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

Competing interests

LMS is co-founder of Sophia Genetics and Levitas Bio and consultant for several companies on genetic analysis.

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 20 R/2017 (to LV, SH, LMS and AT), the Emerson foundation grant 643577 (to LV and LMS) and the German Bundesministerium für Bildung und Forschung (BMBF) through the Juniorverbund in der Systemmedizin 'LeukoSyStem' (FKZ 01ZX1911D to LV, SH and SR). Contributions by SR were further supported by Emmy Noether Fellowship RA 3166/1-1 (DFG). Contributions by CP were supported by a Max-Eder Grant (German Cancer Aid 70111531). Contributions by DN, JCJ, WKH

661 and TB were supported by the Gutermuth Foundation, the H.W. & J. Hector fund, Baden-Württemberg,
662 and the Dr. Rolf M. Schwiete Fund, Mannheim. DN is an endowed professor of the Deutsche José
663 Carreras Leukämie Stiftung (DJCLS H 03/01).
664

665 **Author contributions**

666 LV developed MutaSeq and performed single-cell RNA-seq experiments with assistance by JM, DRL
667 and CST. PH performed sample characterizations, single-cell culture experiments and FACS sorting
668 with support by MP, AD, SR and supervision by SH. BAS and LV developed mitoClone and analyzed the
669 data with contributions from RF. LV, SH, AT and LMS conceived the study. LV and BAS wrote the
670 manuscript with contributions from PHM, SH, SR and LMS. All other authors were involved in the
671 collection and initial characterization of samples. All authors have read and commented on the
672 manuscript.
673

674 **References**

675 1. Clevers, H. What is an adult stem cell? *Science (80-.).* **350**, 1319–1320 (2015).
676 2. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328–
677 37 (2013).
678 3. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
679 4. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia.
680 *Nature* **506**, 328–333 (2014).
681 5. Greaves, M. Leukaemia 'firsts' in cancer research and treatment. *Nat. Rev. Cancer* **16**, 163–172
682 (2016).
683 6. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J.*
684 *Med.* **371**, 2488–2498 (2014).
685 7. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA
686 Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
687 8. Bowman, R. L., Busque, L. & Levine, R. L. Clonal Hematopoiesis and Evolution to Hematopoietic
688 Malignancies. *Cell Stem Cell* **22**, 157–170 (2018).
689 9. Trumpp, A. & Wiestler, O. D. Mechanisms of Disease: cancer stem cells--targeting the evil twin. *Nat.*
690 *Clin. Pract. Oncol.* **5**, 337–347 (2008).
691 10. Essers, M. A. G. & Trumpp, A. Targeting leukemic stem cells by breaking their dormancy. *Mol.*
692 *Oncol.* **4**, 443–450 (2010).
693 11. Pollyea, D. A. & Jordan, C. T. Therapeutic targeting of acute myeloid leukemia stem cells. *Blood* **129**,
694 1627–1635 (2017).
695 12. Hosen, N. *et al.* CD96 is a leukemic stem cell-specific marker in human acute myeloid leukemia.
696 *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11008–13 (2007).
697 13. Majeti, R. *et al.* Dysregulated gene expression networks in human acute myelogenous leukemia
698 stem cells. *Proc. Natl. Acad. Sci.* **106**, 3396–3401 (2009).
699 14. Majeti, R. *et al.* CD47 Is an Adverse Prognostic Factor and Therapeutic Antibody Target on Human
700 Acute Myeloid Leukemia Stem Cells. *Cell* **138**, 286–299 (2009).

701 15. Jan, M. *et al.* Prospective separation of normal and leukemic stem cells based on differential
702 expression of TIM3, a human acute myeloid leukemia stem cell marker. *Proc. Natl. Acad. Sci. U. S. A.*
703 **108**, 5009–14 (2011).
704 16. Eppert, K. *et al.* Stem cell gene expression programs influence clinical outcome in human leukemia.
705 *Nat. Med.* **17**, 1086–1093 (2011).
706 17. Ng, S. W. K. *et al.* A 17-gene stemness score for rapid determination of risk in acute leukaemia.
707 *Nature* **540**, 433–437 (2016).
708 18. Pabst, C. *et al.* GPR56 identifies primary human acute myeloid leukemia cells with high
709 repopulating potential in vivo. *Blood* **127**, 2018–2027 (2016).
710 19. Raffel, S. *et al.* BCAT1 restricts αKG levels in AML stem cells leading to IDHmut-like DNA
711 hypermethylation. *Nature* **551**, 384–388 (2017).
712 20. Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the
713 elderly. *Blood* **130**, 742–752 (2017).
714 21. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J.*
715 *Med.* **374**, 2209–2221 (2016).
716 22. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat.*
717 *Cell Biol.* **19**, 271–281 (2017).
718 23. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*
719 **163**, 1663–1677 (2015).
720 24. Tusi, B. K. *et al.* Population snapshots predict early haematopoietic and erythroid hierarchies.
721 *Nature* **555**, 54–60 (2018).
722 25. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem
723 cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
724 26. Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-
725 cell RNA-seq. *Science (80-.).* **360**, 331–335 (2018).
726 27. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression
727 and Immunity. *Cell* **176**, 1265-1281.e24 (2019).
728 28. Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes.
729 *Nature* **571**, 355–360 (2019).
730 29. Petti, A. A. *et al.* A general approach for detecting expressed mutations in AML cells using single
731 cell RNA-sequencing. *Nat. Commun.* **10**, 3660 (2019).
732 30. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat.*
733 *Methods* **12**, 519–522 (2015).
734 31. Rodriguez-Meira, A. *et al.* Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-
735 Cell Mutational Analysis and Parallel RNA Sequencing. *Mol. Cell* **73**, 1292-1305.e8 (2019).
736 32. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and
737 transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, (2015).
738 33. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell
739 Genomics. *Cell* **176**, 1325-1339.e22 (2019).
740 34. Xu, J. *et al.* Single-cell lineage tracing by endogenous mutations enriched in transposase accessible

741 mitochondrial DNA. *Elife* **8**, 1–14 (2019).
742 35. Coban-Akdemir, Z. *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease
743 Traits by Potential Gain-of-Function Alleles. *Am. J. Hum. Genet.* **103**, 171–187 (2018).
744 36. Malikic, S. *et al.* PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction
745 via integrative use of single-cell and bulk sequencing data. *Genome Res.* **29**, 1860–1877 (2019).
746 37. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–
747 D947 (2019).
748 38. Assi, S. A. *et al.* Subtype-specific regulatory network rewiring in acute myeloid leukemia. *Nat.*
749 *Genet.* **51**, 151–162 (2019).
750 39. Shlush, L. I. *et al.* Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature*
751 **547**, 104–108 (2017).
752 40. Izzo, F. *et al.* DNA methylation disruption reshapes the hematopoietic differentiation landscape.
753 *Nat. Genet.* **52**, 378–387 (2020).
754 41. Pina, C., May, G., Soneji, S., Hong, D. & Enver, T. MLLT3 Regulates Early Human Erythroid and
755 Megakaryocytic Cell Fate. *Cell Stem Cell* **2**, 264–273 (2008).
756 42. Uckelmann, H. J. *et al.* Therapeutic targeting of preleukemia cells in a mouse model of NPM1
757 mutant acute myeloid leukemia. *Science (80-.).* **367**, 586–590 (2020).
758 43. Brunetti, L. *et al.* Mutant NPM1 Maintains the Leukemic State through HOX Expression. *Cancer Cell*
759 **34**, 499-512.e9 (2018).
760 44. Vassiliou, G. S. *et al.* Mutant nucleophosmin and cooperating pathways drive leukemia initiation
761 and progression in mice. *Nat. Genet.* **43**, 470–475 (2011).
762 45. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell
763 development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
764 46. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based
765 knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* **44**, D1023–D1031 (2016).
766 47. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin
767 profiling. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0645-6.
768 48. Schallmoser, K. *et al.* Rapid Large-Scale Expansion of Functional Mesenchymal Stem Cells from
769 Unmanipulated Bone Marrow Without Animal Serum. *Tissue Eng. Part C Methods* **14**, 185–196
770 (2008).
771 49. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
772 50. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous
773 cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
774 51. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-
775 throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).
776 52. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115–e115
777 (2012).
778 53. Östergård, P. R. J. A fast algorithm for the maximum clique problem. *Discret. Appl. Math.* **120**, 197–
779 207 (2002).
780 54. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat.*

781 *Methods* **10**, 1096–1098 (2013).
782 55. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–81
783 (2014).
784 56. Hennig, B. P. *et al.* Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification
785 and Tagmentation Protocol. *G3 (Bethesda).* **8**, 79–89 (2018).
786 57. Gerstung, M. *et al.* Reliable detection of subclonal single-nucleotide variants in tumour cell
787 populations. *Nat. Commun.* **3**, 811 (2012).
788 58. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526**, 68–74
789 (2015).
790 59. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes
791 using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
792 60. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
793 61. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat.*
794 *Biotechnol.* **37**, 38–47 (2019).
795 62. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-
796 sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–
797 427 (2018).
798 63. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and
799 characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
800 64. Wickham, H. *ggplot2: elegant graphics for data analysis*. (Springer New York, 2009).
801 65. Velten, L., Story, B. A. & Steinmetz, L. M. Dataset for the manuscript 'Identification of leukemic and
802 pre-leukemic stem cells by clonal tracking from single-cell transcriptomics'. (2021)
803 doi:10.6084/m9.figshare.12382685.v1.
804 66. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-
805 phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
806 67. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem
807 cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
808 68. Pelechano, V., Wilkening, S., Järvelin, A. I., Tekkedil, M. M. & Steinmetz, L. M. Genome-wide
809 polyadenylation site mapping. *Methods Enzymol.* **513**, 271–96 (2012).
810
811

Figures and Figure legends

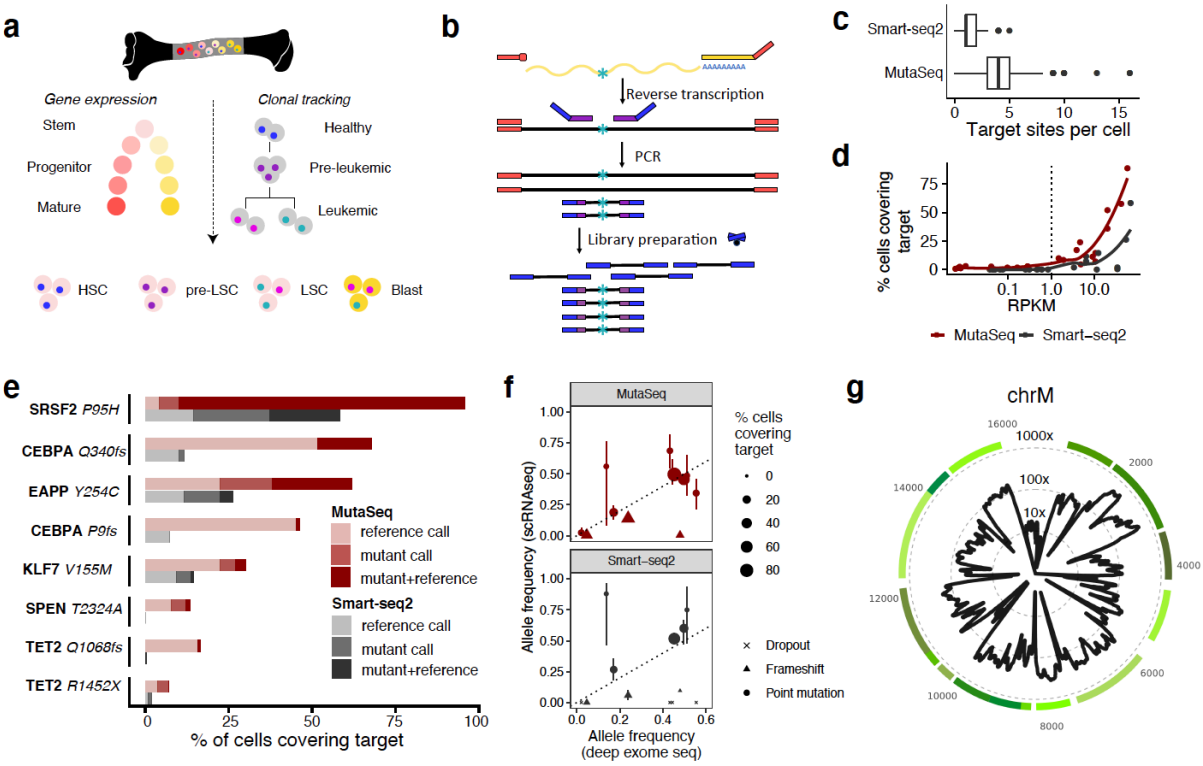


Figure 1. MutaSeq for high-quality single-cell RNA-seq data with clonal information. See also Figure S1.

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 Methods section on '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 (Table S1) 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) or Smart-seq2 (grey dots). Dot size indicates coverage at target site. Point shape indicates the type of mutation.

g. Coverage of the mitochondrial genome in MutaSeq data. See Figure S1j for a comparison across methods. Green line segments correspond to genes in the mitochondrial genome.

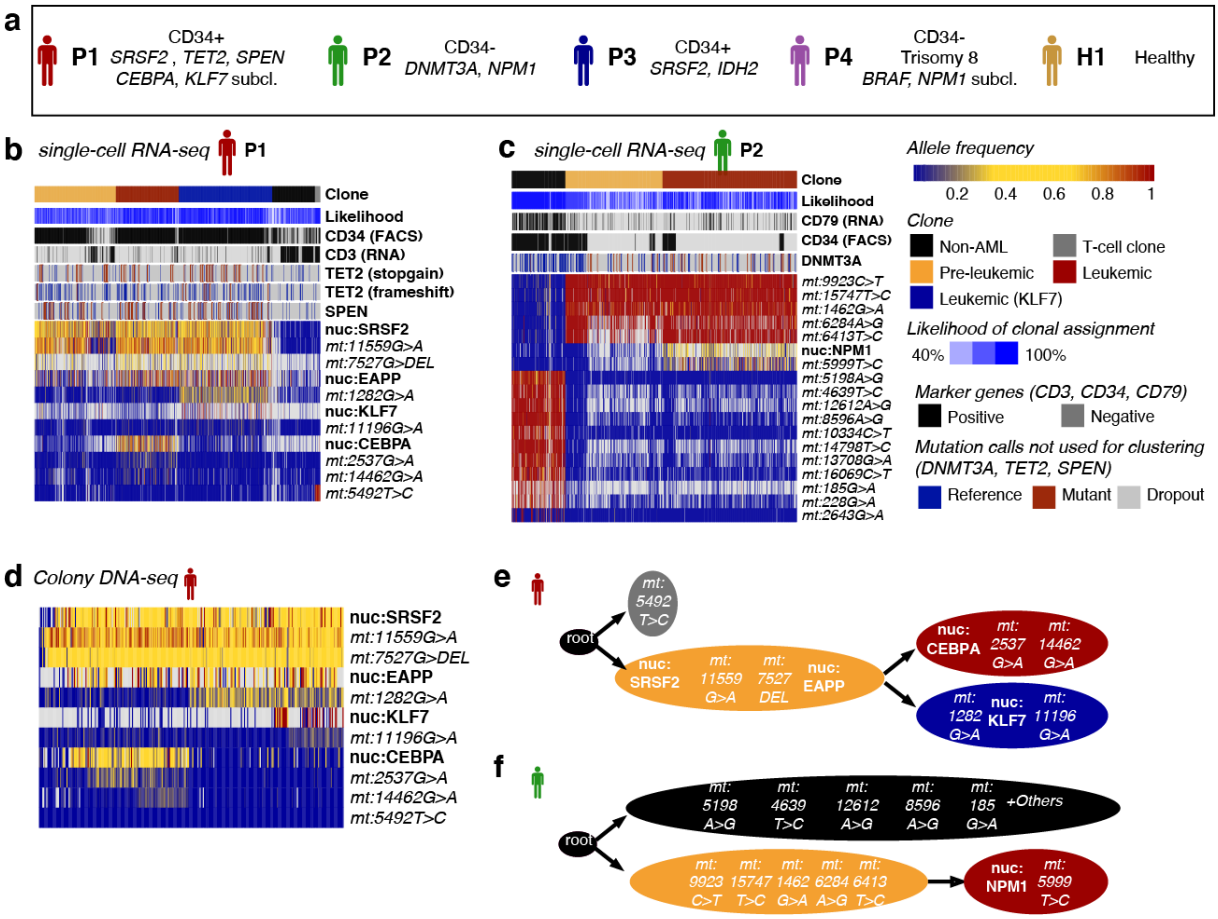


Figure 2. Mitochondrial mutations serve as high-confidence clonal markers in AML. See also Figures S2-S6.

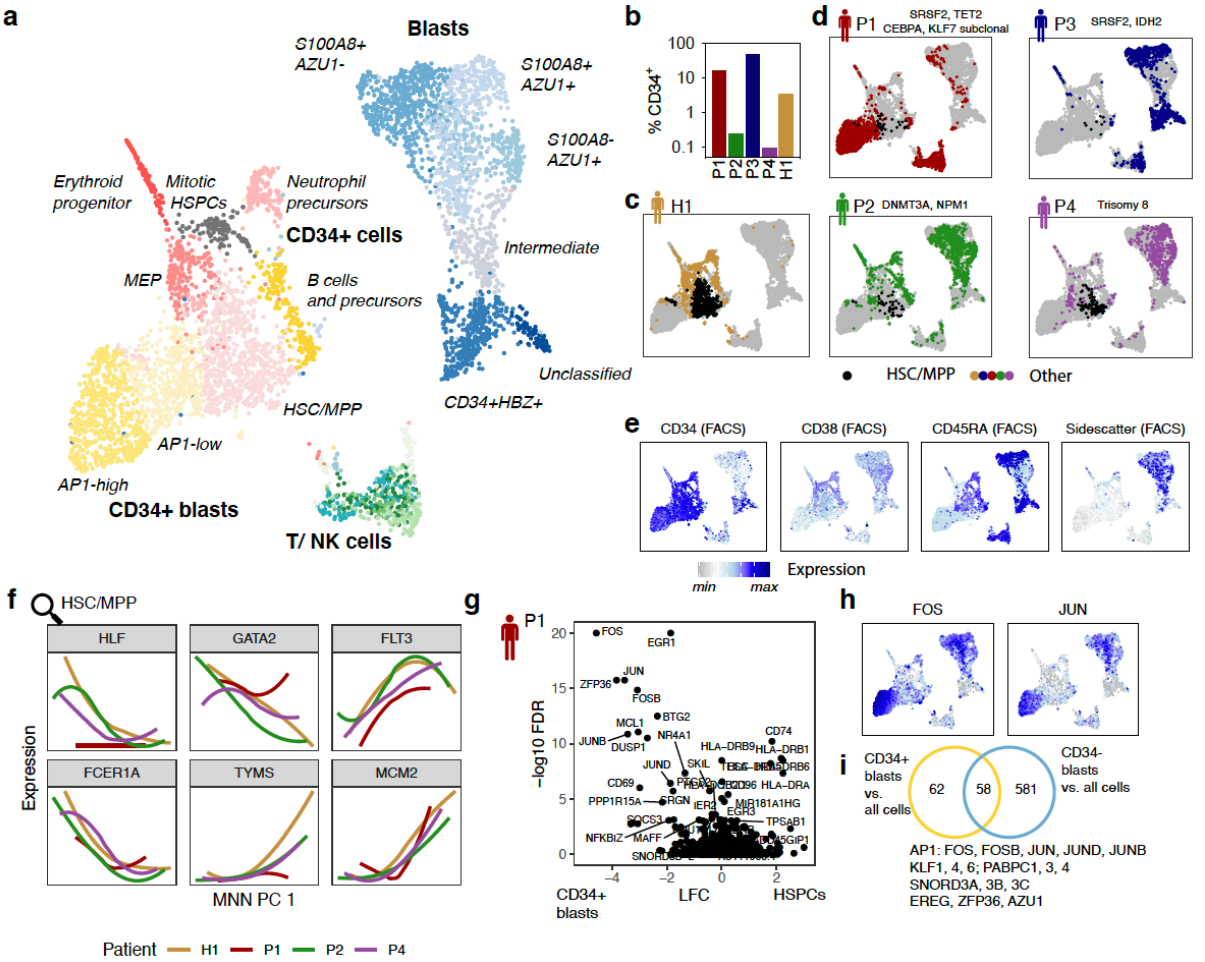
a. Overview over the patients. CD34+ and CD34- indicate the dominant surface phenotype of the leukemic blasts, see Figures S2 and 3b for quantification. For each patient, genes containing mutations are printed in italic, see Table S1 for a complete list. Subcl., sub-clonal mutation.

b. Heatmap depicting variant allele frequencies (color coded, see right of Figure 2c for a key) observed in single-cell RNA-seq data of n=1430 cells from P1. Grey indicates missing values. Cells and mutations are arranged according to the clustering result obtained by PhISCS[36] 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 *DNMT3A*, *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 (see figure S5b). 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.

d. Heatmap depicting variant allele frequencies (color coded, see legend to the right of Figure 2c) observed in targeted DNA-seq data from n=288 single-cell derived colonies from P1.

855 e. PhISCS[36] was run on the mutational data from P1 to reconstruct an clonal hierarchy, see
856 Methods section "Analysis of mitochondrial mutations and reconstruction of clonal
857 hierarchies". Take note that while the order of mutations is based on the PhISCS model, the
858 grouping of mutations into clones is based on an arbitrary cutoff to provide a useful clustering
859 for further analyses. See "Clone" in Figure 2c legend for color codes.
860 f. Like e, but for P2.
861

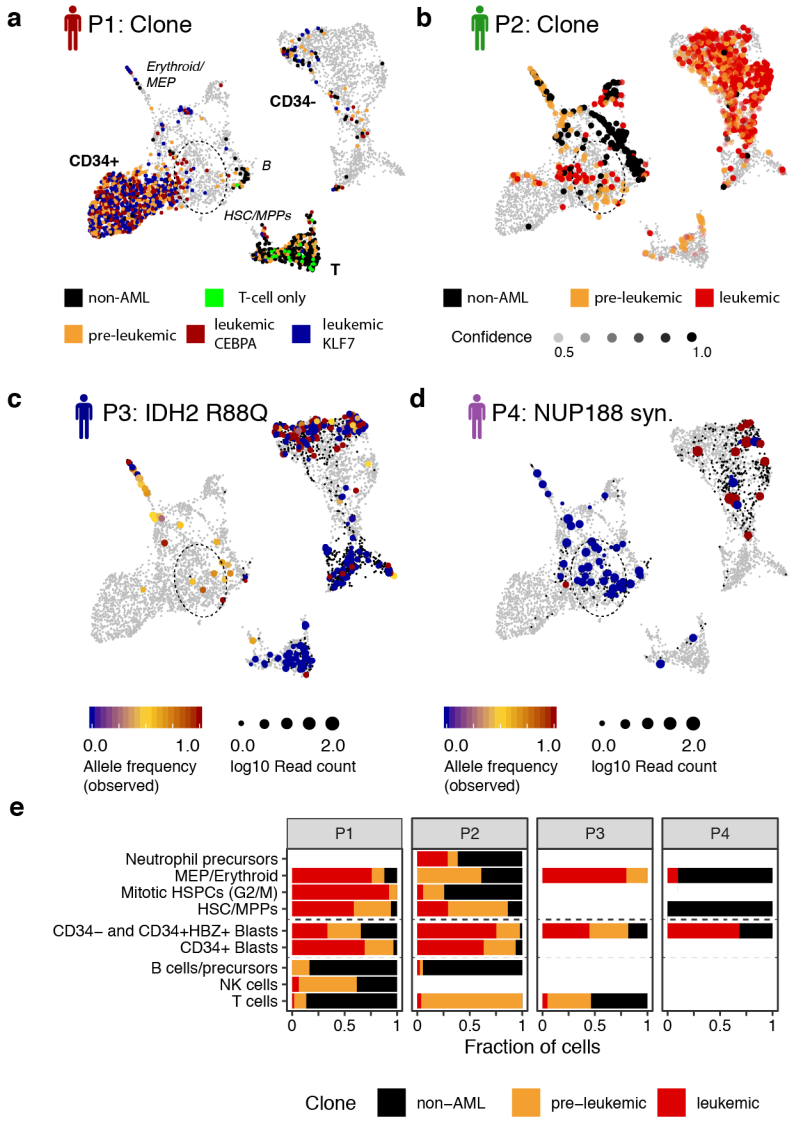


862

863 **Figure 3. An overview of cell types observed in bone marrow of healthy and leukemic**
864 **individuals.** See also Figures S2 and S7.

865 a. Data from the five individuals (Figure 1a) were integrated using scanorama[59] and visualized
866 in two dimensions using uMAP[61]. Clusters are color-coded. N = 5228 cells.
867 b. Fraction of CD34+ cells in total bone marrow from the five individuals. CD34+ cells were
868 enriched during FACS sorting, see Figure S2.
869 c. CD34+ cells from a healthy individual[22] highlighted on the uMAP. Black dots correspond to
870 cells from the HSC/MPP cluster.
871 d. Cells from each patient were highlighted separately on the uMAP. Black dots correspond to
872 cells from the HSC/MPP cluster.
873 e. Logicle-transformed expression of key FACS markers highlighted on the uMAP, see also Figure
874 S7a.

875 f. Data from n= 667 cells from the HSC/MPP cluster were integrated across individuals using
876 MNN[62]. The smoothened expression of several marker genes of healthy HSC/MPP subsets[22] is
877 plotted over the first dimension of variability identified by MNN. See also Figure S7f.
878 g. Volcano plot of the log10 expression change in n = 569 AP1-high CD34+ blasts vs. n = 667
879 HSC/MPP-like cells, plotted against corrected p-values from MAST, using a model that
880 accounts for differences in library quality and patient identity / batch (see Methods, section
881 "Single-cell gene expression data analysis" for detail). AP1-high CD34+ blasts were chosen for
882 this comparison since AP1-low blasts, in terms of all marker genes, appear to constitute an
883 intermediate state between Healthy-like HSC/MPPs and AP1-high blasts.
884 h. Log-normalized expression of FOS and JUN on the uMAP from panel a. See panel e for a color
885 scale.
886 i. Venn diagram displaying genes with significant overexpression in AP1-high CD34+ blasts and
887 CD34- blasts, compared to all other cells from the dataset.



888

889 **Figure 4: Effects of (pre-)leukemic mutations on cellular differentiation.** See also Figure S8.

a. Clonal identity of the cells from P1 highlighted on the uMAP; see also Figure 3a and Figure 2b,e. Grey 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; see also Figure 3a and Figure 2c,f.

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 Figure S8a 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, see Figure S2. See Figure S8a 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 Figure 2 were used. For P3, cell 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 Figure S8b for absolute numbers and S8c,d for a quantitative analysis.

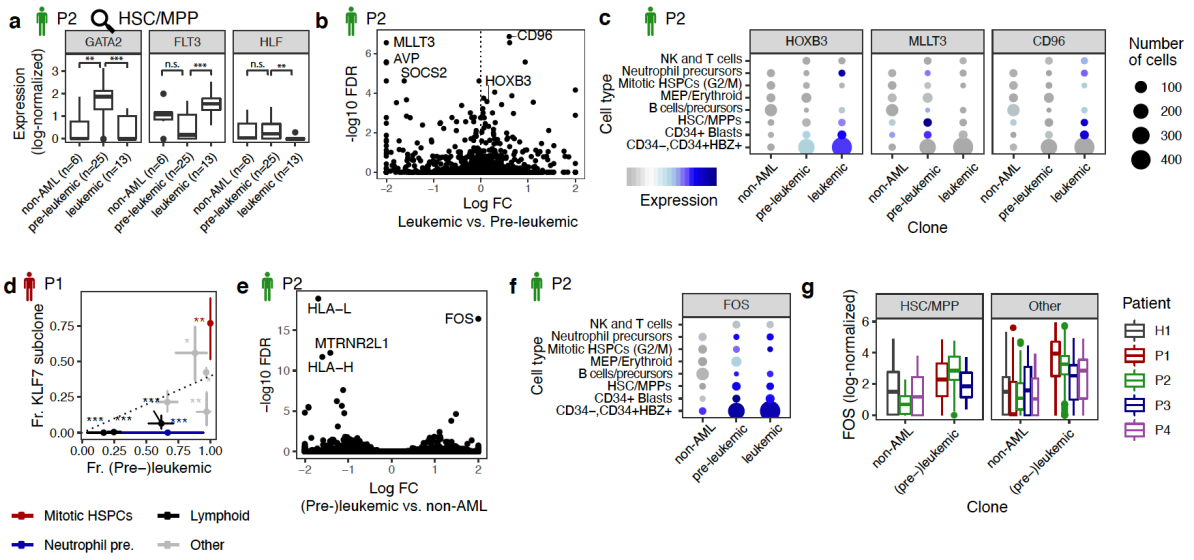


Figure 5: Molecular consequences of leukemic and pre-leukemic mutations. See also Figure S8.

a. Log-normalized expression of GATA2, FLT3 and HLF in HSC/MPPs from P2, stratified by clonal identity. See methods, section *Single-cell gene expression data analysis* for detail on data normalization. Asterisk indicate significance from a Wilcoxon test, ***: $p < 0.001$, **: $p<0.01$, n.s.: not significant.

916 b. Volcano plot of the log10 expression change in pre-leukemic (n=55) vs. leukemic (n=50)
917 CD34+ cells of P2, plotted against p-values from MAST, using a model that accounts for
918 differences in cell type and library quality (see Methods for detail). Only the following CD34+
919 cell types from Figure 3a were included in the test: HSC/MPP, CD34+ Blasts (both subsets),
920 Neutrophil precursors, and MEPs.
921 c. Dot plot comparing the expression of relevant genes across non-leukemic, pre-leukemic and
922 leukemic cells in the different cell types. Symbol size scales with the number of cells per cell
923 type.
924 d. Fraction of (pre-)leukemic in relation to the fraction of cells from the KLF7 clone in various
925 cell types in P1. Dotted line indicates the mean ratio across all cells, error bars denote 95%
926 confidence intervals from a beta distribution, and asterisk indicate significant deviation from
927 the mean ratio, as follows: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. p-values were computed from
928 the quantiles of a beta distribution against the null hypothesis that the true ratio is the mean
929 ratio observed across all cells. See also Figure S8c.
930 e. Volcano plot of the log10 expression change in (pre-)leukemic (n=105) vs. non-leukemic
931 (n=41) CD34+ cells of P3, plotted against p-values from MAST, using a model that accounts for
932 differences in cell type and library quality (see Methods for detail). Only the following cell
933 types from Figure 3a were included in the test: HSC/MPP-like, CD34+ Blasts (both subsets),
934 Neutrophil precursors, and MEPs.
935 f. Dot plot comparing the expression of FOS across non-leukemic, pre-leukemic and leukemic
936 cells in the different cell types. Dot size and color represent the quantity of cells and gene
937 expression level, respectively (see legend in Figure 5c)
938 g. Boxplots comparing the log-normalized expression levels of FOS between cells with evidence
939 of originating from the non-leukemic clone(s), and cells with evidence of originating from the
940 leukemic or pre-leukemic clones. Cells were assigned to clones as in Figure 4e. See methods,
941 section *Single-cell gene expression data analysis* for detail on data normalization.
942

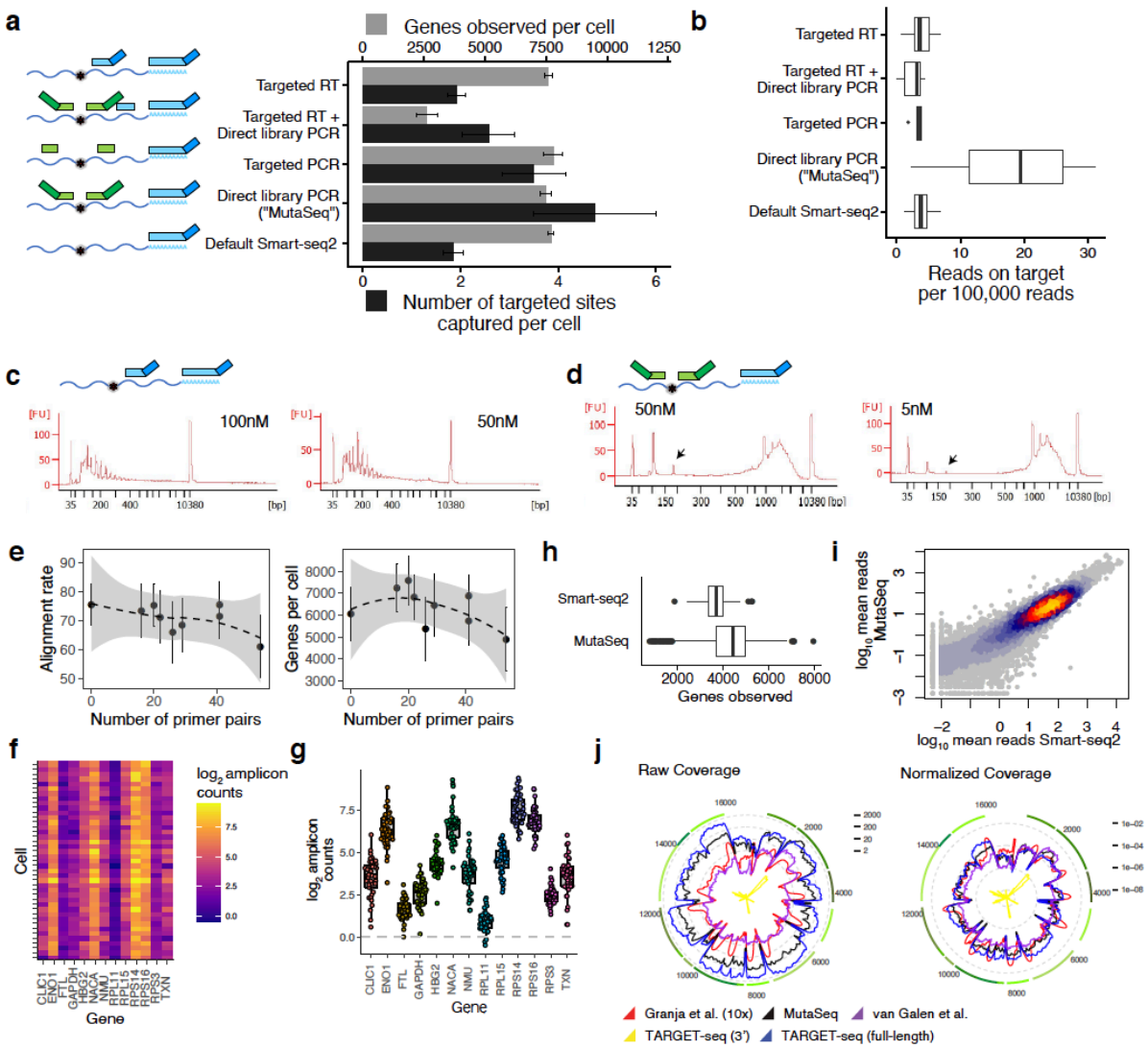


Figure S1. Development of MutaSeq. See also Figure 1.

a. Comparison of different strategies for targeting 14 genomic sites of interest during Smart-seq2 library preparation; see Table S5 for primers used. In the 'targeted RT' protocol, a reverse transcription primer carrying the ISPCR sequence was placed downstream of the sites of interest. In the 'targeted RT+ direct library PCR' protocol, a targeted RT primer without ISPCR sequence was used in conjunction with targeting primers included during PCR. In the 'targeted PCR' protocol, PCR primers were used to generate amplicons of 250-350 bases, whereas in the 'direct library PCR' protocol, shorter amplicons were used and primers were fused to Nextera sequencing adapters; see also Figure 1b. 8-16 K562 cells were sequenced with each protocol and the number of genes observed per cell as well as the number of target sites covered was quantified. Error bars indicate the standard error of the mean. For all analysis in this panel, reads were down-sampled to 250,000 reads per cell prior to alignment to account for differences in coverage.

b. Boxplot comparing the mean number of reads per target in the different protocols. See Methods section on 'Data Visualization' for a definition of boxplot elements.

959 c. Bioanalyzer traces for a protocol targeting sites of interest during RT only. Concentrations
960 correspond to the final concentration of targeting primers in the RT reaction.
961 d. Bioanalyzer traces for a protocol targeting sites of interest during library amplification PCR
962 (MutaSeq). Arrows highlight the MutaSeq amplicons. Concentrations correspond to the final
963 concentration of targeting primers in the PCR reaction.
964 e. Primer sets developed for the MutaSeq patients (Table S5) were combined in all possible
965 combinations (i.e. P1+P2 primers, P1+P3+P4 primers, etc.) and libraries were generated from
966 n=8 K562 cells each to evaluate the effect of multiplexing primers on alignment rate (left) and
967 number of genes observed (right).
968 f. Primer pairs were designed surrounding randomly selected sites on 13 highly-expressed
969 genes in K562 cells (Table S5). The MutaSeq protocol was then performed using these primers
970 on n=48 K562 cells. For each gene, the number of reads from MutaSeq amplicons (i.e. complete
971 matches) is shown, after subtracting the average coverage of the surrounding areas outside of
972 the targeted site (i.e. potential background signal). Seven cells with poor alignment rates
973 (below 50%) were removed.
974 g. Amplicon counts for 13 genes across 41 cells is shown as boxplots. The points in the overlaid
975 beeswarm plot represent cells. Same underlying data as used in Figure S1f.
976 h. Number of genes observed per cell, across n=206 (Smart-seq2) or n=658 CD34+ (MutaSeq)
977 cells. See Methods section on 'Data Visualization' for a definition of boxplot elements.
978 i. Mean gene expression levels measured by Smart-seq2 are compared to mean gene expression
979 levels measured by MutaSeq. Color reflects point density.
980 j. Logarithmic coverage of the mitochondrial genome compared between different
981 methods[27,31,66]. For the plot on the right, coverage was normalized to the number of reads
982 aligning to the transcriptome.

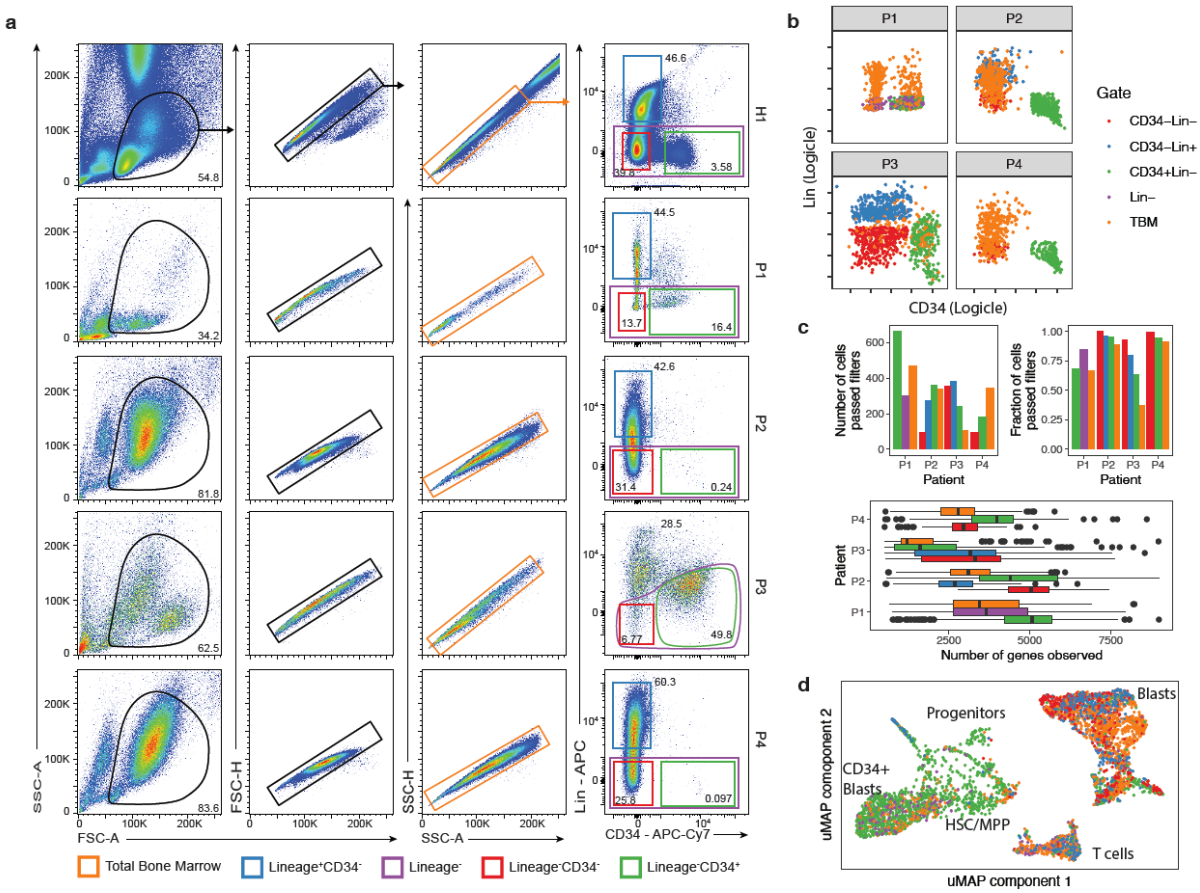


Figure S2. FACS sorting schemes and quality control of single-cell RNA-seq data. See also Figure 2+3.

a. Gating schemes used for the different patients.
b. Index values of cells included into the final data set
c. Top left: number of cells from the various gates included into the final data set; see panel a/b for a color scheme. Top right: fraction of cells passing filters, stratified by patient and gate. Bottom: box plots depicting the number of genes observed in cells from the final data set. Cells with more than 500 genes observed were included.
d. uMAP plot of cells from all individuals, with cells color coded according to their sorting gate. See panel a/b for a color scheme, and main Figures 3a-e and S7a for a more detailed description of the uMAP.

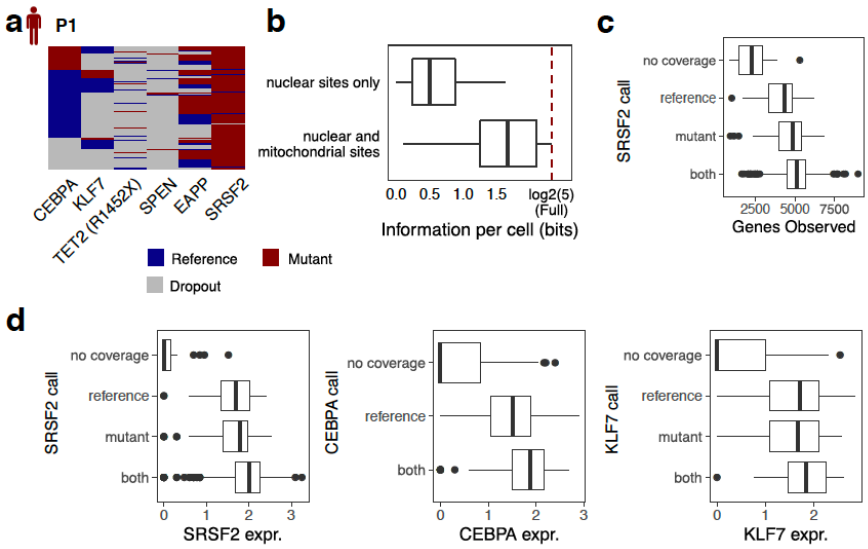


Figure S3. Evaluation of nuclear markers for clonal tracking in single-cell RNA-seq data. See also Figure 2.

a. Heatmap depicting mutation calls of nuclear genomic mutations only for P1. Clustering was performed as described in the methods, section *Analysis of mitochondrial mutations and reconstruction of clonal hierarchies*, except that here, only nuclear mutations were included. Rows represent cells.

b. The full specification of clonal identities in case of P1 requires log2(5) bits of information, since there are five main clones (Figure 2e). For each cell, the information available from nuclear genomic sites only, or both nuclear and mitochondrial sites, was quantified as described in the Methods, section *Analysis of mitochondrial mutations and reconstruction of clonal hierarchies*.

c. Box plots evaluating the extent to which differences in library quality affect measurements of the mutational status of target genes; data from P1 is shown.

d. Like panel c, but investigating the effect of the expression of a targeted gene on the ability to measure its mutational status.

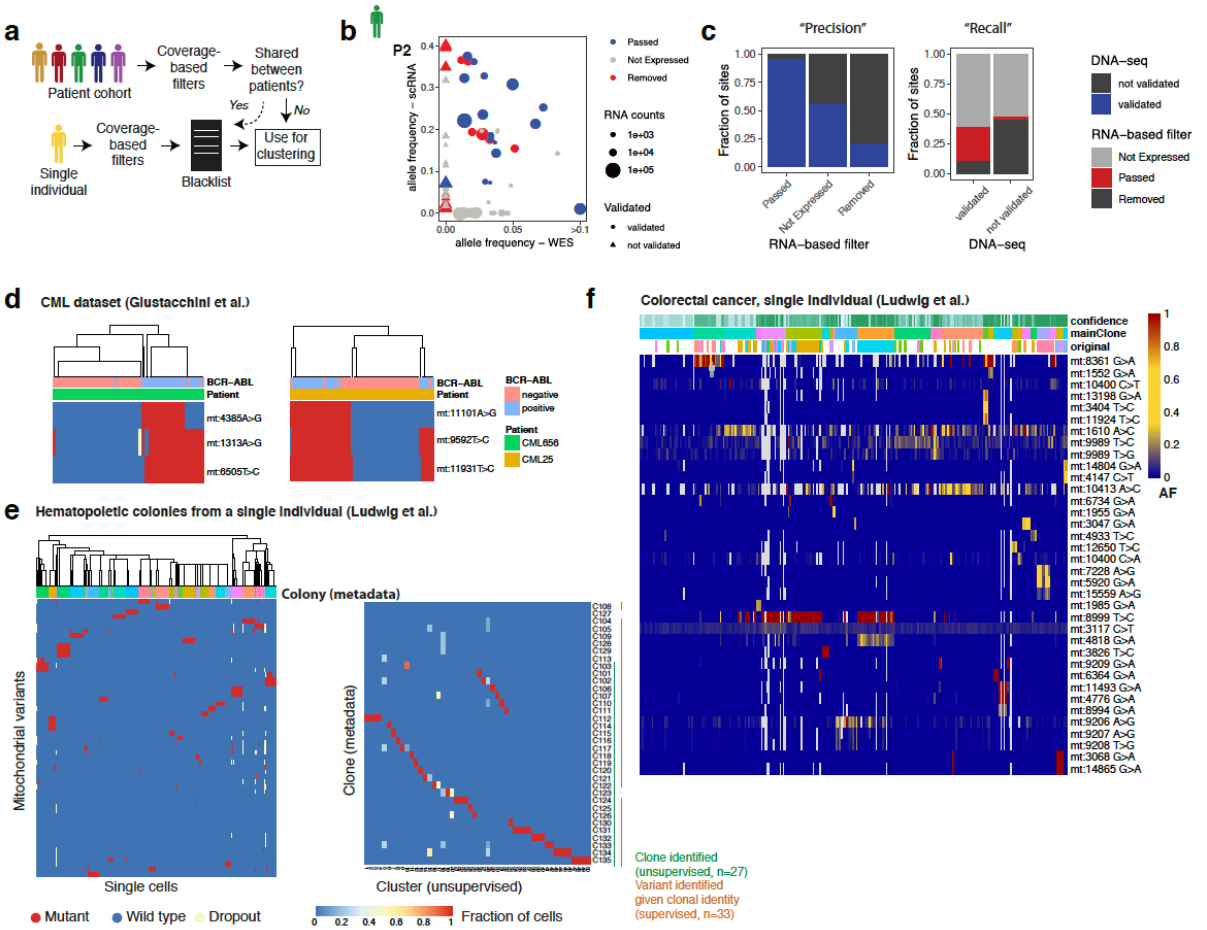


Figure S4. Calling of mitochondrial somatic variants in the absence of a DNA-based reference. See also Figure 2. For an implementation and for reproducing the computations, see the package vignettes of the *mitoClone* package.

a. Overview of the computational strategy used. In the case of data from a group of individuals, sites were filtered based on coverage and annotated as 'mutant' if a specified fraction of reads deviates from the reference allele. Sites were then excluded as likely RNA editing events if the same mutation was observed in more than one individual. Alternatively, in the case of data from a single individual, similar coverage based filtered were applied and data was then filtered against a blacklist created from a cohort.

b. Allele frequencies and coverage of mitochondrial mutations from P2 in single-cell RNA-seq data compared to whole exome sequencing data (WES). Sites with less than 10 reads per cell in RNA-seq were classified as 'not expressed'. Variants that were observed in WES were classified as 'validated' (circles), other variants were classified as 'not validated' (triangles). The low correlation between the two datasets is likely due to different starting cell populations (WES: Total bone marrow, single cell RNA-seq: enriched for CD34+ cells), and data are only used for qualitative statements (presence/absence of mutations).

1027 c. Bar charts summarizing the classifications from panel b. Left, mutation sites are split by their
1028 label based on the *mitoClone* pipeline; right, sites are split by whether they were detected in
1029 WES.
1030 d. De novo variant calling and clustering of a CML patient dataset. Data from ref. [67] were
1031 processed and clustered with the *mitoClone* package. The same variant filtering approach used
1032 on the patients from our study was used. Thereby, two patients with substantial mitochondrial
1033 variability were identified and in both cases clones associated with the BCR-ABL mutation
1034 were resolved in an unsupervised manner. The analysis by ref. [33] had missed one of these
1035 patients, did not achieve an unsupervised separation of BCR-ABL+ and BCR-ABL- cells in either
1036 case (Figure 7G in ref. [33]), and instead relied on stratifying cells by the existing BCR-ABL label
1037 (Figure 7J in ref. [33]).
1038 e. De novo variant calling and clustering of single cells from hematopoietic colonies derived from
1039 a single individual. Data from Figure 5 of ref. [33] were processed and clustered with the
1040 *mitoClone* package. Left panel shows unsupervised clustering of mutations identified by the
1041 *mitoClone* package, right panel quantitatively compares unsupervised clustering and colony
1042 labels. 27 of the colonies were identified in an unsupervised manner. The analysis by ref. [33]
1043 had identified approximately half that number by unsupervised analyses (their Figure 5E), and
1044 using supervised methods identified mitochondrial mutations associated with 33 clones (their
1045 Figure 5H).
1046 f. De novo variant calling and clustering of single cells from a single colorectal cancer patient.
1047 Data from Figure 7 of ref. [33] were processed and clustered with the *mitoClone* package.
1048 Clustering structure obtained by PhISCS is shown and compared to the clustering presented
1049 in ref. [33] (row labeled 'original'), which was based on variant filtering using a DNA-seq based
1050 reference. Despite the different filtering approaches, our unsupervised clustering separated
1051 the clusters identified by Ludwig et al. and identified additional variability.

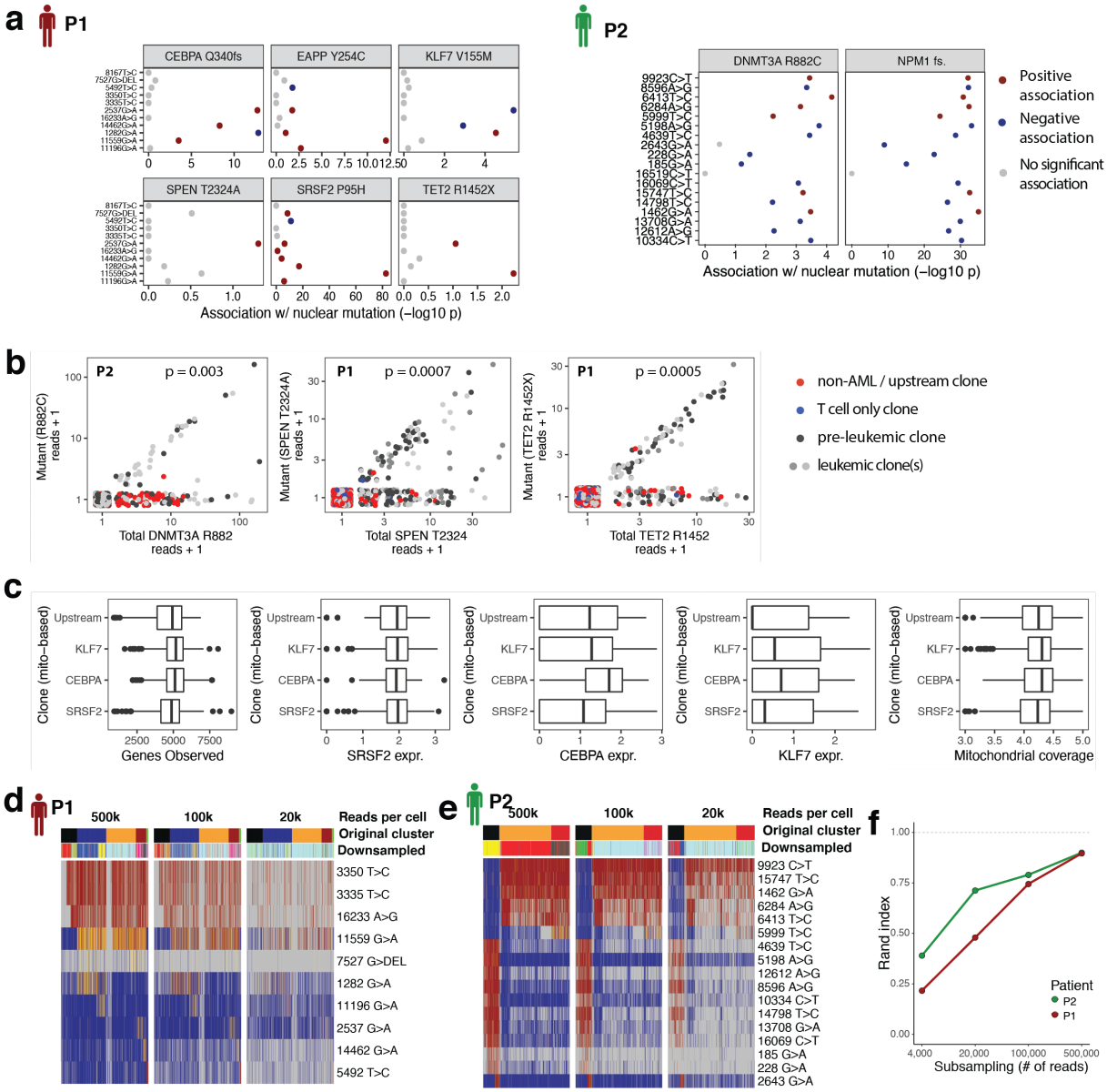


Figure S5. Evaluation of mitochondrial markers for clonal tracking in single-cell RNA-seq data. See also Figure 2.

a. Association between various mutations (y axis) and nuclear mutations (panels) across n=1430 cells from P1 or n=1066 cells from P2. P-values are from a two-sided Fisher test.

b. Association between the lowly covered mutations in DNMT3A, SPEN and TET2 with clonal identity. Scatter plot depicts total coverage on the site of interest (x axis) and the number of mutant reads (y axis) across n=1066 cells from P2 (left panel) or n=1430 cells from P1 (central and right panel). Note that jitter was added in the x and y direction to avoid overplotting. P values are from a chi-square test comparing a model where the probability of detecting at least one mutant read was modelled as a function of total coverage (null model), or a function of total coverage and identity as a non-AML/upstream clone (alternative model).

c. Box plots evaluating the extent to which differences in marker gene expression and library quality affect clonal assignments when using mitochondrial marker mutations; data from P1

is shown. Clone 'SRSF2' refers to the pre-leukemic clone from main figure 2 while clones 'KLF7' and 'CEBPA' refer to the leukemic clones.

d. Effect of read depth on mitochondrial clusters. Clusters obtained from mitochondrial sites only were computed at full read depth (row "Original clusters") and are compared to clusters obtained using the same methodology from data were single cells were down-sampled to read depths of 500k, 100k, or 20k per cell ("Downsampled"). Data from Patient 1 is shown.

e. Like panel c, but for patient 2 (P2).

f. Original clustering result and down-sampled clustering result are compared quantitatively using the Rand index.

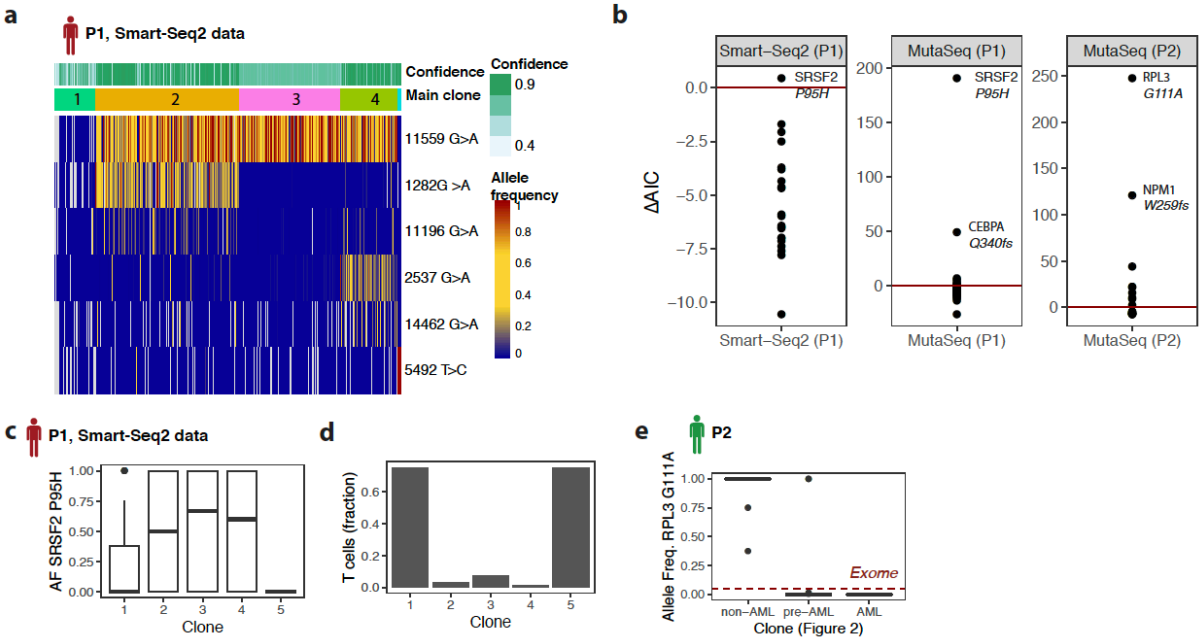


Figure S6. De novo calling and characterization of clones. See also Figure 2. For an implementation and for reproducing the computations, see the package vignettes of the *mitoClone* package.

a. Unsupervised clustering of mitochondrial mutations identified from a Smart-seq2 dataset of n=672 cells from patient P1.

b. De novo identification of nuclear somatic variants associated with the clonal labels from panel a. Difference in Aikake's Information Criterion (AIC) is shown for a comparison between a model where allele frequencies are the same across all cells, and a model where allele frequencies differ between clones. Red line highlights the intercept. See Methods section *Analysis of mitochondrial mutations and reconstruction of clonal hierarchies*.

c. Boxplot of single-cell allele frequencies for the SRSF2 P95H mutation summarized between clones (Smart-seq2 data).

d. Bar plot of contribution of clones to T cells identified from unsupervised clustering of gene expression data (see also Figure 3).

e. Boxplot of single-cell allele frequencies for the RPL3 mutation (COSV53365368) summarized between clones (MutaSeq data). Red dashed line highlights the allele frequency of the mutation identified in exome sequencing.

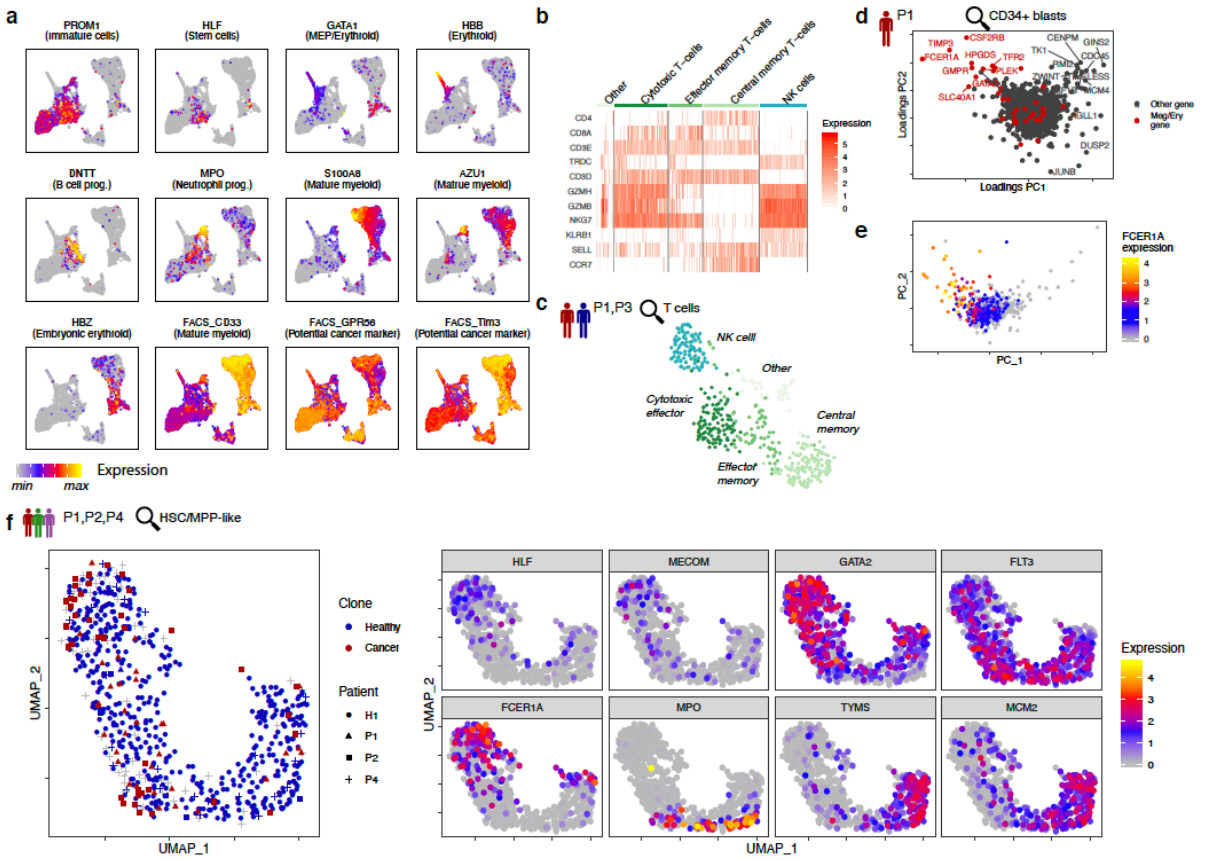


Figure S7. Analysis of single-cell gene expression data. See also Figure 3.

a. Expression of marker genes on the uMAP from Figure 3. Top rows: log-normalized mRNA expression. Bottom row (labels preceded FACS_): Logicle transformed FACS index values.

b. Heatmap depicting the expression of selected marker genes across T-cells from patients P1 and P3. Columns are cells.

c. uMAP representation of all T-cells from patients P1 and P3. Data from these cells only were integrated using MNN[62] and visualized in two dimensions using uMAP. Patients P2 and P4 were omitted from this analysis due to an insufficient number of T cells. Colors denote cluster identity.

d. Loadings plot of a principal component analysis of all CD34+ blasts from Patient P1. Genes associated with erythroid or megakaryocytic priming [22] are highlighted in red.

e. Scores plot of a principal component analysis of all CD34+ blasts from Patient P1. Expression levels of the FCER1A gene are color-coded.

f. uMAP representation of all cells with a healthy HSPC-like gene expression signature from patients P1, P2 and P4. These include both healthy and pre-(leukaemic) clones, see figure 4. Data from these cells only were integrated using MNN[62] and visualized in two dimensions using uMAP. Left panel: clonal identity is highlighted, using the same strategy as in Figure 4e. Right panels: Point color represents the expression of genes involved in differentiation (MPO for myeloid differentiation, FCER1A for erythroid/megakaryocytic differentiation) and cell cycle (TYMS, MCM2).

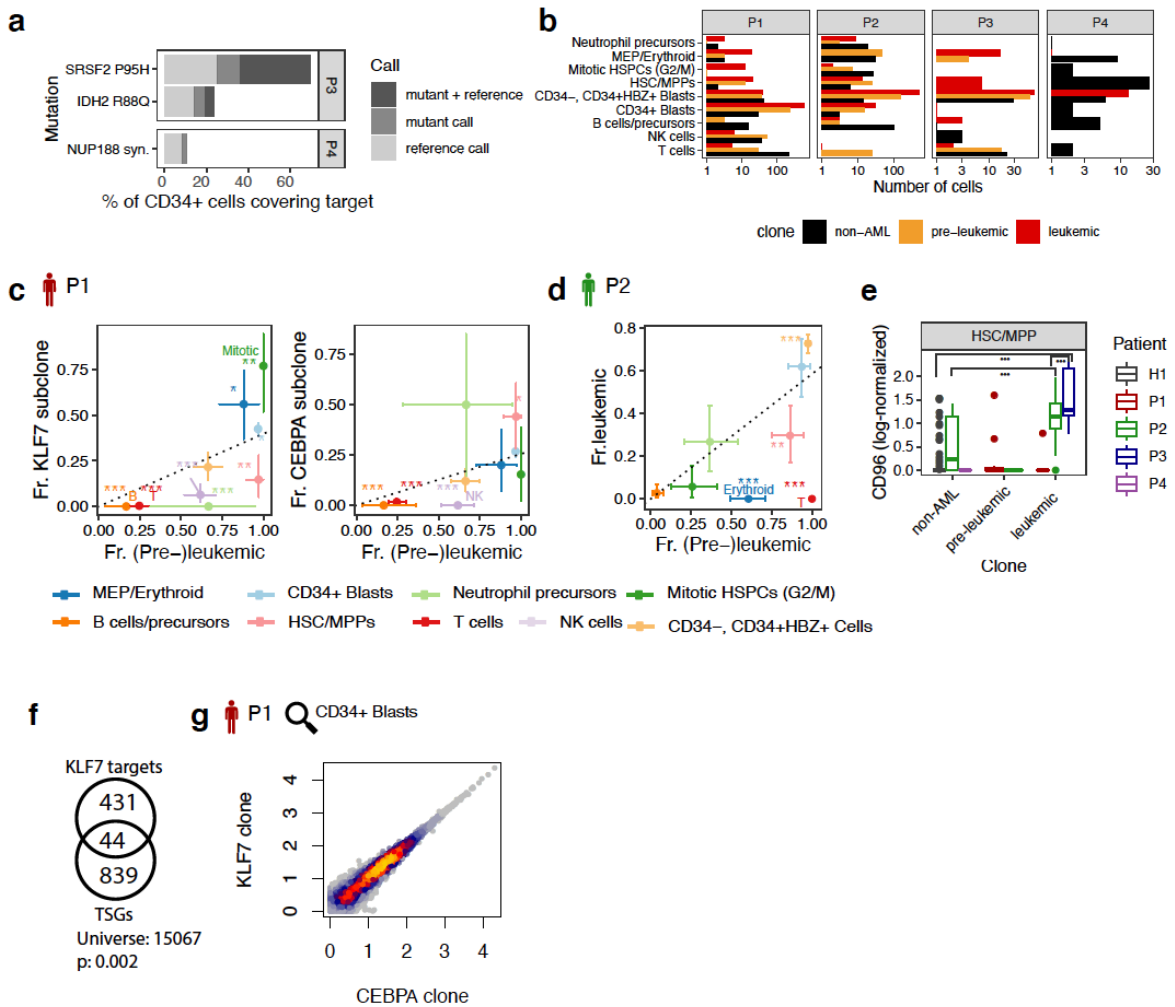


Figure S8. Analysis of single-cell clonal tracking data. See also Figures 4 and 5.

a. Bar chart depicting the percentages of cells with coverage of the mutations used for annotating clones in P3 and P4.

b. Bar chart depicting the absolute cell numbers of the different clones in the different cell populations. Cells were assigned to clones as in Figure 4e.

c. Fraction of (pre-)leukemic cells in relation to the fraction of cells from the sub-clones for 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 were computed from the quantiles of a beta distribution against the null hypothesis that the true ratio is the mean ratio observed across all cells. Only cells with a confident assignment to clones (likelihood > 0.8, see Methods section 'Analysis of mitochondrial mutations and reconstruction of clonal hierarchies') are included.

d. Same as Figure S8c investigating instead the second patient (P2).

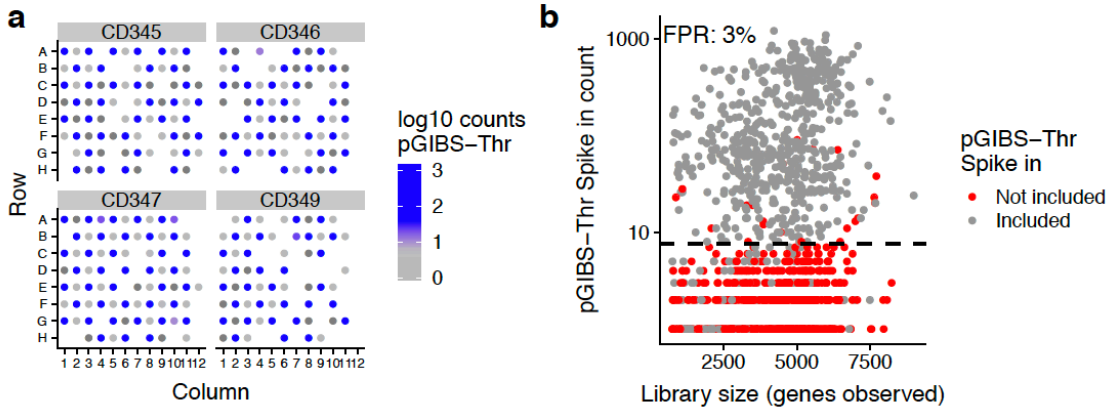
e. Boxplots comparing the normalized, scaled expression levels of CD96 between cells with evidence of originating from the non-leukemic clone(s), and cells with evidence of originating from the leukemic or pre-leukemic clones. Cells were assigned to clones as in Figure 4e. Asterisk denote p-values from a Wilcoxon test for relevant comparisons. ***: $p < 0.001$

1132 f. Venn diagram depicting the number of open chromatin regions containing a KLF7 binding site,
1133 the number of open chromatin regions proximal to tumor suppressor genes from[46], and their
1134 overlap. N=15067 open chromatin regions were identified from single-cell ATAC-seq data of
1135 human CD34+ bone marrow cells[45].
1136 g. Scatter plot comparing the log10 mean gene expression levels in CD34+ blasts from the
1137 CEBPA-mutated clone, and CD34+ blasts from the KLF7 mutated clone. N=15,451 genes with
1138 a mean expression of at least 1 in either population are shown. Color represents point density.

	with mitoClone package					no mitoClone package	
	MutaSeq	SmartSeq	TARGET-seq Full l.	TARGET-seq 3'	GoT-seq	SmartSeq	ATAC-seq
Identification of clones (de novo)	✓✓*	✓✓*	0*	✗	✗	✗	✓✓*
Association of nuclear and mitochondrial variants	✓✓*	0*	0*	✗	✗	✗	✗
Identification of clones (known variants)	✓	0	✓✓	✓✓	✓	0	✗
Differential expression testing between clones	✓✓*	✓✓*	✓✓	✓✓	✗	✗	✗
Throughput	✓	✓	✗	✓	✓✓	✓	✓✓

* If natural mitochondrial somatic variability is present.

1139
1140 **Figure S9. Overview of capabilities of MutaSeq compared to related methods.** Full-l.: full-length.
1141 SmartSeq2: ref. [33,54]. TARGET-seq: ref. [31]. GoT-seq: ref. [28]. "0" means theoretically possible, but
1142 unproven and/or very limited.
1143



1144
1145 **Figure S10. Estimation of MutaSeq's false positive rate.**
1146 a. To estimate MutaSeq's false positive rate, a polyadenylated in vitro transcript, pGIBS-Thr[68],
1147 was added to every second well (A1, A3, B2, B4, etc.) during the P1 experiment. Abundance of
1148 the pGIBS-Thr spike-in across wells from four representative plates is shown.
1149 b. Estimation of the false positive rate using the pGIBS-Thr spike-in. Dashed bold line indicates
1150 the threshold used for classifying a site as dropout.
1151
1152 **Supplementary Materials**

1153 Table S1. Exome sequencing results. Columns contain mutation location, reference and alternate
1154 alleles, consequence for protein sequence, gene biotype, and allele frequencies in their respective
1155 patients' cancer (af_tumor) and healthy (af_normal) exomes. Patients are separated by tab.
1156 Table S2. Genes with significant (false discovery rate <0.001) overexpression in the various cell
1157 populations. For DE testing, MAST[63] with library quality and patient as covariates was used.
1158 Table S3. Results from differential expression test underlying figures 3g, 5b, 5e
1159 Table S4. Antibodies used for flow cytometry.
1160 Table S5. Primers used for targeting genes of interest with MutaSeq.