nature immunology

RESOURCE

https://doi.org/10.1038/s41590-021-01059-0

Check for updates

OPEN

Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states

Sergio Triana[1,2,3,21], Dominik Vonficht[4,5,6,21], Lea Jopp-Saile[4,5,6,7,8,21], Simon Raffel[9], Raphael Lutz[4,5,9], Daniel Leonce[10], Magdalena Antes[4,5,6,9], Pablo Hernández-Malmierca[4,5,6], Diana Ordoñez-Rueda[11], Beáta Ramasz[11], Tobias Boch[12], Johann-Christoph Jann[12], Daniel Nowak[12], Wolf-Karsten Hofmann[12], Carsten Müller-Tidow[9], Daniel Hübschmann[4,5,13,14,15], Theodore Alexandrov[1,16], Vladimir Benes[17], Andreas Trumpp[4,5,14], Malte Paulsen[11,18], Lars Velten[3,19] ✉ and Simon Haas[4,5,7,8,14,20] ✉

Single-cell genomics technology has transformed our understanding of complex cellular systems. However, excessive cost and a lack of strategies for the purification of newly identified cell types impede their functional characterization and large-scale profiling. Here, we have generated high-content single-cell proteo-genomic reference maps of human blood and bone marrow that quantitatively link the expression of up to 197 surface markers to cellular identities and biological processes across all main hematopoietic cell types in healthy aging and leukemia. These reference maps enable the automatic design of cost-effective high-throughput cytometry schemes that outperform state-of-the-art approaches, accurately reflect complex topologies of cellular systems and permit the purification of precisely defined cell states. The systematic integration of cytometry and proteo-genomic data enables the functional capacities of precisely mapped cell states to be measured at the single-cell level. Our study serves as an accessible resource and paves the way for a data-driven era in cytometry.

Single-cell transcriptomic technologies have revolutionized our understanding of tissues[1–3]. The systematic construction of whole-organ and whole-organism single-cell atlases has revealed an unanticipated diversity of cell types and cell states, and has provided detailed insights into cellular development and differentiation processes[4–7]. However, strategies for the prospective isolation of cell populations newly identified by single-cell genomics are needed to enable their functional characterization or therapeutic use. Furthermore, single-cell genomics technologies remain cost-intense and scale poorly, impeding their integration into clinical routine.

Unlike single-cell transcriptomics, flow cytometry offers a massive throughput in terms of samples and cells, is commonly used in routine clinical diagnostics[8] and remains unrivaled in the ability to prospectively isolate live populations of interest for downstream applications. However, flow cytometry provides low-dimensional measurements and relies on predefined sets of surface markers and gating strategies that have evolved historically in a process of trial and error. Hence, single-cell transcriptomics (scRNA-seq) approaches have demonstrated that flow cytometry gating schemes frequently yield impure or heterogeneous populations[9,10], and flow strategies for the precise identification of cell types defined by scRNA-seq are lacking. Conversely, the precision and efficiency of commonly used cytometry gating schemes are largely unknown, and the exact importance of many surface markers remains unclear. Together, these findings highlight a disconnect between single-cell genomics-based molecular cell type maps and data generated by widely used cytometry assays.

The differentiation of hematopoietic stem cells (HSCs) in the bone marrow (BM) constitutes a particularly striking example of this disconnect[11–14]. The classical model of hematopoiesis, which is based mainly on populations defined by flow cytometry[15–17], has

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Collaboration for Joint PhD degree between European Molecular Biology Laboratory and Heidelberg University, Faculty of Biosciences, Heidelberg, Germany. [3]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [4]Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), Heidelberg, Germany. [5]Division of Stem Cells and Cancer, Deutsches Krebsforschungszentrum (DKFZ) and DKFZ–ZMBH Alliance, Heidelberg, Germany. [6]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [7]Berlin Institute of Health (BIH), Charité Universitätsmedizin Berlin, Berlin, Germany. [8]Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin Institute for Medical Systems Biology, Berlin, Germany. [9]Department of Internal Medicine V, Heidelberg University Hospital, Heidelberg, Germany. [10]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [11]Flow Cytometry Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany. [12]Department of Hematology and Oncology, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany. [13]Computational Oncology, Molecular Diagnostics Program, National Center for Tumor diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany. [14]German Cancer Consortium (DKTK), Heidelberg, Germany. [15]Department of Pediatric Immunology, Hematology and Oncology, University Hospital Heidelberg, Heidelberg, Germany. [16]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA. [17]Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany. [18]Novo Nordisk Foundation Center for Stem Cell Biology, DanStem. Faculty of Health and Medical Sciences Blegdamsvej, Copenhagen, Denmark. [19]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [20]Charité-Universitätsmedizin, Berlin, Germany. [21]These authors contributed equally: Sergio Triana, Dominik Vonficht, Lea Jopp-Saile. ✉e-mail: lars.velten@crg.eu; simon.haas@bih-charite.de

recently been challenged in several aspects by single-cell transcriptomic[9,10,18–20], functional[21,22] and lineage tracing[23] approaches. These studies revealed that hematopoietic lineage commitment occurs earlier than previously anticipated, that putative oligopotent progenitors isolated by fluorescence activated cell sorting (FACS) consist of heterogeneous mixtures of progenitor populations and that lineage commitment is represented most accurately by a continuous process of differentiation trajectories rather than by a stepwise differentiation series of discrete progenitor populations[12–14,24]. The frequency of functionally oligopotent progenitors in immunophenotypic hematopoietic stem and progenitor cell (HSPC) gates remains controversial[9,25,26]. These discrepancies have contributed to conflicting results between studies that employ scRNA-seq for the definition of progenitor populations[9,10,18,19,27] and studies that use FACS[15,16,28]. As a consequence, flow-based assays that accurately reflect the molecular and cellular complexity of the hematopoietic system are urgently needed.

Recently, methods to simultaneously measure mRNA and surface protein expression in single cells have been developed[29,30]. Here, we demonstrate that ultrahigh content single-cell proteo-genomic reference maps, alongside appropriate computational tools, can be used to systematically design and analyze cytometry assays that accurately reflect scRNA-seq-based molecular tissue maps at the level of cell types and differentiation states. For this purpose, we have generated proteo-genomic datasets encompassing 97–197 surface markers across 122,004 cells representing the cellular landscape of young, aged and leukemic human BM and blood, as well as all states of HSC differentiation. We demonstrate how such data can be used in an unbiased manner to evaluate and automatically design cytometry gating schemes for individual populations and entire biological systems without previous knowledge. We show that, compared with existing approaches, such optimized schemes are superior in the identification of cell types and more accurately reflect molecular cell states. Projecting datasets from malignant hematopoiesis on our reference atlases enables the fine-mapping of the exact stage of differentiation arrest in leukemias, the identification of leukemia-specific surface markers and an unsupervised classification of disease states. Finally, we demonstrate how such data resources can be used to project low-dimensional cytometry data on single-cell genomic atlases to enable functional analysis of precisely defined states of cellular differentiation. Our data resource and bioinformatic advances enable the efficient identification and isolation of any molecularly defined cell state from blood and BM while laying the grounds for reconciling flow cytometry and single-cell genomics data across human tissues.

Results

A single-cell proteo-genomic reference map of BM. To establish a comprehensive single-cell transcriptomic and surface protein expression map of human BM, we performed a series of Abseq experiments in which mononuclear BM cells from hip aspirates were labeled with 97–197 oligo-tagged antibodies, followed by targeted or whole transcriptome scRNA-seq on the BD Rhapsody platform (Fig. 1a). For targeted single-cell transcriptome profiling, we established a custom panel, consisting of 462 mRNAs covering all HSPC differentiation stages, cell type identity genes, mRNAs of surface receptors and additional genes that permit the characterization of cellular states. These genes were selected systematically to capture all relevant layers of RNA expression heterogeneity observed in this system (Supplementary Note 1 and Supplementary Table 1). Whole transcriptome single-cell proteo-genomics confirmed that no populations were missed due to the targeted nature of the assay (Supplementary Note 2). Using this panel, in combination with 97 surface markers (Supplementary Table 2), we analyzed the BM of three young healthy donors, three aged healthy donors and three acute myeloid leukemia (AML) patients at diagnosis (Fig. 1a, Extended Data Fig. 1 and Supplementary Table 3). For samples from healthy donors, $CD34^+$ cells were enriched to enable a detailed study of HSC differentiation (Extended Data Fig. 2). For samples from AML patients, $CD3^+$ cells were enriched in some cases to ensure sufficient coverage of T cells.

Since single-cell proteo-genomic approaches are not commonly performed at this level of antibody multiplexing, we designed a series of control experiments. First, we performed matched Abseq experiments in the presence or absence of antibodies to ensure that highly multiplex antibody stains do not effect the transcriptome of single cells (Supplementary Note 3). We further performed a series of Abseq experiments on fresh and frozen samples to demonstrate that the freeze–thawing process has no great impact on the data (Supplementary Note 3). Finally, we evaluated the sequencing requirements for optimal cell type classification in high-parametric single-cell proteo-genomic experiments (Supplementary Note 4). In the main reference data set, 70,017 high-quality BM cells were profiled with combined RNA and high-parametric surface protein information, and an average of ~7,500 surface molecules per cell were detected (Extended Data Fig. 3). Following data integration across experiments and measurement modalities, we identified 45 cell types and cell stages covering the vast majority of previously described hematopoietic cell types of the BM and peripheral blood (PB), including all stages of HSC differentiation in the $CD34^+$ compartment, all T cell and natural killer (NK) cell populations of the $CD3^+$ and $CD56^+$ compartments, several dendritic cell and monocyte subpopulations from the $CD33^+$ compartment and all main B cell differentiation states across $CD10^+$, $CD19^+$ and $CD38^{high}$ compartments (Fig. 1b,c, Supplementary Note 5 and Supplementary Table 4). In addition, poorly characterized populations, such as cytotoxic $CD4^+$ T cells and mesenchymal stem or stromal cells (MSCs) are covered. Cells from young and aged BM occupied the same cell states in all individuals, whereas cell states in AML differed (Fig. 1b and see below). Importantly, the combined RNA and surface protein information provided higher resolution and revealed cell types that are not readily identified by one of the individual data layers alone (Supplementary Note 6).

Besides our main reference dataset, we generated 'query' single-cell proteo-genomic datasets, which are displayed in the context of the main reference (Supplementary Note 7). These include, first, the analyses of healthy BM and matched PB samples using a 197-plex antibody panel to query the expression of additional surface markers in the context of our reference (Extended Data Fig. 4 and Supplementary Table 2). Second, the analyses of healthy BM analyzed with a 97-plex antibody panel in combination with whole transcriptome profiling to query any gene's expression in the space defined by our reference (Supplementary Note 2). Third, the profiling of the $CD34^+CD38^-$ BM compartment with a 97-plex antibody panel to provide higher resolution of immature HSPCs (see below and Extended Data Fig. 9c,d) and fourth, a cohort of 12 AML patients (see below and Fig. 4). To make our comprehensive resource accessible, we developed the Abseq App, a web-based application that permits visualization of gene and surface marker expression, differential expression testing and the data-driven identification of gating schemes across all datasets presented in this manuscript. A demonstration video of the app is available in the supplement (Supplementary Video 1). The Abseq App is accessible at: https://abseqapp.shiny.embl.de/.

A directory of the biological importance of surface markers. While surface markers are widely used in immunology, stem-cell biology and cancer research to identify cell types, cell stages and biological processes, the exact importance of individual markers frequently remains ambiguous. To link surface marker expression quantitatively with biological processes, we assigned each cell in our data set to its respective cell type, and determined its differentiation stage, its stemness score, its cytotoxicity score and its current

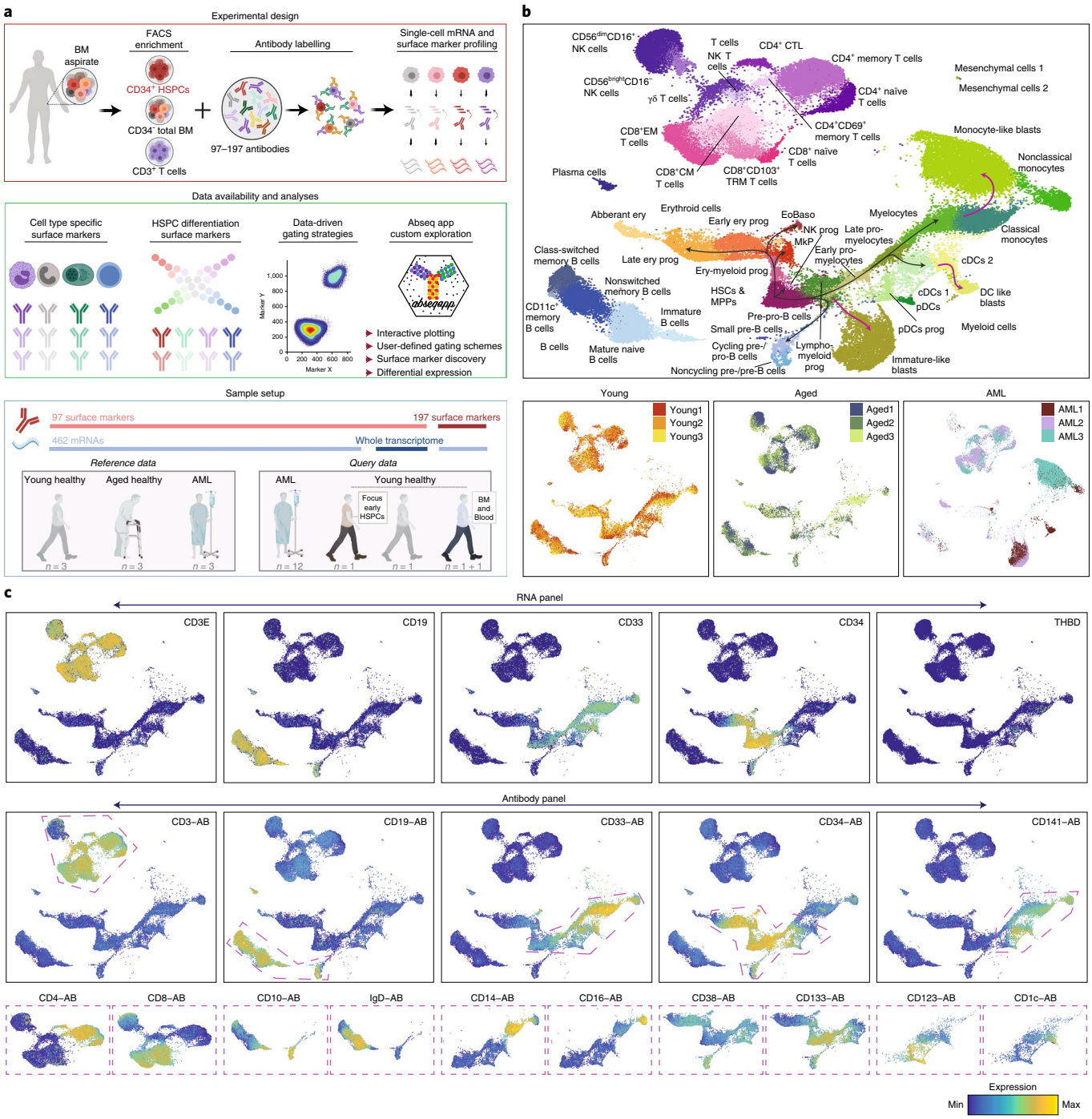


Fig. 1 | A comprehensive single-cell proteo-genomic map of young, aged and malignant BM. a, Overview of the study. See Methods and main text for details. **b**, Top: UMAP display of single-cell proteo-genomics data of human BM from healthy young, healthy aged and AML patients ($n = 70{,}017$ single cells, 97 surface markers), integrated across $n = 9$ samples and data modalities. Clusters are color-coded. ery, erythroid; prog, progenitor. Bottom: UMAPs highlighting sample identities. See Supplementary Note 5 for details of cluster annotation. The whole transcriptome Abseq data is presented in Supplementary Note 2, the Abseq experiments with measurements of 197 surface markers are presented in Extended Data Fig. 4. **c**, Normalized expression of selected mRNAs and surface proteins highlighted on the UMAP space from **b**. Top: expression of mRNAs encoding surface markers widely used to identify main cell types. Middle: expression of the corresponding surface proteins. Bottom: expression of markers widely used to stratify main cell types into subtypes. Only the parts of the UMAPs highlighted by dashed polygons in the middle row are shown. For all data shown throughout the manuscript, BM mononuclear cells from iliac crest aspirations from healthy adult donors or AML patients were used unless stated otherwise.

cell cycle phase as well as technical covariates (see Methods and below). Moreover, we included covariates representing unknown biological processes that were defined in an unsupervised manner using a factor model. Nontechnical covariates were not affected by marker expression level (Extended Data Fig. 5a and Methods). For each surface marker, we then quantified the fraction of variance of expression that is determined by any of these processes (Fig. 2a). This model identified markers that represent cell type identities or

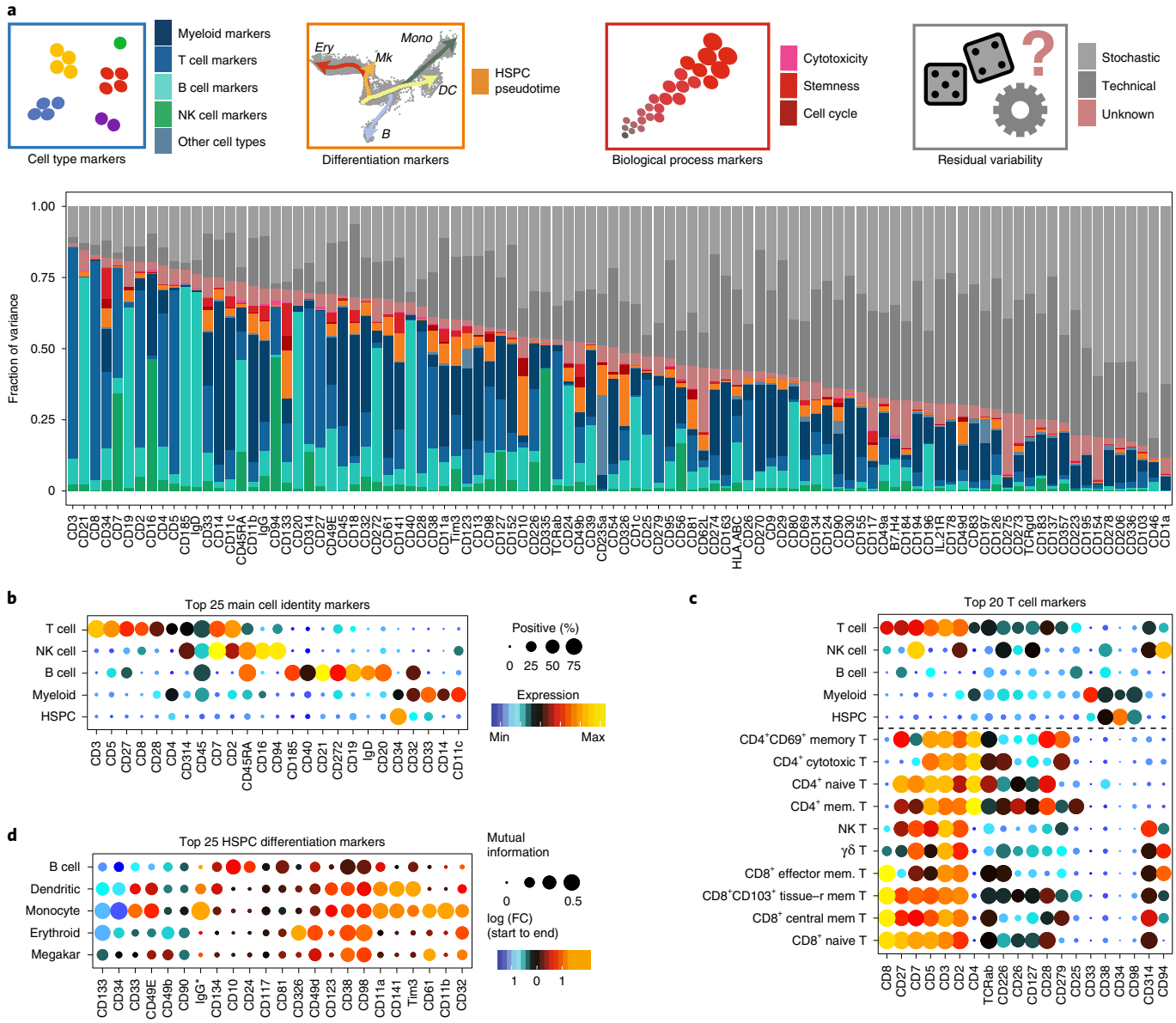


Fig. 2 | Association of surface marker expression with cell type identities, cellular differentiation and biological processes. a, For each surface marker measured in our 97-plex Abseq data, the fraction of variance explained by different covariates (colored insets in top row) is displayed. For this, every single cell from healthy young individuals ($n = 3$ samples, 28,031 single cells) was assigned to a cell type identity (blue inset, see Fig. 1b), and cytotoxicity, stemness and cell cycle scores (red inset, see Extended Data Fig. 5e) as well as technical covariate scores were determined. Additionally, pseudotime analyses were used to assign differentiation scores to HSPCs (orange inset, see Fig. 3a). These covariates were then used to model surface marker expression in a linear model. The fraction of variance explained by each of the processes was quantified. See Methods, section Modeling variance in surface marker expression for details. **b**, Cell type identity markers. Dot plot depicting the expression of the 25 surface markers with the highest fraction of variance explained by cell type across main populations. Colors indicate mean normalized expression, point size indicates the fraction of cells positive for the marker. Automatic thresholding was used to identify positive cells, see Methods, section Thresholding of surface marker expression for details. **c**, T cell subtype markers. The expression of the 20 surface markers with the highest fraction of variance explained by T cell subtype is displayed, legend as in **b**. mem, memory; tissue-r, tissue-resident. **d**, HSPC differentiation markers. Megakar, megakaryocytic. Dot plot depicting expression changes of markers across pseudotime in CD34+ HSPCs. Color indicates logarithmic fold change (FC) between the start and the end of each pseudotime trajectory. Point size indicates the mutual information in natural units of information between pseudotime and marker expression. The 25 surface markers with the highest fraction of variance explained by pseudotime covariates are displayed.

differentiation stages, as well as stemness, cytotoxicity and cell cycle properties (Fig. 2b–d and Extended Data Fig. 5b–f).

To characterize new markers identified by this analysis, we focused initially on the evaluation of surface molecules that specifically mark distinct stages of HSC differentiation, since a lack of specific markers currently impedes the accurate representation of lineage commitment by flow cytometry[9,10,18,21,27]. For this purpose, we performed pseudotime analyses within the CD34+ HSPC compartment and identified surface markers that correlate with the progression of HSCs towards erythroid, megakaryocytic, monocyte, conventional dendritic cell or B cell differentiation trajectories (Methods; Figs. 2d and 3a and Extended Data Fig. 5g).

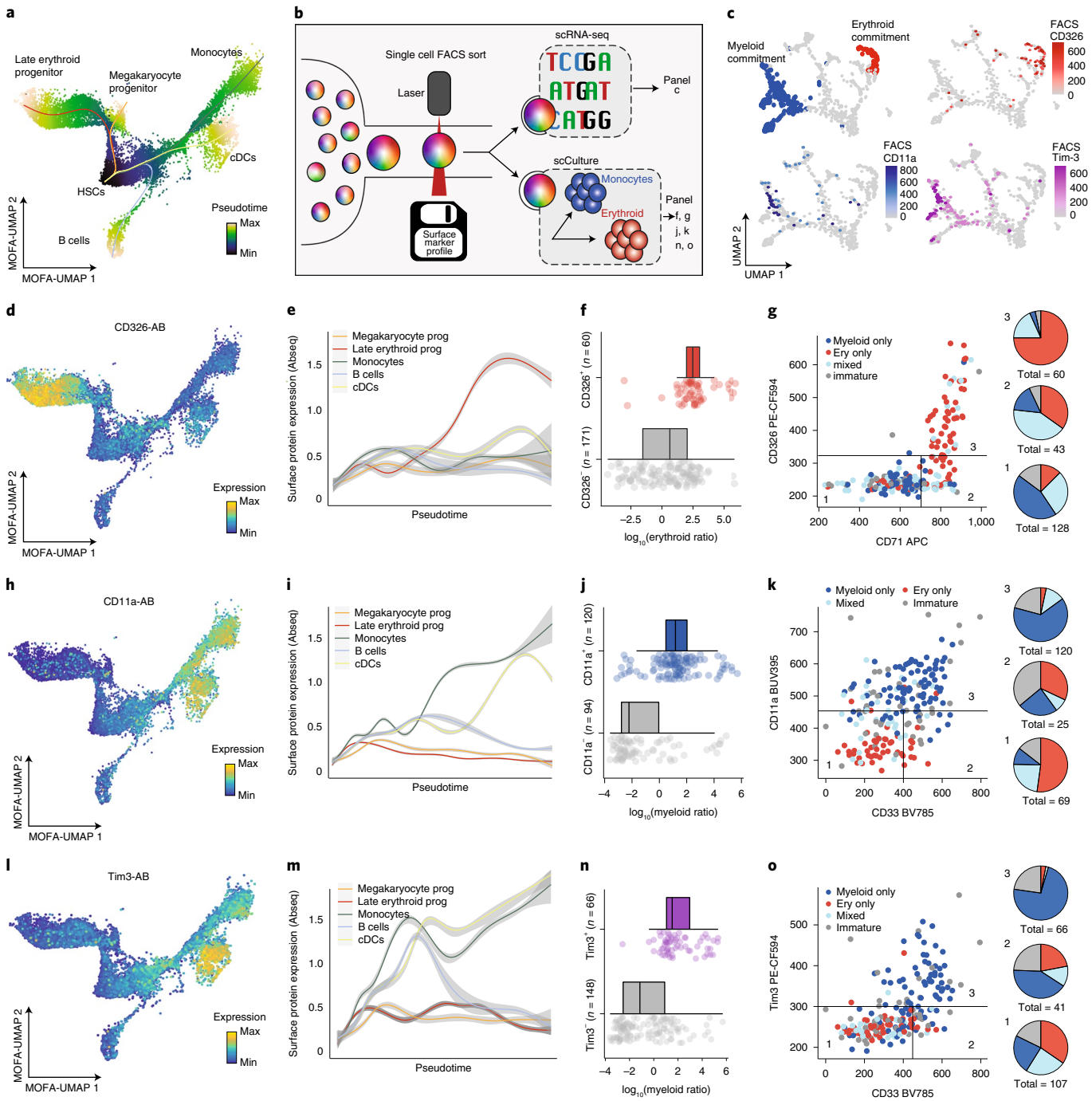


Fig. 3 | Validation of novel stage-specific HSPC differentiation markers. a, UMAP plot depicting CD34⁺ HSPCs and their pseudotime scores along five differentiation trajectories, see Methods, section Pseudotime analysis. The normalized pseudotime score across all lineages is color-coded. **b**, Scheme illustrating the experiments performed to validate the importance of selected markers. See main text and Supplementary Note 8 for details. **c**, UMAP display of mRNA expression of $n = 630$ CD34⁺ cells from a single-cell Smart-seq2 experiment where surface markers were recorded using FACS. For a detailed description of the experiment, see Supplementary Note 8. Upper left panel: cells with myeloid and erythroid gene expression signatures are highlighted on the UMAP. Remaining panels: surface protein expression (FACS data) of indicated markers is shown. **d**, UMAP display highlighting the normalized CD326 surface protein expression (Abseq data). **e**, Line plots depicting normalized CD326 surface protein expression (Abseq data) smoothed over the different pseudotime trajectories illustrated in **a**. Error ribbon indicates 95% confidence interval from the smoothing GAM model. **f**, Boxplots depicting the ratio in erythroid cells produced in single-cell cultures in relation to the CD326 expression of the founder cell ($n = 231$ single-cell derived colonies). See Methods, section Data visualization for a definition of boxplot elements. **g**, Left panel: scatter plots depicting the differentiation potential of single founder cells in relation to their CD326 and CD71 surface expression. The founder cell potential was categorized by its ability to give rise to (red) erythroid only progeny, (skyblue) a mix of erythroid, myeloid or any other progeny, (blue) only myeloid progeny or (gray) remaining, immature cells. Right panel: founder cells were subset according to their CD326 and CD71 surface expression status and relative fractions of their respective potential are summarized as pie charts. **h**–**o**, Analysis of CD11a and Tim3. **h**–**k** as in **d**–**g** except that CD11a is shown in the UMAP (**h**), line plot (**i**), boxplot (**j**) and scatter plot (**k**). **l**–**o**, Panels are analogous to **d**–**g**, except that Tim3 expression is shown in the UMAP (**l**) line plot (**m**), boxplot (**n**) and scatter plot (**o**). For scatter plots in **k** and **o**, CD11a or Tim3 expression was plotted against the myeloid differentiation marker CD33. For **j**,**k**,**n**,**o**, $n = 214$ single-cell derived colonies.

Of note, the monocyte trajectory also includes neutrophil progenitor stages, but mature neutrophils are not included in the datasets due to the use of density gradient centrifugation of samples. Moreover, trajectory analyses were not performed for plasmacytoid dendritic and eosinophil/basophil lineages due to a low number of intermediate cells impeding the unanimous identification of branch points. Pseudotime analyses quantified the exact expression dynamics of many well-established markers, such as CD38 as a pandifferentiation marker, as well as CD10 and CD11c as early B cell and monocyte-dendritic cell lineage commitment markers, respectively (Fig. 2d and Extended Data Fig. 6a). Importantly, our analyses revealed new surface markers that specifically demarcate distinct stages of lineage commitment, including CD326, CD11a and Tim3 (Figs. 2d and 3). To confirm the high specificity of these markers for erythroid and myeloid commitment, respectively, we used FACS-based indexing of surface markers coupled to single-cell RNA-seq ('index scRNA-seq', see also Supplementary Note 8), or coupled to single-cell cultures ('index cultures') (Fig. 3b). As suggested by our proteo-genomic single-cell data, CD326 expression was associated with molecular priming and functional commitment into the erythroid lineage (Fig. 3c–g and Extended Data Fig. 6b,c). By contrast, Tim3 and CD11a were identified as pan-myeloid differentiation markers and were associated with transcriptomic priming and functional commitment into the myeloid lineage (Fig. 3c,h–o and Extended Data Fig. 6c). Finally, CD98 was identified as a new pandifferentiation marker of HSCs, which we confirmed by classical flow cytometry (Fig. 2d and Extended Data Fig. 6d–h). Beyond the progression of HSCs to lineage-committed cells, we also analyzed the surface marker dynamics throughout B cell differentiation, allowing us to identify markers specific to their lineage commitment, maturation, isotype switching and final plasma cell generation (Extended Data Fig. 6i–p).

Our model provides a global and quantitative understanding of how well cell type identities, differentiation stages and biological processes are related to the expression of individual surface markers. A comprehensive overview of surface markers associated with these processes is depicted in the supplement (Supplementary Data 1 and Extended Data Fig. 5).

Surface protein expression in healthy aging and cancer. To investigate surface protein expression throughout healthy aging, we compared Abseq data of BM from young and aged healthy individuals. These analyses revealed that the expression of surface molecules was highly similar across all BM populations between age groups (Fig. 4a,b and Supplementary Data 1), suggesting unexpectedly stable and highly regulated patterns of surface protein expression that are affected only modestly by aging. While cell type frequencies were also affected only modestly by aging, a substantial accumulation of cytotoxic effector CD8$^+$ T cells was observed[31] (Extended Data Fig. 7a). Moreover, the expression of several immune regulatory molecules showed age-related changes in surface presentation, including the death receptor FAS (CD95), the poliovirus receptor (CD155) and the ICOS ligand (CD275) (Fig. 4b). In particular, naive CD8$^+$ and CD4$^+$ T cell subsets displayed an aging-associated decline in surface expression of CD27, a costimulatory molecule required for generation and maintenance of long-term T cell immunity[32] (Fig. 4b,c). Together, these analyses suggest that the overall pattern of surface protein expression is widely maintained upon healthy aging, whereas specific changes, most prominently in the surface presentation of immune regulatory molecules, occur.

We next explored surface marker remodeling in AML—a blood cancer characterized by the accumulation of immature, dysfunctional myeloid progenitors, also called blasts. While the cellular BM of healthy donors displayed highly similar topologies across six individuals, initial analysis of three AML patients demonstrated that leukemic cells showed patient-specific alterations and a large degree of interpatient variability (Fig. 1b). To develop a generically applicable workflow to interpret data from hematological diseases in the context of our reference, we generated single-cell proteo-genomics datasets from a total of 15 AML patients, covering six t(15;17) translocated acute promyelocytic leukemias and nine normal karyotype AMLs with *NPM1* mutations, of which four patients carried an additional *FLT3* internal tandem duplication (Supplementary Table 3). While an unsupervised integration of these data highlighted primarily patient-to-patient variability (Extended Data Fig. 7b), projecting cells onto our healthy reference enabled a fine-mapping of the differentiation stages of leukemia cells (Fig. 4d and Supplementary Note 7). Unsupervised clustering of patients on the basis of relative abundancies of differentiation stages revealed three main categories: 'monocytic AMLs' that displayed an extensive accumulation of blasts with classical monocyte phenotype, acute promyelocytic leukemias that were blocked in early and late promyelocyte states, and 'immature AMLs' that showed high numbers of immature blasts resembling HSC, multipotent progenitors (MPP), early lymphomyeloid progenitor and early promyelocyte states (Fig. 4e,f). In general, leukemic blasts retained many features reminiscent of the cell stage they were blocked in (Extended Data Fig. 7c–e). Accordingly, differential expression analyses revealed that many surface markers that distinguish the different AML states also mark their corresponding healthy counterparts, such as CD133 for immature AMLs or CD14 and CD11b for monocytic AMLs (Fig. 4g). This also translated into differential surface expression of potential drug targets, such as PD-L1 (CD274) and CTLA4 (CD152) (Fig. 4h and Extended Data Fig. 7f), suggesting that the myeloid differentiation program of the AML might be essential in the treatment choice of targeted immune therapies.

By contrast, differential analyses between AML and healthy cells from the same differentiation stage revealed markers specifically overexpressed in leukemic cells (Fig. 4i, Extended Data Fig. 7c and Supplementary Data 2). Interestingly, these analyses readily identified several previously described leukemia stem-cell markers, including CD25, Tim3, CD123 and CD45RA[33], supporting the validity of our approach. Quantifying the degree of interpatient heterogeneity of each marker while accounting for cell state revealed that many known leukemia stem-cell markers vary strongly in their expression between patients (Fig. 4i). Together, this workflow of projection to a well-annotated healthy reference in combination with cell-state-specific differential expression testing might become a standard in scRNA-seq analyses of hematological diseases. Our computational routines are available online at https://git.embl.de/triana/nrn.

Data-driven flow cytometry for immunology. Gating strategies for flow cytometry have evolved historically in a process of trial and error. In particular, the isolation of rare and poorly characterized cell subsets using flow cytometry remains challenging, whereas commonly used gating schemes are not necessarily optimal in purity (precision) and efficiency (recall). To tackle these problems, we explored different machine learning approaches for the data-driven definition of gating schemes. For all populations in our dataset, gating schemes defined by machine learning approaches provided higher precision (purity) when compared with classical gating schemes from the literature (Fig. 5a, Extended Data Fig. 8a–d and Supplementary Table 5). While different machine learning methods tested achieved similar purities, gates defined by the hypergate algorithm[34] offered a higher recall (Fig. 5a and Extended Data Fig. 8a–d).

To validate and demonstrate this approach, we focused on determining new gating strategies for rare and poorly characterized BM cell types, such as cytotoxic CD4$^+$ T cells (Fig. 5b) and MSCs (Fig. 5g). Cytotoxic CD4$^+$ T cells represent a rare T cell population characterized by the expression of cytotoxicity genes typically

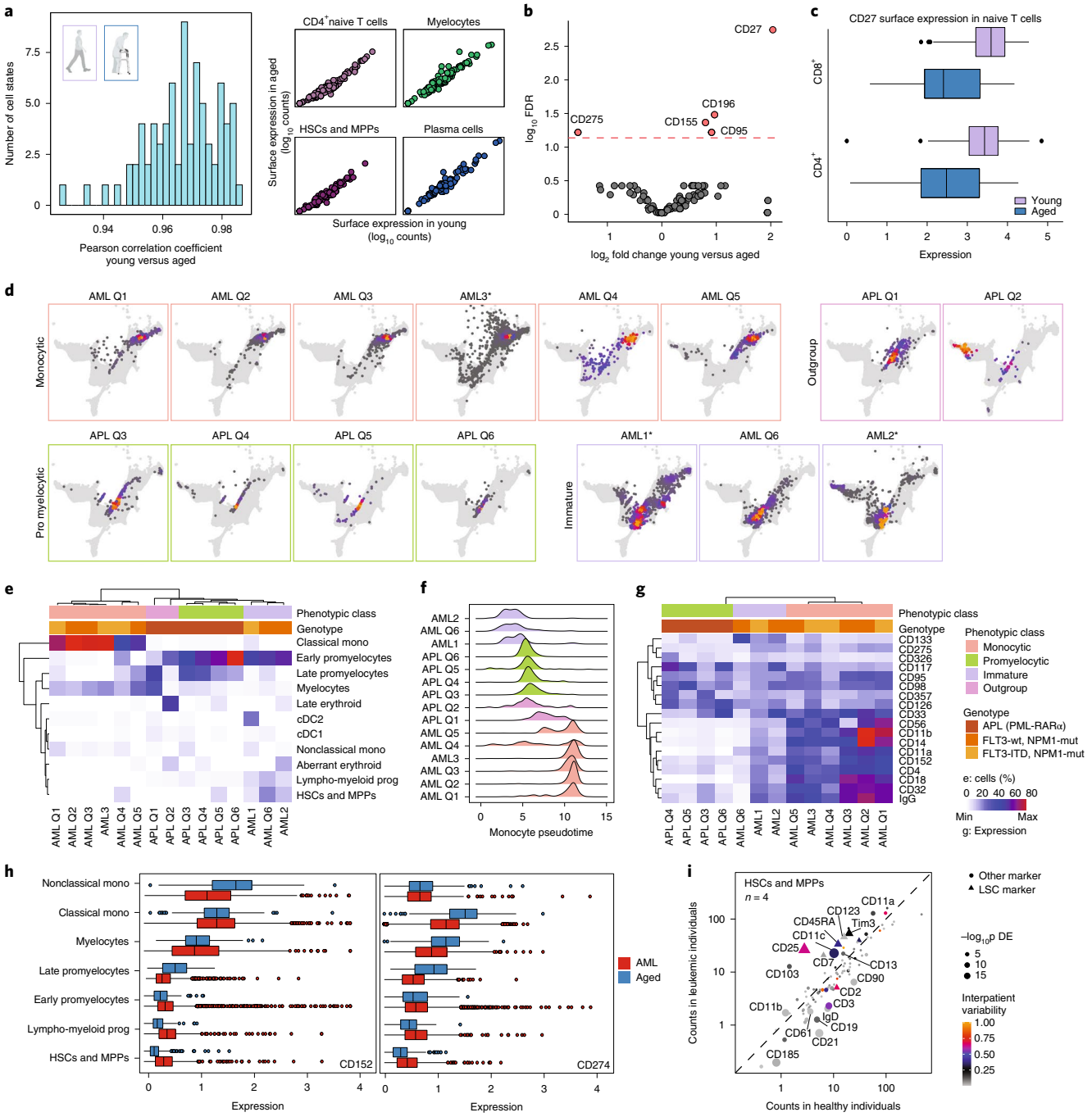


Fig. 4 | Adaptation of surface protein expression in healthy aging and cancer. a, Correlation of surface marker expression between matched cell types from aged and young BM donors. For each cell type, mean surface marker expression across all cells was computed, separately for all 'young' and 'aged' samples. Left panel: histogram of Pearson correlation coefficients. Right panel: sample scatter plots depicting the mean surface expression of all measured markers in indicated cell types. **b**, Volcano plot depicting $\log_2$ fold change and false discovery rate (FDR) for a test for differential surface marker expression between cells from young and aged individuals, while accounting for cell types as covariates. See Methods for details. **c**, Boxplots depicting CD27 surface expression in naïve T cell populations from young and aged individuals. Sample size is provided as Figure Source Data. See Methods, section Data visualization for a definition of boxplot elements. **d**, Projection of AML samples onto healthy reference. See Supplementary Note 7 for details. **e**, Clustering of leukemia samples by their projected cell type composition. Lymphoid cells are excluded from the clustering. **f**, Density plots of monocyte pseudotime, resulting from projection on the healthy reference. See Methods for details. **g**, Heatmap depicting surface markers with differential expression between the phenotypic classes defined in **e**. The eight markers with the most significant *P* values from DESeq2 were selected for each comparison between classes. Average expression across all nonlymphoid cells is shown. ITD, internal tandem duplication; mut, mutation; wt, wild type. **h**, Surface expression of immunotherapy targets CTLA4 (CD152) and PD-L1 (CD274) in different myeloid compartments of healthy donors and AMLs. Sample size is provided as Figure Source Data. **i**, Scatter plot depicting the average expression of all surface markers in healthy HSCs and MPPs (*x* axis) and leukemic stem cells (LSC) projecting to the HSC and MPP cell state (*y* axis). Cells from four patients where the HSC/MPP class was covered with more than 20 cells are included (AML1, AML2, AML3 and AML Q6). *P* values for differential expression were computed using DESeq2 and are encoded in the symbol size, and previously described LSC markers are depicted as a triangle. Interpatient variability is color-coded, see Methods, for details. See also Supplementary Data 2.

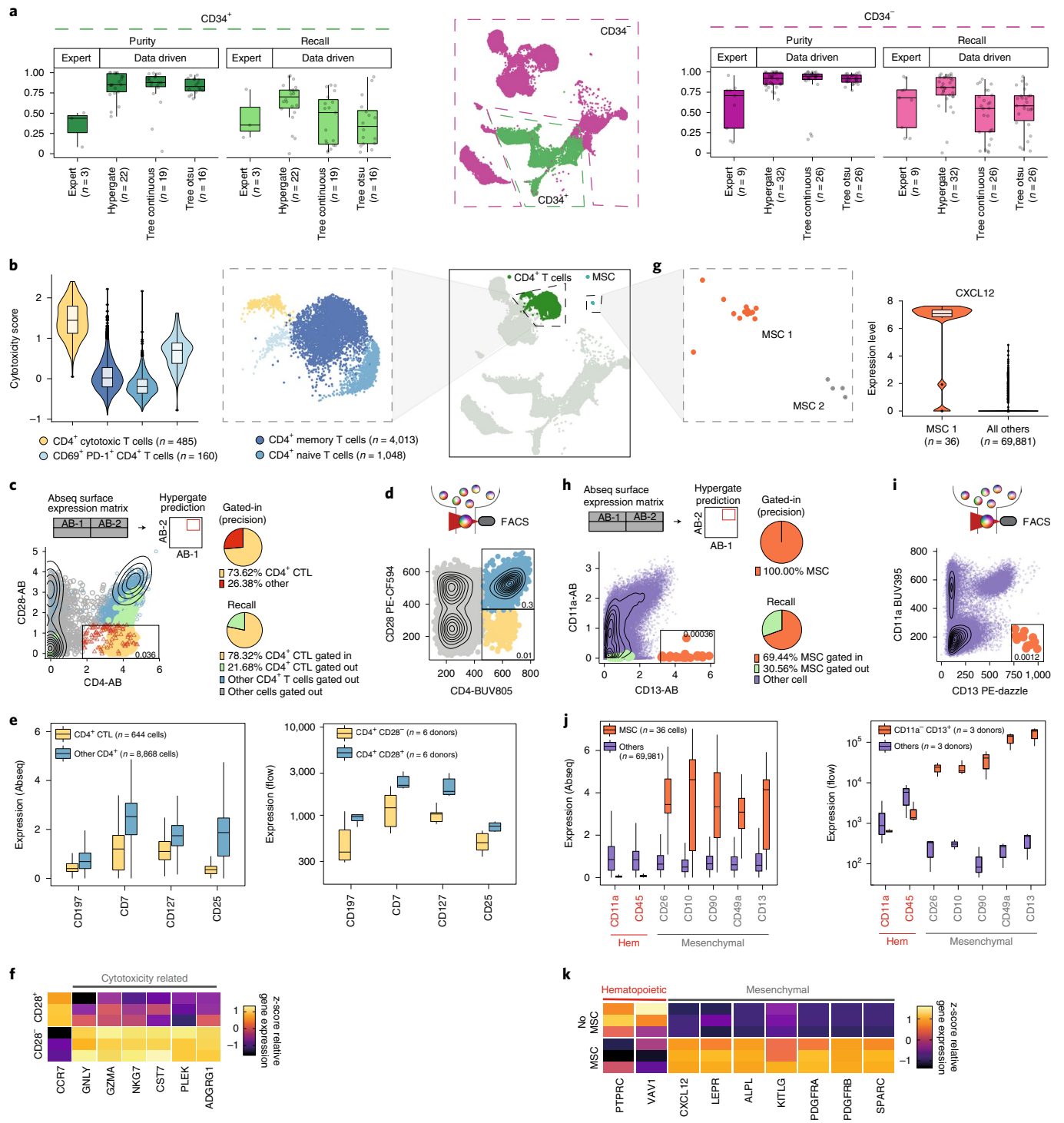


Fig. 5 | Data-driven definition of gating schemes for rare cell types. Boxplot sample sizes are provided in the figure. See Methods, section Data visualization for a definition of boxplot elements. **a**, Purity and recall of published or data-driven gating schemes for cell populations within CD34⁺ and CD34⁻ compartments, see also Extended Data Fig. 8. **b**, Different CD4⁺ T cell subsets are highlighted (central and right panels) and the corresponding distributions of cytotoxicity scores for every subset are displayed (left panel). **c**, Hypergate[34] was used to identify a gating scheme for the isolation of cytotoxic CD4⁺ T cells. The suggested gate is highlighted on a scatter plot of CD4 and CD28 expression as identified from pregated CD45⁺ CD3⁺ Abseq data. Pie charts indicate precision and recall. **d**, FACS plot displaying the expression of CD4 and CD28 on pregated CD45⁺ CD3⁺ cells, and respective gates. **e**, Boxplot depicting the expression of surface markers with differential expression between CD4⁺ cytotoxic T cells and other CD4⁺ subsets, as identified from Abseq data (left panel) and validated with FACS using the gating strategy from **d** (right panel). **f**, Heatmap depicting gene expression of cytotoxicity-related genes in FACS-sorted CD4⁺ CD28⁻ and CD4⁺ CD28⁺ cells, as quantified by qPCR ($n = 3$ patients). **g**–**j**, Analogous to **b**–**e**. MSCs were identified via high CXCL12 expression (**g**) and a CD11a⁻CD13⁺ gate on total BM cells was predicted for the isolation of CXCL12⁺ mesenchymal stem cells (**h**), which was confirmed using flow cytometry (**i**). **j**, Confirmation of differentially expressed surface markers on MSCs, derived from Abseq data, by flow cytometry. **k**, Heatmap depicting gene expression of common hematopoietic and MSC signature genes in FACS-sorted CD11a⁻CD13⁺ MSCs and total BM cells outside the gate, as quantified by qPCR ($n = 3$ patients).

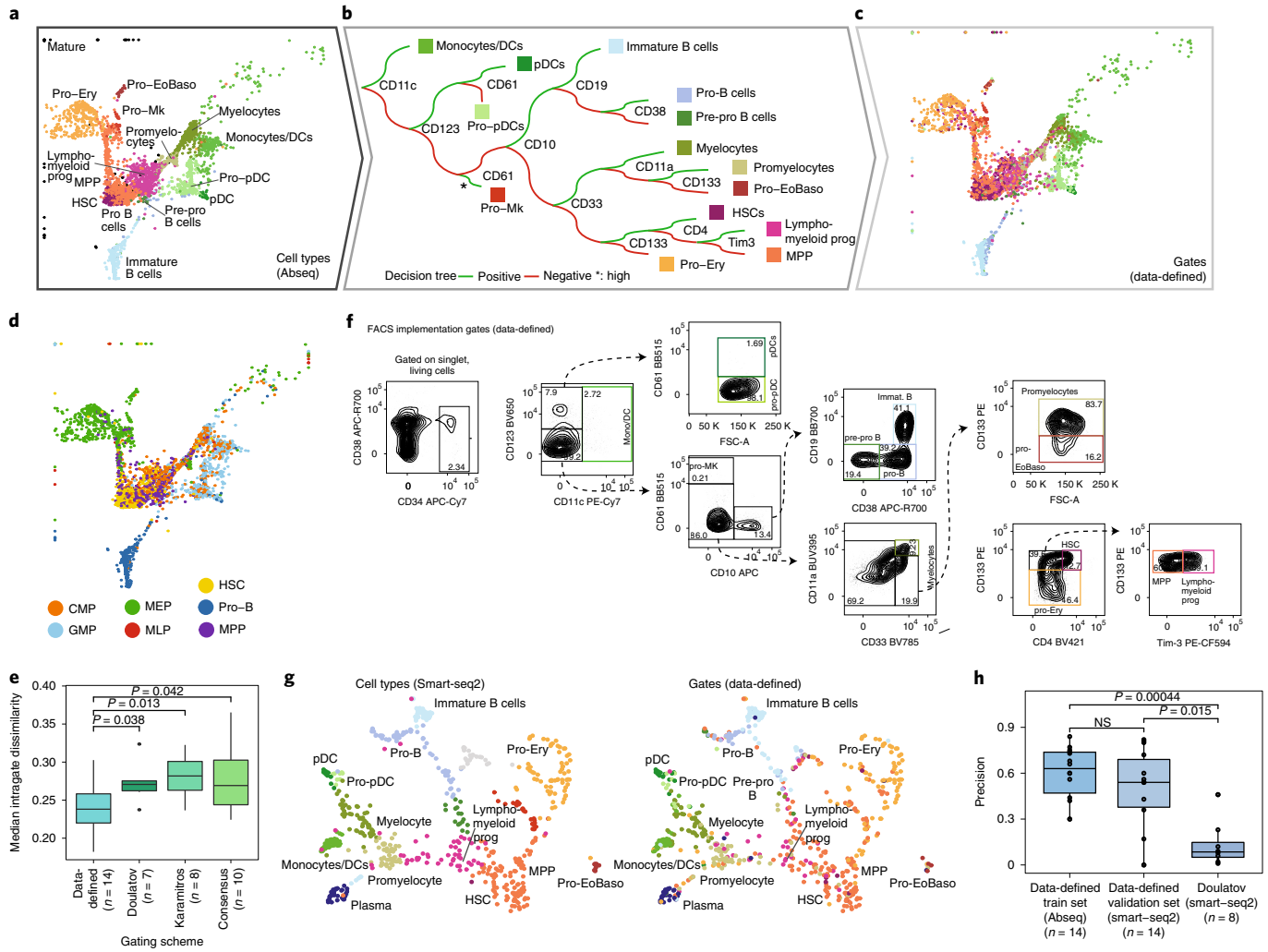


Fig. 6 | Data-driven definition of gating schemes for HSPCs. a, UMAP depicting all CD34⁺ HSPCs cells from one healthy young individual. See **b** for color scheme. **b**, Decision tree using surface marker expression from the Abseq data to classify cells into cell types. See Methods and main text for details. **c**, UMAP highlighting cell type classification obtained from the decision tree. Colors correspond to 'gates' applied to the expression levels of the 12 markers shown in **b**, not gene expression clusters. **d**, UMAP highlighting classification obtained from a decision tree recapitulating the classical gating scheme used in the field[17]. Since CD135 was not part of the Abseq panel, the expression of *FLT3* was smoothed using MAGIC[48]. **e**, Boxplot depicting the intragate dissimilarity for cell classification with panels from Doulatov et al.[17], the gating scheme from Karamitros et al.[25], a 'consensus gating' scheme (*see* Extended Data Fig. 9) and the data-driven gating scheme (**c**). Intragate dissimilarity is defined as one minus the average Pearson correlation of normalized gene and surface antigen expression values of all cells within the gate. *P* values are from a two-sided Wilcoxon test. Sample size is shown in the figure. See Methods, section Data visualization for a definition of boxplot elements. **f**, Implementation of FACS gating scheme from **b**. **g**, UMAP display of mRNA expression of $n = 630$ CD34⁺ HSPCs from an indexed single-cell Smart-seq2 experiment where the expression of relevant surface markers was recorded using FACS. Left panel: color indicates gene expression cluster, see Supplementary Note 8 for details. Right panel: color indicates classification by the FACS scheme from **f**. **h**, Precision of the classification scheme shown in **b**, computed on the training data (Abseq) and the test data (Smart-seq2). Precision was computed per gate as the fraction of correctly classified cells. For comparison with the Doulatov gating scheme, the dataset from Velten et al.[9] was used. NS, not significant. *P* values are from a two-sided Wilcoxon test. Sample size is shown in the figure.

observed in their well-characterized CD8⁺ T cell counterparts[35]. While this cell type has been suggested to be involved in several physiological and pathophysiological processes, no coherent gating strategy for their prospective isolation exists[36]. Hypergate suggested that cytotoxic CD4⁺ T cells display an immunophenotype of CD4⁺CD28⁻, and differential expression analyses of surface markers revealed that cytotoxic CD4⁺ T cells express significantly lower levels of CD7, CD25, CD127 and CD197 when compared with other CD4⁺ T cell subsets (Fig. 5b–e). Flow cytometric analyses of CD4⁺CD28⁻ T cells confirmed the expected immunophenotype in BM from healthy donors and patients with different hematological cancers, suggesting a robust and efficient prospective isolation of this rare cell type (Fig. 5d and Extended Data Fig. 8e). Finally, FACS-based sorting of CD4⁺CD28⁻ T cells followed by gene expression analysis confirmed the expression of cytotoxicity genes in this population (Fig. 5f).

MSCs constitute a rare and heterogeneous group of cells in the BM[37,38]. While ex vivo expanded MSCs have been phenotyped extensively, primary human MSCs remain poorly characterized, in particular due to their extremely low frequency. In our dataset, we captured a small number of heterogeneous MSCs, with one subset (MSC-1) expressing high levels of the key BM-homing cytokine CXCL12 (Fig. 5g). Hypergate suggested *CXCL12*-expressing MSCs to be isolated most efficiently by expression of CD13 and absence of

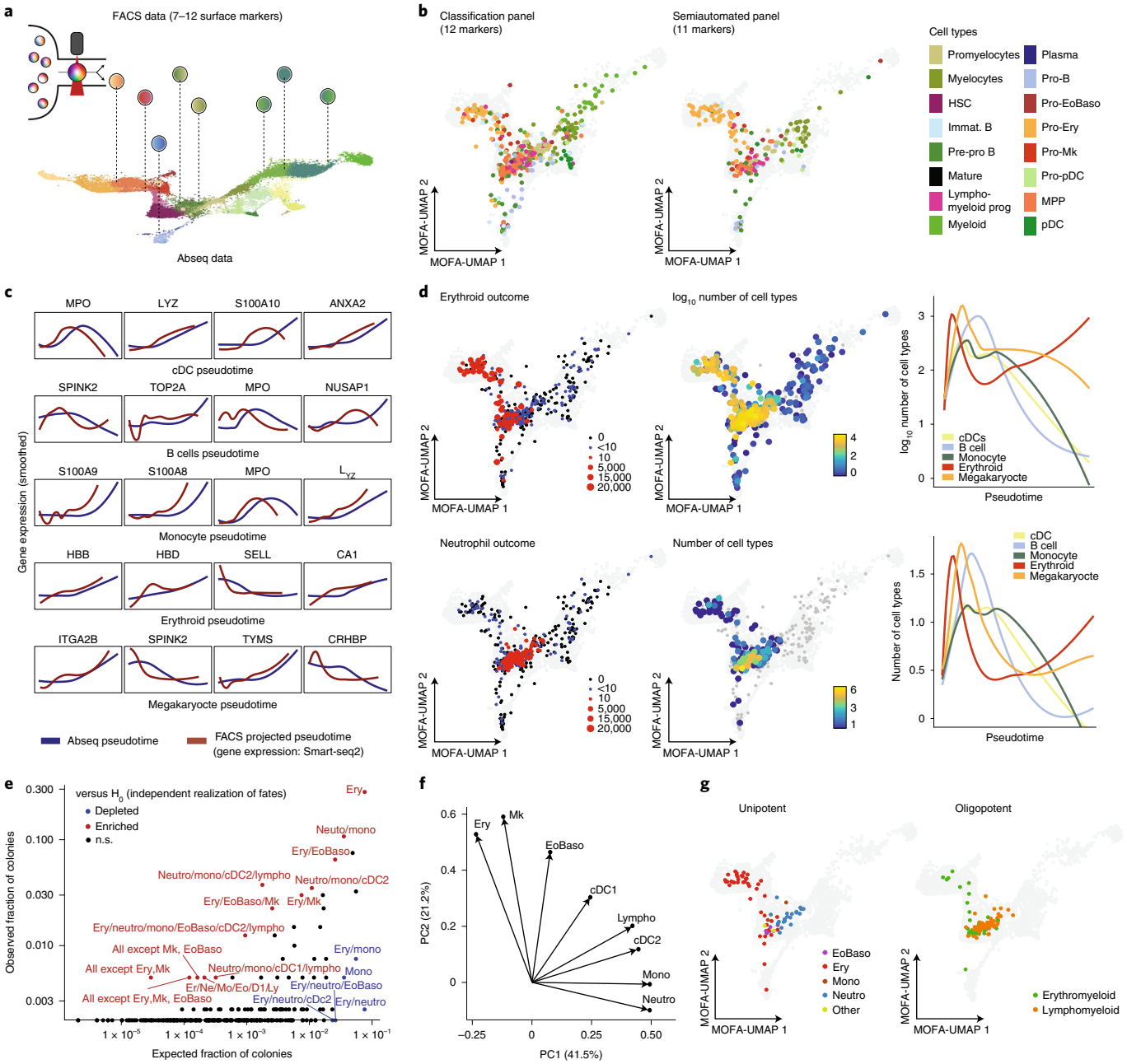


Fig. 7 | Systematic integration of single-cell genomics, flow cytometry and functional data. a, Illustration of the concept. **b**, Projection of indexed Smart-seq2 data onto a reference UMAP. Single cells with recorded FACS measurements of surface markers were subjected to Smart-Seq2 based scRNA-seq. FACS measurements of surface markers were used to project cells onto the UMAP (Methods). Colors denote cell type identified from RNA-seq. See Supplementary Table 6 for composition of the FACS panels. **c**, FACS-based projection of indexed Smart-seq2 data onto reference pseudotime trajectories. Line plots depict the RNA expression of differentiation markers smoothed over projected pseudotime values (red). For comparison, expression values determined from Abseq data are shown (blue). The selected genes correspond to the five genes with the strongest statistical association with the respective trajectory. **d**, Projection of indexed single-cell culture data onto a reference UMAP. Single cells with available FACS measurements of 12 surface markers were projected onto the UMAP defined by Abseq. Single cells were seeded into culture medium supporting the formation of erythroid, megakaryocytic and distinct myeloid cell types. UMAPs highlight the ability of single cells to give rise to erythroid cells and neutrophils, colony size and total number of cell types per colony. Colony and total number of cell types per colony are also plotted against projected pseudotime. **e**, Analysis of cell type combinations in $n = 397$ colonies. For any combination of Erythroid (Ery), Neutrophil (Neutro), Monocytic (Mono), Eosinophil or Basophil (EoBaso), Lymphoid (Lympho), Megakaryocytic (Mk) and Dendritic (cDC1 and cDC2) potential, the scatter plot depicts the fraction of colonies containing this exact combination of cell types (*y* axis) and the theoretical fraction of colonies containing the same combination under the assumption that cell fates are independently realized with the same marginal probabilities (*x* axis). Significance from a binomial test is color-coded. n.s., not significant. These analyses do not exclude that other combinations of fates are not biologically selected as well; that is, absence of evidence does not constitute evidence for absence. **f**, Principal component analysis of colony compositions. PC, principal component. **g**, Distribution of colonies with frequent combinations of cells types in the projected UMAP space. Erythromyeloid, exclusively EoBaso, Mk and/or Ery cells; Lymphomyeloid, all other combinations.

CD11a (Fig. 5h). Indeed, flow cytometric analyses of CD13$^+$CD11a$^-$ MSCs validated the immunophenotype suggested by our Abseq data and confirmed known and new MSC surface markers identified by our approach (Fig. 5i,j and Extended Data Fig. 8f). Moreover, FACS-based isolation of CD13$^+$CD11a$^-$ cells followed by transcriptomic analyses revealed a high enrichment of *CXCL12* and other key MSC signature genes (Fig. 5k).

Together, these analyses demonstrate the utility of our approach for deriving gating schemes from data and mapping the surface marker expression of poorly characterized populations. In combination with our single-cell proteo-genomic reference map, the Abseq App allows users to define new data-driven gating schemes for any population of interest.

A data-defined gating scheme for human hematopoiesis. Gating schemes for complex biological systems, such as the HSPC compartment, are improving steadily. However, there is strong evidence from single-cell transcriptomics[9,10,18,19], lineage tracing[22,23] and single-cell functional experiments[21] that even the most advanced gating schemes do not recapitulate the molecular and cellular heterogeneity observed by single-cell genomics approaches. This has contributed to several misconceptions in the understanding of the hematopoietic system, most notably incorrect assumptions on the purity of cell populations and inconsistent views on lineage commitment hierarchies[11–14].

To generate flow cytometric gating schemes that most adequately reflect the transcriptomic states associated with HSC differentiation, we used the Abseq dataset of CD34$^+$ cells from one BM sample ('Young1') to train a decision tree. Thereby, we obtained a gating scheme that uses 12 surface markers to define 14 leaves representing molecularly defined cell states with high precision (Fig. 6a–c). The data-derived scheme excelled in the identification of lineage-committed progenitors—a principal shortcoming of many current gating strategies (Fig. 6a–c)[9,10,21,22]. Importantly, cell populations defined by the data-defined gating scheme were transcriptionally more homogenous, compared with a widely used gating scheme[17] (Fig. 6d,e), a state-of-the-art gating scheme focusing on lymphomyeloid differentiation[25] (Fig. 6e and Extended Data Fig. 9a–d) and a 'consensus gating' scheme generated in silico to combine the latter with a scheme focusing on erythroid-myeloid differentiation[26] (Fig. 6e and Extended Data Fig. 9b). Of note, individual populations from the data-defined scheme displayed a functional output comparable with that of populations of the 'consensus gating' scheme, while the data-defined scheme overall provided a higher level of information on functional lineage commitment (Extended Data Fig. 9e,f).

To validate this new gating scheme, we implemented the suggested surface marker panel in a classical flow cytometry setup and performed Smart-seq2-based single-cell RNA-seq while simultaneously recording surface marker expression (index scRNA-seq) (Fig. 6f,g and Supplementary Note 8). This approach demonstrated that the new gating strategy efficiently separated molecularly defined cell states (Fig. 6g). Quantitatively, the data-defined gating scheme performed equally well at resolving molecularly defined cell states on the Abseq training data as on the Smart-seq2 validation data, and significantly outperformed the expert-defined gating scheme (Fig. 6h). A limitation of the low cellular throughput of the Smart-seq2 analysis is that the signature-based identification might result in the 'over-identification' of certain cell states. Together, our results demonstrate that high-content single-cell proteo-genomic maps can be used to derive data-defined cytometry panels that describe the molecular states of complex biological systems with high accuracy. Moreover, our gating scheme permits a faithful identification and prospective isolation of transcriptomically defined progenitor states in the human hematopoietic hierarchy using cost-effective flow cytometry.

Mapping flow cytometry data on single-cell reference maps. While classical FACS gating strategies are of great use for the prospective isolation and characterization of populations, single-cell genomics studies revealed that differentiation processes, including the first steps of hematopoiesis, are represented most accurately by a continuous process[9,18,20,27,39]. To complement the approach based on discrete gates, we propose here that high-dimensional flow cytometry data can be used to place single cells into the continuous space of hematopoietic differentiation spanned by single-cell proteo-genomics exploiting shared surface markers (Fig. 7a). Based on the observation that surface marker expressions in flow cytometry and Abseq follow similar distributions (Extended Data Fig. 10a), we developed a new projection algorithm termed nearest rank neighbors (NRN https://git.embl.de/triana/nrn/; see Methods). Given an identical starting population, NRN employs sample ranks to transform surface marker expression of FACS and Abseq data to the same scale, followed by k-nearest neighbors-based projection into a space defined by the proteo-genomic single-cell data. We tested NRN on FACS-indexed Smart-seq2 datasets using the classification panel developed in Fig. 6 (12 markers) and a semiautomated panel based on our Abseq data to better resolve erythromyeloid lineages (11 markers; Supplementary Note 8). We evaluated the performance of NRN using a variety of methods. First, cell types molecularly defined by Smart-seq2 were placed correctly on the Abseq uniform manifold approximation and projection (UMAP) (Fig. 7b). For most molecularly defined cell types, the accuracy of the projection using the flow cytometry data was close to the performance of data integration using whole transcriptome data with a state-of-the-art algorithm (Extended Data Fig. 10b–d). Most importantly, the projections closely reflected the gradual progression of cells through pseudotime, as confirmed by the expression dynamics of key lineage genes from our FACS-indexed Smart-seq2 data (Fig. 7c). This suggests that NRN, in combination with high-quality reference datasets, can be used to study the continuous nature of cellular differentiation processes by flow cytometry.

A key limitation of single-cell genomics remains the lack of insight into functional differentiation capacities of cells. We therefore evaluated whether NRN can be used to interpret functional single-cell data in the context of single-cell genomic reference maps. For this purpose, we performed single-cell culture assays, while recording surface markers of our data-defined gating scheme from Fig. 6, followed by data integration using our Abseq data via NRN. As expected, cells with the highest proliferative capacity and lineage potency were placed in the phenotypic HSC and MPP compartments, and HSPCs placed along the transcriptomically defined differentiation trajectories continuously increased the relative generation of cells of the respective lineage (Fig. 7d). Functionally unipotent progenitors cells were observed along the respective transcriptomic trajectories, but were also present in the phenotypic HSC/MPP compartment (Fig. 7d,g), in line with previous findings on early lineage commitment of HSPCs[9,10,21]. By contrast, oligopotent cells with distinct combinations of cell fates were enriched specifically in the HSC/MPP compartment (Fig. 7d,g). Some of these fate combinations, in particular combinations of erythroid, megakaryocytic and eosinophilic/basophilic fates, and combinations of lymphoid, neutrophilic, monocytic and dendritic fates, co-occurred more frequently than expected by chance (Fig. 7e,f), in line with most recent findings on routes of lineage segregation[9,18,40,41]. Despite strong associations between surface phenotype, transcriptome and function, cells with a highly similar phenotype can give rise to different combinations of lineages (Fig. 7g). This observation suggests a role of stochasticity in the process of lineage commitment, or hints towards layers of cell fate regulation not observed in the transcriptome. Together, our observations confirm that hematopoietic lineage commitment occurs predominantly continuously along the routes predicted by the transcriptome, with an early

primary erythromyeloid versus lymphomyeloid split[9,10,18,21,40,41] and might help reconciling discrepancies in the interpretation of previous studies.

In summary, our data resource, alongside the NRN algorithm, enables accurate integration of flow data with single-cell genomics data. This permits the charting of continuous processes by flow cytometry and the mapping of single-cell functional data into the single-cell genomics space.

Discussion

In this study, we have demonstrated the power of single-cell proteo-genomic reference maps for the design and analysis of cytometry experiments. We have introduced a map of human blood and BM spanning the expression of 97–197 surface markers across 45 cell types and stages of HSC differentiation, healthy ageing and leukemia. Our dataset is carefully annotated and will serve as a key resource for hematology and immunology.

While cytometry experiments remain the workhorse of immunology, stem-cell biology and hematology, recent single-cell atlas projects have revealed that current cytometry setups do not accurately reflect the full complexity of biological systems[10,42]. For the first time, we have exploited single-cell proteo-genomic data to systematically design and interpret flow cytometry experiments that mirror most accurately the cellular heterogeneity observed by single-cell transcriptomics. Unlike approaches based on index sorting[9,10,43,44], single-cell proteo-genomics has a sufficient throughput to enable the profiling of entire tissues or organs, and at the same time covers up to several hundred surface markers. Unlike single-cell RNA-seq data, antibody tag counts reflect the true distribution of surface marker expression, enabling a quantitative integration of cell atlas data with FACS. Building on these unique properties of our reference map, we have automated the design of gating schemes for the isolation of rare cell types, devised a gating strategy that reflects the molecular routes of HSC differentiation and demonstrated the direct interpretation of flow cytometry data in the context of our reference.

These advances enable a functional characterization of molecularly defined cell states and thereby directly affect HSC research. There is a growing consensus in the field that lineage commitment occurs early from primed HSCs, that not all progenitor cells in the classical megakaryocyte-erythrocyte progenitor/granulocyte-macrophage progenitor (MEP/GMP) gates are functionally oligopotent and that the main branches of the hematopoietic system are a *GATA2*-positive branch of erythroid, megakaryocytic and eosinophil/basophil/mast cell progenitors, as well as a *GATA2*-negative branch of lymphomyeloid progenitors, including the progenitors of monocytes, neutrophils and dendritic cells[9,18,19,27,40,41,45]. Due to a lack of better alternatives, many functional studies still use the classical gating scheme alongside the outdated concept of 'common myeloid progenitors'[15,16,28]. Here, we introduce and validate a flow cytometry scheme that allows the prospective isolation of molecularly homogeneous progenitor populations. We have used this scheme to show that transcriptional lineage priming impacts on cellular fate in vitro[9,21], thereby contributing further evidence for the revised model of hematopoiesis. In the future, a wider use of this scheme has the potential to avoid conflicting results stemming from imprecisely defined populations.

Furthermore, these advances enable the rapid profiling of blood formation and other BM phenotypes while offering a resolution comparable with that of single-cell genomics. Recently, BM phenotypes of disease, ranging from sickle cell disease[46] to leukemia[47] have been investigated using scRNA-seq. However, due to economic and experimental hurdles, the throughput of these studies has remained restricted to maximally tens of patients. Accordingly, the ability to associate patient genotypes with phenotypes is thereby highly limited, and these assays have not been translated to diagnostic routines. Our new gating schemes and analytical strategies are widely applicable to profile aberrations encountered in disease, both in research and, ultimately, in clinical diagnostics.

Although we have demonstrated the implementation of data-driven design and analysis strategies for cytometry assays in the context of BM, conceptually the approach presented here can be applied to any organ of interest. Thereby, it has the potential to enable the precise isolation and routine profiling of myriad cell types discovered by recent single-cell atlas projects.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41590-021-01059-0.

Received: 16 February 2021; Accepted: 24 September 2021;
Published online: 22 November 2021

References

1. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
2. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
3. Giladi, A. & Amit, I. Single-cell genomics: a stepping stone for future immunology discoveries. *Cell* **172**, 14–21 (2018).
4. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
5. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107.e17 (2018); erratum 173, 1307 (2018).
6. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
7. Baccin, C. et al. Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat. Cell Biol.* **22**, 38–48 (2020).
8. Van Dongen, J. J. M. et al. EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia* **26**, 1908–1975 (2012).
9. Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
10. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
11. Loughran, S. J., Haas, S., Wilkinson, A. C., Klein, A. M. & Brand, M. Lineage commitment of hematopoietic stem cells and progenitors: insights from recent single cell and lineage tracing technologies. *Exp. Hematol.* **88**, 1–6 (2020).
12. Haas, S., Trumpp, A. & Milsom, M. D. Causes and consequences of hematopoietic stem cell heterogeneity. *Cell Stem Cell* **22**, 627–638 (2018).
13. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018).
14. Jacobsen, S. E. W. & Nerlov, C. Haematopoiesis in the era of advanced single-cell technologies. *Nat. Cell Biol.* **21**, 2–8 (2019).
15. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000).
16. Kondo, M., Weissman, I. L. & Akashi, K. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell* **91**, 661–672 (1997).
17. Doulatov, S. et al. Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat. Immunol.* **11**, 585–593 (2010).
18. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
19. Giladi, A. et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).
20. Nestorowa, S. et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
21. Notta, F. et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
22. Perié, L., Duffy, K. R., Kok, L., De Boer, R. J. & Schumacher, T. N. The branching point in erythro-myeloid differentiation. *Cell* **163**, 1655–1662 (2015).
23. Rodriguez-Fraticelli, A. E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).

24. Haas, S. Hematopoietic stem cells in health and disease—insights from single-cell multi-omic approaches. *Curr. Stem Cell Rep.* **6**, 67–76 (2020).
25. Karamitros, D. et al. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells article. *Nat. Immunol.* **19**, 85–97 (2018).
26. Psaila, B. et al. Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol.* **17**, 83 (2016).
27. Pellin, D. et al. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019).
28. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
29. Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. Abseq: ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci. Rep.* **7**, 44447 (2017).
30. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
31. Fagnoni, F. F. et al. Expansion of cytotoxic $CD8^+$ $CD28^-$ T cells in healthy ageing people, including centenarians. *Immunology* **88**, 501–507 (1996).
32. Peters, M. J. et al. The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, 8570 (2015).
33. Hanekamp, D., Cloos, J. & Schuurhuis, G. J. Leukemic stem cells: identification and clinical application. *Int. J. Hematol.* **105**, 549–557 (2017).
34. Becht, E. et al. Reverse-engineering flow-cytometry gating strategies for phenotypic labelling and high-performance cell sorting. *Bioinformatics* **35**, 301–308 (2019).
35. Szabo, P. A. et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* **10**, 4706 (2019).
36. Takeuchi, A. & Saito, T. CD4 CTL, a cytotoxic subset of CD4+ T cells, their differentiation and function. *Front. Immunol.* **8**, 194 (2017).
37. Al-Sabah, J., Baccin, C. & Haas, S. Single-cell and spatial transcriptomics approaches of the bone marrow microenvironment. *Curr. Opin. Oncol.* **32**, 146–153 (2020).
38. Frenette, P. S., Pinho, S., Lucas, D. & Scheiermann, C. Mesenchymal stem cell: keystone of the hematopoietic stem cell niche and a stepping-stone for regenerative medicine. *Annu. Rev. Immunol.* **31**, 285–316 (2013).
39. Macaulay, I. C. et al. Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.* **14**, 966–977 (2016).
40. Drissen, R., Thongjuea, S., Theilgaard-Mönch, K. & Nerlov, C. Identification of two distinct pathways of human myelopoiesis. *Sci. Immunol.* **4**, eaau7148 (2019).
41. Görgens, A. et al. Multipotent hematopoietic progenitors divide asymmetrically to create progenitors of the lymphomyeloid and erythromyeloid lineages. *Stem Cell Rep.* **3**, 1058–1072 (2014).
42. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).
43. Baron, C. S. et al. Cell type purification by single-cell transcriptome-trained sorting. *Cell* **179**, 527–542.e19 (2019).
44. Wilson, A. et al. Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118–1129 (2008).
45. Zheng, S., Papalexi, E., Butler, A., Stephenson, W. & Satija, R. Molecular transitions in early progenitors during human cord blood hematopoiesis. *Mol. Syst. Biol.* **14**, e8041 (2018).
46. Hua, P. et al. Single-cell analysis of bone marrow–derived $CD34^+$ cells from children with sickle cell disease and thalassemia. *Blood* **134**, 2111–2115 (2019).
47. van Galen, P. et al. Single-cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281.e24 (2019).
48. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729.e27 (2018).

Methods

All reagents and antibodies used are listed in Supplementary Tables 1 (primers for targeted transcriptomics), 2 (Abseq antibodies) and 6 (all other reagents, oligonucleotides, equipment and software).

Human samples. BM samples from healthy and diseased donors were obtained at the University clinics in Heidelberg and Mannheim after informed written consent using ethic application numbers S480/2011 and S-693/2018. For demographic characteristics on sample donors, see Supplementary Table 3. BM aspirates were collected from iliac crest. Healthy BM donors received financial compensation in some cases. For BM, mononuclear cells were isolated by Ficoll (GE Healthcare) density gradient centrifugation and stored in liquid nitrogen until further use. All experiments involving human samples were approved by the ethics committee of the University Hospital Heidelberg and were in accordance with the Declaration of Helsinki.

Cell sorting for Abseq. Human BM samples were thawed in a water bath at 37 °C and transferred dropwise into RPMI-1640 10% FCS. Cells were centrifuged for 5 min at 350 and washed once with RPMI-1640 10% FCS. Cells were resuspended in FACS buffer (FB) (PBS 5% FCS 0.5 mM EDTA) containing CD34-PE and CD3 PE-Cy7 and FcR blocking reagent (Miltenyi) and incubated for 15 min at 4 °C. Cells were washed with FB and resuspended in 1 ml FB, followed by addition of 1 µl CellEvent Caspase-3/7 Green (ThermoFisher) and 1 µl 4,6-diamidino-2-phenylindole (DAPI) (ThermoFisher) to the cell suspension. After 3 min incubation at room temperature, cells were filtered through a 40 µm cell strainer. Singlet, CaspaseGreen$^-$ DAPI$^-$ total BM and singlet, CaspaseGreen$^-$ DAPI$^-$ CD34$^+$ (HSPCs) as well as singlet, CaspaseGreen$^-$ DAPI$^-$ CD3$^+$ (T cells) cells were sorted on an Aria Fusion II cell sorter (BD). In general, the entire CD34$^+$ fraction from one thawed vial was sorted (~2×10^4) and combined with 1×10^5 CD34$^-$ total BM cells (see also Extended Data Fig. 2). In CD3$^+$ T cell-enriched AML samples, 2×10^4 CD3$^+$ T cells were mixed with the CD34$^+$ HSPC fraction and combined with 1×10^5 CD34$^-$ total BM cells. For the generation of the AML query datasets, 2×10^4 live total BM cells from each of 12 different AML samples were sorted. In case of the CD34$^+$ immature HSPCs enrichment experiment, healthy adult human BM cells were stained with anti-human CD34, CD38, CD45RA, CD10 and fixable viability dye efluor506 and 5×10^3 were sorted from each of four different gates (CD34$^+$CD38$^+$CD45RA$^-$, CD34$^+$CD38$^+$CD45RA$^+$, CD34$^+$CD38$^-$CD45RA$^-$, CD34$^+$CD38$^-$CD45RA$^+$). In cases where different biological samples or sorted populations were combined in the same run, cells of interest were sorted and labeled by cell hashing antibodies before surface labeling and single-cell capture as described in Abseq surface labeling, single-cell capture and library preparation.

Cell sorting for gene expression analysis and flow cytometry. Human BM samples were thawed as described above. For dead cell exclusion and blocking of nonspecific binding, fixable viability dye efluor506 (ThermoFisher) and FcR blocking reagent (Miltenyi) were used in all staining solutions. Cells were generally stained for 15 min at 4 °C and then washed once with FB, resuspended in 1 ml FB and filtered through a 40 µm cell strainer. For cytotoxic CD4$^+$ T cell sorting, cells were stained in FB containing anti-CD3, CD4, CD7, CD28, CD45RA, CD45 and CD127 surface antibodies. Singlet, live, CD45$^+$, CD3$^+$ cells were gated and CD4$^+$CD28$^-$ or CD4$^+$CD28$^+$ cells were sorted and processed as described below. For MSC gene expression analysis, cells were stained in FB containing anti-CD10, CD11a, CD13, CD26, CD31, CD45, CD49a, CD90, CD105, CD146 and CD271 surface antibodies. Singlet, live, CD11a$^-$CD13$^+$ MSCs or all cells outside this gate were sorted. Cells were sorted on either FACSAria Fusion or FACSAria II equipped with 100 µm nozzles, respectively.

For flow cytometric analysis, human BM samples were processed as described above. For analysis of cytotoxic CD4 T cells across hematopoietic malignancies, cells were stained with anti-CD3, CD4, CD7, CD25, CD28, CD45RA, CD45, CD69 and CD127 surface antibodies. For analysis of CD98 expression in hematopoietic stem and progenitors, cells were stained with anti-human CD4, CD10, CD11a, CD34, CD38, CD45RA, CD49f, CD90, CD98, CD133 and Tim3 antibodies. For analysis of CD326 surface expression in comparison with CD71 and CD41, healthy adult human BM was stained with anti-human CD34, CD38, CD41, CD44, CD45RA, CD49b, CD49d, CD71, CD90 and CD123 antibodies. All experiments were measured on BD FACSFortessa flow cytometers, equipped with five lasers.

Panel design for targeted transcriptomics. Panel design is described in Supplementary Note 1. In short, we used a human cell atlas reference and followed the method described by Schraivogel et al. for target gene selection[49].

Abseq surface labeling, single-cell capture and library preparation. Abseq surface antibody libraries (Supplementary Table 2) were pipetted 24 h before experiments. For most antibodies, 1 µl was used for surface library preparation. Antibodies recognizing epitopes with well-known high surface expression were further diluted in PBS and 1 µl was added to the surface library (for example HLA ABC, CD45, CD11a). Sorted cells (around 1.2×10^5–1.4×10^5; described in Cell sorting for Abseq) were centrifuged 5 min at 350*g* and resuspended in the surface library mix (around 100 µl for the 97 Ab panel, 200 µl for the 197 Ab panel).

In cases where different biological samples or sorted populations were combined in the same run, sorted cells were labeled individually with oligonucleotide coupled cell hashing antibodies (BD single-cell multiplexing kit) for 25 min on ice, washed three times in all, each followed by 5 min centrifugation at 350*g* and then pooled and then subjected to Abseq cell surface labeling. Cells were then labeled for 30 min at 4 °C and washed three times in all, each followed by 5 min centrifugation at 350*g*. Cells were resuspended in sample buffer (BD Rhapsody Cartridge reagent kit) and between 1×10^4 and 2×10^4 cells were captured with the BD Rhapsody single-cell system following the manufacturer's instructions[50]. Antibody tag libraries, multiplexing libraries and targeted mRNA gene expression libraries were generated following manufacturer instructions. For mRNA libraries, the targeted panel (Supplementary Table 1) or the whole transcriptome analysis library preparation protocol was used according to the manufacturer's instructions (BD). Resulting libraries were quality checked by Qubit and Bioanalyzer, pooled and sequenced using NextSeq500 or Illumina Novaseq S2 (Illumina; high-output mode).

Single-cell index cell cultures. Two days before index sorting, irradiated MS-5 feeder cells were plated at a density of 1×10^4 cells per well into 96-well flat-bottom cell culture plates in αlpha-minimal essential medium with ribo-and deoxynucleosides (ThermoFisher) containing 10% FCS (Gibco), glutamine (2 mM) (ThermoFisher), penicillin/streptomycin (100 U ml^{-1}) (ThermoFisher) and sodium pyruvate (2 mM) (Gibco). Several hours before index sorting, the medium was replaced by 100 µl H5100 medium (StemCell Technologies) containing glutamine (2 mM) (ThermoFisher), penicillin/streptomycin (100 U ml^{-1}) (ThermoFisher), hydrocortisone (1 nM) (StemCell Technologies), SCF (20 ng ml^{-1}), FLT3-L (100 ng ml^{-1}), TPO (50 ng ml^{-1}), IL-3 (20 ng ml^{-1}), IL-5 (20 ng ml^{-1}), IL-6 (20 ng ml^{-1}), IL-7 (20 ng ml^{-1}), IL-11 (20 ng ml^{-1}), G-CSF (20 ng ml^{-1}), GM-CSF (20 ng ml^{-1}), M-CSF (20 ng ml^{-1}) (all Preprotech) and EPO (3 U ml^{-1}) (R&DSystems). Two BM samples from the same donor were thawed and washed as described above. The first sample was subsequently resuspended in 100 µl FB containing anti-human CD4, CD10 (BioLegend), CD11a, CD11c, CD19, CD33, CD34, CD38, CD61, CD123, CD133 and Tim3 antibodies (Classification panel), whereas the second sample was stained with anti-human CD11a, CD33, CD34 (Biolegend), CD38, CD49b, CD61, CD71, CD123, CD133, CD326 and FcεR1A (eBioscience) (Semiautomated panel). In another experiment, cells were labeled with anti-human CD11a, CD71, CD45RA, CD44, CD135, Tim3 (Biolegend), CD90, CD326, CD41 (BioLegend), CD123 (ThermoFisher), CD10, CD38 and CD34 (BioLegend) antibodies (Consensus panel). All antibody clones for flow cytometry matched clones from Abseq experiments and were purchased from BD, except otherwise indicated. For dead cell exclusion and blocking of nonspecific binding, fixable viability dye efluor506 (ThermoFisher) and FcR blocking reagent (Miltenyi) were included in both staining solutions. After staining for 15 min at 4 °C, cells were washed with FB, resuspended in 1 ml FB and filtered through a 40 µm cell strainer. For both assays, 480 single, live CD34$^+$ cells were FACS indexed and sorted into the feeder cell containing 96-well plates as described above. Cells were incubated at 37 °C, 5% CO_2 for 16–19 days. To analyze clonal output, cells were harvested and transferred to 96-well V bottom plates, washed with FB and resuspended in 10 µl FB containing anti-human CD1c (Biolegend), CD14, CD19 (Biolegend), CD34 (Biolegend), CD41a (Biolegend), CD45, CD56, CD66b, CD123, CD235a, CD303, CD141, CD370 (Biolegend) and FcεR1a (eBioscience). For dead cell exclusion and blocking of nonspecific binding, fixable viability dye efluor506 (ThermoFisher) and FcR blocking reagent (Miltenyi) were included in the staining solution. After staining for 15 min at 4 °C, cells were washed with FB and resuspended in 100 µl FB and filtered through a 40 µm cell strainer. Cells were analyzed on a LSRII (BD) flow cytometer. Erythroid lineage output was determined via CD235$^+$ expression, which was concomitant with the downregulation of CD45 expression (CD45$^-$CD235$^+$). Myeloid lineages were defined via CD66b and CD14 antibodies (CD235$^-$CD45$^+$CD66b$^+$ or CD235$^-$CD45$^+$CD14$^+$). Dendritic cell lineages were defined via CD1c, CD141, CD370, CD303 and CD123 expression. Lymphoid cell lineages were defined via CD19 and CD56 expression. Megakaryocyte output was determined via CD41a expression, Eosinophil/basophil output was determined via FcεR1a expression. Generally, only wells that contained more than ten CD45$^+$CD235$^-$ or CD45$^+$CD235$^+$ or CD45$^+$CD235$^-$ cells were considered during analysis if not stated otherwise. For calculation of erythroid ratios, the count of all generated erythroid cells was divided by the sum of all other generated cells. Myeloid ratios were determined by dividing the sum of generated myeloid and dendritic cells by the sum of all other generated cells.

Single-cell index RNA-sequencing. For single-cell index RNA-sequencing, cells from the same samples that were prepared for single-cell cell index cultures were used. Hardshell 96-well polymerase chain reaction (PCR) plates (Bio-Rad) were prefilled with 4 µl lysis buffer containing 1 µl RNase inhibitor (40 U ml^{-1}, Takara), 1.9 µl Triton X-100 (0.2%, Sigma), 1 µl oligo $dT_{30}VN$ (10 µM, Sigma) and dNTPs (10 mM, ThermoFisher). Cells were FACS indexed, sorted into lysis buffer and snap frozen on dry ice. For cell lysis, plates were incubated for 5 min at 10 °C, followed by incubation for 3 min at 72 °C in a thermocycler (PCRMax). For reverse transcription, 0.25 µl RNase inhibitor (40 U ml^{-1}, Takara), 0.5 µl DTT (20 mM, Takara) 0.2 µl template switching oligonucleotides (50 µM, IDT), 1.05 µl H_2O (Ambion), 2 µl Smartscribe buffer (5×, Takara) and 1 µl Smartscribe

(100 U ml^{-1}, Takara) was added to each well. Reverse transcription was performed by incubating plates for 90 min at 42 °C, followed by ten cycles of 2 min at 50 °C, 2 min 42 °C, followed by 10 min at 72 °C followed by 4 °C storage. To amplify cDNA, 12.5 µl KAPA HiFi HotStart (Roche), 0.25 µl ISPCR primer (10 µM, Sigma) and 2.25 µl H_2O was added to each well. Plates were incubated for 3 min at 98 °C, 23 cycles of 20 s at 98 °C, 15 s at 67 °C, 6 min at 72 °C followed by one stage for 5 min at 72 °C, followed by final storage at 4 °C. cDNA was then cleaned up using an equal volume (25 µl) of SPRIselect beads (Beckman) and tagmented using homemade Tn5[51]. Resulting libraries were quality checked by Qubit and Bioanalyzer, pooled and sequenced using all lanes in an Illlumina Hiseq 4000.

Real-time-quantitative PCR. For real-time-quantitative PCR (RT-qPCR) analysis, cells of interest were sorted directly into RNA lysis buffer (Arcturus PicoPure RNA Isolation Kit, Life Technologies, Invitrogen), snap frozen and stored at −80 °C or processed directly for cDNA synthesis using SuperScript VILO cDNA synthesis kit (Invitrogen) according to the manufacturer's instructions. Depending on the sorted cell number, cDNA was further diluted 1:5–1:10 in RNase-free water and 6 µl was mixed in technical triplicates in 384-well plates with 0.5 µl of forward and reverse primer (10 µM) and 7 µl PowerUP SybrGreen Mastermix (Thermo Fisher). Program: 50 °C for 2 min, 95 °C for 10 min and 40 cycles of 95 °C for 15 s, 60 °C 1 min. Primers were designed to be intron spanning whenever possible using PrimerBlast (National Center for Biotechnology Information) and purchased from Sigma Aldrich (purification: desalting). Experiments were performed on the ViiA7 System (Applied Biosystems) and analysis of gene amplification curves was performed using the Quant StudioTM Real-Time PCR Software v.1.3 (Applied Biosystems). RNA expression was normalized to the housekeepers glyceraldeyde-3-phosphate dehydrogenase and beta actin for gene expression analysis. Relative expression levels ($2^{-\Delta Ct}$, ΔCt = (geometric mean Housekeeper Ct) − (gene of interest Ct)) of replicates were $\log_{10}$ transformed and z-scored. Primers used in this study can be found in Supplementary Table 6.

Analysis of Abseq data. Fastq files were processed via the standard Rhapsody analysis pipeline (BD Biosciences) on Seven Bridges (https://www.sevenbridges.com) according to the manufacturer's recommendations. The resulting unique molecular identifier (UMI) count matrices were imported into R (v.3.6.2) and processed with the R package Seurat (v.3.1.3 and 3.2.0)[52]. To account for differences in sequencing depth across cells, both layers were normalized independently using Seurat defaults. RNA UMI counts were log-normalized, while antibody UMI counts were centered using log ratio normalization to account for unspecific binding background signal. Subsequently, both normalized matrices were concatenated and integration across patients was performed using Scanorama[53]. The resulting corrected counts were used for visualization and clustering analysis. Nonintegrated, raw counts were used for differential expression testing.

Multiomics factor analysis integration, clustering and identification of cell type markers. Following integration, we removed genes and surface markers with variance near to zero using the caret package[54] and used MOFA to perform data integration across modalities[55]. A total of 30 multiomics factor analysis (MOFA) factors were used as a starting point, with a drop factor threshold of 0.001. The resulting MOFA dimensions were used to construct a shared nearest neighbor graph and modularity-based clustering using the Louvain algorithm was performed. Finally, UMAP visualization was calculated using 30 neighboring points for the local approximation of the manifold structure. Marker genes and surface markers for every cell type were identified by comparing the expression of each in a given cluster against the rest of the cells using the receiver operating characteristic test. To evaluate, which genes classify a cell type, cell type specific markers were selected as those with the highest classification power defined by the area under the receiver operating characteristic curve.

Processing of Smart-seq2 data. Count matrices were generated using pseudoalignment with Kallisto[56] using the GRCh38 human reference genome as implemented in the Scater package v.1.14.6 (ref. [57]). Gene level expression counts were imported into Seurat. Low-quality cells were removed on the basis of the percentage of mitochondrial RNA reads (>20%) and number of detected genes (<1,000). The remaining data were further processed using Seurat. Data was log-normalized and scaled. The top 2,000 highly variable genes were used for clustering and UMAP calculation. Cells were then annotated as described in Supplementary Note 8.

Abseq App web application. Differential expression, data visualization and gating scheme calculation can be performed in the Abseq App shiny web application (https://abseqapp.shiny.embl.de/). The application was written in R and relies on the packages shiny and aws.s3. A demonstration video of the app is included as Supplementary Video 1.

Pseudotime analysis. To reconstruct possible cell lineages from our single-cell gene expression data, data from individual samples were subset to include only the cell types from the CD34$^+$ hematopoietic stem and progenitor compartment. MOFA–UMAP embedding was then used as input for pseudotime analysis by slingshot[58]. The HSC cluster was used as a start cluster, and myelocytes, class switched memory B cells, late erythroid progenitors, megakaryocyte progenitors and conventional dendritic cell compartments as the end clusters. The genes that significantly changed through pseudotime were determined by fitting a generalized additive model (GAM) for each gene, using the TradeSeq package[59].

Modeling variance in surface marker expression. To attribute the variance in surface marker expression to biological processes, we used the variancePartition package[60] on log-transformed antibody read count data. As covariates, we used cell type annotation (for all cells except CD34$^+$ HSPCs), splines with three degrees of freedom fitted through pseudotime (for CD34$^+$ HSPCs, Pseudotime analysis), cell cycle scores (calculated using Seurat package defaults), scores for cytotoxicity and stemness (calculated using the gene lists in Supplementary Table 7 and the Seurat function AddModuleScore()), as well as technical covariates (number of genes observed, number of surface markers observed, reads on surface markers, reads on genes). To also account for variance explained by any hypothetical processes not in this predefined list, we additionally performed a factor analysis of the entire dataset (RNA plus surface markers) while accounting for the known covariates using ZiNB-WAVE[61]. We ran ZiNB-WAVE with four unknown factors on the concatenated mRNA and surface marker expression matrices while using a gene level-covariate specifying whether each row in the matrix is an mRNA or surface marker. The unknown factors explained only a very small part of the variance, and appeared to capture mostly differentiation processes not optimally explained by the pseudotime. Of note, markers with low absolute expression are more strongly subject to stochastic expression or measurement noise, while markers that are expressed by many different cell types are more strongly subject to technical effects, such as differences in single-cell library quality, likely due to the absence of true biological variability for these markers (Extended Data Fig. 5a). Other covariates are not affected by the expression level of the markers.

Projection on a reference atlas. The projection on the reference dataset is described in Supplementary Note 7. In short, we used scMAP to calculate nearest neighbors and thereby determined cell type label, MOFA–UMAP coordinates and pseudotime value.

Differential expression testing between experimental groups and estimation of interpatient variability. For comparing surface protein abundance between young and aged healthy as well as leukemic individuals, antibody tag read counts were summed at the level of cell types for each experimental batch (that is, donor). Differential expression testing was then performed for these pseudobulks using DESeq2 (ref. [62]), either separately for each cell type (Fig. 4i, Extended Data Fig. 7c and Supplementary Data 2), or jointly across all cells while accounting for cell type as a covariate (Fig. 4b). For cell-type-specific comparisons, only samples for which the respective cell type was covered with at least 20 cells were included. When comparing leukemic with healthy individuals, age and gender were used as additional covariates. Unlike single-cell specific methods, DESeq2 estimates the variance in gene expression between experimental replicates to separate signal from noise while using a negative binomial distribution that is sufficiently generic to capture the count-nature data of antibody-based pseudobulk expression values.

To estimate the degree of interpatient variability of surface marker abundance while accounting for cell state differences, we trained random forest classifiers to predict the experimental batch (that is, donor) from gene expression separately for each cell state. The feature importance score from these classifiers was then scaled from zero to one and used to estimate interpatient variability.

Changes in cell type abundance between experimental groups. To identify cell types that change in abundance between young and aged individuals (Extended Data Fig. 7a), we considered the following: first, different amounts of CD34$^+$, CD3$^+$ and total BM cells were sorted. Hence, frequencies were always computed within the respective gate (for example, for CD8$^+$ effector T cells, the frequency among CD3$^+$ T cells was computed). We then compared the following statistical models of observed cell type frequency $_{pi}$ in individual i:

$$M_0 : p_i \sim \text{Binom}\,(q) \quad \text{with } q \sim \text{Beta}\,(1, 1)$$

$$M_1 : p_i \sim \text{Binom}\left(q_{C(i)}\right) \quad \text{with } q_{C(i)} \sim \text{Beta}\,(1, 1)$$

Here $C(i)$ indicates if individual i is young or old.

Finally, we sought to distinguish between a model where cell type frequencies change as a function of age, and a model where cell type frequencies are simply highly variable between individuals, with no relationship to age:

$$M_2 : p_i \sim \text{Binom}\,(q_i) \quad \text{with } q_i \sim \text{Beta}\,(1, 1)$$

We compared the M1 and M2 models to M0 using a Bayesian strategy termed leave-one-out information criterion[63] to identify cell types with high evidence for between-group and interindividual variability, respectively.

Thresholding of surface marker expression. For every sample separately, thresholds were calculated using the normalized antibody counts to distinguish

marker-positive from marker-negative cells. For this we implemented the Otsu algorithm as described by Otsu[64].

Data-driven identification of gating schemes. To account for the CD34+ FACS enrichment of HSPCs performed in our samples, we divided the BM cells into CD34+ and CD34− subsets. For individual cell type gating scheme calculation, we compared three different methods. The first two methods are based on a decision tree using either the continuous normalized surface marker expression matrix named 'Tree continuous', or a transformed Boolean matrix ('Tree Otsu'). For the latter method, a cutoff for each antibody was calculated using the histogram-based Otsu algorithm as described above and the matrix was binarized accordingly. In both cases, the tree was determined using the package Rpart and, if needed, pruned to the maximum number of required surface markers. The third method is based on the Hypergate algorithm[34]. For this, we used the target population as the Hypergate gate vector input, calculated the predicted gating scheme and calculated the channel contributions of each surface marker using a beta of 1. Afterwards we used the contributions to optimize the predicted gating scheme to only include the maximum number of surface markers selected. Furthermore, we gated the samples using a canonical gating scheme (Expert) reported in the literature to predict gatings for some of the cell types present in the BM (Supplementary Table 5). For this, we used the Otsu threshold to split each population into marker-positive and negative populations. For each gate, the following metrics were calculated: first, the purity (Pr), that is, the proportion of target cells in the final gate and second, the recall (Rc), that is, the proportion of target cells gated from their original total population.

For the simultaneous gating calculation of all cells from the HSC and progenitor compartment, we selected the cells from BM (Young1) with a CD34 surface expression higher than 0.95. Subsequently, we downsampled cells to the same number of cells across populations. Subsequently, we calculated the decision tree with the Rpart package, using the 'continuous' approach defined above.

The NRN algorithm for integrating FACS and single-cell genomics data. To project flow cytometric measurements of surface protein abundance from CD34+ cells onto the single-cell reference, we initially subset the single-cell reference to exclude CD34− cells, and flow cytometry data was transformed using the "logicle" transform using FlowJo (v.10.7.1). Subsequently, the expression of each surface marker was normalized separately both in the flow cytometry and in the Abseq dataset using a rank-based approach. In particular, sample ranks were computed and divided by the total number of samples, that is, data was mapped to a scale from 0 to 1 where 0 indicates lowest expression within the dataset, and 1 indicates highest expression. Within this normalized gene expression space, the cosine distance between any cell from the Abseq (reference) dataset and the FACS (query) dataset was computed, the four nearest reference neighbors of every query cell were identified and the average position of these neighbors in UMAP and pseudotime space was computed using scmap[65]. Subsequently, the average Euclidean distance of the reference neighbors in MOFA space was computed to identify cells with inconsistent mapping results. These cells were later removed by applying a user-defined threshold (here, 8). In the case of the Smart-seq2 dataset, a total of 75 cells were thereby removed from the analyses.

Data visualization for a definition of boxplot elements. All plots were generated using the ggplot2 (v.3.2.1) package in R 3.6.2, GraphPad Prism (v.8 and v.9.1 for MacOS) or FlowJo (v.10.7.1, BD). Boxplots are defined as follows: the middle line corresponds to the median; the lower and upper hinges correspond to first and third quartiles, respectively; the upper whisker extends from the hinge to the largest value no further than 1.5× the interquartile range (or the distance between the first and third quartiles) from the hinge and the lower whisker extends from the hinge to the smallest value at most 1.5× the interquartile range of the hinge. Data beyond the end of the whiskers are called 'outlying' points and are plotted individually.

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

Data availability

Data is available for interactive browsing at https://abseqapp.shiny.embl.de. Datasets including raw and integrated gene expression data, cell type annotation, metadata and dimensionality reduction are available as Seurat v.3 objects through figshare: https://figshare.com/projects/Single-cell_proteo-genomic_reference_maps_of_the_human_hematopoietic_system/94469. FACS data are provided through figshare: https://figshare.com/projects/Supplementary_data_FACS_data_from_Single-cell_proteo-genomic_reference_maps_of_the_human_hematopoietic_system/122716. Fastq files are available from the European Genome-Phenome Archive under accession number EGAS00001005593. Source data are provided with this paper.

Code availability

The implementation of the NRN algorithm and vignettes describing the workflow for projecting single-cell RNA-seq data on the reference are available at https://git.embl.de/triana/nrn.

References

49. Schraivogel, D. et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
50. Erickson, J. R. et al. AbSeq protocol using the nano-well cartridge-based rhapsody platform to generate protein and transcript expression data on the single-cell level. *STAR Protoc.* **1**, 100092 (2020).
51. Hennig, B. P. et al. Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3 (Bethesda)* **8**, 79–89 (2018).
52. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
53. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
54. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).
55. Argelaguet, R. et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
56. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
57. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).
58. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
59. Van den Berge, K. et al. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
60. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinf.* **17**, 483 (2016).
61. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).
62. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
63. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).
64. Otsu, N. Threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
65. Kiselev, V. Y., Yiu, A. & Hemberg, M. Scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).

Acknowledgements

We thank V. Lopez-Salmeron, V. Ramani, E. Kowalczyk and W. Keilholz from BD Biosciences/Multiomics for providing oligo-labeled antibodies and for their support in the implementation of the Rhapsody platform. We would like to thank members of the Haas, Velten, Trumpp and Steinmetz laboratories for helpful discussions. Moreover, we thank members of the DKFZ flow cytometry and the EMBL genomics core facility for support. This work was supported financially by the Emerson foundation grant 643577 (to L.V.), grant PID2019-108082GA-I00 from the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI/FEDER, UE), the German Bundesministerium für Bildung und Forschung (BMBF) through the Juniorverbund in der Systemmedizin 'LeukoSyStem' (FKZ 01ZX1911D to L.V., S.H. and S.R.), SFB873, FOR2674 and FOR2033 funded by the Deutsche Forschungsgemeinschaft (DFG), the SyTASC consortium (Deutsche Krebshilfe), The Darwin Trust of Edinburgh (to S.T.), the ERC Consolidator Grant METACELL (773089) (to T.A.), the Dietmar Hopp Foundation (all to A.T.) and the José Carreras Foundation for Leukemia Research (grant no. DCJLS 20 R/2017 to L.V., A.T. and S.H.). L.V. acknowledges the support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and 'the CERCA Programme/Generalitat de Catalunya. D.N. is an endowed professor of the Deutsche José 641Carreras Leukämie Stiftung (DJCLS H 03/01). Contributions by D.N., J.-C.J., W.-K.H. and T.B. were supported by the Gutermuth Foundation, the H.W. & J. Hector fund, Baden-Württemberg. Figure 1a and Supplementary Note Fig. 3f were created at BioRender.com.

Author contributions

S.H., L.V. and M.P. conceived the study with help from D.H., A.T. and V.B. D.V., S.T. and M.P. performed the single-cell proteo-genomics experiments with help from D.L. and V.B. D.V. performed the experimental validations, established new experimental gating schemes and performed functional experiments with help from M.A. and P.H-M. S.T., L.J-S. and L.V. performed bioinformatics analyses with conceptional input from D.V., M.P. and S.H. S.T. developed the Abseq App. S.T. and L.V. established the NRN algorithm. S.H. supervised the experimental work with conceptional input from L.V. L.V. supervised the bioinformatics analyses with conceptional input from S.H. T.A. cosupervised S.T. M.P., D.O-R. and B.R. provided assistance in cell sorting and single-cell work-flows. S.R., R.L., T.B., J-C.J, D.N., W-K.H. and C.M-T. provided clinical samples and conceptional input on data interpretation. S.H., L.V., S.T., L.J-S. and D.V. wrote the manuscript and prepared figures. All authors have carefully read the manuscript.

Funding
Open access funding provided by Deutsches Krebsforschungszentrum (DKFZ).

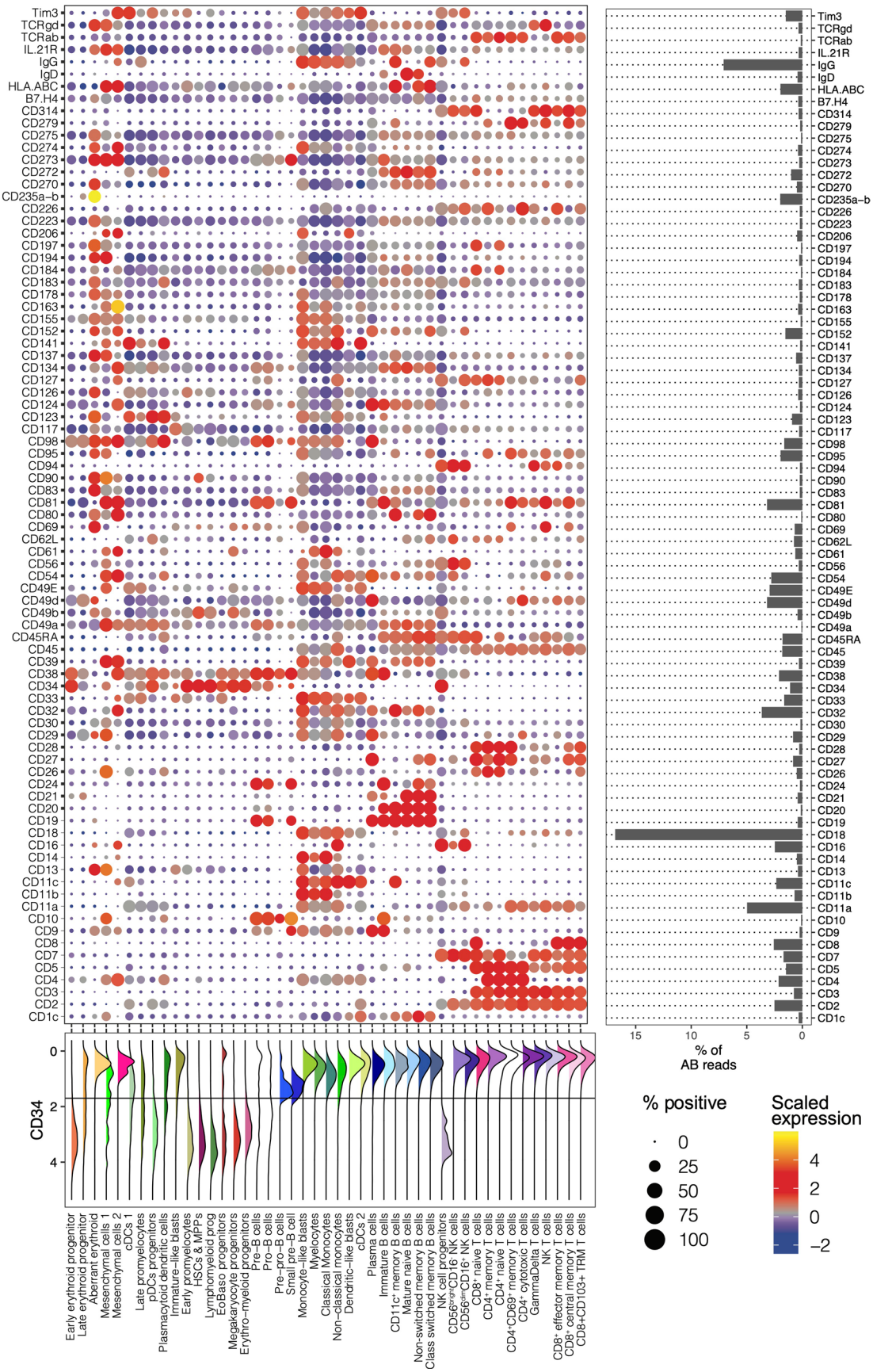
Competing interests
The oligo-coupled antibodies used in this study were a gift from BD Biosciences. The authors declare no other relevant conflicts of interest.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41590-021-01059-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41590-021-01059-0.

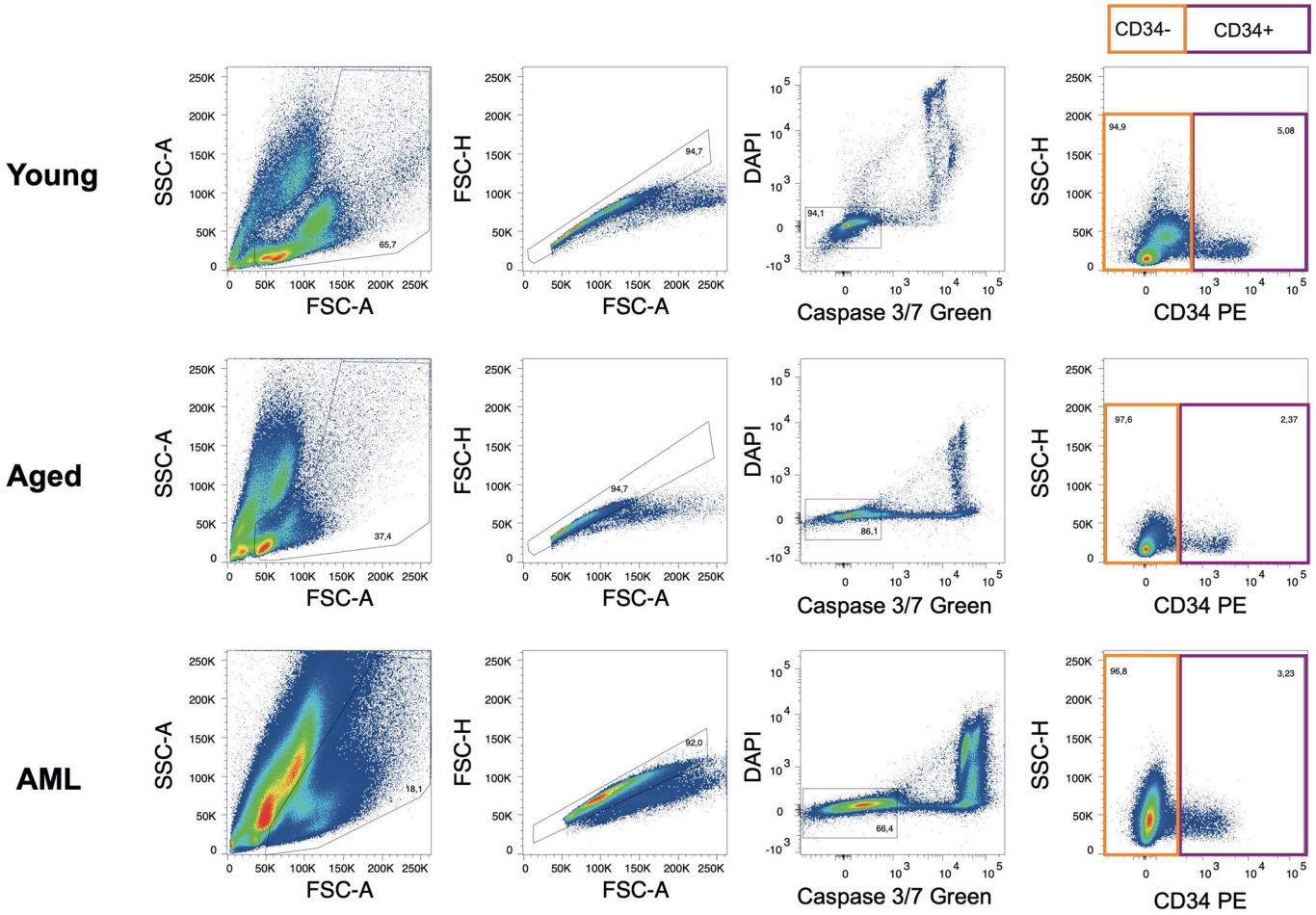
Correspondence and requests for materials should be addressed to Lars Velten or Simon Haas.

Peer review information *Nature Immunology* thanks the anonymous reviewers for their contribution to the peer review of this work. Zoltan Fehervari was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. Peer reviewer reports are available.

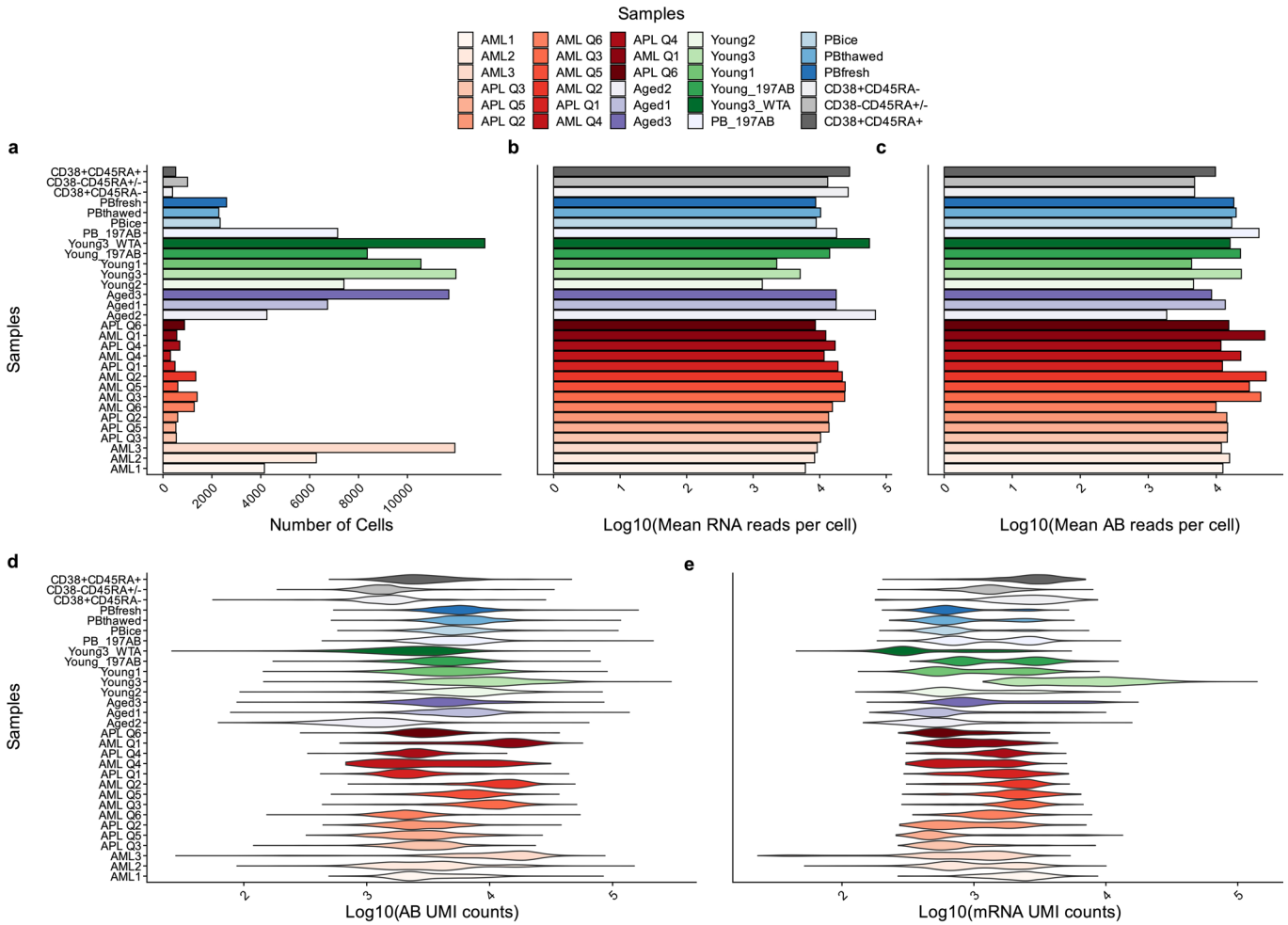


Extended Data Fig. 1 | See next page for caption.

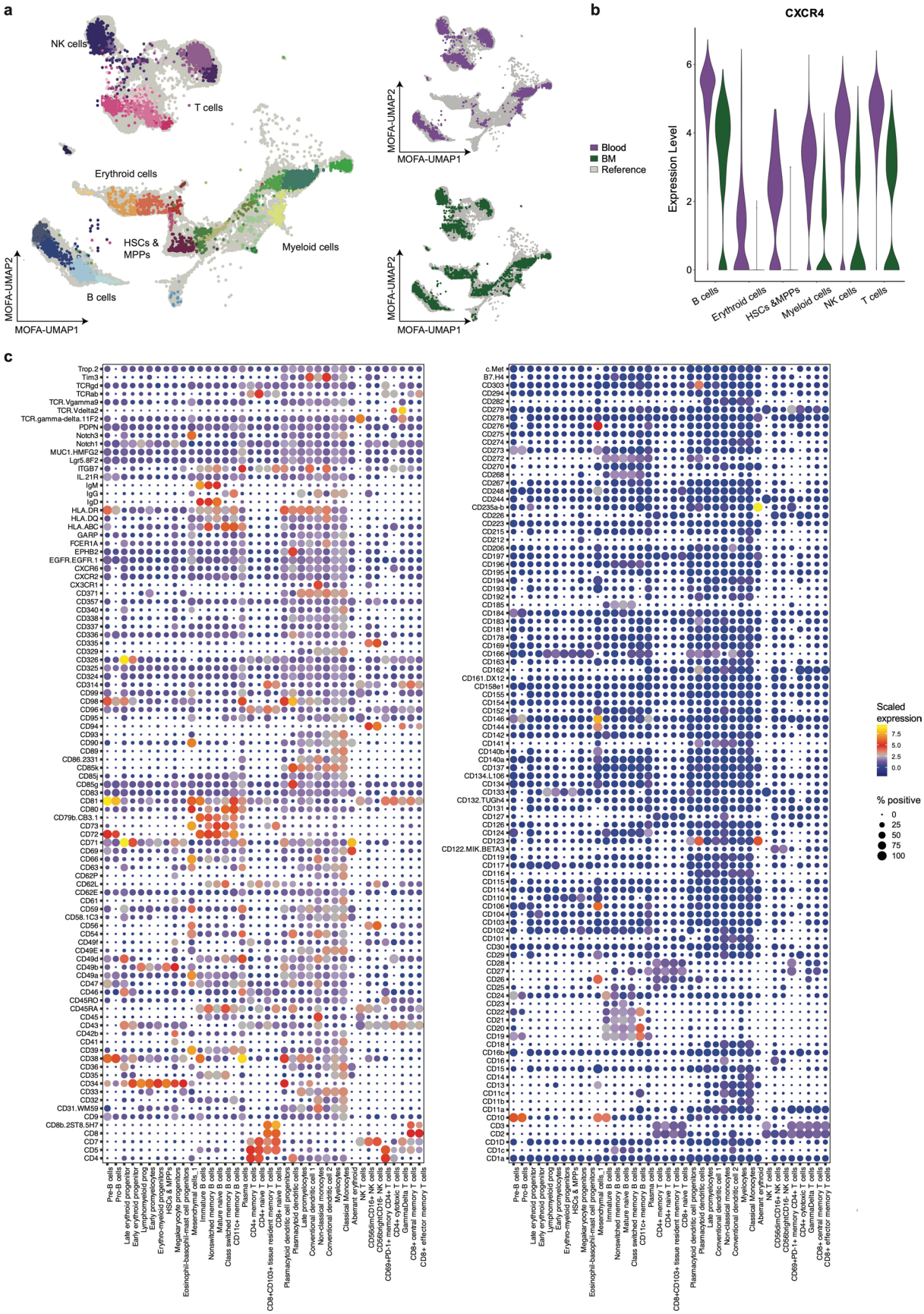
Extended Data Fig. 1 | A proteo-genomic single-cell map of 97 surface markers in human bone marrow. *Related to* Fig. 1. Dot plot depicting the expression of all surface markers by cell type. Color indicates mean normalized expression, point size indicates the fraction of cells positive for the marker. Automatic thresholding was used to identify positive cells, see Methods, section *'Thresholding of surface marker expression'* for details. The panel on the right depicts the fraction of total reads obtained for each marker as a proxy for absolute expression levels. Bottom panel illustrates the distribution of CD34 + expression across populations, similar plots can be generated for any marker using the Abseq App.



Extended Data Fig. 2 | Representative gating schemes used for the enrichment of CD34+ cells. Related to Fig. 1. For additional information on cell sorting setups, see Methods, section 'Cell sorting for Abseq'.

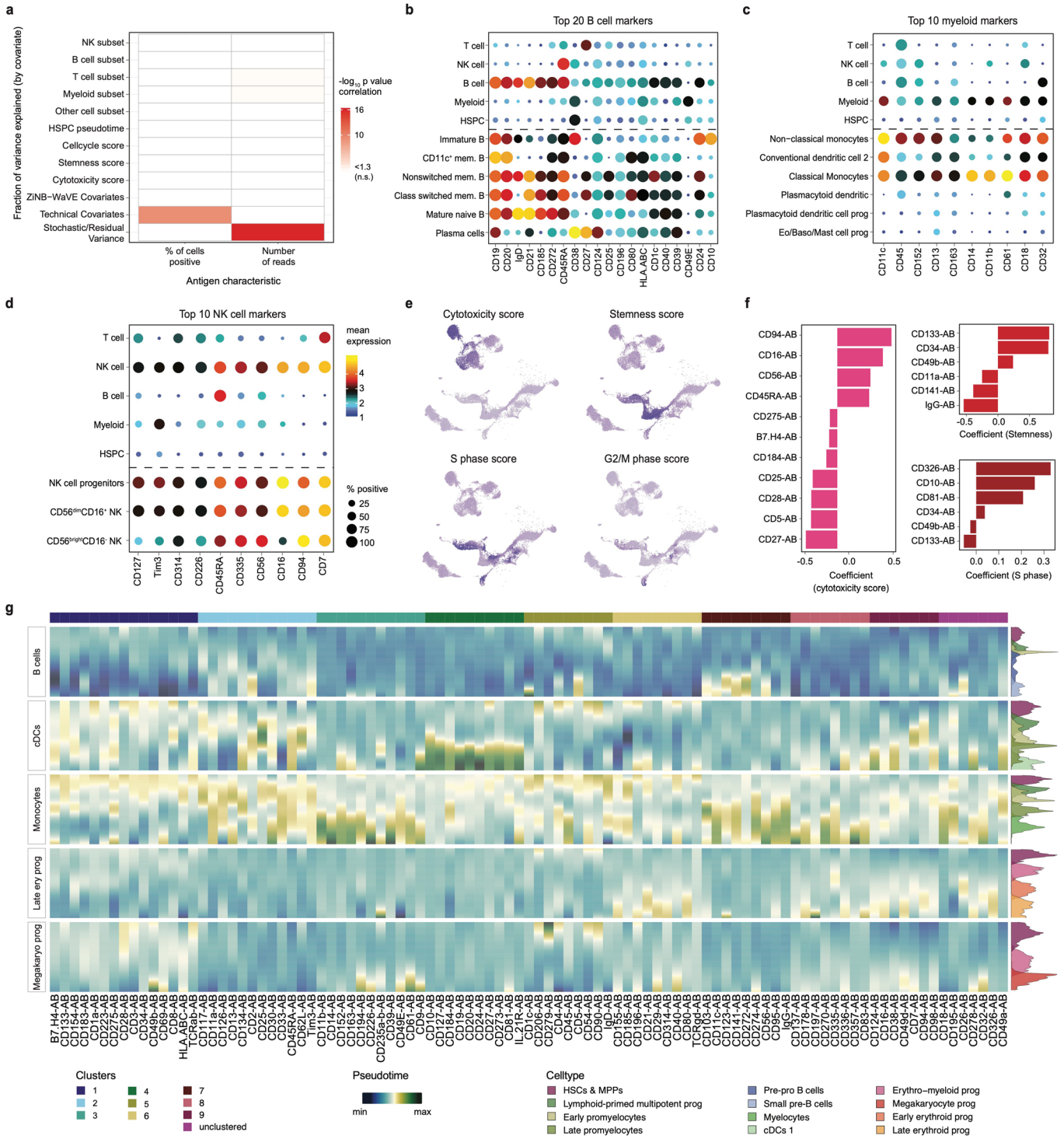


Extended Data Fig. 3 | Sequencing statistics. Related to Fig. 1. Plots depict **a**. The number of cells passing filters. Note that samples AML Q1-Q6 and APQ1–6 were multiplexed (hashed) into one experiment. **b**, **c**. The sequencing depth on the surface and mRNA level and **d**, **e**. The number of surface and mRNA molecules per cell observed. Note that targeted mRNA sequencing was performed as described in the main text.

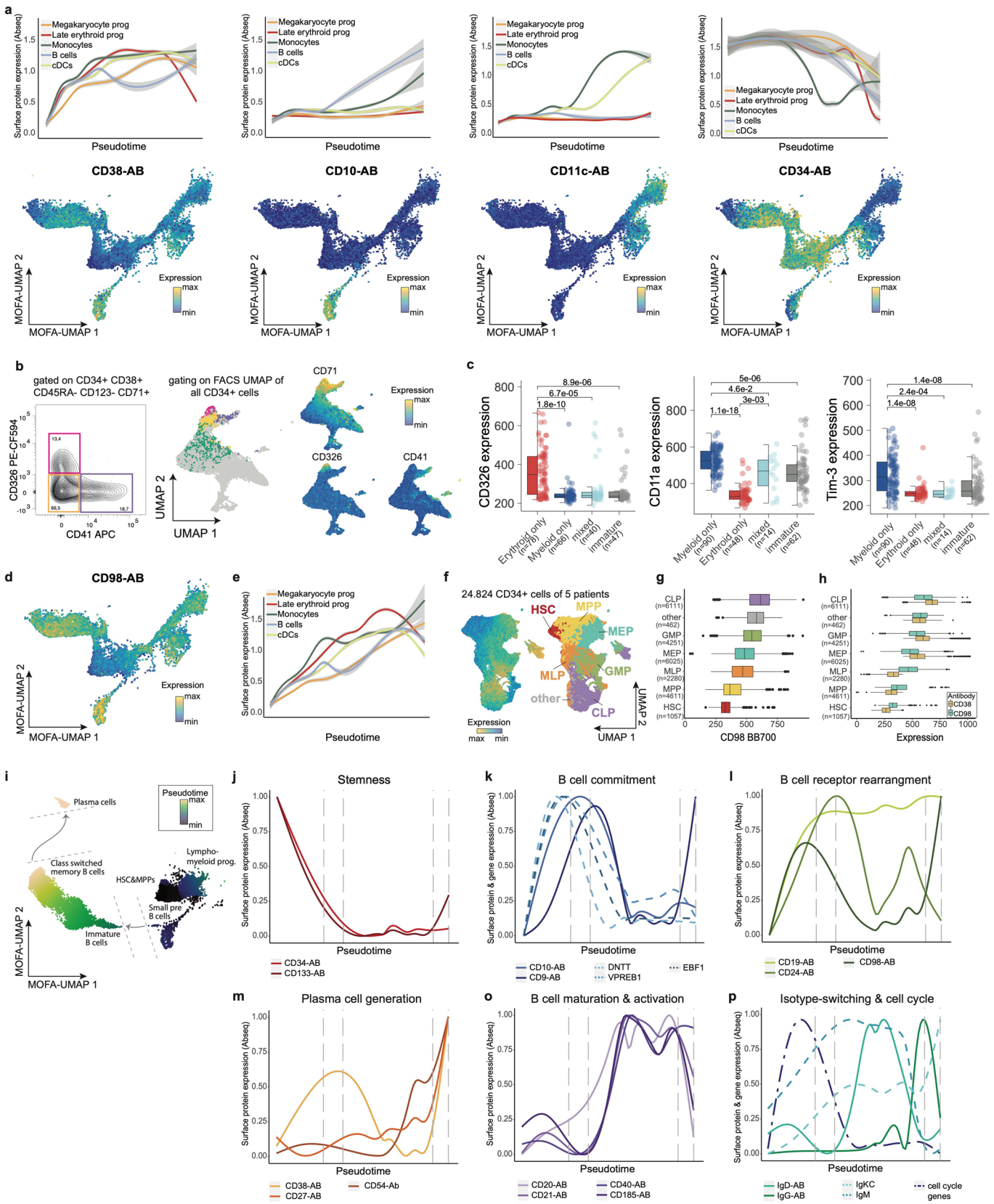


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | A single-cell proteo-genomic map of 197 surface markers in human bone marrow and blood. *Related to* Fig. 1. **a**. Left: UMAP projection on the original coordinate system from the healthy dataset (see Supplementary Note 7). Cells are colored by the mapped cell type. Right: UMAP colored by sample origin (blood and bone marrow). **b**. Violin plot depicting the expression of the bone marrow homing receptor CXCR4 on matching cell types of the blood and bone marrow. **c**. Dot plot depicting the expression of all surface markers by cell type. Color indicates mean normalized expression, point size indicates the fraction of cells positive for the marker. Automatic thresholding was used to identify positive cells, see Methods, *section 'Thresholding of surface marker expression'* for detail.

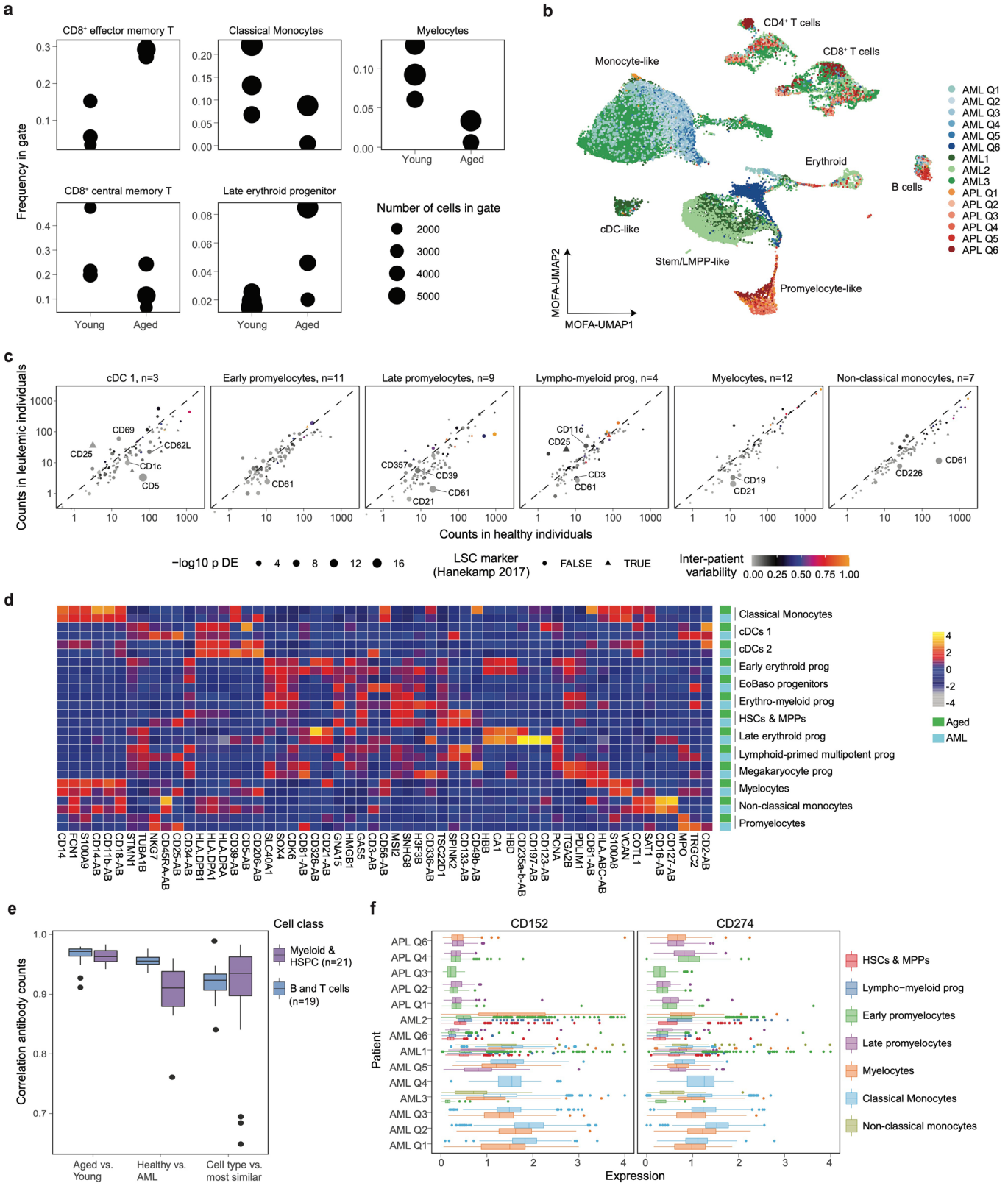


Extended Data Fig. 5 | Markers of cell types and biological processes. *Related to* Fig. 2. **a**. Heatmap investigating if the fraction of variance explained by the different covariates is correlated to antigen-level technical covariates. P values were calculated from Pearson correlation using a one-sided test based on the t-distribution. **b**-**d**. Dot plot depicting the expression of the 10–20 surface markers with the highest fraction of variance explained by B cell subtype *(b)*, myeloid subtype *(c)* and NK cell subtype *(d)*. Color indicates mean normalized expression, point size indicates the fraction of cells positive for the marker. Automatic thresholding was used to identify positive cells, see Methods, section *'Thresholding of surface marker expression'* for details. **e**. UMAPs highlighting the scores for various biological processes, as computed using the gene lists from Supplementary Table 7. **f**. Bar charts depicting the markers with the highest fraction of variance explained by cytotoxicity score (pink), stemness score (red) and S-phase score (dark red), and the corresponding model coefficients. See Supplementary Table 7 for the gene lists used for calculating these scores. **g**. Pseudotime of all 97 surface proteins for the five trajectories (B cells, cDCs, Monocytes, Late erythroid progenitor and Megakaryocyte progenitor). Markers were clustered according to their expression pattern using tradeseq (van den Berge, 2020). The density plots indicate the differentiation stages along the pseudotime.



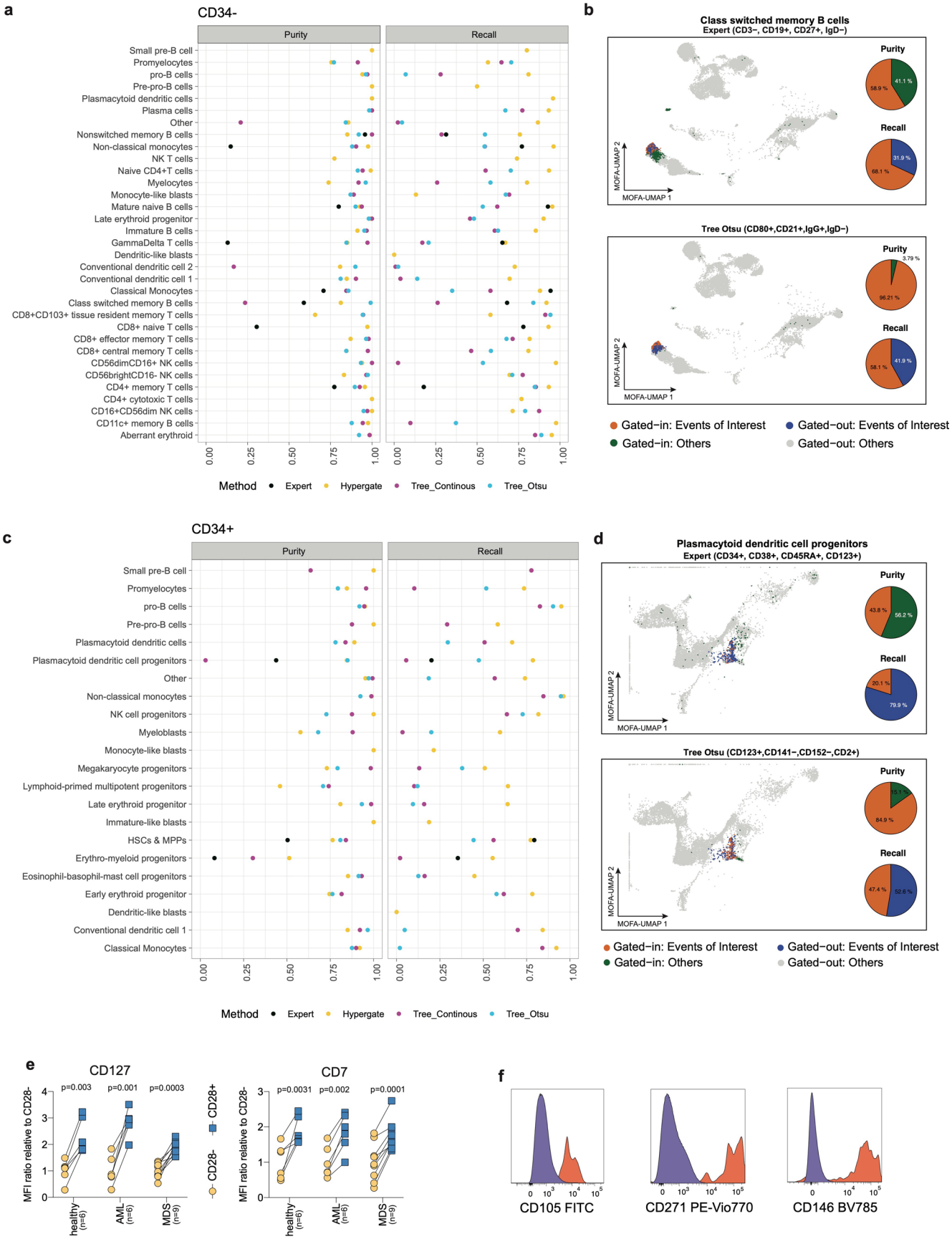
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Surface markers associated with HSC and B cell differentiation. Related to Figs. 2 and 3. See methods, section *Data visualization* for a definition of boxplot elements. **a**. Top: Line of surface protein expression smoothened over pseudotime (see Fig. 3a). Error ribbon indicates 95% confidence interval from the smoothing GAM model. Bottom: UMAP display of marker expression in CD34 + HSPCs. **b**. Left: Gating strategy for subsetting CD71 + erythroid/megakaryocytic HSPCs into CD41 + megakaryocyte and CD326 + erythroid progenitors. Right: UMAP display of flow cytometric data from CD34 + cells from a healthy donor analyzed with a 12-color FACS panel for erythroid/megakaryocytic differentiation (Supplementary Table 6). Feature plots of CD71, CD326 and CD41 expression highlight the bifurcation within CD71 + HSPCs. **c**. Culture outcome categories described in Fig. 3g were analyzed with regards to their CD326, CD11a or Tim3 surface expression. A two-sided Wilcoxon rank sum test was used for comparison of individual groups and significance levels between groups. P-values were adjusted for multiple comparisons using the Holm method. **d**, **e**. Like Fig. 3d, e, except that CD98 expression is shown. **f**. UMAP display of flow cytometric data from CD34 + cells from five healthy donors analyzed with a 12-color FACS stem and progenitor panel (Supplementary Table 6). Left: shows CD98 surface expression, right panel shows assignment of individual gates to the UMAP according, as follows: HSC: CD34 + CD38-CD45RA-CD90 + ; MPP: CD34 + CD38-CD45RA-CD90-; MLP: CD34 + CD38-CD45RA + ; MEP: CD34 + CD38 + CD10-CD45RA-; GMP: CD34 + CD38 + CD10-CD45RA + ; CLP: CD34 + CD38 + CD10 + CD45RA + . **g**. Boxplots showing CD98 expression in individual cell populations mentioned in *f*. **h**. Boxplots showing co-expression of CD98 and CD38 markers. **i**. Like Fig. 3a, UMAP depicting the pseudotime score along the B cell differentiation trajectory emanating from CD34 + HSCs & MPPs and Lymphomyeloid progenitors. **j**–**p**. Line plots depicting surface expression representative for different biological processes smoothened over the B cell pseudotime trajectory.



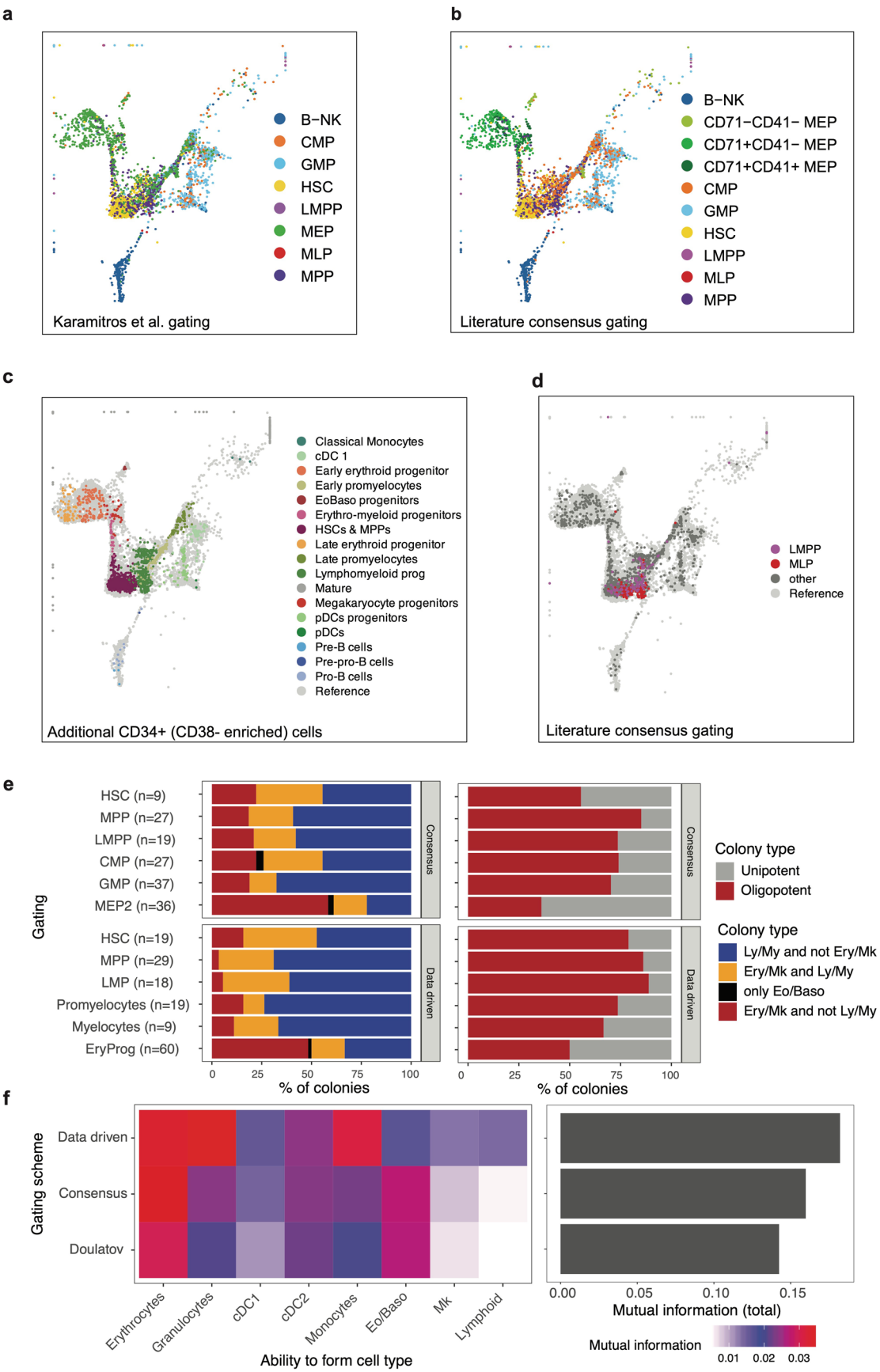
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Changes in surface protein expression and cell type abundance induced by ageing and leukemia. Related to Fig. 4. **a**. Frequency of selected cell types in young and aged individuals. Only cell types with the highest significant changes are shown, see Methods, section '*Changes in cell type abundance between experimental groups*'. **b**. UMAP display of all AML patients. Data were integrated using scanorama and MOFA (see Method *'Data analysis of Abseq data'* and *'MOFA integration, Clustering, and identification of cell type markers'*). **c**. For every myeloid cell state with sufficient representation of ≥ 20 cells in at least three patients, surface marker expression between AML (x-axis) and healthy individuals (y-axis) is compared. AML cell types were defined using a projection as in Fig. 4d, e. P-values for differential expression were computed using DESeq2 and encoded in the symbol size. Inter-patient variability is color-coded (n = number of patients included), see Methods, section '*Differential expression testing between experimental groups and estimation of inter-patient variability*' and Supplementary Data 2. **d**. Heatmap depicting cell state specific gene expression in leukemic and healthy individuals. Five most significantly overexpressed markers were identified for each cell state, using only leukemic cells. The expression of all markers selected is shown and compared to their expression in the corresponding healthy cell states. **e**. Correlations of surface marker expression are shown for matching cell types from young versus aged individuals, from healthy individuals versus AML patients, and for cell types versus the transcriptomically most similar cell type available in the dataset. See Methods, section *Data visualization for a definition of boxplot elements*. **f**. Boxplot depicting the expression of CD152 and CD274 in different cell states from different patients. Only populations covered with ≥ 50 cells per patient are included (Fig. 4h) and see source data (*Source Data* Extended Data Fig. 7) for sample size.



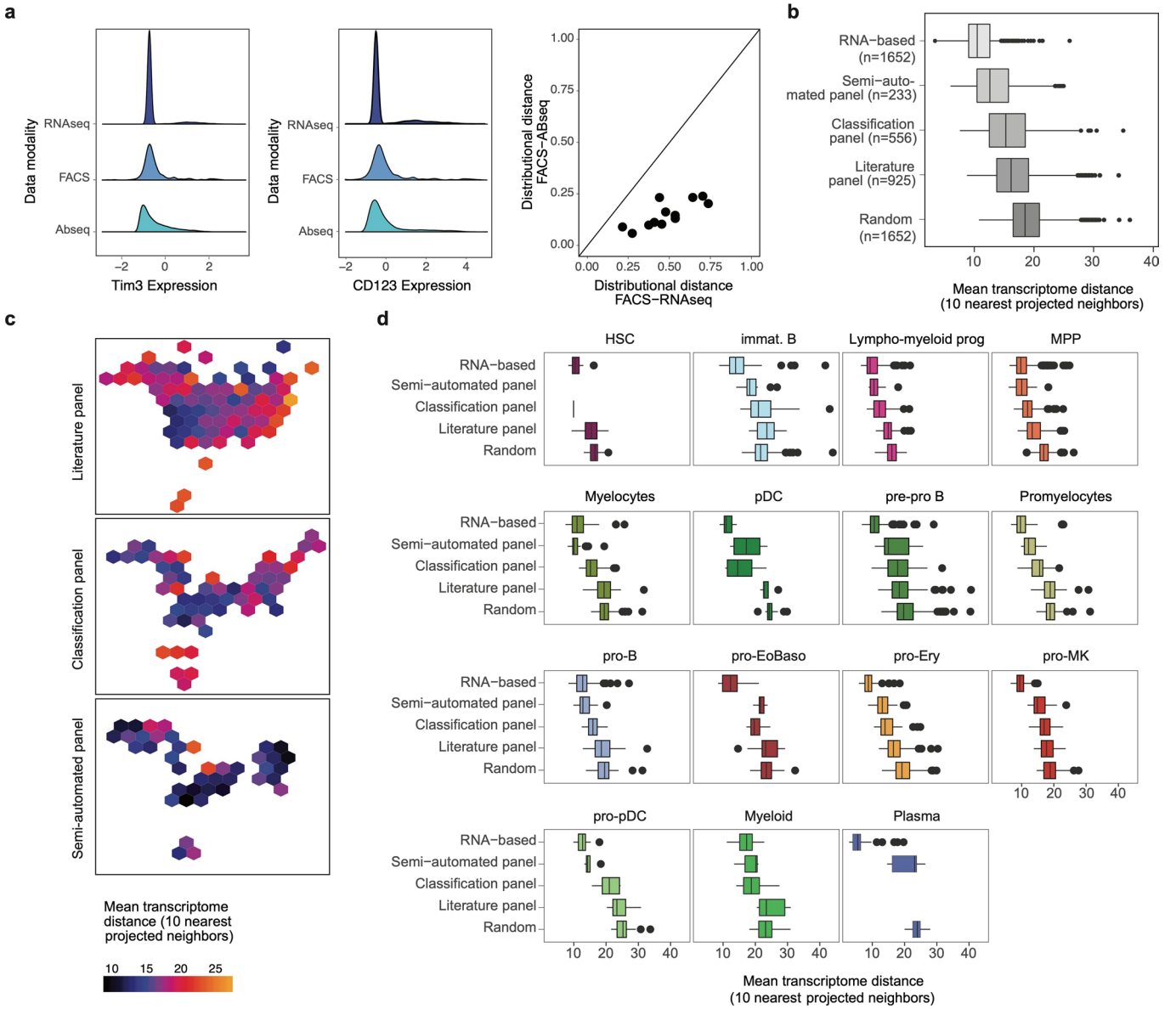
Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Comparison of data-defined and state-of-the-art (expert-defined) gating schemes. *Related to* Fig. 5. **a**. Performance of different methods for the definition of gates of CD34- populations. Gates for each cell type were defined from CD34- Abseq data as follows: Black dots correspond to gates identified from literature (Supplementary Table 5). Yellow dots correspond to gates that were set using the hypergate algorithm (Becht et al., 2019). Light blue and violet dots correspond to gates that were set using a decision tree with or without predefined thresholds, respectively. See also Methods. For each gating scheme, precision (purity) and recall were calculated. **b**. Automated and expert-defined gates of class switched memory B cells. Orange and blue dots on the UMAP correspond to class switched memory B cells located within and outside of the selected gate, respectively (that is true positives and false negatives). Green and gray dots correspond to other cells located inside and outside the gate, respectively (that is false positives and true negatives). Pie charts indicate precision and recall. Top: Shows an expert-defined state of the art gating scheme (CD3-CD19 + CD27 + IgD-). Bottom: Shows a data-defined gating scheme (CD80 + CD21 + IgG+IgD-). **c**. Like *a*, except that CD34 + populations are shown. **d**. Like *b*, except that gating schemes to define pDC progenitors are shown. **e**. Paired scatter plot depicting the mean fluorescence intensities (MFI) of CD127 and CD7 in CD4 + CD28- cytotoxic CD4 + T cells (yellow) and CD4 + CD28 + other CD4 + T cells (blue) in BM samples from healthy, AML and MDS patients. n = 6, 6 and 9 patients in the respective groups. **f**. Representative FACS histograms showing surface expression of well-known MSC surface markers. No significance = ns, $P < 0.05$ *, $P < 0.01$ **, $P < 0.001$ ***, $P < 0.0001$ ****. CD4 + CD28- and CD4 + CD28 + paired cell populations within the same BM donors from different disease entities were compared using paired two-tailed t-test. P-values were adjusted for multiple comparisons using the Bonferroni method.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Evaluation of different gating schemes. Related to Fig. 6. **a**. UMAP highlighting classification obtained from the gating scheme described by Karamitros et al., 2018, that is HSC: CD34 + CD38-CD10-CD45RA-CD90 + ; MPP: CD34 + CD38-CD10-CD45RA-CD90-; LMPP:CD34 + CD38-CD10-CD45RA + ; MLP: CD34 + CD38-CD10 + ; MEP: CD34 + CD38 + CD10-CD45RA-CD123-; CMP: CD34 + CD38 + CD10-CD45RA-CD123 + ; GMP: CD34 + CD38 + CD10-CD45RA + CD123 + ; B-NK: CD34 + CD38 + CD10 + . **b**. UMAP highlighting classification obtained from a consensus scheme combining the schemes of Doulatov et al., Karamitros et al. and Psaila et al., HSC: CD34 + CD38-CD10-CD45RA-CD90 + ; MPP:CD34 + CD38-CD10-CD45RA-CD90-; LMPP:CD34 + CD38-CD10-CD45RA + ; MLP: CD34 + CD38-CD10 + ; CD71-CD41- MEP: CD34 + CD38 + CD10-CD45RA-*FLT3-ITGA2B-TFRC*-; CD71 + CD41- MEP: CD34 + CD38 + CD10-CD45RA-*FLT3-ITGA2B-TFRC* + ; CD71 + CD41 + MEP: CD34 + CD38 + CD10-CD45RA-*FLT3-ITGA2B* + ; CMP: CD34 + CD38 + CD10-CD45RA-*FLT3* + ; GMP: CD34 + CD38 + CD10-CD45RA + ; B-NK: CD34 + CD38 + CD10 + . The marker CD135, CD41, CD71 were not part of the 97 Abseq panel. The expression of the corresponding genes, *FLT3*, *ITGA2B* and *TFRC*, were smoothened using MAGIC respectively (van Dijk et al., 2018). **c**. UMAP of additional CD34 + cells with specific enrichment of CD34 + CD38- cells, projected on the original coordinate system, colored by mapped cell types **d**. Same as *c* but colored by immunophenotypic classification obtained from a consensus scheme recapitulating the scheme of Karamitros et al. and Psaila et al. (see above). **e**. Separation of functional potential by the data-driven and the literature 'consensus gating' scheme. Single cells were sorted according to the two gating schemes and cultured for 19 days. Colonies were scored as Ery/Mk if they contained at least 5 erythroid or megakaryocytic cells, and as Ly/My if they contained at least 5 cells of types Neutrophil, cDC, Monocyte, or B/NK. Unipotent: Only one of these cell types was formed with at least 5 cells; oligopotent: At least two of these cell types were formed. Only gates for which at least 9 colonies were observed are shown. **f**. Mutual information (in nats) between the gate identity and the ability to form any of the cell types, or the total mutual information across all cell types.



Extended Data Fig. 10 | Projection and classification of cytometry data using a single-cell proteo-genomic reference. Related to Fig. 7. **a**. Distribution of normalized, scaled expression values of Tim3 (left panel) and CD123 (central panel) measured by scRNA-seq, Abseq, and FACS. Right panel: Scatter plot depicts the dissimilarity between the distribution of expression values measured by FACS, and the distribution measured by scRNA-seq (x-axis) or Abseq (y-axis) as quantified using Kolmogorov-Smirnov distance. Data for all markers included in the panel from main Fig. 6f is shown. **b**–**d**. Comparison of data integration strategies. Smart-seq2 data and Abseq data were integrated with five different strategies. RNA-based: Integration by Seurat v3, based on gene expression (transcriptome). Random: Random selection of ten nearest neighbors. Others: Surface marker-based integration using NRN, using defined sets of surface markers (Classification panel, Semi-automated panel: see Supplementary Table 6. Literature panel: CD34, CD38, CD45RA, CD90, CD10, CD135/*FLT3*, CD49f). For every cell projected on the UMAP, the ten nearest neighbors in projected UMAP space were identified. Subsequently, the mean Euclidean distance between their location in a gene expression-based PCA space (Smart-seq2) was computed. Sample size n = 1652. **b**. Boxplot summarizing the distance across data integration strategies. See figure for sample size. See Methods, section '*Data visualization* for a definition of boxplot elements'. **c**. Hexagonal plot summarizing the projection accuracy for different regions of the UMAP. **d**. Boxplots stratified by cell type demonstrate that projection using the semiautomated panel performs close to an RNA-based integration in most cases. See panel b for sample size.

nature research

Corresponding author(s): Lars Velten
Simon Haas

Last updated by author(s): Sep 13, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Raw data was processed using the standard Rhapsody analysis pipeline (BD Biosciences) on Seven Bridges (https://www.sevenbridges.com) according to the manufacturer's recommendations. Described in detail in the Methods, section "Seven Bridges processing for Abseq data". Smart-seq2 data was processed using Kallisto, see Methods section, "Processing of Smart-seq2 data" For the acquisition of flow cytometry data, BD FACSDiva was used. For the acqusition of qPCR data, a ViiA7 System (Applied Biosystems) was used.
Data analysis	Data were analyzed using R v.3.6.2 , Seurat v. 3.1.3 and v.3.2.0, MOFA v.1.3.1, caret v. 6.0-84, zinbwave v.1.8.0, scater 1.14.6, slingshot v.1.4.0, variancePartition v1.16.1, DESeq2 v1.26.0, randomForest_4.6-14, rpart_4.1-15, BIOMOD 1.1-7.04, Hypergate 0.8.3, scmap 1.8.0, Tradeseq 1.6.0 , ggplot2 v3.2.1 and rstatix 0.7.0.999. Python package scanorama v.16. A full description of data analysis is contained in the Methods, all sections starting with "Data analysis of Abseq data". Custom code is available at https://git.embl.de/triana/nrn A custom web app was built additionally using packages shiny 1.6.0 and aws.S3 0.3.21 and was also used for data analysis (https://abseqapp.shiny.embl.de/) Flow cytometry data was analyzed using FlowJo v.10.7.1. ViiA7 System and corresponding software (v1.6.1) was used to analyze RT-qPCR data. In some cases, GraphPad Prism v8 and v9.1 was used for graphical representations and statistics. Microsoft Excel for mac v 16.16.27 or R Studio Server 1.2.5033 were used for data analysis in some cases.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is available for interactive browsing at https://abseqapp.shiny.embl.de . Datasets including raw and integrated gene expression data, cell type annotation, metadata and dimensionality reduction are available as Seurat v3 objects through figshare: https://figshare.com/projects/Single-cell_proteo-genomic_reference_maps_of_the_human_hematopoietic_system/94469
Relevant flow cytometry and cell sorting FCS files are available through figshare:https://figshare.com/projects/Supplementary_data_FACS_data_from_Single-cell_proteo-genomic_reference_maps_of_the_human_hematopoietic_system/122716. Raw data from analyzed single cell index cultures are available upon request. Fastq files are available under accession number EGAS00001005593

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to pre-determine sample size, but an approximate calculation based on experience suggested that typically, 50 cells are enough to robustly define a cell type and hence 60,000 CD34- and 11,000 CD34+ cells in the main reference dataset would be sufficient to identify populations present at <0.1% of total or <0.5% of CD34+ BM, respectively. In practice, obtained sample size was also affected by the fraction of high quality cells that passed quality metrics and in the case of CD34+ cells the availability of biological material. In practice, the smallest population we identified in the main dataset was the Mesenchymal cell 2 population, which was covered with 11 cells in the final dataset (0.02% of total BM); the smallest CD34+ hematopoietic population were the putative NK cell progenitors, covered with 30 cells (0.2% of CD34+). Hence in practice given our dataset size (n=70017 cells), very small populations can robustly be identified.
Data exclusions	a) Abseq: For targeted scRNA-seq no filtering was performed for the WTA sample cells with < 500 detected genes and >30% mitochondrial counts were discarded. Such quality control steps are customary in the field and were set after inspection of the data, and not pre-established. b) indexed Smart-seq2: All cells with mitochondrial reads > 20% were excluded. Second, we limited the acceptable numbers of detected genes. Cells with < 1000 detected genes were discarded. Such quality control steps are customary in the field and were pre-established. c) single-cell index culture readouts: generally, only wells that contained more than 10 CD45+ CD235- or CD45+ CD235+ or CD45+ CD235- cells were considered during analysis.
Replication	3 independent bone marrow donors were assayed per experimental group (healthy young, healthy aged, AML). All findings on healthy young and healthy old bone marrow donors were successfully replicated. 12 more individual AML/APL patients were used to make in depth statements about disease states. Sample metadata can be found in Supplementary Table 3. For RT-qPCR sorts, cells from at least three individual human BM donors were sorted individually and gene expression was analyzed in technical triplicates. RT-qPCR experiments were only performed once, due to sample availability. For FACS based index sorting, both index single-cell cultures and index RNA-seq was performed with cells from the same donor. A second single-cell index culture experiment (Figure S9 e,f) succesfully replicated data shown in Figure 3 (Figure 3 f,j,n).
Randomization	Not relevant (no treatment groups)
Blinding	Blinding during single cell data generation was not performed as the same standard Abseq protocol and library preparation protocol was followed for each of the bone marrow samples. FACS analysis required grouping of data and sample identities could not be blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

For a list of oligo coupled antibodies, see Supplementary Table S2

FACS antibodies (Epitope+Fluorochrome, Vendor + Catalog number, RRID, clone, dilution):

ABseq cell sorting
Anti-CD34 PE BD Biosciences Cat# 555822, RRID:AB_396151 ,1:30
DAPI BD Biosciences Cat# 564907, RRID:AB_2869624 ,1:1000
Caspase 3/7 probe Thermo Fisher Scientific Cat# C10423 ,1:500
Anti-CD3 PE-Cy7 BD Biosciences Cat# 560910, RRID:AB_10563409 ,SK7 ,1:50
Anti-CD45RA BB515 BD Biosciences Cat# 564552, RRID:AB_2738841 ,HI100 ,1:50
Anti-CD38 APC-R700 BD Biosciences Cat# 564979, RRID:AB_2744373 ,HIT2 ,1:200
Anti-CD10 BV421 BioLegend Cat# 312218, RRID:AB_2561833 ,HI10a ,1:50
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000

Classification panel
Anti-CD4 BV421 BD Biosciences Cat# 565997, RRID:AB_2739448 ,SK3 ,1:20
Anti-CD11a BUV395 BD Biosciences Cat# 745986, RRID:AB_2743392 ,HI111 ,1:100
Anti-CD33 BV785 BD Biosciences Cat# 740974, RRID:AB_2740599 ,WM53 ,1:50
Anti-CD123 BV650 BD Biosciences Cat# 563405, RRID:AB_2738185 ,7G3 ,1:50
Anti-CD19 BB700 BD Biosciences Cat# 566396, RRID:AB_2744310 ,SJ25C1 ,1:50
Anti-CD61 BB515 BD Biosciences Cat# 565123, RRID:AB_2739075 ,VI-PL2 ,1:50
Anti-CD10 APC BioLegend Cat# 312209, RRID:AB_314920 ,HI10a ,1:50
Anti-CD38 APC-R700 BD Biosciences Cat# 564979, RRID:AB_2744373 ,HIT2 ,1:200
Anti-CD34 APC-Cy7 BioLegend Cat# 343514, RRID:AB_1877168 ,581 ,1:30
Anti-CD133 PE BD Biosciences Cat# 566593, RRID:AB_2744281 ,W6B3C1 ,1:30
Anti-CD11c PE-Cy7 BD Biosciences Cat# 561356, RRID:AB_10611859 ,B-ly6 ,1:50
Anti-Tim-3 PE-CF594 BD Biosciences Cat# 565560, RRID:AB_2744371 ,7D3 ,1:30
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:500

Semi-automated panel
Anti-CD49b BV421 BD Biosciences Cat# 743201, RRID:AB_2871492 ,12F1 ,1:50
Anti-CD326 PE-CF594 BD Biosciences Cat# 565399, RRID:AB_2739219 ,EBA-1 ,1:50
Anti-CD71 APC BioLegend Cat# 334107, RRID:AB_10916388 ,CY1G4 ,1:200
Anti- FcER1A PE-Cy7 Thermo Fisher Scientific Cat# 25-5899-42, RRID:AB_2573495 ,AER-37 ,1:50
Anti-CD11a BUV395 BD Biosciences Cat# 745986, RRID:AB_2743392 ,HI111 ,1:100
Anti-CD33 BV785 BD Biosciences Cat# 740974, RRID:AB_2740599 ,WM53 ,1:50
Anti-CD123 BV650 BD Biosciences Cat# 563405, RRID:AB_2738185 ,7G3 ,1:50
Anti-CD61 BB515 BD Biosciences Cat# 565123, RRID:AB_2739075 ,VI-PL2 ,1:50
Anti-CD38 APC-R700 BD Biosciences Cat# 564979, RRID:AB_2744373 ,HIT2 ,1:200
Anti-CD34 APC-Cy7 BioLegend Cat# 343514, RRID:AB_1877168 ,581 ,1:30
Anti-CD133 PE BD Biosciences Cat# 566593, RRID:AB_2744281 ,W6B3C1 ,1:30
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000

Stem and progenitor panel
Anti-CD4 BV421 BD Biosciences Cat# 565997, RRID:AB_2739448 ,SK3 ,1:20
Anti-CD45RA BB515 BD Biosciences Cat# 564552, RRID:AB_2738841 ,HI100 ,1:50
Anti-CD98 BB700 BD Biosciences Cat# 746147, RRID:AB_2743507 ,UM7F8 ,1:50
Anti-CD90 PE BD Biosciences Cat# 555596, RRID:AB_395970 ,5E10 ,1:30
Anti-CD49f PE-Cy7 Thermo Fisher Scientific Cat# 12-0495-82, RRID:AB_891474 ,eBioGoH3 ,1:50
Anti-CD38 APC-R700 BD Biosciences Cat# 564979, RRID:AB_2744373 ,HIT2 ,1:200
Anti-CD34 APC-Cy7 BioLegend Cat# 343514, RRID:AB_1877168 ,581 ,1:30
Anti-CD133 BV650 BD Biosciences Cat# 747642, RRID:AB_2744206 ,W6B3C1 ,1:30
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000
Anti-CD11a BUV395 BD Biosciences Cat# 745986, RRID:AB_2743392 ,HI111 ,1:100
Anti-Tim-3 PE-CF594 BD Biosciences Cat# 565560, RRID:AB_2744371 ,7D3 ,1:30
Anti-CD10 APC BioLegend Cat# 312209, RRID:AB_314920 ,HI10a ,1:50

sc-index culture readouts

Anti-CD303 BUV395 BD Biosciences Cat# 747999, RRID:AB_2872460 ,V24-785 ,1:200
Anti-CD14 BUV805 BD Biosciences Cat# 612902, RRID:AB_2870189 ,M5E2 ,1:150
Anti-CD141 BV421 BioLegend Cat# 344114, RRID:AB_2563858 ,M80 ,1:200
Anti-CD19 BV785 BioLegend Cat# 302240, RRID:AB_2563442 ,HIB19 ,1:200
Anti-CD56 BV785 BD Biosciences Cat# 564058, RRID:AB_2738569 ,NCAM16.2 ,1:200
Anti-CD41a FITC Thermo Fisher Scientific Cat# 11-0419-42, RRID:AB_10718234 ,HIP8 ,1:200
Anti-CD66b PerCP-Cy5.5 BioLegend Cat# 305107, RRID:AB_2077856 ,G10F5 ,1:200
Anti-CD370 PE BioLegend Cat# 353804, RRID:AB_10965546 ,8F9 ,1:200
Anti-CD1c PE-Dazzle 594 BioLegend Cat# 331532, RRID:AB_2565293 ,L161 ,1:200
Anti-CD235a APC Thermo Fisher Scientific Cat# 17-9987-42, RRID:AB_2043823 ,HIR2 ,1:200
Anti CD45 APC-R700 BD Biosciences Cat# 566041, RRID:AB_2744399 ,HI30 ,1:200
Anti-CD123 BV650 BD Biosciences Cat# 563405, RRID:AB_2738185 ,7G3 ,1:200
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000
Anti-CD34 APC-Cy7 BioLegend Cat# 343514, RRID:AB_1877168 ,581 ,1:200
Anti- FcER1A PE-Cy7 Thermo Fisher Scientific Cat# 25-5899-42, RRID:AB_2573495 ,AER-37 ,1:200
Anti- CD11b BUV615 BD Biosciences Cat# 751140, RRID:AB_2875166 ,M1/70 ,1:200
Anti- CD1c APC-Cy7 BioLegend Cat# 331520, RRID:AB_10644008 ,L161 ,1:200
Anti- CD56 PE-Dazzle BioLegend Cat# 362544, RRID:AB_2565922 ,5.1H11 ,1:200

Cytotoxic CD4+ T cell analysis
Anti-CD3 BUV395 BD Biosciences Cat# 563546, RRID:AB_2744387 ,UCHT1 ,1:50
Anti-CD25 BUV737 D Biosciences Cat# 564385, RRID:AB_2744342 ,2A3 ,1:50
Anti-CD4 BUV805 BD Biosciences Cat# 612887, RRID:AB_2870176 ,SK3 ,1:50
Anti-CD197 Pacific Blue BioLegend Cat# 353210, RRID:AB_10918984 ,G043H7 ,1:50
Anti-CD7 BV711 BD Biosciences Cat# 564018, RRID:AB_2738544 ,M-T701 ,1:50
Anti-CD45RO FITC BioLegend Cat# 304242, RRID:AB_2564159 ,UCHL1 ,1:50
Anti-CD28 PE-CF594 BD Biosciences Cat# 562323, RRID:AB_11153681 ,CD28.2 ,1:50
Anti-CD127 PE-Cy7 Thermo Fisher Scientific Cat# 25-1278-42, RRID:AB_1659672 ,eBioRDR5 ,1:50
Anti-CD45RA APC Thermo Fisher Scientific Cat# 17-0458-41, RRID:AB_1944379 ,HI100 ,1:50
Anti-CD45 Alexa Fluor 700 BioLegend Cat# 304024, RRID:AB_493761 ,HI30 ,1:50
Anti-CD69 APC-Cy7 BD Biosciences Cat# 560912, RRID:AB_10563414 ,FN50 ,1:100
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000

MSC analysis
Anti-CD10 BV421 BioLegend Cat# 312209, RRID:AB_314920 ,HI10a ,1:50
Anti-CD146 BV785 BD Biosciences Cat# 743303, RRID:AB_274141 ,P1H12 ,1:50
Anti-CD105 FITC BioLegend Cat# 323204, RRID:AB_755956 ,43A3 ,1:50
Anti-CD31 BB700 BD Biosciences Cat# 566563, RRID:AB_2744362 ,WM59 ,1:100
Anti-CD49a PE BD Biosciences Cat# 559596, RRID:AB_397288 ,SR84 ,1:20
Anti-CD13 PE-Dazzle 594 BioLegend Cat# 301719, RRID:AB_2616763 ,WM15 ,1:50
Anti-CD271 PEVio770 Miltenyi Biotec Cat# 130-113-422, RRID:AB_2733220 ,ME20.4-1.H4 ,1:100
Anti-CD26 APC BD Biosciences Cat# 563670, RRID:AB_2738363 ,M-A261 ,1:50
Anti-CD45 APC R700 BD Biosciences Cat# 566041, RRID:AB_2744399 ,HI30 ,1:100
Anti-CD90 APC-Cy7 BioLegend Cat# 328132, RRID:AB_2566341 ,5E10 ,1:100
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:50
Anti-CD11a BUV395 BD Biosciences Cat# 745986, RRID:AB_2743392 ,HI111 ,1:100

Erythroid/Megakaryocyte differentiation panel
Anti-CD38 BUV563 BD Biosciences Cat# 741446, RRID:AB_2870920 ,HB7 ,1:200
Anti-CD71 BUV805 BD Biosciences Cat# 749294, RRID:AB_2873669 ,M-A712 ,1:200
Anti-CD49b BV421 BD Biosciences Cat# 743201, RRID:AB_2871492 ,12F1 ,1:50
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000
Anti-CD44 BV650 BD Biosciences Cat# 743665, RRID:AB_2871540 ,L178 ,1:200
Anti-CD49d BV711 BD Biosciences Cat# 563177, RRID:AB_2738049 ,9F10 ,1:50
Anti-CD45RA BB515 BD Biosciences Cat# 564552, RRID:AB_2738841 ,HI100 ,1:100
Anti-CD90 PE BD Biosciences Cat# 555596, RRID:AB_395970 ,5E10 ,1:30
Anti-CD326 PE-CF594 BD Biosciences Cat# 565399, RRID:AB_2739219 ,EBA-1 ,1:50
Anti-CD123 PE-Cy7 Thermo Fisher Scientific Cat# 25-1239-42, RRID:AB_1257136 ,6H6 ,1:50
Anti-CD41 APC Thermo Fisher Scientific Cat# 17-0419-42, RRID:AB_2573144 ,HIP8 ,1:50
Anti-CD34 APC-Cy7 BioLegend Cat# 343514, RRID:AB_1877168 ,581 ,1:50

Consensus panel
Anti-CD11a BUV395 BD Biosciences Cat# 745986, RRID:AB_2743392 ,HI111 ,1:50
Anti-CD71 BUV805 BD Biosciences Cat# 749294, RRID:AB_2873669 ,M-A712 ,1:400
Anti-CD45RA BV421 BD Biosciences Cat# 562885, RRID:AB_2737864 ,HI100 ,1:50
Fixable viability dye efluor506 Thermo Fisher Scientific Cat# 65-0866-14 ,1:1000
Anti-CD44 BV650 BD Biosciences Cat# 743665, RRID:AB_2871540 ,L178 ,1:300
Anti-CD135 BV711 BD Biosciences Cat# 563908, RRID:AB_2738479 ,4G8 ,1:20
Anti-Tim3 FITC BioLegend Cat# 345021, RRID:AB_2563936 ,F38-2E2 ,1:30
Anti-CD90 PE BD Biosciences Cat# 555596, RRID:AB_395970 ,5E10 ,1:30
Anti-CD326 PE-CF594 BD Biosciences Cat# 565399, RRID:AB_2739219 ,EBA-1 ,1:50
Anti-CD41 PE-Cy5 BioLegend Cat# 303708, RRID:AB_314378 ,HIP8 ,1:50
Anti-CD123 PE-Cy7 Thermo Fisher Scientific Cat# 25-1239-42, RRID:AB_1257136 ,6H6 ,1:30
Anti-CD10 APC BioLegend Cat# 312209, RRID:AB_314920 ,HI10a ,1:30
Anti-CD38 APC-R700 BD Biosciences Cat# 564979, RRID:AB_2744373 ,HIT2 ,1:150
Anti-CD34 APC-Cy7 BioLegend Cat# 343514, RRID:AB_1877168 ,581 ,1:50

Validation

All antibodies used are commercially available, broadly established and validated by the respecitve manufacturers as indicated on the websites (See RRIDS above for respective websites for each antibody). In addition, used antibodies are used routinely in our laborarotory with reproducible results.

Human research participants

Policy information about studies involving human research participants

Population characteristics

The age of young healthy donors ranged between 25 and 34. Aged healthy donors ranged between 59 and 69 years. AML patients ranged between 44 and 78 years. APL patients ranged between 32 and 70 years. Both female and male BM donors were included in every group.
Healthy young and aged patients showed normal BM cellularity and were devoid of any known mutations.
AML/APL patient samples used in this study were taken at initial diagnosis. AML Patients had normal karyotypes. Following list states diseased patient metadata in detail:

AML1	FLT3-ITD,NPM1-mut
AML2	FLT3-wt,NPM1-mut
AML3	FLT3-ITD,NPM1-mut
AML Q4	FLT3-ITD,NPM1-mut
AML Q1	FLT3-ITD,NPM1-mut
AML Q3	FLT3-wt,NPM1-mut
AML Q6	FLT3-wt,NPM1-mut
AML Q2	FLT3-wt,NPM1-mut
AML Q5	FLT3-wt,NPM1-mut
APL Q5	APL t(15;17)
APL Q3	APL t(15;17)
APL Q6	APL t(15;17)
APL Q4	APL t(15;17)
APL Q2	APL t(15;17)
APL Q1	APL t(15;17)

Recruitment

Young healthy donors and aged healthy donor BM samples were obtained from patients that are without any clinical signs of disease. AML patient samples were obtained at initial diagnosis and had normal karyotypes. Rare samples of healthy young and aged donors were included with respect to sample availability.

Ethics oversight

Bone marrow (BM) samples from healthy and diseased donors were obtained at the University clinics in Heidelberg and Mannheim after informed written consent using ethic application numbers S480/2011 and S-693/2018. All experiments involving human samples were approved by the ethics committee of the University Hospital Heidelberg and was in accordance with the Declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

After obtaining informed written consent, healthy or diseased bone marrow samples were collected from iliac crest bone marrow aspirations. Peripheral blood was obtained from venipuncture. Mononuclear cells from bone marrow and blood were isolated by density gradient centrifugation, frozen and stored in liquid nitrogen until further use.
For sample preparation, samples were thawed in a water bath at 37°C and transferred dropwise into RPMI-1640 10% FCS. Cells were centrifuged for 5 min at 350g and washed once with RPMI-1640 10% FCS. Cells were then resuspended in FACS buffer (FB, PBS 5% FCS 0.5 mM EDTA) containing fluorochrome conjugated antibodies, dead cell exclusion dye and Fc-receptor blocking solution. Cell suspensions were incubated for 15 min at 4°C in the dark. Cell suspensions were then washed with FB and resuspended in 0,2- 1 ml FB and were inserted into the respective analyzer or cell sorter.

Instrument

All flow cytometric analyses were performed using BD Fortessa or LSRII flow cytometers. Cell sorting was done using BD Aria Aria II and BD Aria Fusion sorters equipped with 100 µm or 130 µm nozzle and sorting was performed in 4-way purity or single cell purity modes.

Software	BD FACSDiva and FlowJo v10.7.1 were used throughout the study. In some cases, logicle transformed FCS data using a built-in FlowJo function was exported and plotted in R using the ggplot2 v3.2.1. package.
Cell population abundance	Purity in post sort fractions was not directly determined. Post-sort, single-cell RNA seq and RT-qPCRs were performed, which gave detailed insights into the biology of sorted cell populations.
Gating strategy	FSC-SSC gates were set so that FSC low and SSC high cells were excluded, following by singlet gating using FSC-A vs. FSC-H. After doublet exlcusion, dead cells, which are efluor506 high were excluded according to the manufacters's instructions and the indicated gating strategies (see respective figures that show flow cytometry data, i.e Fig. 5 and Fig.6) were followed. During Abseq cell sorts, a combination of Caspase 3/7 and DAPI was used for dead cell exclusion and is shown in Supplemental Figure 2.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.