MutaSeq reveals the transcriptomic consequences of clonal evolution in acute myeloid leukemia

Running title: Mutation tracking by single-cell RNA-seq

Lars Velten[1,*,§], Benjamin A. Story[1,2,*], Pablo Hernandez-Malmierca[3,4,*], Jennifer Milbank[1], Malte Paulsen[5], Christoph Lutz[6], Daniel Nowak[7], Johann-Christoph Jann[7], Caroline Pabst[6], Tobias Boch[3,4,7], Wolf-Karsten Hofmann[7], Carsten Müller-Tidow[6], Simon Raffel[3,4,6], Andreas Trumpp[3,4,8,§], Simon Haas[3,4, §], and Lars M. Steinmetz[1,9,10, §]

1. European Molecular Biology Laboratory (EMBL), Genome Biology Unit, 69117 Heidelberg, Germany

2. Swiss Federal Institute of Technology (ETH) Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, Basel, 4058, Switzerland

3. Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), 69120 Heidelberg, Germany

4. Division of Stem Cells and Cancer, Deutsches Krebsforschungszentrum (DKFZ) and DKFZ-ZMBH Alliance, 69120 Heidelberg, Germany

5. European Molecular Biology Laboratory (EMBL), Flow Cytometry Core Facility, 69117 Heidelberg, Germany

6. Department of Internal Medicine V, Hematology, Oncology and Rheumatology, University of Heidelberg, 69120 Heidelberg, Germany

7. Department of Hematology and Oncology, Medical Faculty Mannheim, Heidelberg University, 68167 Mannheim, Germany.

8. German Cancer Consortium (DKTK), 69120 Heidelberg, Germany

9. Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

10. Stanford Genome Technology Center, Palo Alto, California 94304, USA

* These authors contributed equally

§ These authors jointly supervised this work

Correspondence should be addressed to lars.velten@embl.de or a.trumpp@dkfz.de or s.haas@dkfz.de or larsms@embl.de

Abstract: The step-wise acquisition of genetic abnormalities in cancer is thought to represent a major driver of disease initiation, relapse and therapy resistance. Acute myeloid leukemia (AML) represents a prime example of an aggressive cancer that develops in a multi-step manner from multipotent hematopoietic progenitors via pre-leukemic intermediates to leukemic cells. While bulk and single-cell genomics provide powerful tools to study the phylogenetics of cancer evolution, the specific transcriptomic changes induced by the accumulation of mutations remain largely unexplored. Here, we introduce MutaSeq, a combined single-cell genetic and transcriptomics platform for the identification of molecular consequences of cancer evolution. Through in-depth profiling of an AML patient, we demonstrate that MutaSeq is capable of: (1) fine-mapping clonal and developmental hierarchies (2) quantifying the ability of leukemic and pre-leukemic clones to give rise to mature lineages and (3) identifying surface markers and mRNA transcripts specific to pre-leukemic, leukemic, and residual healthy cells. The experimental and analytical approach presented here is broadly applicable to other types of cancer, and can help identify targets for eradicating both pre-cancerous and cancerous reservoirs of relapse.

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

Introduction

Tumorigenesis is a multistep process, where an accumulation of mutations frequently drives initiation and subsequent progression from a benign neoplasm to an invasive carcinoma and finally to a therapy-resistant disease. Ultimately, an elevated clonal heterogeneity is associated with a worse clinical outcome across cancer types (Andor *et al*, 2016; Maley *et al*, 2006). Moreover, a large fraction of cancer deaths may not be caused by the dominant clone found at diagnosis, but rather by clones that were initially of low abundance but emerge and progress to a highly aggressive state during therapy (Mcgranahan & Swanton, 2017). A better understanding of the molecular consequences of pre-malignant, malignant and subclonal mutations would enable the identification of common pathways associated with disease progression.

This process is exemplified by the etiology of acute myeloid leukemia (AML). In 10-20% of healthy individuals >70 years of age, the acquisition of pre-leukemic mutations in hematopoietic stem cells (HSCs) results in the dominance of a small number of HSC-derived clones (Jaiswal *et al*, 2014). This process, known as clonal hematopoiesis of indeterminate potential (CHIP), is associated with an increased risk of cardiovascular mortality and the development of hematological cancers, such as AML (Jaiswal *et al*, 2014, 2017). Pre-leukemic stem cells are capable of giving rise to healthy blood and immune cells, but frequently display skewed lineage output and decreased differentiation rates (Chan & Majeti, 2013; Shlush *et al*, 2014; Pronier *et al*, 2011). Additional mutations can cause a complete block in differentiation and thereby result in the malignant accumulation of clonal myeloid progenitor cells arrested in their ability to differentiate into mature blood cells. These highly proliferative 'blasts' are unable to efficiently self-renew, but are thought to depend on leukemic stem cells (LSCs) for their continuous production (Meacham & Morrison, 2013; Kreso & Dick, 2014). Blasts, LSCs, and pre-LSCs may be capable to survive chemotherapy and to evolve into more aggressive clones that drive relapse (Ding *et al*, 2012; Shlush *et al*, 2017; Hou *et al*, 2012). Accordingly, chemotherapy usually results in remission, but due to frequent recurrence 5-year overall survival rates are less than 5% for patients >65 years of age (Thein *et al*, 2013). A highly specific strategy to distinguish between residual healthy, pre-leukemic, and leukemic stem- and progenitor cells is required to determine the molecular consequences of clonal evolution and to identify drug targets active in all potential reservoirs of relapse.

Single-cell genomics approaches can potentially provide both the single-cell genetic information needed to distinguish between (sub)clones, as well as the single-cell transcriptomic information needed to unravel the molecular consequences of clonal evolution and to identify differentiation stages. Published work has made use of single-cell RNA-seq data to distinguish between healthy and cancer cells based on copy number variation or a single point mutation (Giustacchini *et al*, 2017; Filbin *et al*, 2018; Tirosh *et al*, 2016; Patel *et al*, 2014). However, a multiplex readout of mutations is required to distinguish between healthy HSCs, pre-LSCs and various leukemic clones. Here, we introduce an experimental and analytical pipeline for multiplex mutation tracking by targeted amplification in single cell RNA-sequencing ("MutaSeq") and demonstrate its utility for the identification of molecular differences between healthy, pre-leukemic and leukemic stem cells as well as their differential ability to give rise to various hematopoietic lineages.

Results

Herein, we develop a method capable of generating high quality transcriptomic data in combination with multiplex mutational information from individual cells. We first evaluated different strategies for targeting mutations as part of a low-cost single-cell RNA-sequencing (scRNA-seq) workflow based on a highly sensitive, modified version of Smart-Seq2 (SmartSeq.HSC from Velten *et al.,* 2017). In particular, we designed primers to target genomic sites during reverse transcription or cDNA amplification steps. We found that inclusion of targeting primers during reverse transcription frequently resulted in the formation of undesired byproducts (Figure EV1a-d). By contrast, when sites of interest were targeted during cDNA amplification, high quality transcriptome data was robustly obtained and the average number of target sites captured per cell was doubled compared to a non-targeted approach (Figure EV1a,b,e). While amplicon length had little effect on target capture rate, the use of short (90-145bp) targeted amplicons during cDNA amplification and the direct incorporation of sequencing adapters (Filbin *et al*, 2018) increased the number of reads on target (Figure EV1b). We termed this protocol *multiplex mutational tracking by targeted amplification of cDNA in single-cell RNA-sequencing* (MutaSeq) (Figure 1a). An automated pipeline for primer design that minimizes off-target sites and potential primer-dimer formation is available at https://git.embl.de/velten/PrimerDesign. With homemade Tn5 transposase (Hennig *et al*, 2018), the cost of library preparation in MutaSeq is below US$4 per cell in a 96-well format, and around $2 in a 384-well format.

To demonstrate the potential of MutaSeq in AML, we performed an in-depth investigation of the genetic and non-genetic heterogeneity of the bone marrow of a 78-year old patient who presented with an AML with normal karyotype, surface marker expression of CD34, CD13, CD38 and HLA-DR and a blast count of 45%. We identified somatic variants present in CD34+ bone marrow cells from this patient by deep exome sequencing in comparison to a non-hematopoietic germline control. This revealed a missense mutation in the splice factor *SRSF2*, two frameshift mutations in the methylcytosine dioxygenase *TET2,* and eleven additional nonsynonymous mutations present with high (near 50%) allele frequency (Figure 1b). Moreover, three nonsynonymous mutations were found at an allele frequency of 15-25%, including a frameshift mutation in the myeloid lineage transcription factor *CEBPA* and a missense mutation in *KLF7*. Twenty-seven mutations, including all coding mutations present at an allele frequency of >10%, were targeted by MutaSeq (table EV1, EV2). We then systematically compared the performance of MutaSeq and non-targeted Smart-Seq2 in a dataset of 658 (MutaSeq) and 206 (Smart-Seq2) bone marrow derived CD34+ cells from that patient. In line with the initial protocol comparison, MutaSeq increased the number of target sites covered per single cell from a median of 1 to a median of 4 (Figure 1c). Importantly, the transcriptome data from both methods covered a median of >3500 genes per CD34+ cell and were highly correlated between methods (Pearson R = 0.9), suggesting that MutaSeq provides increased coverage of mutational sites, while retaining excellent transcriptomic data quality (Figure 1c,d). For genes expressed at intermediate or high levels, MutaSeq recapitulated the variant allele frequencies measured by exome sequencing more accurately than Smart-Seq2 (Figure 1e,f). Mutations occurring in lowly or non-expressed genes (RPKM < 1) could not be called by either method. In case of frameshift mutations, coverage of the mutant alleles was generally decreased, likely as a consequence of nonsense mediated decay (Figure 1f). The three key target sites in *SRSF2*, *CEBPA* and *KLF7* were efficiently captured in 30 to 96% of cells and hence at 1.6 to 5.8 times the number achieved by the classical Smart-Seq2 approach (Figure 1g). Together, these results demonstrate the ability of MutaSeq to efficiently cover genomic target sites in expressed genes during single-cell RNA-sequencing experiments.

Given the information from deep exome sequencing (Figure 1b), the *KLF7* and *CEBPA* mutations could co-occur in a single subclone, or demarcate two independent subclones. To unambiguously identify the number of subclones and their evolutionary hierarchy, we clustered the mutational profile established by MutaSeq on single cells and used SCITE (Jahn *et al*, 2016) to infer a likely clonal hierarchy. These results demonstrate that most covered mutations co-occur within single cells, whereas the *KLF7* mutation and the *CEBPA* mutation were mutually exclusive if taking an empirically determined false positive rate of 3% into account (Figure 2a, and see methods/figure EV3). Additionally, a large fraction of cells carried the persistent *SRSF2* mutation, without mutation calls in *KLF7* and *CEBPA*. To confirm these results, we expanded 278 colonies from single CD34+ leukemic cells and performed targeted sequencing of all 27 mutations from clonal DNA (see methods and Figure EV1f). The result confirmed the presence of a founding *SRSF2* clone and two additional distinct subclones marked by *CEBPA* and *KLF7* mutations, as indicated by MutaSeq (Figure 2b,c). Therefore, the mutational sites highly covered by single-cell RNA sequencing (*SRSF2*, *CEBPA* and *KLF7*) are ideal markers for these clones. Together, these analyses demonstrate that MutaSeq is a powerful tool to confidently identify subclonal structures at the single-cell level.

It has previously been shown that pre-leukemic mutations arise early in AML evolution, likely in healthy hematopietic stem cells (HSCs), and can therefore be detected in all blood and immune lineages. Subsequent mutations often result in differentiation blocks and transformation into leukemic stem cells (Shlush *et al*, 2014). We envisioned that MutaSeq would be ideal for classifying mutations as leukaemic or pre-leukaemic, since it captures both cell type and mutational information. We therefore generated a dataset of 1430 bone marrow cells from the same patient, with enrichment for lineage-negative (Lin-) and CD34+ compartments (Figure EV1g), and FACS-based quantification ('indexing') of multiple differentiation markers (Lin, CD33, CD34, CD38, CD45RA, CD90, CD135) or putative leukemic stem cell markers (Tim3, Gpr56). In these data, we identified a large and heterogeneous cluster of CD34+ leukemic cells (see below) as well as 8 additional CD34- clusters (Figure 3a, b). Based on the analysis of marker gene expression and differential expression tests (Table EV3, Figure 3b,c, EV2), the CD34- clusters were identified as Neutrophils, B-cells, two clusters of NK cells distinguished by the expression of *KLRB1* (CD161) (Kurioka *et al*, 2018), and four clusters of T cells, (CD45RA-*CCR7+SELL+* central memory T cells, CD45RA-*CCR7-SELL-NKG7+* effector memory T cells, Lin+CD45RA+*CD8A+CD8B+* cytotoxic effector T cells, and a CD4-CD8-CD45RA+*CD3+* subset of T cells that expressed the T cell receptor delta chain and other markers of CD16+ γδ T cells described by Ryan *et al*, 2016). Together, these results underline the ability of MutaSeq to distinguish between highly related cell types, such as T and NK cell subsets.

We next mapped the contribution of each individual sub-clone to the mature immune cell populations (Figure 3d,e). *SRSF2* mutations were observed in all lineages with allele frequencies in neutrophils and NK cells being only slightly reduced when compared to the CD34+ precursor and blast cluster. In contrast, *SRSF2* allele frequencies were strongly depleted in B cells and particularly T cells. Together, these data suggest that the *SRSF2* mutation is a pre-leukemic event that has likely been acquired during adulthood when B cell, and in particular T cell, generation is attenuated, while other lineages maintain high turn-over rates. The high penetrance of the mutation into the neutrophil and NK cell lineages, suggests that the pre-leukemic *SRSF2*-mutated clone has effectively taken over hematopoiesis. Importantly, these analyses account for differential allelic drop-out rates as a function of library quality and target gene expression (see methods, Figure EV3). Sequencing of DNA of sorted mature immune lineages confirmed a high abundance of *SRSF2* mutant cells in B cells, NK cells and Neutrophils (Figure 3e, inlay).

The *CEBPA* mutation did occur at low frequency in the mature lineages, but was observed in about 30% of CD34+ cells (Figure 3f). These data suggest that the clonal divergence into the *CEBPA*-mutated subclone is associated with a block in differentiation and a transition from a pre-leukemic to a leukemic state. In contrast, the *KLF7* mutation was observed in all lineages with no evidence for depletion or enrichment in terms of its frequency relative to *SRSF2* mutations, thereby qualifying this subclone as preleukemic. These analyses demonstrate the ability of MutaSeq to distinguish between pre-leukemic and leukemic clones, and link additional subclonal mutations to the acquisition of a differentiation block.

To characterize the cellular heterogeneity of the CD34+ compartment, we first projected our data on single-cell RNA-seq data from healthy human hematopoiesis and used the STEMNET algorithm to estimate the degree of priming into myeloid, lymphoid, megakaryocyte and erythroid lineage (Velten *et al*, 2017) (Figure 4a). We found that a number of cells resembled healthy myeloid, lymphoid, megakaryocyte and erythroid progenitors. We did not observe a significant association of subclonal mutations with altered degrees of lineage priming (Figure 4b); however, the fractions of cells with neutrophil and lymphoid priming were overall significantly decreased in this AML patient, while the fractions of low-primed, megakaryocyte-primed and eosinophil/basophil primed cells were increased (Figure 4c). Like in healthy reference individuals, CD38 expression was correlated with the overall degree of healthy lineage priming (Figure 4d and Velten *et al*, 2017), and megakaryocyte/erythroid progenitors displayed a normal CD45RA-CD135- surface phenotype. Interestingly, most leukemic cells were CD33-CD45RA+CD135+/-, whereas phenotypic common myeloid progenitors or CD33+ cells were virtually absent. These analyses demonstrate the ability of MutaSeq to identify developmental hierarchies in leukemic stem and progenitor cells.

Beyond serving as markers for stages of differentiation, FACS-based indexing of surface markers permits the investigation of differences in surface marker expression between pre-leukemic and leukemic CD34+ cells. We found that surface expression of GPR56, which has previously been suggested to enrich for highly aggressive leukemic stem cells (Pabst *et al*, 2016), was significantly associated with an increased abundance of *CEBPA* mutated leukemic cells (Figure 5a,b). In contrast, *KLF7* mutant cells were enriched in the CD45RA^{low} population. Importantly, in all these analyses covariates affecting allelic dropout rates were taken into account (see methods). This demonstrates the ability of MutaSeq to correlate both transcriptomic and clonal information with surface marker expression, opening the possibility for systematic characterization of cancer stem cell markers at the single-cell level.

Finally, we investigated the effects of leukemic and pre-leukemic mutations on gene expression. Interestingly, when comparing residual healthy CD34+ cells with CD34+ cells that acquired the pre-leukemic *SRSF2* mutation, the *FIS1* gene was most significantly upregulated in the mutant cells. *FIS1* activation has previously been shown to be critical for the transformation of healthy HSCs to LSCs (Pei *et al*, 2018) (Figure 5c). The subsequently acquired leukemic *CEBPA* mutation caused expression changes in several genes (Figure 5d). For example, tumor suppressor genes (TSGs) were significantly enriched among transcripts downregulated in the CEBPA mutant clone ($p = 0.03$, hypergeometric test). These included *MLL3* (*KMT2C*) and the nuclear orphan receptor NR4A1, both playing important roles in the development of AML and can mediate impairment of HSPC differentiation on their own (Chen *et al*, 2014; Wenzl *et al*, 2015). The gene most strongly upregulated in the *CEBPA*-mutant clone was *PTRF* (*CAVIN1*), which has previously been found to be one of few genes consistently upregulated across AML genotypes (Lee *et al*, 2006). Principal component analysis confirmed that the *CEBPA* mutation substantially affected gene expression in the CD34+ compartment (Figure 5a). By contrast, the effects of the *KLF7* mutation were subtle, in line with its minor impact on HSPC differentiation capacity. However, we did observe a small but significant increase in the fraction of *KLF7*-mutated cells in G2/M phase of the mitotic cell cycle, suggesting that the *KLF7* mutation may increase the proliferation rate of CD34+ cells (Figure 5a).

Together, these analyses demonstrate that MutaSeq can be used to interrogate the molecular and cellular consequences of clonal evolution during carcinogenesis.

Discussion

Here we introduce MutaSeq, an experimental and analytical pipeline that combines high quality transcriptomic profiling with multiplex mutational mapping of single cells. MutaSeq is capable of simultaneously determining subclonal structures, identifying subclone-specific gene expression patterns, and mapping developmental hierarchies in leukemic stem and progenitor cells. This makes MutaSeq a powerful tool for the identification of the molecular consequences of clonal evolution in cancer, and will help prioritize drug targets that are active in both cancerous and precancerous cells.

In the AML patient studied here, we have found that initial pre-leukemic *SRSF2*, *TET2* and *SPEN* mutations resulted in the clonal expansion of developmentally immature cells with skewed lineage priming but a high net contribution to all blood and immune lineages except T cells. On a molecular level, the transition from healthy to pre-leukemic CD34+ cells was accompanied by increased expression of *FIS1*, a process that has recently been characterized as crucial for LSC maintenance (Pei *et al*, 2018). Clonal evolution subsequently resulted in the expansion of two subclones in addition to the remaining founder clone. Of these, only one subclone (mutation in the myeloid transcription factor *CEBPA*) was associated with a strict block in differentiation. Concomitantly, the leukemic *CEBPA*-clone down-regulated several tumor suppressor genes, including genes whose loss has previously been linked to the acquisition of a differentiation block and leukemogenesis (Chen *et al*, 2014; Wenzl *et al*, 2015). Together, these results demonstrate the ability of MutaSeq to identify how clonal evolution gives rise to intra-patient heterogeneity and differentiation blocks in leukemogenesis.

In addition, MutaSeq allows the combined simultaneous transcriptional and mutational profiling of single cells with the recording of candidate surface markers by flow cytometry. In our study we tested whether distinct differentiation markers and proposed LSC markers associate with subclonal divergence. Indeed, GPR56, which has previously been suggested to serve as a LSC marker (Pabst *et al*, 2016), was specifically upregulated in the CEBPA-mutant subclone, and therefore upon transition from a pre-leukemic to an leukemic state. This demonstrates the use of MutaSeq to identify or characterize potential LSC makers at the single-cell and subclonal level.

MutaSeq has distinct advantages and disadvantages compared to other methods for mutational profiling during single-cell RNA-seq currently proposed on preprint servers. GoT-Seq (Nam *et al,* biorxiv) performs targeted amplifications from cDNA libraries obtained from droplet-based single-cell RNA-seq, and therefore is biased towards mutations near the cDNA 3' or 5' end. Moreover, it provides a decreased molecular sensitivity compared to MutaSeq (median 4420 genes detected per CD34+ bone marrow cell, compared to ~2300 in GoT-Seq). For mutations in three highly expressed genes, GoT-Seq enabled mutation calls in >60% of CD34+ bone marrow cells, whereas MutaSeq achieved a similar coverage for genes of substantially lower expression (*SRSF2*, *CEBPA* and *EAPP* are expressed at 15% to 45% the level *CALR*). GoT-Seq has a more streamlined workflow but unlike

MutaSeq is incompatible with FACS index-recording of surface markers. TARGET-Seq (Rodriguez-Meira *et al,* biorxiv) performs amplification from genomic DNA that is released into the lysis buffer through a protease digestion step. This offers obvious advantages in the detection of lowly expressed genes, intergenic mutations, or genes subject to nonsense mediated decay. On the downside, the TARGET-Seq protocol requires two additional PCR steps subsequent to the initial pre-amplification and thereby approximately doubles the hands-on time required. In contrast, in the MutaSeq workflow, amplification of the target is achieved as part of the default library preparation protocol. Alternatively, mutations can also be called from single-cell RNA-seq data without targeted amplification (Petti *et al*, biorxiv). However, a low coverage of mutations was thereby achieved, with maximally 26% of cells covering the best-covered mutation across five separate experiments. The advantage of this workflow is that it can be applied to existing data, but insights on clonal identity remain restricted. Together, the four protocols proposed serve distinct requirements, with MutaSeq offering an excellent balance between mutation detection rate and throughput.

In conclusion, MutaSeq is a highly powerful method for identifying the molecular consequences of clonal evolution in cancer. The proposed experimental design, methodology, and data analysis workflow of our study can serve as a blueprint for future large-scale efforts to characterize cellular and molecular properties of LSCs and pre-LSCs and may also be adapted to solid cancers.

Material and Methods

Patient and sample collection

The AML sample was collected from a diagnostic bone marrow aspiration from a 78-year-old individual at the University hospital in Heidelberg, Germany, after obtaining informed written consent. Bone marrow mononuclear cells were isolated by density gradient centrifugation and stored in liquid nitrogen until further use. All experiments involving human samples were conducted in compliance with the Declaration of Helsinki and all relevant ethical regulations and were approved by the ethics committee of the medical faculty of the University of Heidelberg.

Deep exome sequencing

DNA was extracted from $9x10^3$ flow sorted CD34+ cells from a bone marrow sample or buccal swab as healthy reference. Sequencing libraries were constructed using the SureSelect XT HS enrichment system (Agilent), and a mean on-exon sequencing coverage of >90x was obtained. Genomic alignments were performed using BWA MEM (Li, 2013) and cancer variants were identified using MuTeCT2 v3.8 (Cibulskis *et al*, 2013), following GATK best practice recommendations. Variants were annotated using ANNOVAR (Wang *et al*, 2010), and all coding variants present at an allele frequency of at least 10%, plus several variants present at a lower allele frequency were selected for targeting in MutaSeq (table EV1).

FACS sorting

Bone marrow mononuclear cells were stained for 30 minutes on ice according to standard protocols. For single cell liquid cultures and MutaSeq, cells were stained with fluorescent-labelled antibodies against lineage markers (CD4, CD8, CD19, CD20, CD41a, CD235a) and additional markers (CD45RA, CD135, GPR56, CD34, CD38, CD90, CD33, Tim3), and sorted according to the gating scheme illustrated in Figure EV1g. For the mature bulk populations sort (Figure 3e), cells were stained with antibodies against CD45, CD3, CD19, CD56, CD33, CD14 and CD15, and live cells were gated as CD45+CD3+ T cells, CD45+CD19+ B cells, CD45+CD19-CD3-CD56+ NK cells, or CD45+CD19-CD3-CD56-CD33+CD15+CD14- Neutrophils. BD FACS Fusion (BD Biosciences) equipped with 405nm, 488nm, 561nm and 640nm lasers were used. A list of all antibodies used can be found in table EV4.

Primer Design

Primers for MutaSeq (Figure 1), for other single-cell targeting protocols tested (Figure EV1), as well as for targeted DNA sequencing were designed using the computational pipeline available at http://git.embl.de/velten/PrimerDesign . For MutaSeq, the refgene transcripts spanning each genomic site of interest were selected as template; if multiple refgene transcripts were found for one site, a consensus transcript containing only exonic sequences present in all variants was created. We then used primer3 (Untergasser *et al*, 2012) to design five possible pairs of primers, with an amplicon length of 90-145bp and a melting temperature of (nominally) 60°C. BLAST was used to remove primer pairs which potentially form off-target amplicons. Then, the pair complementarity (i.e. potential to form dimers) was computed for each possible combination of primers (forward-reverse, forward-forward and reverse-reverse), and the set of primer pairs that minimizes the total sum of complementarities was selected using a stepwise optimization algorithm. Nextera adapters were then added (fwd: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG, rev: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG).

For the protocol that combines targeting during the RT and targeting during the PCR (Figure EV1a, 'Targeted RT + Direct library PCR'), an RT primer was added upstream of the PCR amplicon. For the protocol that only targets during the RT (Figure EV1a, 'Targeted RT'), the same RT primers were fused to the ISPCR oligo sequence (AAGCAGTGGTATCAACGCAGAGT, Picelli *et al*, 2013). For the protocol labeled 'Targeted PCR' in Figure 1b, a primer pair with amplicon length 200-350 and no adapters was used.

For targeted DNA-seq experiments, the genomic sequence surrounding the target was used as template. Inner primers were designed as in the case of MutaSeq, and outer primers surrounding the inner PCR product with an amplicon length of 200-350bp and a nominal annealing temperature of 58°C were added.

A list of all primers used for this study is included in Table EV2.

Single cell RNA sequencing with targeting of genomic sites of interest

MutaSeq is based on the Smart-Seq2 protocol (Picelli *et al*, 2013, 2014) with the modifications introduced by (Velten *et al*, 2017). For lysis, we used 5μL of a buffer containing 0.1μL RNAsin+

(Promega), 0.04µL 10% Triton X-100 (SigmaAldrich), 0.1µL of 100µM Smart-Seq2 Oligo-dT primer (SigmaAldrich), 1µL dNTP mix (10mM each, NEB), and 0.075µL of a 1:1,000,000 dilution of ERCC spike in mix 1 (Ambion). We added an additional polyadenylated in vitro transcript, pGIBS-Thr (Pelechano *et al*, 2012), to every second well (A1, A3, B2, B4 etc.) to control for false positives (see below). A site on pGIBS-Thr was targeted during MutaSeq in the same way as genomic target sites. Plates were snap frozen directly after sorting and later thawed at 10°C in a PCR machine for 5' and denatured at 72°C for 3'. 5µL of a buffer containing 0.25µL RNAsin+, 2µL 5x SMART FS buffer, 0.5µL DTT 20mM, 1µL SmartScribe enzyme (all TaKaRa) and 0.2µL 50µM Smart-Seq2 TSO (Exiqon) were then added and RT was performed for 90' at 42°C, 10 cycles of [50°C, 2' and 42°C, 2'], and enzyme inactivation at 70°C for 15'. Then, we added 15µL PCR mix containing 12.5µL KAPA HiFi HS mastermix (Merck), 0.25µL 10µM Smart-Seq2 ISPCR primer (SigmaAldrich) and 0.5µL of a pool of all targeting primers, present at 1µM each. cDNA amplification was performed by 98°C 3', 21 cycles of [98°C, 20'', 67°C, **60''**, 72°C 6'] and 72°C, 5'. cDNA was the cleaned up using an equal volume (25µL) of CleanPCR beads (CleanNA) and tagmented using homemade Tn5 (Hennig *et al*, 2018).

For the alternatives to MutaSeq displayed in Figure 1b, the same protocol was used only that the annealing time during PCR was reduced to 15'' if no targeting primers were present at PCR stage. For targeting during the RT, 0.0625 µL ('Targeted RT' in Figure EV1a) or 0.35µL ('Targeted RT + Direct library PCR') of a pool of all targeting RT primers, present at 1µM each, was added to the lysis buffer; see Figure EV1c-e for the effect of varying primer concentrations.

Single cell cultures

Bone Marrow mononuclear cells were stained and Lin- or Lin-CD34+ single cells were index-sorted into ultra-low attachment 96-well plates (Corning) containing 100µL StemSpan SFEM media (Stem Cell Technologies). Media was supplemented with penicillin/streptomycin (100ng/mL), L-glutamine (100ng/mL) and the following human cytokines (all from Peprotech): SCF (20ng/mL), Flt3-L (20ng/mL), TPO (50ng/mL), IL-3 (20ng/mL), IL-6 (20ng/mL), G-CSF (20ng/mL), EPO (40ng/mL), IL-5 (20ng/mL), M-CSF (20ng/mL), GM-CSF (50ng/mL). After 21 days at 5% CO_2 and 37°C, colonies were imaged by microscopy, and processed as detailed in the following.

Targeted DNA sequencing by nested PCR amplification

Single cell derived colonies or mature immune populations were transferred into 50µL buffer RLT (Qiagen). Cleanup was performed using CleanPCR beads (CleanNA) at a 1.8x volume ratio and eluted in 20µL 10mM Tris-HCl pH 7.8. 4.5µL were transferred to a PCR plate containing 7.5µL Kapa HiFi HS mastermix and 3 µL of a pool of all outer primers (each primer at 0.5µM) were added, followed by a PCR program of 98°C 3', 30 cycles of [98°C, 20'', 63°C, 60'', 72°C 10''] and 72°C, 5' and subsequent enzymatic cleanup with 2.5µL 10x ExoI buffer, 0.4µL ExoI (NEB) and 0.4µL FastAP (ThermoFisher), 30' incubation at 37°C and 5' inactivation at 95°C. Afterwards, 1µL was transferred to a PCR tube containing 5.9µL water, 7.5µL Kapa HiFi HS mastermix and 0.6µL of a pool of all outer primers (each primer at 0.5µM), followed 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 cleanup as above. 1µL was then transferred to a PCR with nextera indexing primers as described in (Hennig *et al*, 2018) and amplified with 98°C 3', 10 cycles of [98°C, 20'', 60°C, 15'', 72°C 30''] and 72°C, 5'. Sample bioanalyzer traces are shown in Figure EV1f.

Processing of Next Generation Sequencing data

Raw sequencing reads from MutaSeq and Smart-Seq2 experiments were processed using the BBDuk software to trim both the standard Illumina Nextera adapters and the ISPCR adapter. Reads were then mapped to the hg38 human genome (Ensembl release 89), with ERCC and pGIBS-Thr sequences appended, using STAR v2.6 (Dobin et al, 2013), with the outFilterMismatchNmax parameter set to 5. Exonic gene counts were tabulated, keeping only reads that did not overlap with targeted regions, overlapped with only one annotated genomic feature, and were longer than 30 bp. Reads covering the reference and mutant alleles for sites of interest were counted in each single cell using the deepSNV R package (Gerstung et al, 2012), and cells were classified as dropout if less than *k* reads covered a given location, mutant if at least 5% of the base calls covered the mutant allele, and reference otherwise. *k* was set to 10% the average number of reads on the site across all cells. Thereby, the false positive rate of MutaSeq was controlled at 3%, as estimated using the pGIBS-Thr synthetic spike in control included in every second well (Figure EV3a,b).

For DNA sequencing experiments, reads were mapped to the hg38 human genome using bwa mem (v0.7.17) (Li, 2013). Mutation calls were made as described above.

Reconstruction of clonal hierarchies

The *CEBPA* and *KLF7* mutations were mutually exclusive in 109 cells with both alleles covered, and co-occurred in 2 cells. The SCITE model can be used to formally infer a clonal hierarchy by considering both false positive rates and allelic dropout rates (Jahn et al, 2016). We fitted the SCITE model to the data using an FP rate of 0.001 to be maximally conservative with regard to false positives, and learnt an average allelic dropout rate from the data. This consistently resulted in the split displayed in figure 2 across five independent MCMC chains, both for the MutaSeq and for the Colony-Seq data.

Single cell gene expression data analysis

Gene count tables were loaded into R and further processed using the Seurat R package (Butler *et al*, 2018). Cells with less than 500 distinct genes observed and cells with more than 5% of UMIs stemming from mitochondrial genes were removed. During initial clustering, we identified an outlier group of 26 cells that all originated from the bottom rows of various plates; these cells were removed and likely originate from an 'edge effect'. PCA was then performed on significantly variable genes, and the first 9 PCs were selected as input for clustering and t-SNE, based on manual inspection of a PC variance plot ("PC elbow plot"). Clustering was performed using the default method from the Seurat package, with resolution parameter set to 6. While lower resolution parameters caused biologically distinct T cell subsets to be merged into a single cluster, this relatively large parameter resulted in the heterogeneous CD34+ cluster to be split in a large number of subgroups. For the analyses at the cell type level (Figure 3), all CD34+ cells were treated as one group, and marker genes for each population were identified using the FindMarkersAll function and ROC-based test statistics.

To compare CD34+ cells from the patient to CD34+ cells from a healthy individual, we made use of the STEMNET algorithm (Velten *et al*, 2017). In short, STEMNET uses data from 1034 healthy CD34+ hematopoietic stem and progenitor cells to identify genes specific to lineage restricted progenitor populations (Neutrophil, Eosinophil/Basophil, B-cell, Monocyte, Erythroid and Megakaryocyte progenitors). STEMNET then computes the probability that a stem, progenitor or leukemic cell can be assigned to any of these classes. STEMNET thereby places cells with a large similarity to healthy progenitor cells on the corners of a simplex, while any cell that does not resemble a healthy progenitor cell is placed in the centre. The visualization in figure 4a,b results from the projection of this simplex on the unit circle (Velten *et al*, 2017).

Joint analysis of single cell gene expression and mutational data

The probability that MutaSeq covers a target genomic site in a single cell depends on the expression of the gene in that cell (Figure EV3c and see also Figure 1e). For a purely qualitative analysis, cells can be classified as 'mutant' if the mutant allele is observed, and as 'reference' if only the reference allele is observed. However, a reference call can then originate either from a mutant cell with a dropout event or a non-mutant cell. The apparent fraction of non-mutant cells is therefore higher in cells with a high drop-out rate, and correlated with the expression of the target gene (Figure EV3c). The sequencing depth per cell also affects dropout rate and thereby the apparent abundance of mutant cells (Figure EV3c). To account for these confounders, the association between a feature F of interest (e.g. cell type identity in Figure 3e or surface marker expression in Figure 5a) and mutational status was computed using weighted multivariate logistic regression

$$M_s \sim X_{g(s)} + N + F$$

Where $M_s \in \{0,1\}$ is the mutational status at a genomic site s, $X_{g(s)}$ is the log-normalized, scaled expression of the gene spanning s (i.e. the number of reads aligning anywhere in the gene body), and N is the number of genes observed per cell as a measure of library quality. The logarithm of the total reads covering the exact genomic site s was used as weights, since the probability of dropout in either allele was tightly correlated to the number of reads on that site (Figure EV3d). Cells with no coverage of the genomic site of interest were thereby excluded. The resulting models were then compared to a reduced model that omits the feature term F, and p-values were computed using ANOVA and a likelihood ratio test. For the identification of genes with differential expression between clones, expression values were used as the feature F and the same test was applied. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. For all continuous features F, including gene expression values, sample ranks were used.

In Figures 3e, f and 5b, binned estimates of allele frequency were computed as the weighted means of single-cell mutation status, using the same weights introduced above. Non-parametric bootstrapping with 1000 sampling repeats was used to obtain confidence intervals.

Data availability

Raw sequencing data and VCF files will be made available through the European Genome-phenome Archive (EGA) upon publication or earlier upon justified request. Processed data is further available as dataset EV1.

Acknowledgements

We thank Niko Beerenwinkel for discussions, Andreas Gschwind for contributing code and Britta Velten for critical reading of the manuscript. We thank the members of the Steinmetz, Haas and Trumpp 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 José-Carreras Stiftung grant DJCLS 20 R/2017 (to L.V., A.T., S.H. L.S.) and the HIPO2 program of the Heidelberg Center for Personalized Oncology (DKFZ-HIPO) (to S.H., A.T.). Moreover, this work was supported by the FOR 2674 funded by the Deutsche Forschungsgemeinschaft (DFG), the SyTASC consortium funded by the Deutsche Krebshilfe and the Dietmar Hopp Foundation (all to A.T.). Contributions by D.N., J.C.J., W.K.H. and T.B. were supported by the Gutermuth Foundation, the H.W. & J. Hector fund, Baden Wuerttemberg, and the Dr. Rolf M. Schwiete Fund, Mannheim. D.N. is, an endowed Professor of the Deutsche José Carreras Leukaemie Stiftung (DJCLS H 03/01).

Author contributions

LV developed MutaSeq and performed single cell RNA-seq experiments with assistance by JM. PH performed initial sample characterizations, single cell culture experiments and FACS sorting with support by MP. BAS and LV analyzed the data. LV, SH, AT and LMS conceived the study and jointly supervised work on the project. LV, SH, BAS and PH wrote the manuscript with contributions from the other main authors. All other authors were involved in the collection and initial characterization of samples. All authors have read and understood the manuscript.

Conflict of interest

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

References

Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, Ji HP & Maley CC (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22:** 105–113

Butler A, Hoffman P, Smibert P, Papalexi E & Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36:** 411–420

Chan SM & Majeti R (2013) Role of DNMT3A, TET2, and IDH1/2 mutations in pre-leukemic stem cells in acute myeloid leukemia. *Int. J. Hematol.* **98:** 648–657

Chen C, Liu Y, Rappaport AR, Kitzing T, Schultz N, Zhao Z, Shroff AS, Dickins RA, Vakoc CR, Bradner JE, Stock W, LeBeau MM, Shannon KM, Kogan S, Zuber J & Lowe SW (2014) MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. *Cancer Cell* **25:** 652–65

Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES & Getz G (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31:** 213–219

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD, McMichael JF, Wallis JW, Lu C, Shen D, Harris CC, Dooling DJ, Fulton RS, Fulton LL, Chen K, Schmidt H, et al (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481:** 506–10

Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, Neftel C, Frank N, Pelton K, Hebert CM, Haberler C, Yizhak K, Gojo J, Egervari K, Mount C, Galen P van, Bonal DM, Nguyen Q-D, Beck A, Sinai C, et al (2018) Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science (80-.).* **360:** 331–335

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS & Gottardo R (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16:** 278

Giustacchini A, Thongjuea S, Barkas N, Woll PS, Povinelli BJ, Booth CAG, Sopp P, Norfo R, Rodriguez-Meira A, Ashley N, Jamieson L, Vyas P, Anderson K, Segerstolpe Å, Qian H, Olsson-Strömberg U, Mustjoki S, Sandberg R, Jacobsen SEW & Mead AJ (2017) Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23:** 692–702

Hennig BP, Velten L, Racke I, Tu CS, Thoms M, Rybin V, Besir H, Remans K & Steinmetz LM (2018) Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3 (Bethesda).* **8:** 79–89

Hou H-A, Kuo Y-Y, Liu C-Y, Chou W-C, Lee MC, Chen C-Y, Lin L-I, Tseng M-H, Huang C-F, Chiang Y-C, Lee F-Y, Liu M-C, Liu C-W, Tang J-L, Yao M, Huang S-Y, Ko B-S, Hsu S-C, Wu S-J, Tsay W, et al (2012) DNMT3A mutations in acute myeloid leukemia: stability during disease evolution and clinical implications. *Blood* **119:** 559–68

Jahn K, Kuipers J & Beerenwinkel N (2016) Tree inference for single-cell data. *Genome Biol.* **17:** 86

Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman P V., Mar BG, Lindsley RC, Mermel CH, Burtt N, Chavez A, Higgins JM, Moltchanov V, Kuo FC, Kluk MJ, Henderson B, Kinnunen L, Koistinen HA, Ladenvall C, Getz G, Correa A, et al (2014) Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371:** 2488–2498

Jaiswal S, Natarajan P, Silver AJ, Gibson CJ, Bick AG, Shvartz E, McConkey M, Gupta N, Gabriel S, Ardissino D, Baber U, Mehran R, Fuster V, Danesh J, Frossard P, Saleheen D, Melander O, Sukhova GK, Neuberg D, Libby P, et al (2017) Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377:** 111–121

Kreso A & Dick JE (2014) Evolution of the cancer stem cell model. *Cell Stem Cell* **14:** 275–291

Kurioka A, Cosgrove C, Simoni Y, van Wilgenburg B, Geremia A, Björkander S, Sverremark-Ekström E, Thurnheer C, Günthard HF, Khanna N, Walker LJ, Arancibia-Cárcamo C V., Newell EW, Willberg CB & Klenerman P (2018) CD161 Defines a Functionally Distinct Subset of Pro-Inflammatory Natural Killer Cells. *Front. Immunol.* **9:**

Lee S, Chen J, Zhou G, Shi RZ, Bouffard GG, Kocherginsky M, Ge X, Sun M, Jayathilaka N, Kim YC, Emmanuel N, Bohlander SK, Minden M, Kline J, Ozer O, Larson R a, LeBeau MM, Green ED, Trent J, Karrison T, et al (2006) Gene expression profiles in acute myeloid leukemia with common translocations using SAGE. *Proc Natl Acad Sci U S A* **103:** 1030–1035

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Maley CC, Galipeau PC, Finley JC, Wongsurawat VJ, Li X, Sanchez CA, Paulson TG, Blount PL, Risques R-A, Rabinovitch PS & Reid BJ (2006) Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38:** 468–473

Mcgranahan N & Swanton C (2017) Review Clonal Heterogeneity and Tumor Evolution : Past , Present , and the Future. *Cell* **168:** 613–628

Meacham CE & Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. *Nature* **501:** 328–37

Nam AS, Kim K, Chaligne R, Izzo F, Ang C, Omans ND, Taylor J, Pastore A, Alonso A, Mariani M, Cubillos-ruiz JR, Tam W, Hoffman R, Scandura JM, Rabadan R, Abdel-wahab O, Smibert P & Landau DA High Throughput droplet single-cell Genotyping of Transcriptomes (GoT) reveals the cell identity dependency of the impact of somatic mutations.

Pabst C, Bergeron A, Lavallée VP, Yeh J, Gendron P, Norddahl GL, Krosl J, Boivin I, Deneault E, Simard J, Imren S, Boucher G, Eppert K, Herold T, Bohlander SK, Humphries K, Lemieux S, Hébert J, Sauvageau G & Barabé F (2016) GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood* **127:** 2018–2027

Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed B V, Curry WT, Martuza RL, Louis DN, Rozenblatt-Rosen O, Suvà ML, Regev A & Bernstein BE (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344:** 1396–401

Pei S, Minhajuddin M, Adane B, Khan N, Stevens BM, Mack SC, Lai S, Rich JN, Inguva A, Shannon KM, Kim H, Tan AC, Myers JR, Ashton JM, Neff T, Pollyea DA, Smith CA & Jordan CT (2018) AMPK/FIS1-Mediated Mitophagy Is Required for Self-Renewal of Human AML Stem Cells. *Cell Stem Cell* **23:** 86–100.e6

Pelechano V, Wilkening S, Järvelin AI, Tekkedil MM & Steinmetz LM (2012) Genome-wide polyadenylation site mapping. *Methods Enzymol.* **513:** 271–96

Petti AA, Williams SR, Miller CA, Fiddes IT, Sridhar N, Chen DY, Fronick CC, Fulton RS & Church DM (2018) Mutation detection in thousands of acute myeloid leukemia cells using single cell RNA-sequencing.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G & Sandberg R (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10:** 1096–1098

Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S & Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9:** 171–81

Pronier E, Almire C, Mokrani H, Vasanthakumar A, Simon A, da Costa Reis Monte Mor B, Massé A, Le Couédic J-P, Pendino F, Carbonne B, Larghero J, Ravanat J-L, Casadevall N, Bernard OA, Droin N, Solary E, Godley LA, Vainchenker W, Plo I & Delhommeau F (2011) Inhibition of TET2-mediated conversion of 5-methylcytosine to 5-hydroxymethylcytosine disturbs erythroid and granulomonocytic differentiation of human hematopoietic progenitors. *Blood* **118:** 2551–5

Rodriguez-Meira A, Buck G, Clark S, Povinelli BJ, Alcolea V, Louka E, Mcgowan S, Hamblin A, Sousos N & Barkas N Unraveling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA-sequencing.

Ryan PL, Sumaria N, Holland CJ, Bradford CM, Izotova N, Grandjean CL, Jawad AS, Bergmeier LA & Pennington DJ (2016) Heterogeneous yet stable Vδ2(+) T-cell profiles define distinct cytotoxic effector potentials in healthy human individuals. *Proc. Natl. Acad. Sci.* **113:** 14378–14383

Shlush LI, Mitchell A, Heisler L, Abelson S, Ng SWK, Trotman-Grant A, Medeiros JJF, Rao-Bhatia A, Jaciw-Zurakowsky I, Marke R, McLeod JL, Doedens M, Bader G, Voisin V, Xu C, McPherson JD, Hudson TJ, Wang JCY, Minden MD & Dick JE (2017) Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547:** 104–108

Shlush LI, Zandi S, Mitchell A, Chen WC, Brandwein JM, Gupta V, Kennedy JA, Schimmer AD, Schuh AC, Yee KW, McLeod JL, Doedens M, Medeiros JJF, Marke R, Kim HJ, Lee K, McPherson JD, Hudson TJ, Pan-Leukemia Gene Panel Consortium TH, Brown AMK, et al (2014) Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506:** 328–333

Thein MS, Ershler WB, Jemal A, Yates JW & Baer MR (2013) Outcome of older patients with acute myeloid leukemia: an analysis of SEER data over 3 decades. *Cancer* **119:** 2720–7

Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin J-R, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CGK, Kazer SW, et al (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352:** 189–196

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M & Rozen SG (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40:** e115–e115

Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, Hirche C, Lutz C, Buss EC, Nowak D, Boch T, Hofmann W-K, Ho AD, Huber W, Trumpp A, Essers MAG & Steinmetz LM (2017) Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19:** 271–281

Wang K, Li M & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38:** e164–e164

Wenzl K, Troppan K, Neumeister P & Deutsch AJA (2015) The nuclear orphan receptor NR4A1 and NR4A3 as tumor suppressors in hematologic neoplasms. *Curr. Drug Targets* **16:** 38–46

Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO & Botstein D (2002) Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Mol. Biol. Cell* **13:** 1977–2000

Zhao M, Kim P, Mitra R, Zhao J & Zhao Z (2016) TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* **44:** D1023–D1031

Figure legends

Figure 1. MutaSeq permits efficient targeting of mutations during single-cell RNA-sequencing.

a. Overview of the MutaSeq protocol. Targeting primers (orange) are included during the cDNA amplification step of the Smart-Seq2 protocol. Targeting primers are directly fused to illumina library adapters (dark green). Tagmentation introduces the same adapters to the full-length cDNA product.

b. Overview of clonal and subclonal mutations present in the patient studied here. Variant allele frequency was computed from deep exome sequencing with a mean coverage of >90x.

c. Number of target sites and genes covered per cell, across n=206 (Smart-Seq2) or n=658 CD34+ (MutaSeq) leukemic cells.

d. Correlation in mean gene expression, across n=206 (Smart-Seq2) or n=658 CD34+ (MutaSeq) leukemic cells.

e. Scatter plot depicting the relationship between gene expression level and mutation site coverage in Smart-Seq2 (grey dots) and MutaSeq (red dots).

f. Allele frequency estimates derived from deep exome sequencing compared to allele frequency estimates derived from MutaSeq (red dots) or Smart-Seq2 (grey dots)

g. Fractions of cells covering key non-synonymous mutations observed in the patient.

Figure 2. MutaSeq permits the reconstruction of clonal hierarchies.

a. Left panel: Heatmap depicting mutation calls across n=872 cells with coverage of at least two mutation sites. Right panel: Summarized output of SCITE (Jahn *et al*, 2016).

b. Left panel: Heatmap depicting mutation calls across n=206 colonies. Right panel: Summarized output of SCITE.

c. Clone sizes estimated from single-cell MutaSeq, colony DNA-seq, and deep exome seq. For MutaSeq and colony-seq, residual healthy cells were defined as no observed mutation in *SRSF2*, *TET2*, *SPEN* and *EAPP*. Fractions were computed using means of mutational status, weighted by on-site coverage. For exome-seq, residual healthy cells were defined as one minus the mean allele frequency of all clonal mutations.

Figure 3. MutaSeq quantifies the contribution of leukemic clones to mature immune lineages.

a. t-SNE visualization of cell types present in total bone marrow of a leukemia patient. For detail on cell type annotation, see main text and Figure EV2.

b. Surface marker expression (FACS index values) of all cells sequenced.

c. Expression of marker genes for mature lineages superimposed on the t-SNE from panel a, see also figure EV2a.

d. Mutation calls for four key mutations superimposed on the t-SNE from panel a.

e. Quantification of the fraction of cells carrying an *SRSF2* mutation in the various populations. Fractions were computed using the mean of mutational status, weighted by the logarithm of reads spanning the genomic site of interest; error bars indicate 90% confidence intervals estimated from a non-parametric bootstrap. P-values were computed using a weighted generalized linear model that accounts for the potentially confounding variables library size and *SRSF2* gene expression (see methods). Multiple testing correction was performed using the method of Benjamini and Hochberg, and asterisk indicate significance as follows: ***, $p_{adj} < 0.001$; **, $p_{adj} < 0.01$; *, $p_{adj} < 0.1$. Inset: Fraction of *SRSF2* mutated cells estimated from MutaSeq are compared to *SRSF2* allele frequencies measured by targeted DNA-seq of mature T cells, B cells, neutrophils and NK cells (circles), and exome-seq of CD34+ cells (triangle); see panel f for color code.

f. Quantification of the fraction of cells carrying the subclonal *CEBPA* or *KLF7* mutations in relation the fraction of cells carrying the *SRSF2* mutation. Calculations were done as in panel e; B-cells were omitted due to insufficient coverage. p-values were calculated against the null hypothesis that the ratio between subclonal mutation and SRSFS2 mutation was constant in all populations (dotted line).

Figure 4. MutaSeq identifies residual healthy lineage priming in leukemic CD34+ cells.

a. Identification of healthy-like lineage priming in all cells from the heterogeneous LSC/Blast/HSPC cluster (Figure 3a). Data were projected on a dataset of CD34+ cells from a healthy individual (see Velten *et al*, 2017 and methods).

b. Superimposition of mutation calls on the projection from panel a.

c. Quantitative comparison of the number of cells with various directions of lineage priming between a healthy and a leukemic individual. The number of cells with a lineage priming of > -0.7 (triangles in panel b) were compared; p-values and confidence intervals were computed using fisher's test.

d. Superimposition of the direction and degree of healthy-like lineage priming on FACS index values.

Figure 5. MutaSeq identifies the effects of mutations on gene expression in CD34+ cells

a. Tests for correlation between the presence or absence of mutations in the three clonal markers *SRSF2*, *CEBPA* and *KLF7*, and a variety of features including surface marker expression (left panel), scores from principal component analysis (middle panel) and cell cycle scores computed using the default method from Seurat (Butler *et al*, 2018) and the expression of marker genes typically used for this application (Whitfield *et al*, 2002) (right panel). P values were computed by accounting for potentially confounding variables such as expression of the target gene and library size, see methods; multiple testing correction was performed using the method of Benjamini and Hochberg.

b. Estimation of the fraction of cells carrying a mutation in *CEBPA* in 20 equal-sized bins of GPR56 surface expression. Means and 90% confidence intervals were computed as in fugure 3e, see also methods.

c. Identification of genes affected by *SRSF2* mutational status, see methods.

d. Identification of genes affected by *CEBPA* mutational status, see methods. Highlighted in red are tumor suppressor genes from the TSGene 2.0 database (Zhao *et al*, 2016).

Tables & table legends

Table EV1. Identification of somatic variants targeted in MutaSeq.

Table EV2. List of primers used in MutaSeq.

Table EV3. Identification of marker genes for mature lineages.

Table EV4. Antibodies used for flow cytometry.

Dataset EV1. Count tables, mutation calls, and other data from the MutaSeq experiment.

Tables EV1-EV4 are available as excel files accompanying this submission.

Expanded View Figure legends

Figure EV1. Development of MutaSeq

a. Comparison of different strategies of targeting the mutations of interest. 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 1a. 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.

b. Boxplot comparing the mean number of reads per target in the different protocols.

c. Bioanalyzer traces for the 'targeted RT' protocol and varying RT primer concentrations.

d. Bioanalyzer traces for the 'targeted RT + direct library PCR' protocol and varying RT primer concentrations.

e. Bioanalyzer traces for the 'direct library PCR' protocol (MutaSeq) and varying PCR primer concentrations.

f. Representative bioanalyzer traces from the nested protocol for targeted DNA amplification

g. FACS sorting scheme used. 8 plates of Total Bone Marow cells, 4 plates of lineage negative cells and 10 plates of CD34+ cells were processed for MutaSeq.

Figure EV2: Characterization of T and NK cell subtypes

a. Gene expression of select marker genes highlighted on the t-SNE from Figure 1a. Gray: No expression, dark red: low expression, orange: maximal expression.

b. Top differentially expressed genes between T-cell subsets; MAST was used for differential expression testing (Finak *et al*, 2015).

Figure EV3: Analysis of false negatives and false positives

a. Abundance of the pGIBS-Thr spike in across wells from four representative plates. pGIBS-Thr was spiked in to every second well (wells A1, A3, B2, B4, etc).

b. Estimation of the false positive rate using the pGIBS-Thr spike in. Dashed bold line indicates the threshold used for classifying a site as dropout.

c. The fraction of CD34+ cells with a dropout in *SRSF2*, as well as the fraction of non-dropout cells with a mutant allele observed, in relation to the expression of *SRSF2* and the number of genes observed per cell.

d. Number of reads covering the genomic site of interest in the *SRSF2* gene in relation to the fraction of dropouts in the reference and alternative allele.

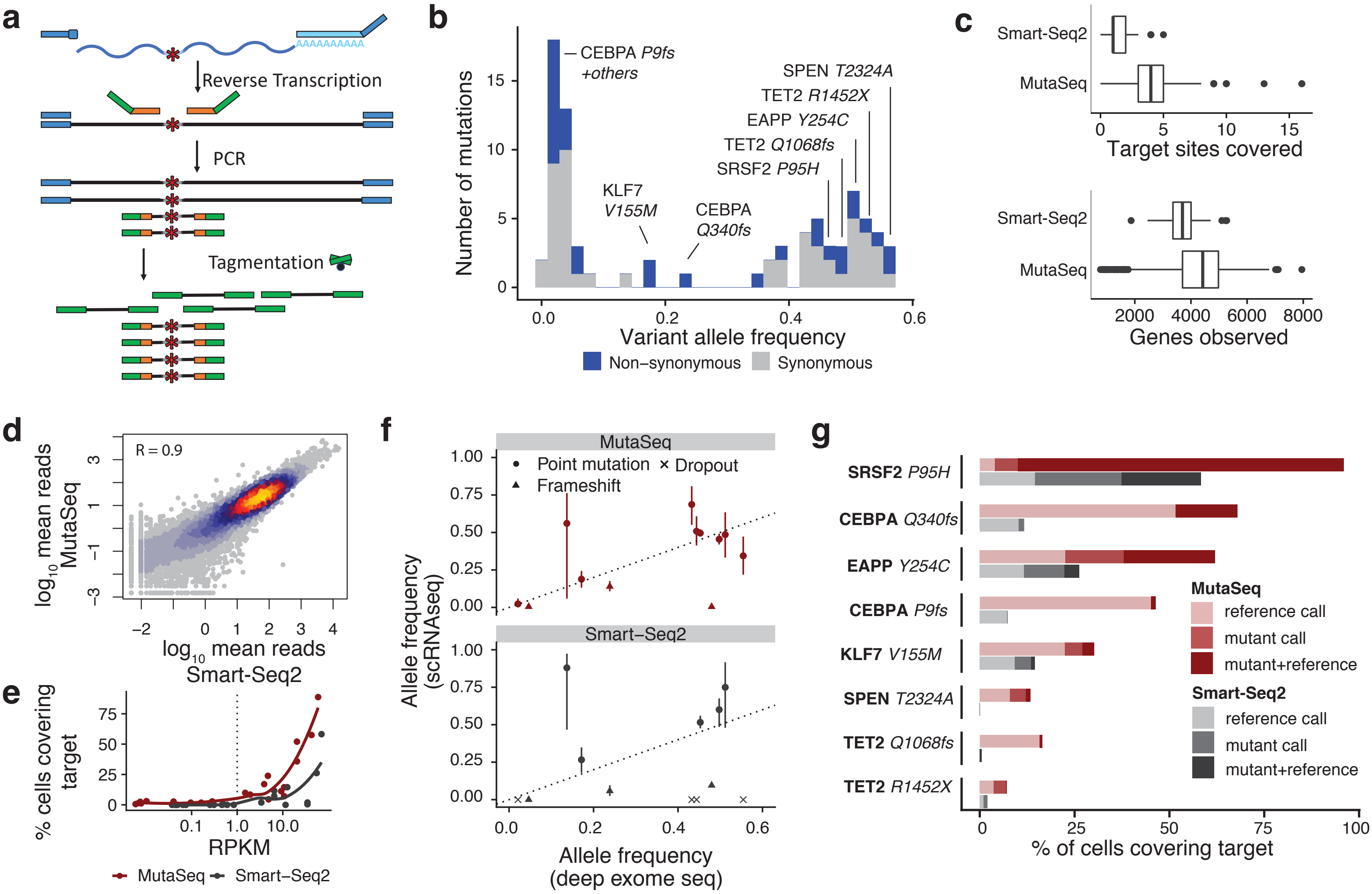
a
AAAAAAAAAA
Reverse Transcription
PCR
Tagmentation
b
Number of mutations
15
10
5
0
CEBPA P9fs
+others
KLF7
V155M
CEBPA
Q340fs
SPEN T2324A
TET2 R1452X
EAPP Y254C
TET2 Q1068fs
SRSF2 P95H
0.0
0.2
0.4
0.6
Variant allele frequency
Non−synonymous
Synonymous
c
Smart-Seq2
MutaSeq
0
5
10
15
Target sites covered
Smart-Seq2
MutaSeq
2000
4000
6000
8000
Genes observed
d
R = 0.9
$\log_{10}$ mean reads MutaSeq
3
1
−1
−3
−2
0
1
2
3
4
$\log_{10}$ mean reads Smart-Seq2
e
% cells covering target
75
50
25
0
0.1
1.0
10.0
RPKM
MutaSeq
Smart−Seq2
f
MutaSeq
Point mutation
Dropout
Frameshift
Smart−Seq2
1.00
0.75
0.50
0.25
0.00
Allele frequency (scRNAseq)
0.0
0.2
0.4
0.6
Allele frequency (deep exome seq)
g
SRSF2 P95H
CEBPA Q340fs
EAPP Y254C
CEBPA P9fs
KLF7 V155M
SPEN T2324A
TET2 Q1068fs
TET2 R1452X
MutaSeq
reference call
mutant call
mutant+reference
Smart-Seq2
reference call
mutant call
mutant+reference
0
25
50
75
100
% of cells covering target

Figure 1

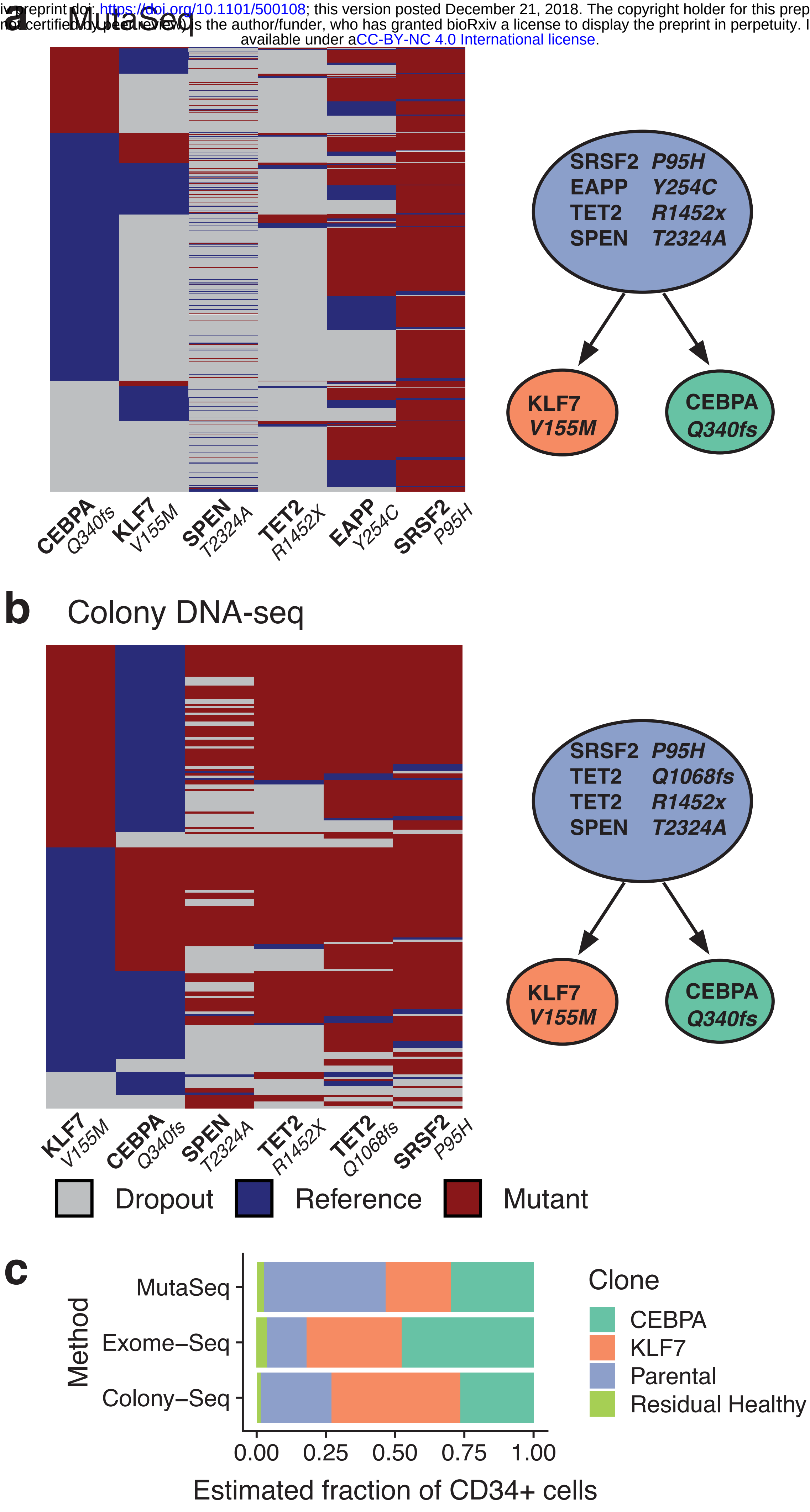
a
MutaSeq
SRSF2 P95H
EAPP Y254C
TET2 R1452x
SPEN T2324A
KLF7 V155M
CEBPA Q340fs
CEBPA Q340fs
KLF7 V155M
SPEN T2324A
TET2 R1452X
EAPP Y254C
SRSF2 P95H
b
Colony DNA-seq
SRSF2 P95H
TET2 Q1068fs
TET2 R1452x
SPEN T2324A
KLF7 V155M
CEBPA Q340fs
KLF7 V155M
CEBPA Q340fs
SPEN T2324A
TET2 R1452X
TET2 Q1068fs
SRSF2 P95H
Dropout
Reference
Mutant
c
Method
MutaSeq
Exome−Seq
Colony−Seq
0.00
0.25
0.50
0.75
1.00
Estimated fraction of CD34+ cells
Clone
CEBPA
KLF7
Parental
Residual Healthy

Figure 2

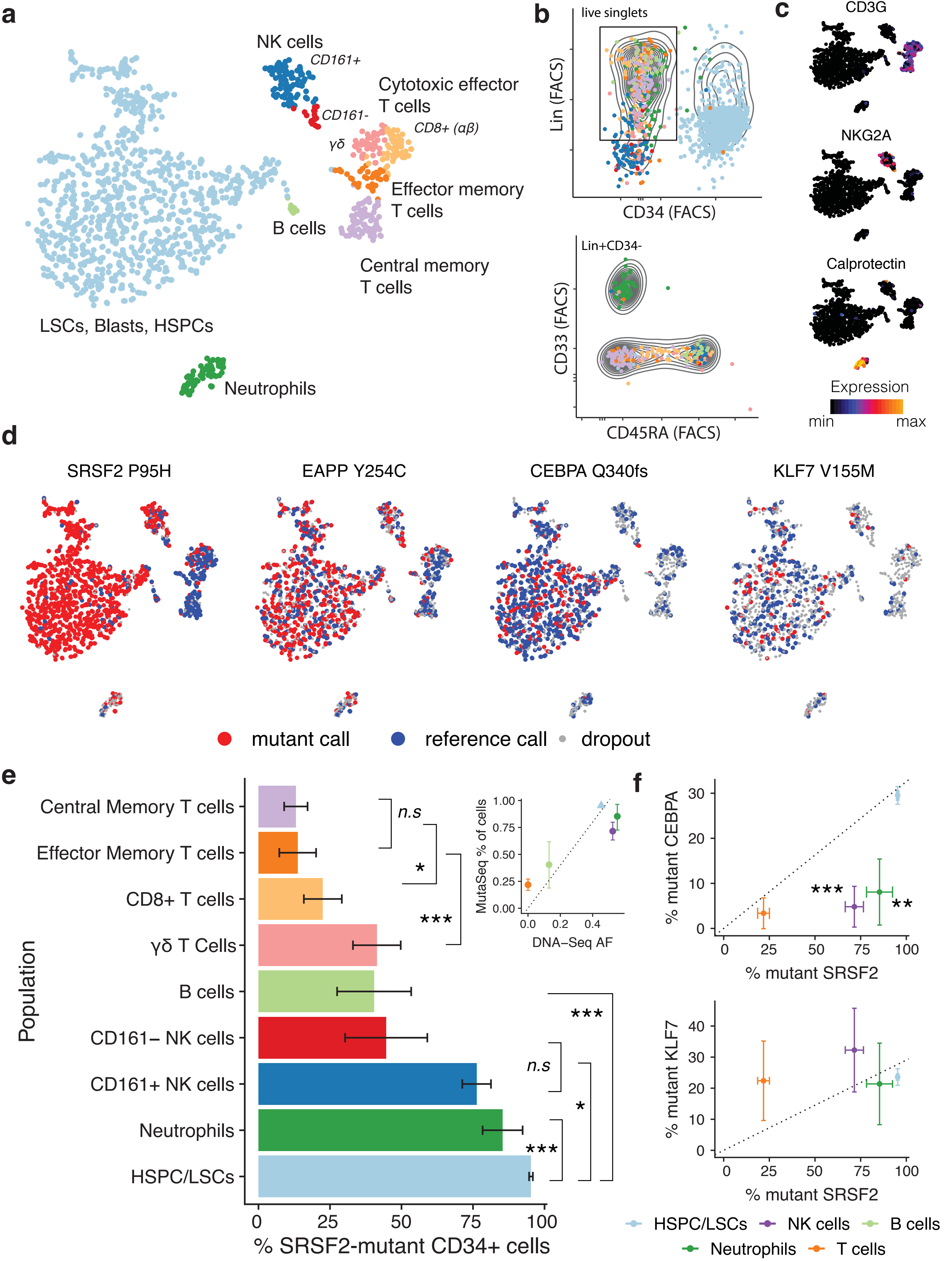
a
NK cells
CD161+
CD161-
Cytotoxic effector T cells
CD8+ (αβ)
γδ
Effector memory T cells
B cells
Central memory T cells
LSCs, Blasts, HSPCs
Neutrophils
b
live singlets
Lin (FACS)
CD34 (FACS)
Lin+CD34-
CD33 (FACS)
CD45RA (FACS)
c
CD3G
NKG2A
Calprotectin
Expression
min
max
d
SRSF2 P95H
EAPP Y254C
CEBPA Q340fs
KLF7 V155M
mutant call
reference call
dropout
e
Central Memory T cells
Effector Memory T cells
CD8+ T cells
γδ T Cells
B cells
CD161− NK cells
CD161+ NK cells
Neutrophils
HSPC/LSCs
Population
n.s
*
***
***
n.s
*
***
0
25
50
75
100
% SRSF2-mutant CD34+ cells
MutaSeq % of cells
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
DNA−Seq AF
f
% mutant CEBPA
30
20
10
0
***
**
% mutant SRSF2
% mutant KLF7
40
30
20
10
0
0
25
50
75
100
% mutant SRSF2
HSPC/LSCs
NK cells
B cells
Neutrophils
T cells

Figure 3

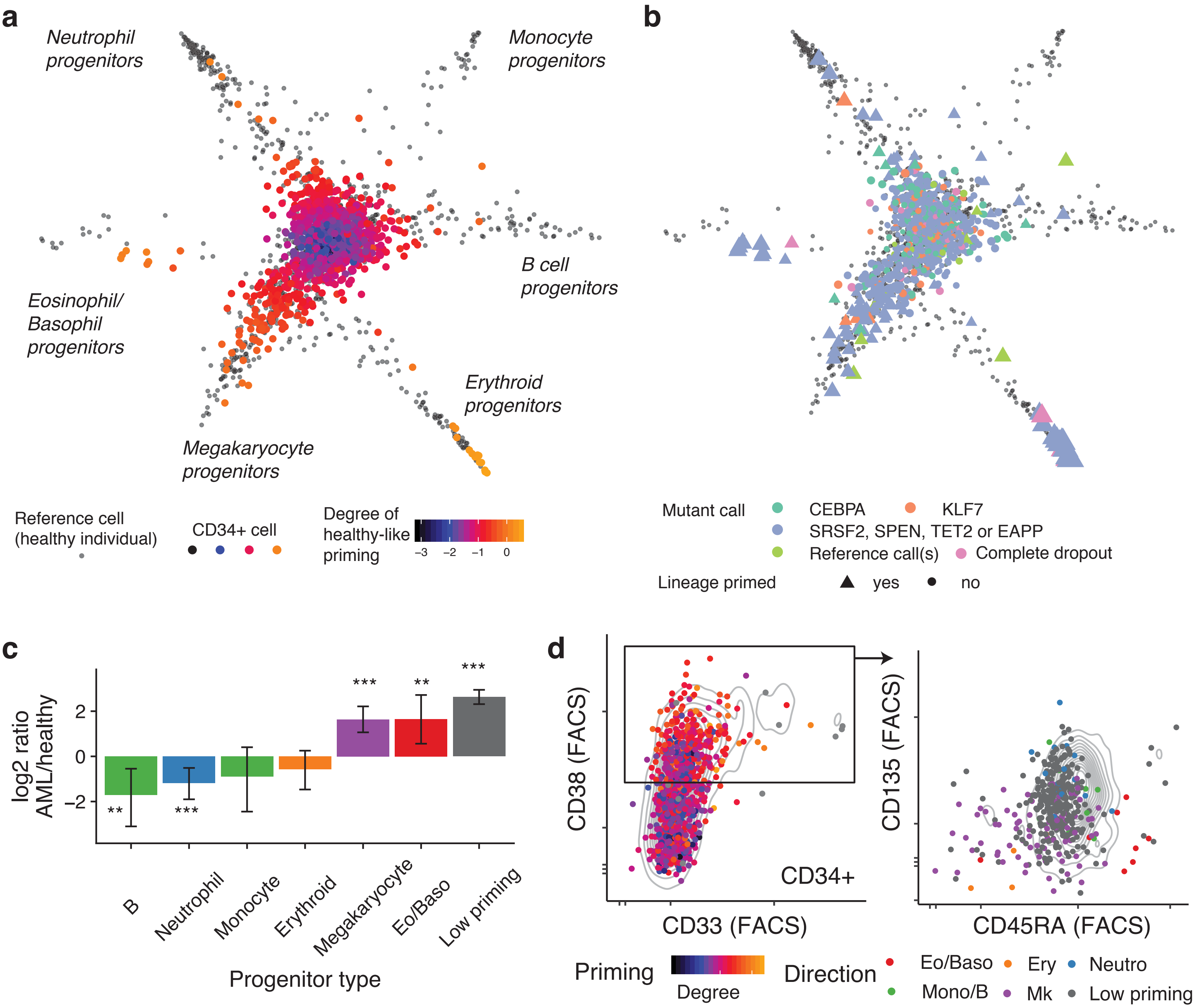
a
Neutrophil progenitors
Monocyte progenitors
Eosinophil/ Basophil progenitors
B cell progenitors
Erythroid progenitors
Megakaryocyte progenitors
Reference cell (healthy individual)
CD34+ cell
Degree of healthy-like priming
−3
−2
−1
0
b
Mutant call
CEBPA
KLF7
SRSF2, SPEN, TET2 or EAPP
Reference call(s)
Complete dropout
Lineage primed
yes
no
c
log2 ratio AML/healthy
2
0
−2
**
***
***
**
***
B
Neutrophil
Monocyte
Erythroid
Megakaryocyte
Eo/Baso
Low priming
Progenitor type
d
CD38 (FACS)
CD33 (FACS)
CD34+
CD135 (FACS)
CD45RA (FACS)
Priming
Degree
Direction
Eo/Baso
Ery
Neutro
Mono/B
Mk
Low priming

Figure 4

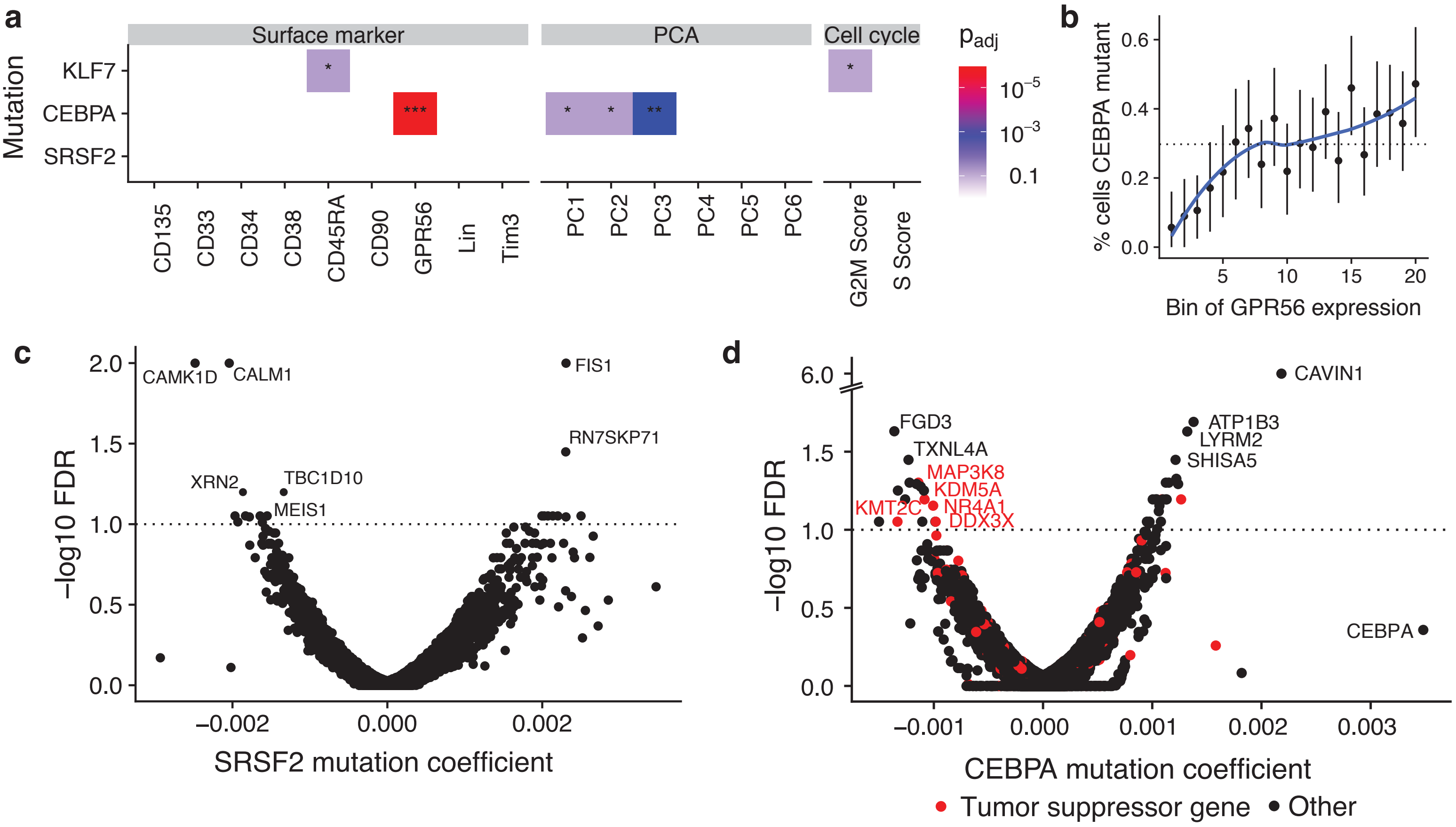
a
Surface marker
PCA
Cell cycle
Mutation
KLF7
CEBPA
SRSF2
*
***
*
*
**
*
CD135
CD33
CD34
CD38
CD45RA
CD90
GPR56
Lin
Tim3
PC1
PC2
PC3
PC4
PC5
PC6
G2M Score
S Score
p_{adj}
10^{-5}
10^{-3}
0.1
b
% cells CEBPA mutant
0.0
0.2
0.4
0.6
5
10
15
20
Bin of GPR56 expression
c
−log10 FDR
2.0
1.5
1.0
0.5
0.0
CAMK1D
CALM1
FIS1
RN7SKP71
XRN2
TBC1D10
MEIS1
−0.002
0.000
0.002
SRSF2 mutation coefficient
d
−log10 FDR
6.0
1.5
1.0
0.5
0.0
CAVIN1
FGD3
TXNL4A
MAP3K8
KDM5A
KMT2C
NR4A1
DDX3X
ATP1B3
LYRM2
SHISA5
CEBPA
−0.001
0.000
0.001
0.002
0.003
CEBPA mutation coefficient
● Tumor suppressor gene
● Other

Figure 5

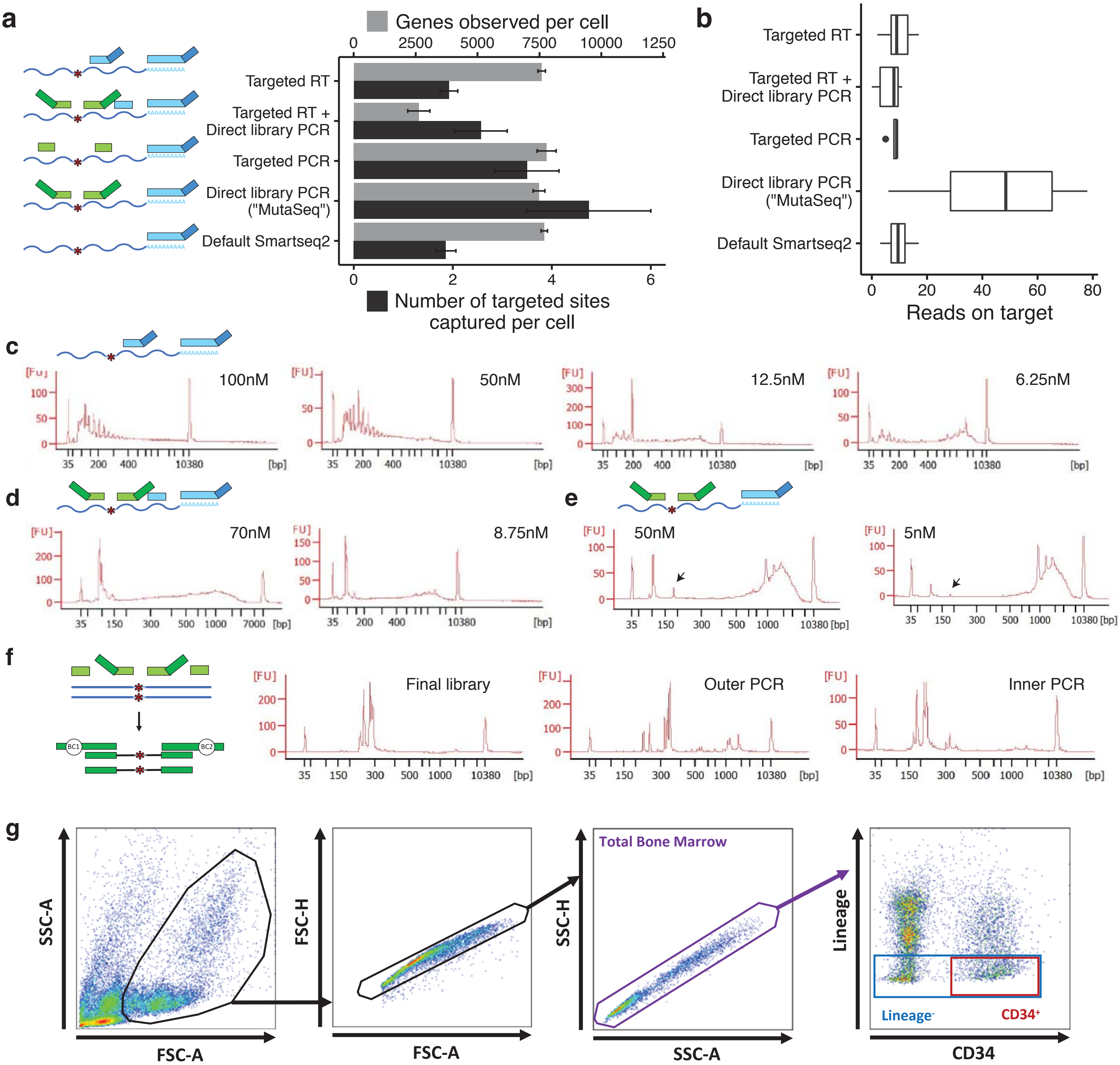
a
Genes observed per cell
0
2500
5000
7500
10000
1250
Targeted RT
Targeted RT +
Direct library PCR
Targeted PCR
Direct library PCR
("MutaSeq")
Default Smartseq2
0
2
4
6
Number of targeted sites
captured per cell
b
Targeted RT
Targeted RT +
Direct library PCR
Targeted PCR
Direct library PCR
("MutaSeq")
Default Smartseq2
0
20
40
60
80
Reads on target
c
100nM
50nM
12.5nM
6.25nM
[FU]
[bp]
35
200
400
10380
d
70nM
8.75nM
e
50nM
5nM
35
150
300
500
1000
7000
10380
f
BC1
BC2
Final library
Outer PCR
Inner PCR
g
SSC-A
FSC-A
FSC-H
FSC-A
Total Bone Marrow
SSC-H
SSC-A
Lineage
CD34
Lineage-
CD34+

Figure EV1

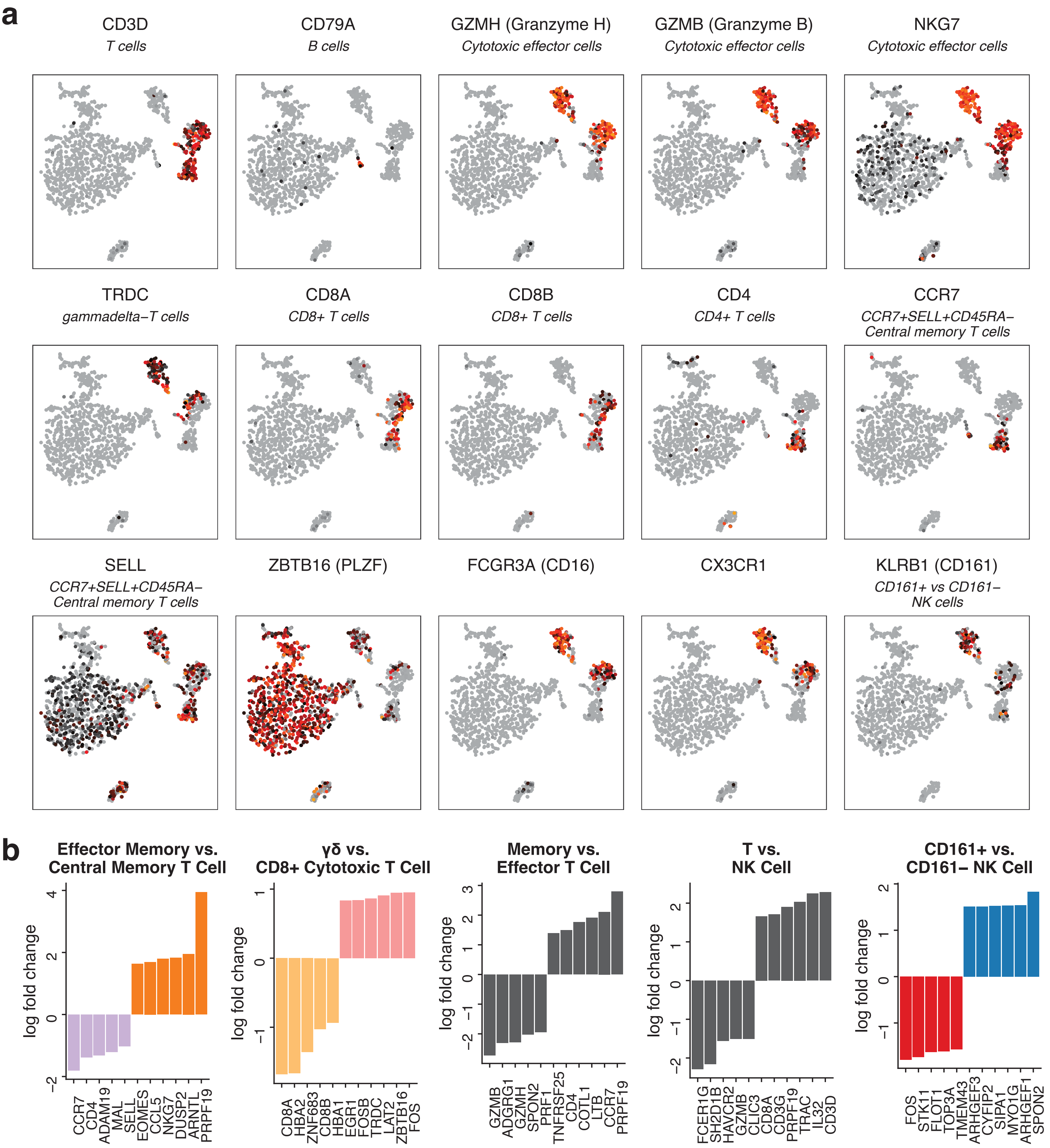
a
CD3D
T cells
CD79A
B cells
GZMH (Granzyme H)
Cytotoxic effector cells
GZMB (Granzyme B)
Cytotoxic effector cells
NKG7
Cytotoxic effector cells
TRDC
gammadelta–T cells
CD8A
CD8+ T cells
CD8B
CD8+ T cells
CD4
CD4+ T cells
CCR7
CCR7+SELL+CD45RA–
Central memory T cells
SELL
CCR7+SELL+CD45RA–
Central memory T cells
ZBTB16 (PLZF)
FCGR3A (CD16)
CX3CR1
KLRB1 (CD161)
CD161+ vs CD161–
NK cells
b
Effector Memory vs.
Central Memory T Cell
log fold change
4
2
0
−2
CCR7
CD4
ADAM19
MAL
SELL
EOMES
CCL5
NKG7
DUSP2
ARNTL
PRPF19
γδ vs.
CD8+ Cytotoxic T Cell
log fold change
1
0
−1
CD8A
HBA2
ZNF683
CD8B
HBA1
EGR1
FOSB
TRDC
LAT2
ZBTB16
FOS
Memory vs.
Effector T Cell
log fold change
3
2
1
0
−1
−2
GZMB
ADGRG1
GZMH
SPON2
PRF1
TNFRSF25
CD4
COTL1
LTB
CCR7
PRPF19
T vs.
NK Cell
log fold change
2
1
0
−1
−2
FCER1G
SH2D1B
HAVCR2
GZMB
CLIC3
CD8A
CD3G
PRPF19
TRAC
IL32
CD3D
CD161+ vs.
CD161– NK Cell
log fold change
2
1
0
−1
FOS
STK11
FLOT1
TOP3A
TMEM43
ARHGEF3
CYFIP2
SIPA1
MYO1G
ARHGEF1
SPON2

Figure EV2

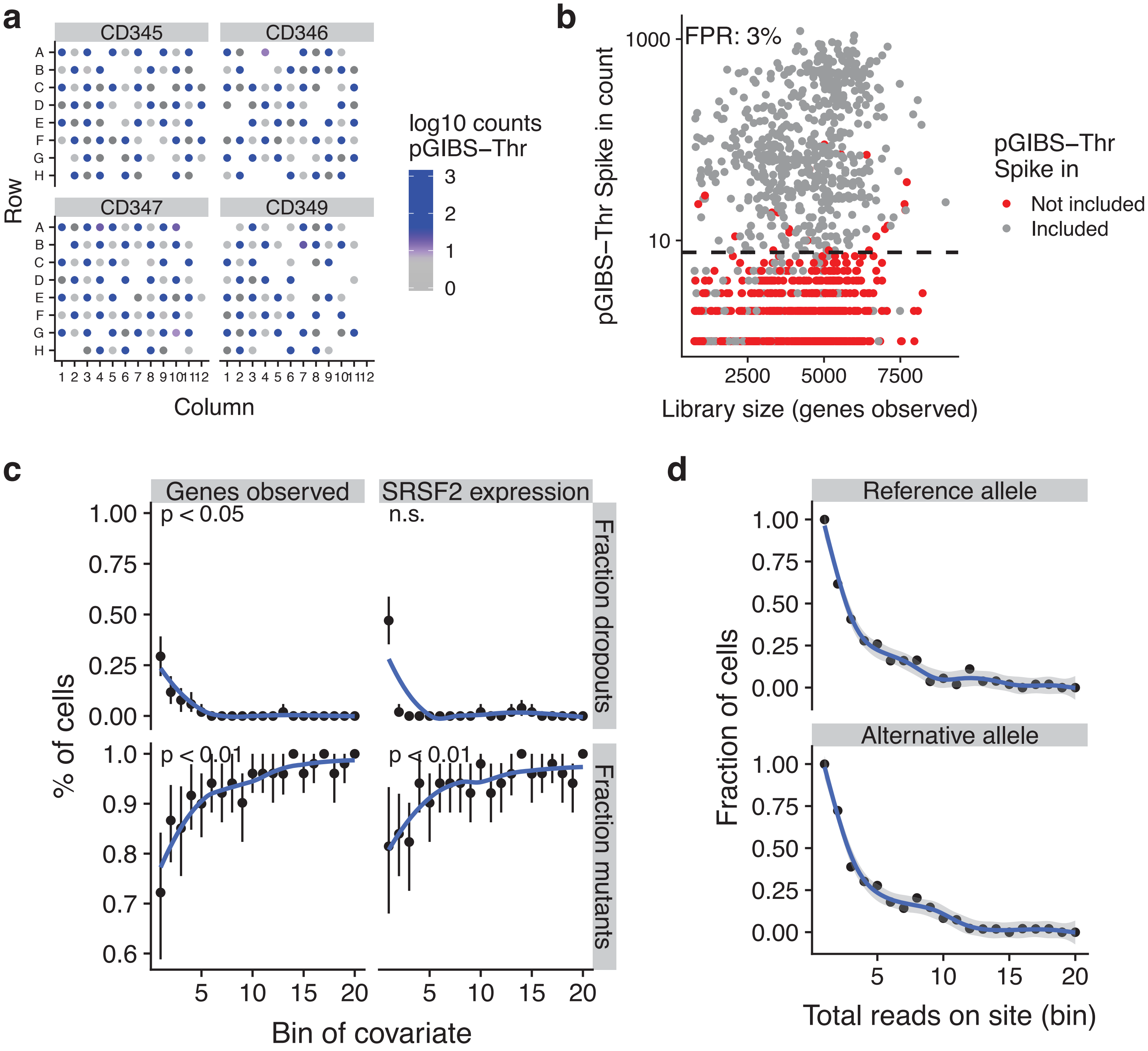
a
CD345
CD346
CD347
CD349
Row
A
B
C
D
E
F
G
H
1 2 3 4 5 6 7 8 9 10 11 12
Column
log10 counts
pGIBS−Thr
3
2
1
0
b
pGIBS−Thr Spike in count
1000
10
FPR: 3%
2500
5000
7500
Library size (genes observed)
pGIBS−Thr
Spike in
● Not included
● Included
c
Genes observed
SRSF2 expression
p < 0.05
n.s.
p < 0.01
p < 0.01
Fraction dropouts
Fraction mutants
% of cells
1.00
0.75
0.50
0.25
0.00
1.0
0.9
0.8
0.7
0.6
5
10
15
20
Bin of covariate
d
Reference allele
Alternative allele
Fraction of cells
1.00
0.75
0.50
0.25
0.00
5
10
15
20
Total reads on site (bin)

Figure EV3