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

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

1. Structural and Computational Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
2. Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, 69120 Heidelberg, Germany
3. Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
4. Heidelberg Institute for Stem Cell Technology and Experimental Medicine (HI-STEM gGmbH), 69120 Heidelberg, Germany
5. Division of Stem Cells and Cancer, Deutsches Krebsforschungszentrum (DKFZ) and DKFZ–ZMBH Alliance, 69120 Heidelberg, Germany
6. Faculty of Biosciences, Heidelberg University, 69117 Heidelberg, Germany
7. Department of Internal Medicine V, Heidelberg University Hospital, 69120 Heidelberg, Germany
8. Genome Biology Unit, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
9. Flow Cytometry Core Facility, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
10. Department of Hematology and Oncology, Medical Faculty Mannheim, University of Heidelberg, 68167 Mannheim, Germany
11. Computational Oncology, Molecular Diagnostics Program, National Center for Tumor diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
12. German Cancer Consortium (DKTK), 69120 Heidelberg, Germany
13. Department of Pediatric Immunology, Hematology and Oncology, University Hospital Heidelberg, 69120 Heidelberg, Germany
14. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, USA
15. Genomics Core Facility, European Molecular Biology Laboratory (EMBL), 69117 Heidelberg, Germany
16. Novo Nordisk Foundation Center for Stem Cell Biology, DanStem. Faculty of Health and Medical Sciences Blegdamsvej, DK-2200 Copenhagen, Denmark
17. Universitat Pompeu Fabra (UPF), Barcelona, Spain
18. Berlin Institute of Health (BIH), 10178 Berlin, Germany
19. Charité-Universitätsmedizin, 10117 Berlin, Germany
20. Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, 10115 Berlin, Germany

*These authors contributed equally

Address correspondence to lars.velten@crg.eu or simon.haas@charite.de

ABSTRACT

Single-cell genomics has transformed our understanding of complex cellular systems. However, excessive costs 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 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 interpretation of functional capacities of such precisely mapped cell states at the single-cell level. Our study serves as an accessible resource and paves the way for a data-driven era in cytometry.

INTRODUCTION

Single-cell transcriptomic technologies have revolutionized our understanding of tissues (Giladi and Amit, 2018; Stuart and Satija, 2019; Tanay and Regev, 2017). 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 (Baccin et al., 2020; Han et al., 2018, 2020; Schaum et al., 2018). 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 the clinical routine.

Unlike single-cell transcriptomics, flow cytometry offers a massive throughput in terms of samples and cells, is commonly used in the clinical routine diagnostics (Van Dongen et al., 2012) 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 (Paul et al., 2015; Velten et al., 2017), 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 significance 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 constitutes a particularly striking example for this disconnect (Haas et al., 2018; Jacobsen and Nerlov, 2019; Laurenti and Göttgens, 2018; Loughran et al., 2020). The classical model of hematopoiesis, which is mainly based on populations defined by flow cytometry (Akashi et al., 2000; Doulatov et al., 2010; Kondo et al., 1997) has recently been challenged in several aspects by single-cell transcriptomic (Giladi et al., 2018; Nestorowa et al., 2016; Paul et al., 2015; Tusi et al., 2018; Velten et al., 2017), functional (Notta et al., 2016; Perié et al., 2015), and lineage-tracing (Rodriguez-Fraticelli et al., 2018) approaches. These studies revealed that HSC commitment occurs earlier than previously anticipated, that putative oligopotent progenitors isolated by FACS consist of heterogeneous mixtures of lineage committed progenitor populations, and that HSC commitment is most accurately represented by a continuous process of

differentiation trajectories rather than by a stepwise differentiation series of discrete progenitor populations (Haas, 2020; Haas et al., 2018; Jacobsen and Nerlov, 2019; Laurenti and Göttgens, 2018). This discrepancy has contributed to conflicting results between studies that employ scRNA-seq for the definition of progenitor populations (Giladi et al., 2018; Paul et al., 2015; Pellin et al., 2019; Tusi et al., 2018; Velten et al., 2017) and studies that use FACS (Akashi et al., 2000; Kondo et al., 1997; Pei et al., 2017). 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 (Shahi et al., 2017; Stoeckius et al., 2017). Here, we demonstrate that ultra-high 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 reference atlases encompassing 97-197 surface markers across 95,463 cells representing all major populations of young, aged and leukemic human bone marrow and blood, as well as all states of hematopoietic stem cell 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 prior knowledge. We show that, compared to existing approaches, such optimized schemes are superior in the identification of cell types and more accurately reflect molecular cell 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 bone marrow while laying the grounds for reconciling flow cytometry and single-cell genomics data across human tissues.

RESULTS

A comprehensive single-cell transcriptomic and surface proteomic map of young, aged and malignant bone marrow

To establish a comprehensive single-cell transcriptomic and surface proteomic map of the human bone marrow (BM), we performed a series of Abseq experiments, in which mononuclear BM cells from hip aspirates were labelled with 97-197 oligo-tagged antibodies, followed by targeted or whole transcriptome scRNAseq on the Rhapsody platform (Figure 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 (see Supplementary Note 1, Supplementary Table 1). 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 (Figures 1a, S1, Supplementary Table 3). In addition, we performed analyses of healthy BM and matched peripheral blood (PB) samples using a 197 plex antibody panel (Figure S2, Supplementary Table 2). For samples from healthy donors, CD34+ cells were enriched to enable a detailed study of hematopoietic stem cell (HSC) lineage commitment. For samples from AML patients, CD3+ cells were enriched in some cases to ensure sufficient coverage of T cells.
Since proteo-genomic single-cell approaches have previously not been performed using such high-parametric settings, we designed a series of control experiments. First, we performed matched Abseq

experiments in the presence or absence of antibodies to ensure that ultra-high plex antibody stains do not impact on the transcriptome of single cells (Supplementary Note 2). Subsequently, we performed whole transcriptome Abseq experiments to confirm that our custom targeted approach accurately captures all cell types and cell stages (Supplementary Note 3). Finally, we evaluated the sequencing requirements for optimal cell type classification in high-parametric proteo-genomic single-cell experiments (Supplementary Note 4). In total, 98,463 high-quality BM or PB cells with combined transcriptomic and high-parametric surface proteomic information were profiled, and an average of ~7500 surface molecules per cell were detected (Figure S3).

Following data integration across experiments and measurement modalities (Methods), we identified 45 cell types and cell stages covering the vast majority of previously described hematopoietic cell types of the bone marrow and peripheral blood, including all stages of HSC differentiation in the CD34+ compartment, all T cell and NK cell populations of the CD3+ and CD56+ compartments, several dendritic cell and monocyte subpopulations from the CD33+ compartment and all B cell differentiation states across CD10+, CD19+ and CD38^{high} compartments (Figure 1b,c Supplementary Note 5, 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 bone marrow occupied the same cell states in all individuals, whereas cell states in AML differed between patients and were exclusive to disease (Figure 1b and see below). Importantly, the combined transcriptomic and surface proteomic information provided higher resolution and revealed cell types that are not readily identified by one of the individual data layers alone (Supplementary Note 6). 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 (Supplemental Video S1). The Abseq-App is accessible at: https://abseqapp.shiny.embl.de/.

Systematic association of surface markers with cell type identities, differentiation stages and biological processes

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 significance of individual markers remains frequently ambiguous. To quantitatively link surface marker expression 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, its current 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. For each surface marker, we then quantified the fraction of variance of expression that is determined by any of these processes (Figure 2a). This model specified the exact significance of known markers and identified a series of novel markers that represent cell type identities or differentiation stages, as well as stemness, cytotoxicity and cell cycle properties (Figures 2b-d, S1, S2 and S4).

We initially focused on the evaluation of surface molecules that specifically mark distinct stages of HSC differentiation, since a lack of specific commitment markers currently impedes the accurate representation of HSC lineage commitment by flow cytometry (Notta et al., 2016; Paul et al., 2015; Pellin et al., 2019; Tusi et al., 2018; Velten et al., 2017). For this purpose, we performed pseudotime

189 analyses within the CD34+ HSPC compartment and identified surface markers that correlate with the
190 progression of HSCs towards erythroid, megakaryocytic, monocyte, dendritic cell or B cell
191 differentiation trajectories (Figure 2d, 3a, see methods). These analyses quantified the exact expression
192 dynamics of many well-established markers, such as CD38 as a pan-differentiation marker, as well as
193 CD10 and CD11c as early B cell and monocyte-dendritic cell lineage commitment marker, respectively
194 (Figures 2d, and S5a). Importantly, our analyses revealed a large number of novel surface markers that
195 specifically demarcate distinct stages of HSC lineage commitment, including CD326, CD11a, Tim3 and
196 CD98 (Figure 2d). To confirm the high specificity of these markers for erythroid, myeloid, lympho-
197 myeloid and isotropic lineage commitment, respectively, we used FACS-based indexing of surface
198 markers coupled to single-cell RNAseq ("index scRNAseq", see also Supplementary Note S7), or
199 coupled to single-cell cultures ("index cultures") (Figure 3b). As suggested by our proteo-genomic
200 single-cell data, CD326 expression was associated with molecular priming and functional commitment
201 into the erythroid lineage (Figure 3c-g). In contrast, Tim3 and CD11a were identified as lympho-myeloid
202 or pan-myeloid differentiation markers and were associated with transcriptomic priming and functional
203 commitment into the myeloid lineage (Figures 3c, h-k, S5b-f). Finally, CD98 was identified as a novel
204 pan-differentiation marker of HSCs, which we confirmed by classical flow cytometry (Figure S5g-k).
205 Beyond the progression of HSCs to lineage committed cells, we also analyzed the surface marker
206 dynamics throughout B cell differentiation, allowing us to identify novel markers specific to their lineage
207 commitment, maturation, isotype switching and final plasma cells generation (Figure S5l-r).
208
209 Together, our model provides a global and quantitative understanding of how well cell type identities,
210 differentiation stages and biological processes are explained by the expression of individual surface
211 markers. A comprehensive overview of surface markers associated with these processes is depicted in
212 the supplement (Supplementary Table 5, Figures S1, S2 and S4).
213
214
215 **Surface proteome adaptation in healthy aging and cancer**
216
217 To investigate the surface proteome throughout healthy aging, we compared Abseq data of bone marrow
218 from young and aged healthy individuals. These analyses revealed that the expression of surface
219 molecules was highly similar across all BM populations between the age groups (Figure 4a, b,
220 Supplementary Table 6), suggesting an unexpectedly stable and highly regulated surface proteome that
221 is only modestly affected by aging. While cell type frequencies were also only modestly affected by
222 aging, a significant accumulation of cytotoxic effector CD8 T cells was observed (S6a, Fagnoni et al.,
223 1996). Moreover, the expression of several immune regulatory molecules showed age-related changes
224 in surface presentation, including the death receptor FAS (CD95), the poliovirus receptor (CD155) and
225 the ICOS ligand (CD275) (Figure 4b). In particular, naive CD8+ and CD4+ T cell subsets displayed an
226 aging-associated decline of CD27 surface expression, a co-stimulatory molecule required for generation
227 and maintenance of long-term T cell immunity (Figure 4b, c, Peters et al., 2015). Together these analyses
228 suggest that the overall surface proteome is widely maintained upon healthy aging, whereas specific
229 changes, most prominently in the surface presentation of immune regulatory molecules, occur.
230 We next explored surface marker remodeling in AML, a blood cancer characterized by the accumulation
231 of immature, dysfunctional myeloid progenitors, also called blasts. Few cells with transcriptomes
232 resembling healthy cells were detected in AML samples, and the majority of leukemic cells clustered
233 apart from the normal differentiation trajectory (Figure 4d). Both inter- and intra-patient heterogeneity
234 of blast cells was observed, with blast cells adopting stem-like (CD34, CD133), LMPP-like (IGHM,
235 DNTT), GMP-like (CD33, AZU1), dendritic cell-like (CD11c) or myelomonocytic (CD14) states.

236 Surface proteomic profiles of these states were highly similar to their healthy counterparts (Figure 4e,
237 S6b), despite substantial differences for individual markers (Figure S6b, c, Supplementary Table 6).
238 Importantly, blasts retained many features reminiscent of the cell stage they were blocked in, including
239 the surface expression of potential drug targets, such as PD-L1 (CD274), CTLA4 (CD152) and Tim3
240 (Figure 4f), suggesting that the myeloid differentiation program of the AML might be essential in the
241 treatment choice of targeted immune therapies.
242
243
244 **Data-driven isolation strategies and immunophenotypic characterization of rare bone marrow cell**
245 **populations**
246
247 Gating strategies for flow cytometry have evolved historically in a process of trial and error. In particular,
248 the isolation of rare and poorly characterized cell subsets using flow cytometry remains challenging,
249 whereas commonly used gating schemes are not necessarily optimal in purity (precision) and efficiency
250 (recall). To tackle these problems, we explored the utility of different machine learning approaches for
251 the data-driven definition of gating schemes. For all populations in our dataset, gating schemes defined
252 by machine learning approaches provided higher precision (purity) if compared to classical gating
253 schemes from literature (Figures S7, Supplementary Table 7); While different machine learning methods
254 tested achieved similar purities, gates defined by the hypergate algorithm (Becht et al., 2019) offered a
255 higher recall (Figure S7).
256 To validate and demonstrate this approach, we focused on determining novel gating strategies for rare
257 and poorly characterized BM cell types, such as cytotoxic CD4 T cells (Figure 5a) and mesenchymal
258 stem or stromal cells (MSCs) (Figure 5g). Cytotoxic CD4 T cells represent a rare T cell population
259 characterized by the expression of cytotoxicity genes typically observed in their well-characterized CD8
260 T cell counterparts (Szabo et al., 2019). While this cell type has been suggested to be involved in several
261 physiological and pathophysiological processes, no coherent gating strategy for their prospective
262 isolation exists (Takeuchi and Saito, 2017). Hypergate suggested cytotoxic CD4 T cells to display an
263 immunophenotype of CD4+CD28-, and differential expression analyses of surface markers revealed that
264 cytotoxic CD4 T cells express significantly lower levels of CD7, CD25, CD127 and CD197 if compared
265 to other CD4 T cell subsets (Figure 5a-d). Flow cytometric analyses of CD4+CD28- T cells confirmed
266 the expected immunophenotype in BM from healthy donors and patients with different hematological
267 cancers, suggesting a robust and efficient prospective isolation of this rare cell type (Figure 5c-e).
268 Finally, FACS-based sorting of CD4+CD28- T cells followed by gene expression analysis confirmed
269 the expression of cytotoxicity genes in this population (Figure 5f).
270 MSCs constitute a rare and heterogeneous group of cells in the bone marrow (Al-Sabah et al., 2020;
271 Frenette et al., 2013). While *ex vivo*-expanded MSCs have been phenotyped extensively, primary human
272 MSCs remain poorly characterized, in particular due to their extremely low frequency. In our dataset,
273 we captured a small amount of heterogeneous MSCs, with one subset (MSC-1) expressing high levels
274 of the key bone marrow-homing cytokine CXCL12 (Figure 5g). Hypergate suggested CXCL12-
275 expressing MSCs to be most efficiently isolated by expression of CD13 and absence of CD11a
276 expression (Figure 5h). Indeed, flow cytometric analyses of CD13+CD11a- MSCs validated the
277 immunophenotype suggested by our Abseq data and confirmed known and novel MSC surface markers
278 identified by our approach (Figure 5i-k). Moreover, FACS-based isolation of CD13+CD11a- cells
279 followed by transcriptomic analyses revealed a high enrichment of CXCL12 and other key MSC
280 signature genes (Figure 5l).
281 Together, these analyses demonstrate the utility of our approach for deriving data-driven gating schemes
282 and mapping the surface proteome of poorly characterized populations. The Abseq-App in combination

with our proteo-genomic reference map allows users to define new data-driven gating schemes for any population of interest.

A fully data-driven gating scheme reflects the molecular routes of human hematopoiesis

Gating schemes for complex biological systems, such as the hematopoietic stem and progenitor cell (HSPC) compartment, are steadily improving. However, there is strong evidence from single-cell transcriptomics (Giladi et al., 2018; Paul et al., 2015; Tusi et al., 2018; Velten et al., 2017), lineage tracing (Perié et al., 2015; Rodriguez-Fraticelli et al., 2018) and single-cell functional experiments (Notta et al., 2016) 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 (Notta et al., 2016; Paul et al., 2015; Velten et al., 2017).

In order to generate flow cytometric gating schemes that reflect most adequately the molecular states associated with hematopoietic stem cell differentiation, we used the surface proteomic 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 16 leaves representing molecularly defined cell states with high precision (Figure 6a-c). The data-derived scheme outperformed a widely used scheme at resolving the different cell states (Figure 6d; Doulatov et al., 2010), particularly in the identification of lineage committed progenitors, a major shortcoming of many current gating strategies (Notta et al., 2016; Paul et al., 2015; Perié et al., 2015; Velten et al., 2017). Importantly, cell populations defined by the data-defined gating scheme were transcriptionally more homogenous, compared to the literature-based scheme (Figure 6e).

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-sequencing while simultaneously recording surface marker expression (index-scRNAseq) (Fig 6f, g, Supplementary Note 7). This approach demonstrated that the new gating strategy efficiently separated molecularly defined cell states (Figure 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 (Figure 6h).

Together, these results demonstrate that high-content single-cell proteo-genomic experiments 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 molecularly defined progenitor states in the human hematopoietic hierarchy using cost-effective flow cytometry.

Systematic integration of single-cell genomics, flow cytometry and functional data via NRN

While classical FACS gating strategies are of great use for the prospective isolation and characterization of populations, single-cell genomics studies revealed that developmental processes, including early hematopoiesis, are most accurately represented by a continuous process, and that subdivisions into discrete populations are arbitrary (Macaulay et al., 2016; Nestorowa et al., 2016; Pellin et al., 2019; Tusi et al., 2018; Velten et al., 2017). To complement the approach based on discrete gates, we here propose that high-dimensional flow cytometry data can be used to place single cells into the continuous space of

hematopoietic differentiation spanned by single-cell genomics exploiting commonly used surface markers (Figure 7a). Based on the observation that surface marker expressions in flow cytometry and Abseq follow similar distributions (Figure S8a), 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 Figure 6 (12 markers) and a semi-automated panel based on our Abseq data to better resolve erythro-myeloid lineages (11 markers, Supplementary Note 7). 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 UMAP (Figure 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 (Figure S8c-e). 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 (Figure 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 insights into functional 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 Figure 6, followed by data integration with 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 (Figure 7d). These data suggest that NRN accurately places indexed single-cell functional data into single-cell proteo-genomics reference maps.

Taken together, 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. We provide a user-friendly implementation of these tools as part of the Abseq-App.

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 bone marrow spanning the expression of 97-197 surface markers across 45 cell types and all stages of hematopoietic stem cell differentiation, healthy ageing, and leukemia. Our dataset is carefully annotated (Supplementary Note 5) and will serve as a key resource for hematology and immunology.

While cytometry experiments remain the working horses 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 (Papalexi and Satija, 2018; Paul et al., 2015). For the first time, we have exploited proteo-genomic single-cell 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 (Baron et al., 2019; Paul et al., 2015; Velten et al., 2017; Wilson et al., 2008), single-cell proteo-genomics has a sufficient throughput to enable the profiling of entire tissues or organs, and at the same time covers hundreds of surface

markers. Unlike single-cell RNA-seq data, antibody tag counts reflect the true distributions 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, we have devised a gating strategy that reflects the molecular routes of hematopoietic stem cell differentiation, and we have 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 impact on hematopoietic stem cell research. There is a growing consensus in the field that lineage commitment occurs early from primed HSCs, that oligopotent progenitors are very rare, 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 lympho-myeloid progenitors (Drissen et al., 2019; Giladi et al., 2018; Görgens et al., 2014; Pellin et al., 2019; Tusi et al., 2018; Velten et al., 2017; Zheng et al., 2018). Due to a lack of better alternatives, many functional studies still use the classical gating scheme alongside the outdated concept of 'common myeloid progenitors'(Akashi et al., 2000; Kondo et al., 1997; Pei et al., 2017). 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 determines cellular fate *in vitro* at a much higher resolution compared to previous studies (Notta et al., 2016; Velten et al., 2017), 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 the use of imprecisely defined populations.

Furthermore, these advances enable the rapid profiling of blood formation and other bone marrow phenotypes while offering a resolution comparable to single-cell genomics. Recently, bone marrow phenotypes of diseases, ranging from sickle cell disease (Hua et al., 2019) to leukemia (van Galen et al., 2019) 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.

While we have demonstrated the implementation of data-driven design and analysis strategies for cytometry assays in the context of bone marrow, 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 the myriad of cell types discovered by recent single-cell atlas projects.

SUPPLEMENTAL MATERIAL

- Methods
- Supplemental Tables S1-S9
- Supplemental Video S1
- Supplementary Notes 1-7

ACKNOWLEDGEMENTS

We thank Vadir Lopez-Salmeron, Vishnu Ramani and Wieland Keilholz from BD Biosciences / Multiomics for providing oligo-labelled antibodies and their support in the implementation of the Rhapsody platform. We would like to thank members of the Haas, Velten, Trumpp and Steinmetz labs for helpful discussions. Moreover, we thank members from the DKFZ flow cytometry and the EMBL genomics core facility for support. This work was financially supported by the Emerson foundation grant 643577 (to LV), the German Bundesministerium für Bildung und Forschung (BMBF) through the Juniorverbund in der Systemmedizin 'LeukoSyStem' (FKZ 01ZX1911D to LV, SH and SR), SFB873, FOR2674 and FOR2033 funded by the Deutsche Forschungsgemeinschaft (DFG), the SyTASC consortium (Deutsche Krebshilfe), the Dietmar Hopp Foundation (all to AT); and the José Carreras Foundation for Leukemia Research (grant no. DCJLS 20R/2017 to LV, AT, and SH). LV acknowledges support of the Spanish Ministry of Science and Innovation to the EMBL partnership, the Centro de Excelencia Severo Ochoa and `the CERCA Programme / Generalitat de Catalunya. DN is an endowed professor of the Deutsche José 641Carreras Leukämie Stiftung (DJCLS H 03/01). Contributions by DN, JCJ, WKH and TB were supported by the Gutermuth Foundation, the H.W. & J. Hector fund, Baden-Württemberg.

CONFLICT OF INTEREST

The oligo-coupled antibodies used in this study were a gift from BD Biosciences. Besides this, the authors declare no relevant conflicts of interest.

CODE AVAILABILITY

The implementation of the NRN algorithm is available at https://git.embl.de/triana/nrn

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

MAIN FIGURES

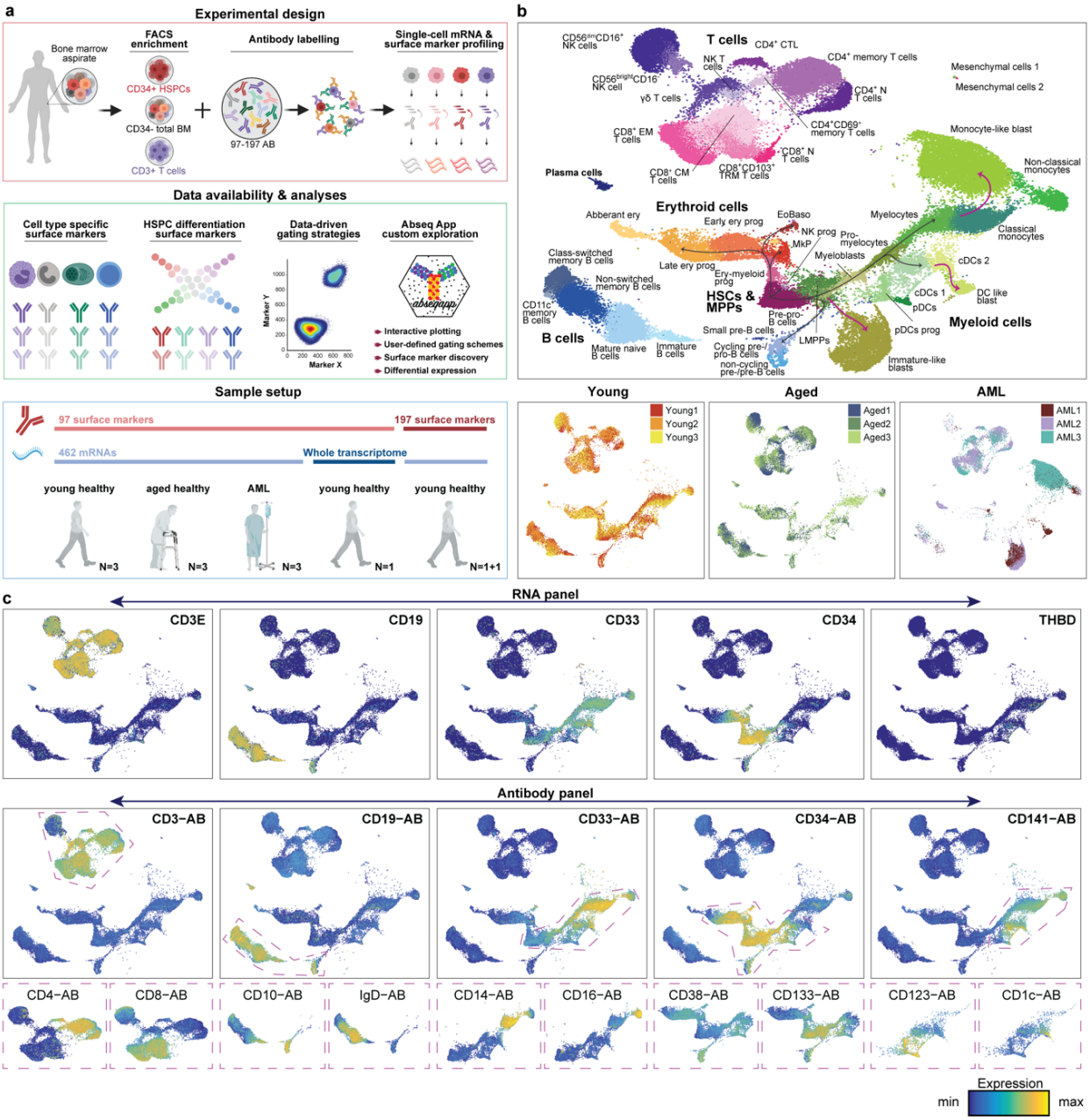


Figure 1. A comprehensive single-cell transcriptomic and surface proteomic map of young, aged and malignant bone marrow. a. Overview of the study. See methods and main text for details. **b.** Top: uMAP display of single-cell proteo-genomics data of human bone marrow from healthy young, healthy old and AML patients (n=70,017 single cells, 97 surface markers), integrated across n=9 samples and data modalities. Clusters are color-coded. Bottom: uMAPs highlighting sample identities. See Supplementary Note 5 for details on cluster annotation. The whole transcriptome Abseq data is presented in Supplementary Note 3, the Abseq experiments with measurements of 197 surface markers are presented in Figure S1. **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 major cell types. Middle: Expression of the corresponding surface proteins. Bottom: Expression of markers widely used to stratify major cell types into subtypes. Only the parts of the uMAPs highlighted by dashed polygons in the middle row are shown.

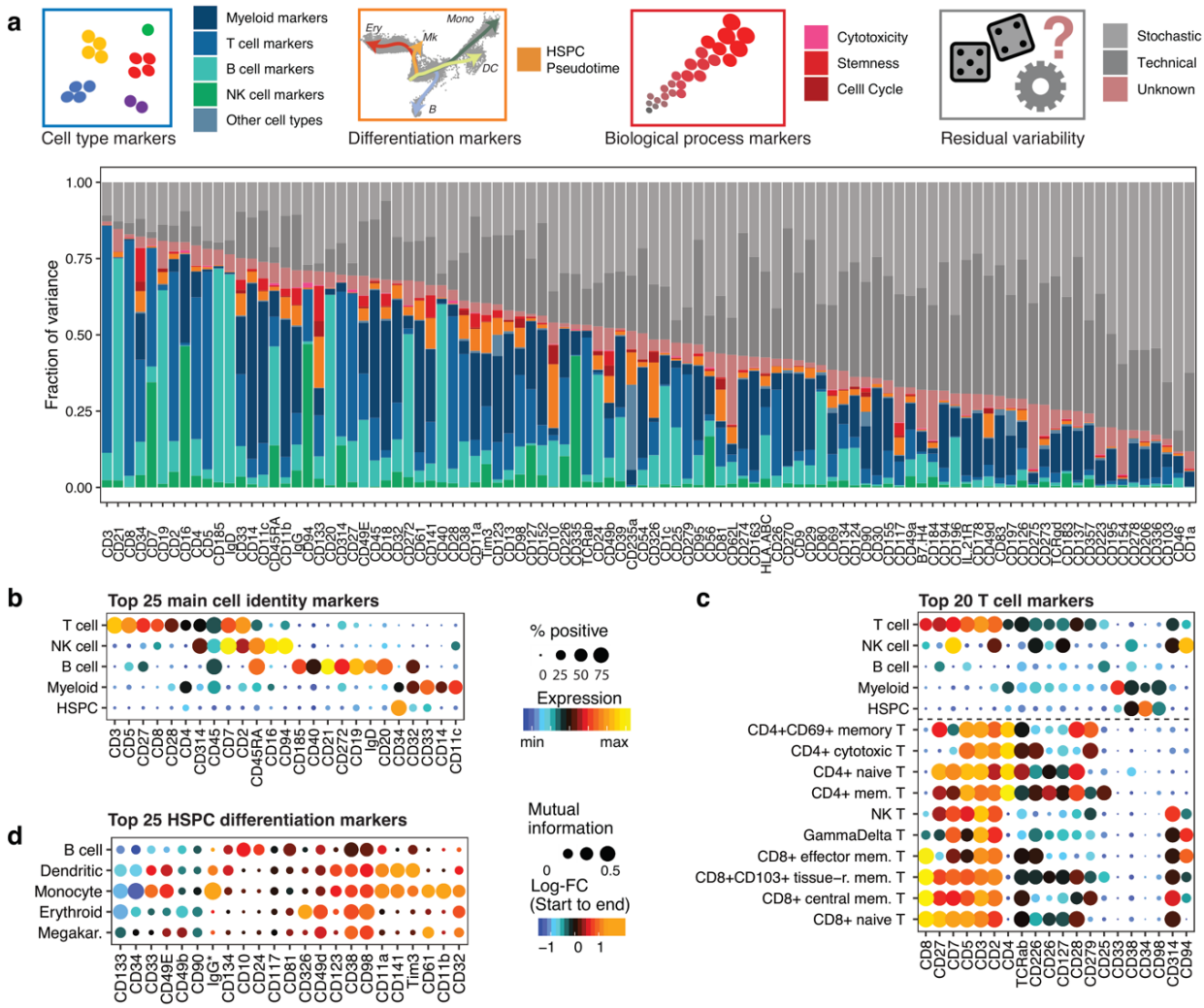


Figure 2: Association of surface marker expression with cell type identities, cellular differentiation, and biological processes. a. For each surface marker measured in our 97-plex Abseq data, the fraction of variance explained by different covariates (colored insets in top row) is displayed. For this, every single cell from healthy young individuals (n=3 samples, 28,031 single cells) was assigned to a cell type identity (blue inset, see Figure 1b), and cytotoxicity, stemness and cell cycle scores (red inset, see Figure S4d), as well as technical covariate scores were determined. Additionally, pseudotime analyses were used to assign differentiation scores to HSPCs (orange inset, see Figure 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 '*Modelling 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 major 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 25 surface markers with the highest fraction of variance explained by T cell subtype is displayed, legends like in *b*. **d.** HSPC differentiation markers. Dot plot depicting expression changes of markers across pseudotime in CD34+ HSPCs. Color indicates logarithmic fold change between the start and the end of each pseudotime trajectory. Point size indicates the mutual information in natural units of information (nats) between pseudotime and marker expression. The 25 surface markers with the highest fraction of variance explained by pseudotime covariates are displayed.

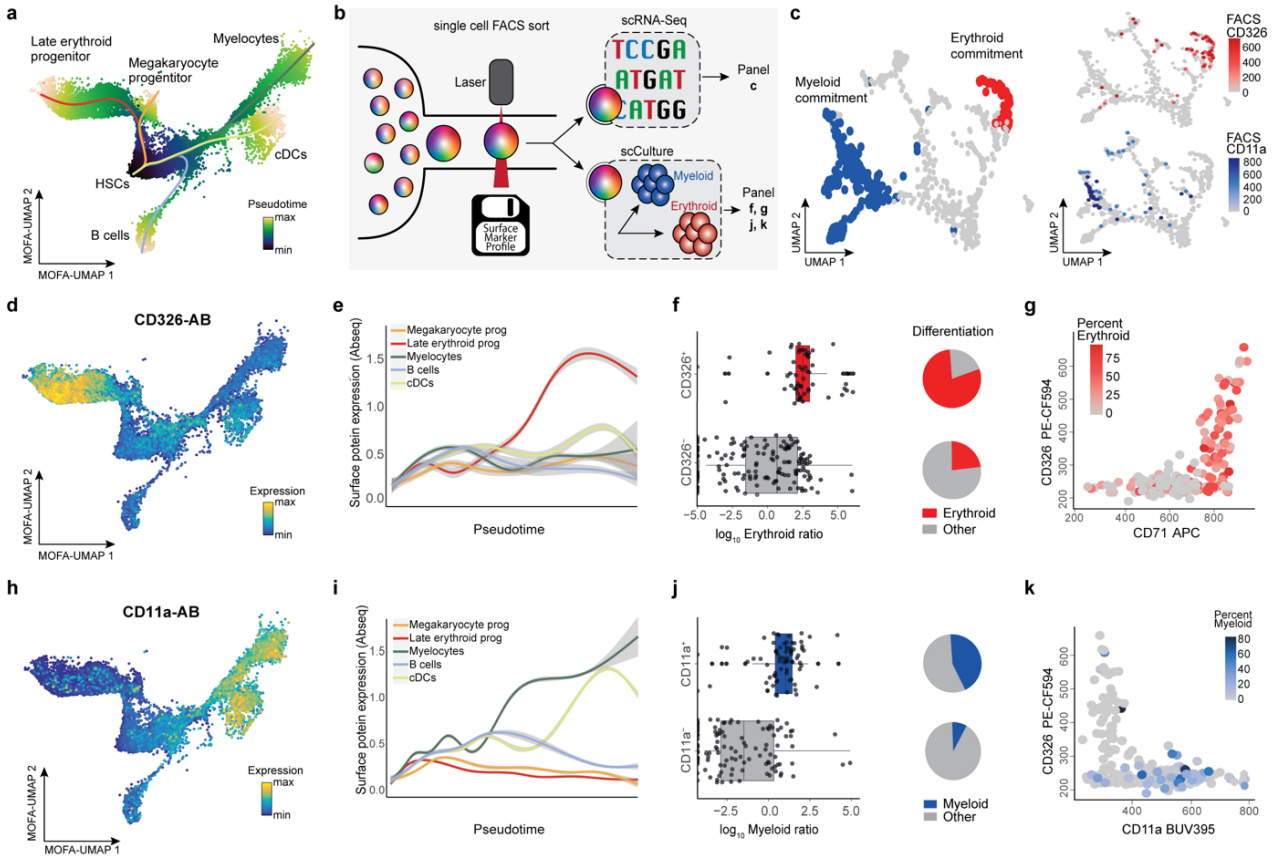


Figure 3. Validation of novel stage-specific HSPC differentiation markers. a. uMAP plot depicting CD34+ HSPCs and their pseudotime scores along five differentiation trajectories, see Methods, section "*Pseudotime analysis*". The normalized pseudotime scores across all lineages is color-coded. **b.** Scheme illustrating the experiments performed to validate the significance of selected markers. See main text and Supplementary Note 7 for details. **c.** uMAP display of mRNA expression of n=630 CD34+ cells from a single-cell Smart-seq2 experiment where the expression of 12 surface markers was recorded using FACS. For a detailed description of the experiment, see Supplementary Note 7. Left panel: Cells with myeloid and erythroid gene expression signatures are highlighted on the uMAP. Right 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) smoothened over the different pseudotime trajectories illustrated in panel a. **f.** Box plots depicting the ratio in erythroid cells produced in single-cell cultures in relation to the CD326 expression of the founder cell (n=213 single cell derived colonies). Pie charts summarize the relative fraction of wells showing erythroid output (CD45-CD235+) within CD326+ or CD326- founder cells **g.** Scatter plots depicting the fraction of erythroid cells produced in single cell cultures in relation to the CD326 and CD71 expression of the founder cell. **h-k.** Like *d-g*, except that CD11a and its relation to the formation of myeloid cells in single-cell cultures is investigated. For myeloid differentiation pie charts, CD45+ CD235- CD66b+ and CD45+ CD235- CD14+ progeny cells were considered.

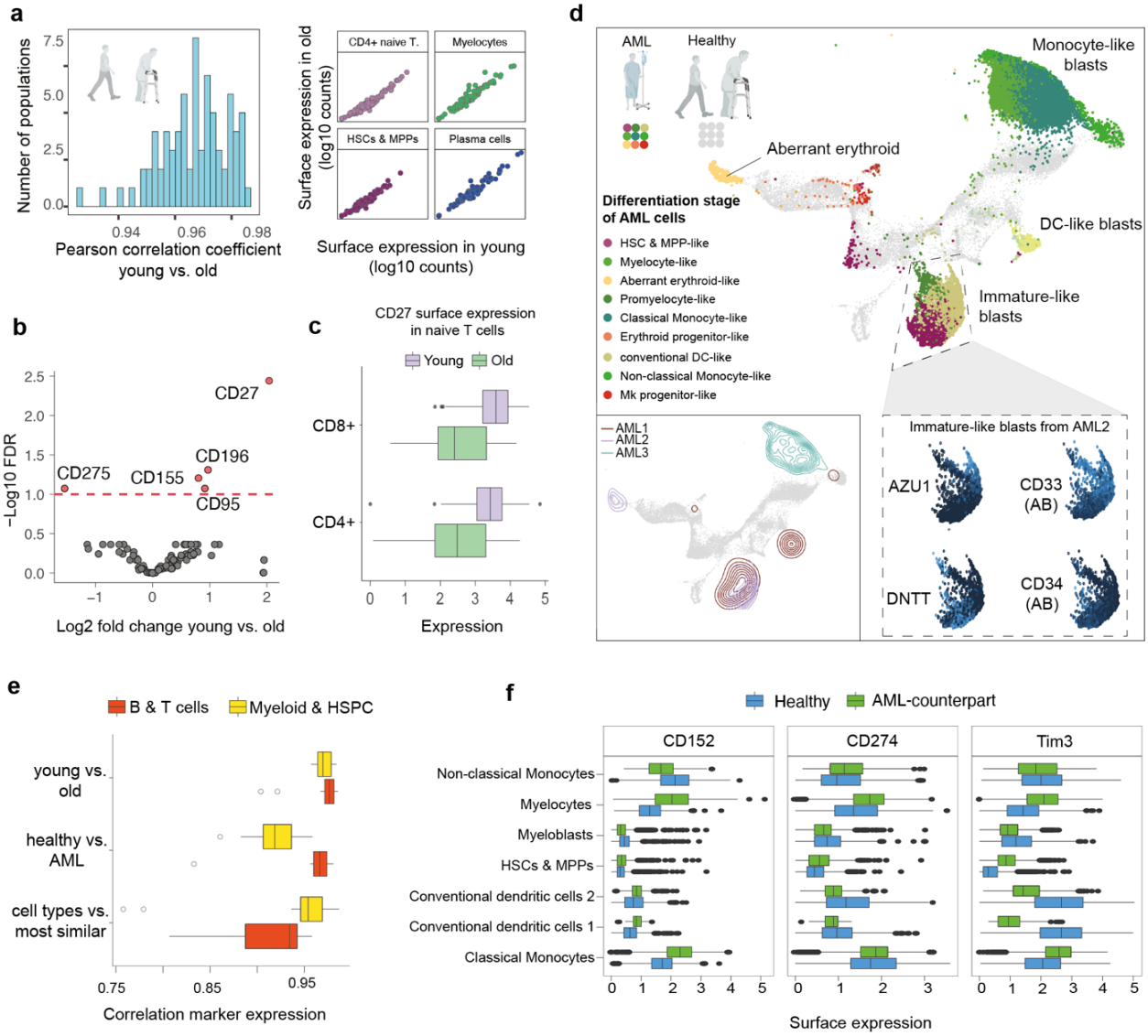


Figure 4. Surface proteome adaptation in healthy aging and cancer. a. Correlation in surface marker expression between matched cell types from old and young bone marrow donors. For each cell type (Figure 1b), mean surface marker expression across all cells was computed, separately for all 'young' and 'old' samples, and the correlation between the two matched cell types was determined. 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, see also Supplementary Table 6. **b.** Volcano plot depicting log fold change and false discovery rate (FDR) for a test for differential surface marker expression between cells from young and old individuals, while accounting for cell types as covariates. See Methods, section "*Differential expression testing between experimental groups*" for details. **c.** Box plots depicting CD27 surface expression in different T cell populations from young and old individuals. **d.** uMAP of cells from AML patients, color-coded according to the cell types from healthy individuals they are most similar to. A random forest classifier was trained to predict cell type labels from expression of the healthy individuals. Subsequently, the classifier was applied to cells from the AML patients to obtain the indicated color-code. Insets depict density of cells from the different AML patients (lower left panel), and expression of relevant markers on the immature-like blasts from patient 2 (lower right panel). **e.** Correlation in surface marker expression between cells from old, young and leukemic individuals. Similar to panel *a*, except that here, a random forest classifier was trained on cells from the healthy individuals to predict cell type labels from gene expression data. This classifier was then applied to all cells from all individuals, and the predicted labels were used.

Correlations are shown for matching cell types from young versus old individuals, from healthy individuals versus AML patients, as well as for cell types versus the transcriptomically most similar cell type available in the dataset. See Supplementary Figure 6b for a complete overview. **f.** Surface expression of immunotherapy targets CTLA-4 (CD152), PD-L1 (CD274) and Tim-3 in different myeloid compartments of healthy donors and the AML subsets that are most closely related to the respective healthy cell type.

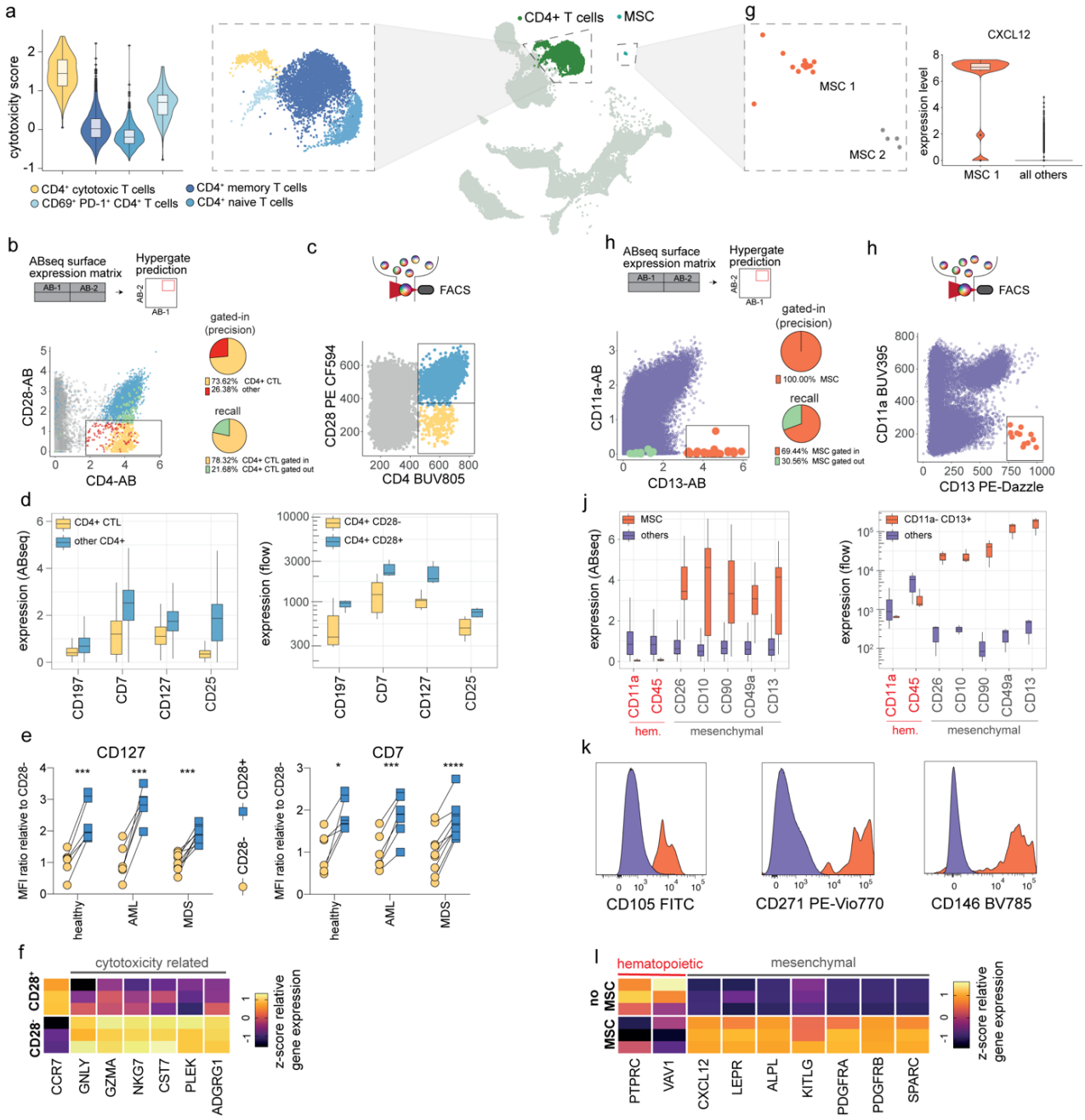


Figure 5. Data-driven definition of gating schemes for rare cell types. a. Different CD4 T cell subsets are highlighted (central and right panels) and the corresponding distributions of cytotoxicity scores for every subset are displayed (left panel). **b.** Hypergate (Becht et al., 2019) was used to identify a gating scheme for the prospective isolation of cytotoxic CD4+ T cells. The suggested gate is highlighted on a plot depicting the surface protein expression of CD4 and CD28 as identified from pregated CD45+CD3+ Abseq data. Yellow and green dots correspond to cytotoxic T cells located within and outside of the selected gate, respectively. Red dots correspond to other cells located inside the selected gate (false positives), blue dots correspond to other CD4+ cells located outside the gate and grey dots to other cells located outside the gate. Pie charts indicate precision and recall. **c.** FACS plot displaying the expression of CD4 and CD28 on pre-gated CD45+CD3+ cells, and respective gates (yellow dots correspond to CD4+CD28- cytotoxic T cells and blue dots to other CD4+CD28+ T cells. **d.** 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 ***c*** (right panel). **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 bone marrow samples from healthy, AML and MDS patients. n= 6, 6 and 9 patients in the respective groups. **f.** Heat map depicting gene expression of cytotoxicity-related genes in FACS sorted CD4+CD28- and CD4+CD28+ cells, as quantified by qPCR. n= 4 patients. **g-j.** Analogous to a-d, but with the identification and use of a CD11a-CD13+ gate for the isolation of CXCL12+ mesenchymal stem cells (MSC). Orange and green dots correspond to MSCs located within and outside of the selected gate, respectively. Purple dots correspond to other cells located outside the gate. **k.** Representative FACS histogram plots showing surface expression of well-known MSC surface markers, which were not contained in the original 97 antibody Abseq panel. ***l.*** Heat map depicting gene expression of common hematopoietic and MSC signature genes in FACS sorted CD11a-CD13+ MSCs and cells outside the gate, as quantified by qPCR. n= 3 patients. 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.

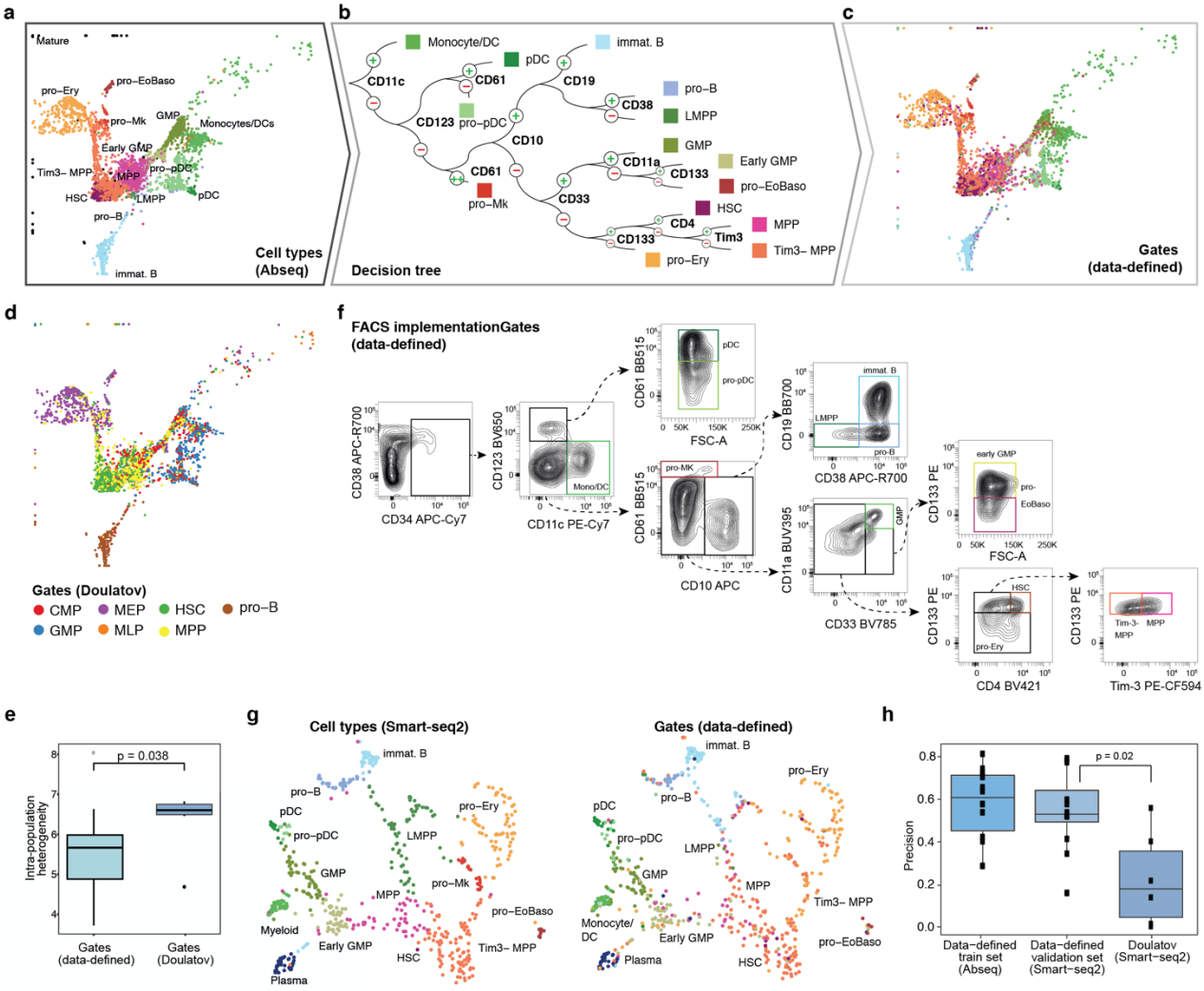
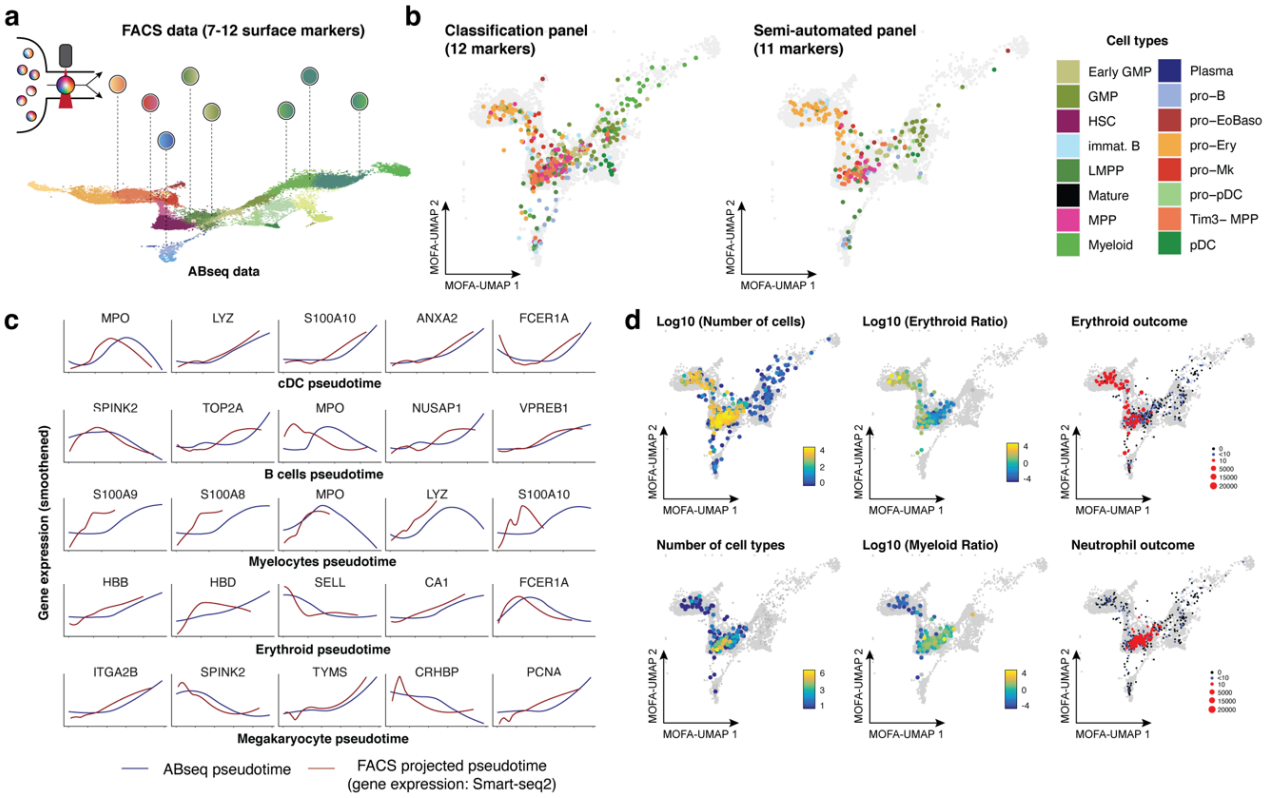


Figure 6. Data-driven definition of gating schemes for hematopoietic stem and progenitor cells. a. uMAP depicting all CD34+ HSPCs cells from one healthy young individual. Clustering and cluster annotation were performed exclusively on this individual to achieve a higher subtype resolution of stem cells ('HSCs'), immature progenitors with myeloid/lymphoid transcriptomic priming ('MPP') and immature progenitors with erythroid/megakaryocytic transcriptomic priming ('Tim3- MPP'). See panel b for a color scheme. **b.** Decision tree using surface marker expression from the Abseq data in order to classify cells into cell types. See Methods, section '*Data-driven identification of gating schemes*' and main text for details. **c.** uMAP highlighting cell type classification obtained from the decision tree. Please take note that colors now correspond to putative 'gates' applied to the expression levels of the 12 markers shown in panel b, and not to cell types defined from single-cell multi-omics data. **d.** uMAP highlighting classification obtained from a decision tree recapitulating the classical gating scheme used in the field (Doulatov et al., 2010), i.e. HSC: CD34+CD38-CD45RA-CD90+, MPP: CD34+CD38-CD45RA-CD90-, MLP: CD34+CD38-CD45RA+, CMP: CD34+CD38+CD10-CD45RA-*Flt3*+, MEP: CD34+CD38+CD10-CD45RA-*Flt3*-, GMP: CD34+CD38+CD10-CD45RA+*Flt3*+, pro-B: CD34+CD38+CD10+. Since CD135 was not part of the Abseq panel, the expression of *Flt3* was smoothened using MAGIC (van Dijk et al., 2018) for this purpose. Automatic thresholding was used to identify marker-positive cells, see Methods, section '*Thresholding of surface marker expression*' for details. **e.** Boxplot depicting the intra-population heterogeneity for each of the classification schemes shown in panels a, c and d. For every population, the median of all Euclidean distances in low dimensional (MOFA) space between any pair of cells was computed – the boxplot is summarizing median distances across populations. Low values are indicative for homogenous populations. P-value is from a two-sided Wilcoxon test. **f.** Implementation of FACS gating scheme suggested by the decision tree from panel b. **g.** uMAP display of mRNA expression of n=630 CD34+ HSPCs from an indexed

single-cell Smart-seq2 experiment where the expression of the 12 surface markers (for the data-defined gating) was recorded using FACS. Left panel: Clusters are highlighted based gene expression, see Supplementary Note 7 for details. Right panel: Classification of the cells based on FACS markers using the data-defined gates shown in panel f. **h.** Precision of the classification scheme shown in panel b, computed on the training data (i.e. the Abseq dataset) and the test data (i.e. the Smart-seq2 dataset). Precision was computed per gate as the fraction of correctly classified cells. For comparison with the Doulatov gating scheme, matching cell types were determined if possible; CMPs and MEPs were excluded from the analysis as no unique match to a molecularly defined cell could be identified, see panel d and Velten et al., 2017; Paul et al. 2015. P-value is from a two-sided Wilcoxon test.



.

Figure 7. Systematic integration of single-cell genomics, flow cytometry and functional data via NRN. a. Illustration of the concept. See main text and methods for details. **b.** Projection of indexed Smart-seq2 data onto a reference uMAP. Single-cells with recorded ('indexed') FACS measurements of surface markers (data-defined classification panel or semi-automated panel) were subjected to Smart-Seq2 based scRNA-seq. The commonly used surface markers were used to project cells via NRN onto the Abseq uMAP (see Methods, section "*The NRN algorithm for integrating FACS and single cell genomics data"* and Supplemental Note 7 for details). Take note that only FACS data was used for the projection in uMAP space, whereas colors depict cell types identified from RNA expression. **c.** Projection of indexed Smart-seq2 data onto reference pseudotime trajectories. The same single cells were projected onto the differentiation trajectories shown in Figure 3a using FACS measurements only. The expression of differentiation markers was then determined from available Smart-seq2 data and smoothened over projected pseudotime values (red lines). For comparison, the expression values of the same genes were determined from Abseq data and smoothened over the reference pseudotime values (blue lines). The selected genes correspond to the five genes with the strongest statistical association to 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 (data-defined classification panel from Figure 5) were projected onto the uMAP defined by Abseq via NRN. Single cells were seeded into culture medium supporting the formation of erythroid, megakaryocytic and distinct myeloid cell types, see Methods, section *'Single-cell index cultures'* for details. Colony size, the total number of cell types per colony, the ratio between erythroid and myeloid cells, as well as the ability of single cells to give rise to erythroid cells and neutrophils were highlighted on the uMAPs.

SUPPLEMENTARY FIGURES

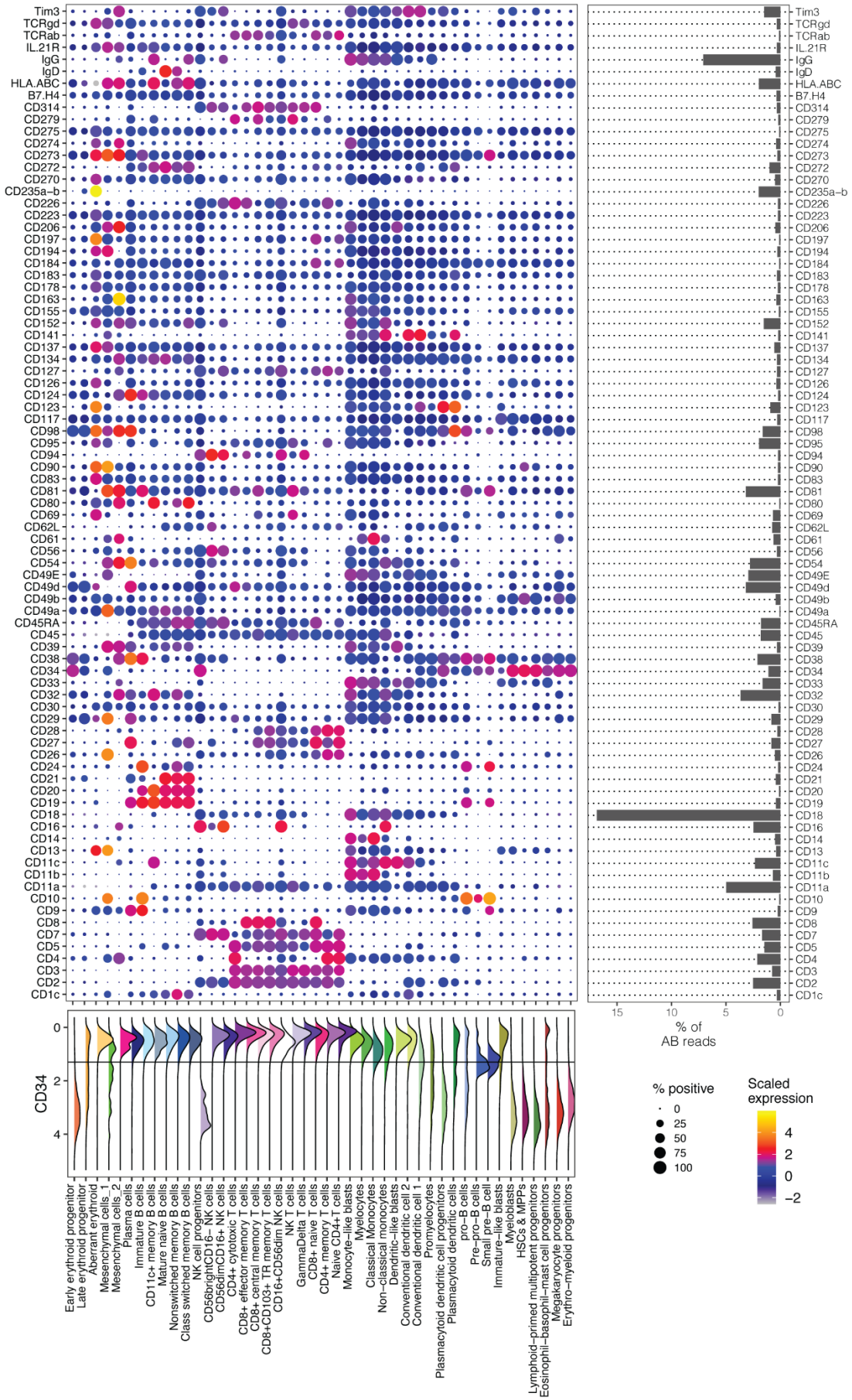


Figure S1. A proteo-genomic single-cell map of 97 surface markers in human bone marrow. Related to Figure 1. Dot plot depicting the expression of all surface markers by cell type. Color indicates mean normalized expression, point size indicates the fraction of cells positive for the marker. Automatic 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.

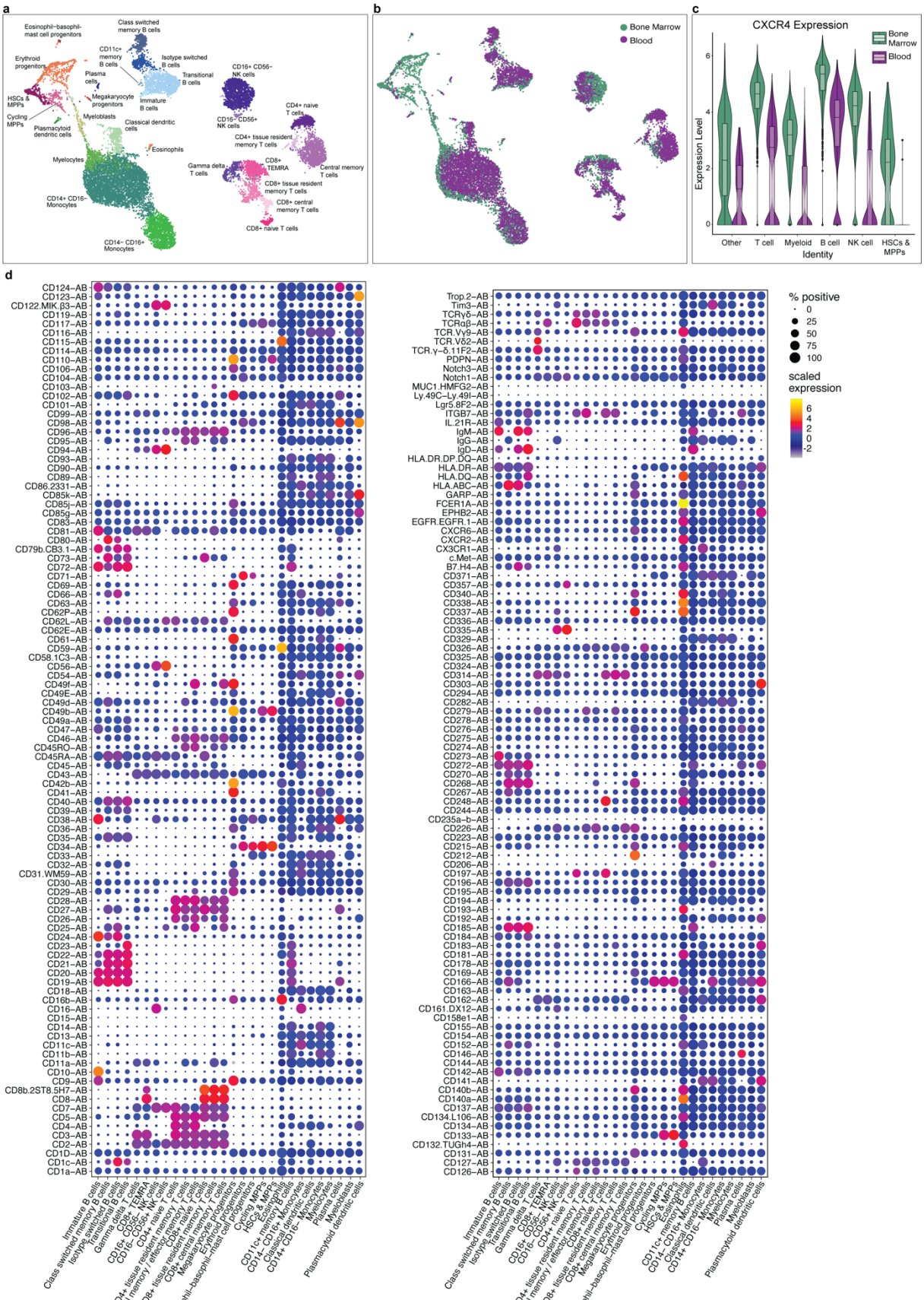


Figure S2. A proteo-genomic single-cell map of 197 surface markers in human bone marrow and blood. Related to Figure 1. **a.** uMAP projection, integrated across two samples from the same individual (blood and BM) and two data modalities (surface markers and RNA). Cells are colored

by cell type and annotation was performed as described in Supplementary Note 5 and Supplementary Table 4. N= 15281 cells. **b.** Split samples (blood and bone marrow) are displayed. **c.** Violin plot depicting the expression of the bone marrow homing receptor CXCR4 on matching cell types of the blood and bone marrow. **d.** 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.

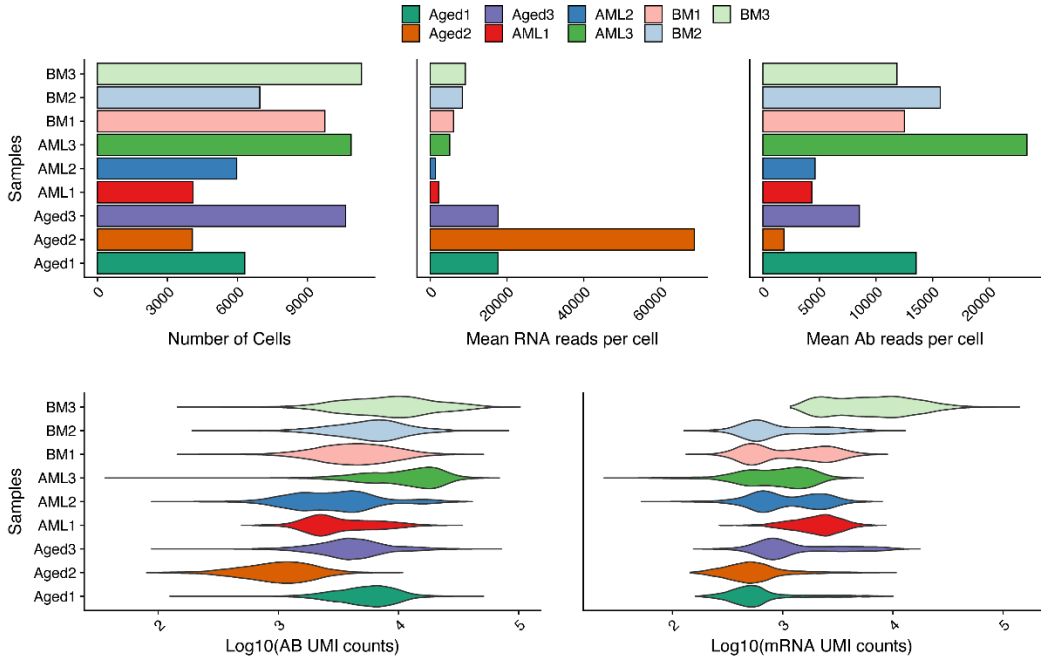


Figure S3. Sequencing statistics. Related to Figure 1. Plots depict the number of cells passing filters, the number of surface molecules per cell observed, the number of mRNA molecules per cell observed, and the sequencing depth on the surface and mRNA level, respectively. Note that targeted mRNA sequencing was performed as described in the main text.

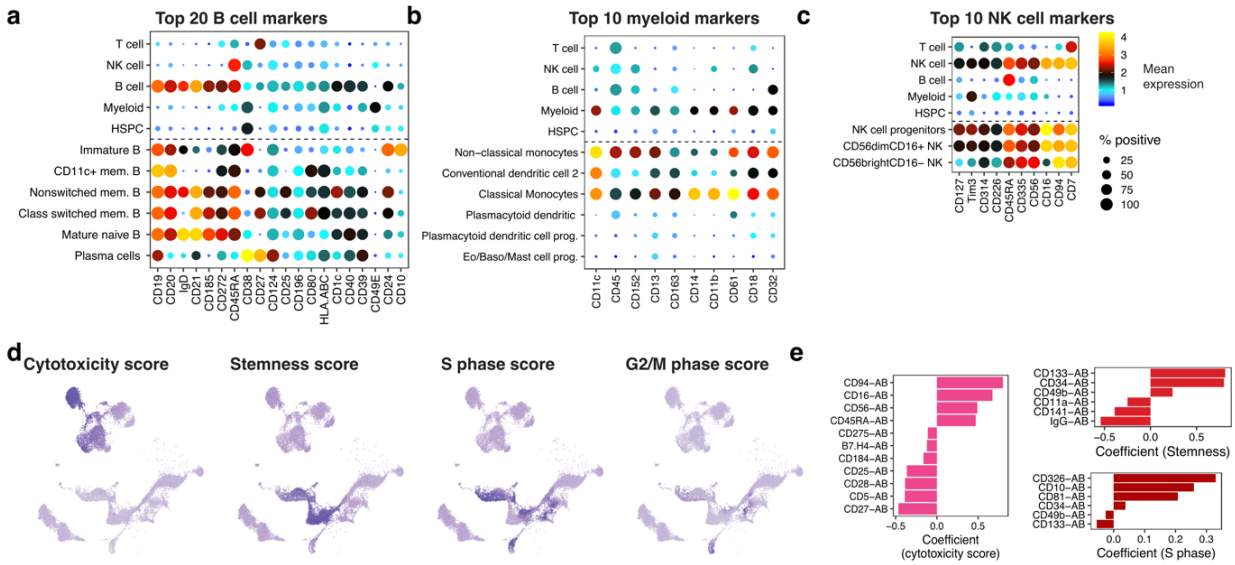


Figure S4. Markers of cell types and biological processes. Related to Figure 2. **a-c.** Dot plot depicting the expression of the 10-20 surface markers with the highest fraction of variance explained by B cell subtype (a), myeloid subtype (b) and NK cell subtype (c). Color indicates mean normalized expression, 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. **d.** uMAPs highlighting the scores for various biological processes, as computed using the gene lists from Supplementary Table 9. **e.** 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 9 for the gene lists used for calculating these scores.

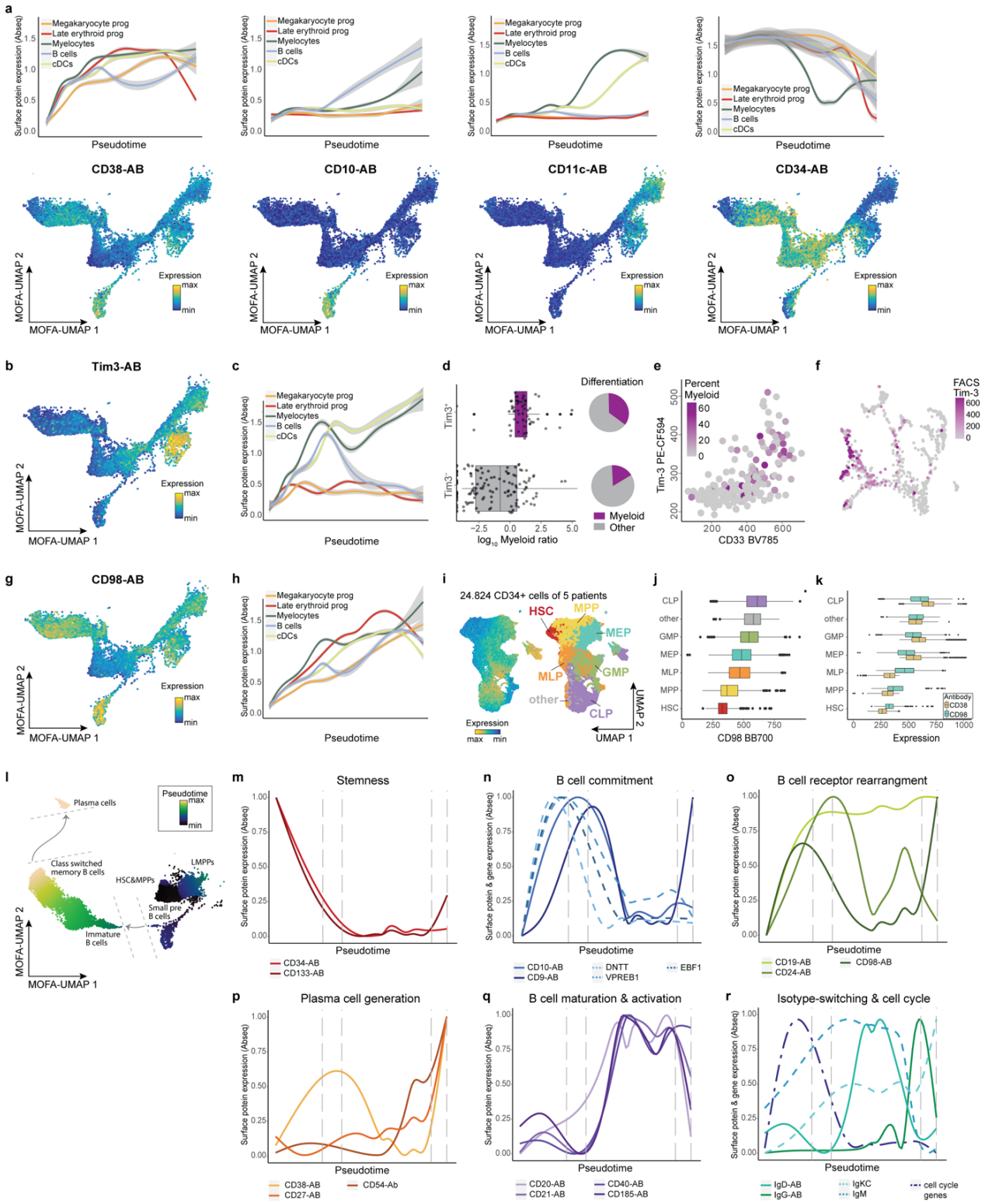


Figure S5. Surface markers associated with HSC and B cell differentiation. Related to Figure 2 and 3. **a.** top panels: Line plots depicting normalized CD38, CD10, CD11c and CD34 surface protein expression (Abseq data) smoothened over the different pseudotime trajectories illustrated in Figure 3a. bottom panels: uMAP display of CD34+ HSPCs, highlighting the surface expression of each corresponding marker. **b.** uMAP display highlighting the normalized Tim3 surface protein expression (Abseq data). **c.** Same as in *a* top panels, Tim3 surface expression is displayed. **d.** Box plots depicting the ratio of myeloid cells produced in single-cell cultures in relation to the Tim3 expression of the founder cell (n=190, single cell derived colonies). Pie charts summarize the relative fraction of wells

showing myeloid output (CD45+ CD235- CD66b+ and CD45+ CD45+ CD235-CD14+) within Tim3+ or Tim3- founder cells. **e.** Scatter plots depicting the fraction of myeloid cells produced in single cell cultures in relation to the CD33 and Tim3 expression of the founder cell. **f.** same uMAP display from Figure 3c; surface protein expression (FACS data) of Tim3 is shown. **g-h.** Like b and c, except CD98 expression is shown. **i.** uMAP display of CD34+ cells from five healthy donors analyzed with a 12-color FACS stem and progenitor panel (see Supplemental Table S8). Left panel shows CD98 surface expression, right panel shows assignment of individual gates to the uMAP following a classical gating strategy, i.e. 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+, other: cells that did not fall into any of the mentioned gates. **j.** Boxplots showing CD98 expression in individual cell populations mentioned in *i*. **k.** Boxplots showing co-expression of CD98 and CD38 surface markers in respective cell populations. **l.** like Figure 3a, uMAP plot depicting the pseudotime score along the B cell differentiation trajectory emanating from CD34+ HSCs & MPPs and LMPPs. **m-r.** Line plots depicting surface protein expression (Abseq data) representative for different indicated biological processes smoothened over the B cell pseudotime trajectory.

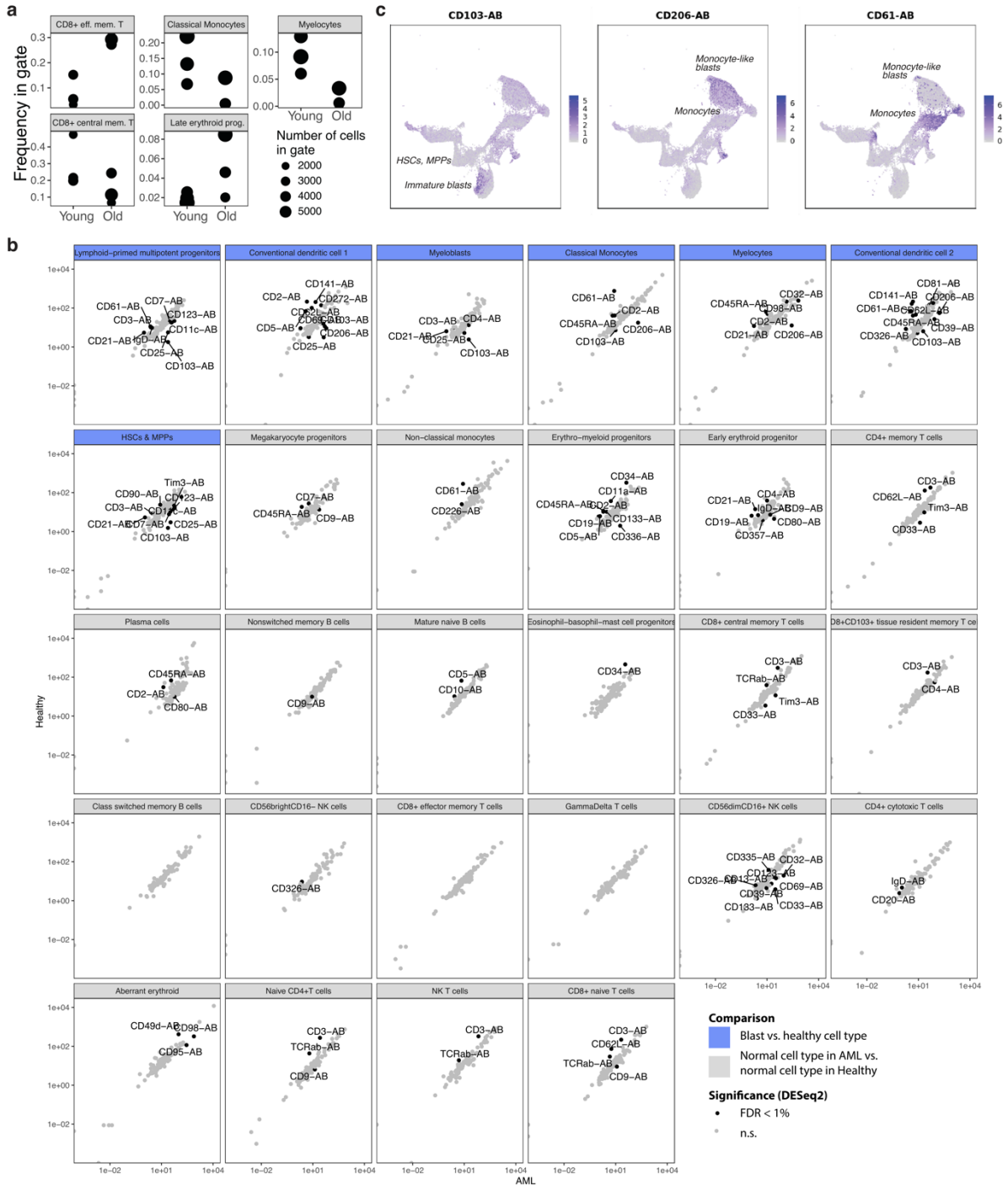


Figure S6. Changes in surface protein expression and cell type abundance induced by ageing and leukemia. Related to Figure 4. **a.** Frequency of selected cell types in young and old individuals. Only the cell types with the most significant changes are shown, see Methods, section '*Changes in cell type abundance between experimental groups*'. **b.** For every population, surface marker expression in AML (x axis) is compared to surface marker expression in healthy individuals (y axis). Cell types were defined using a random forest as described in main Figure 4d. **c.** uMAPs highlighting the expression of CD103, CD206 and CD61 as examples for surface markers with differential expression between blasts and most similar healthy cells.

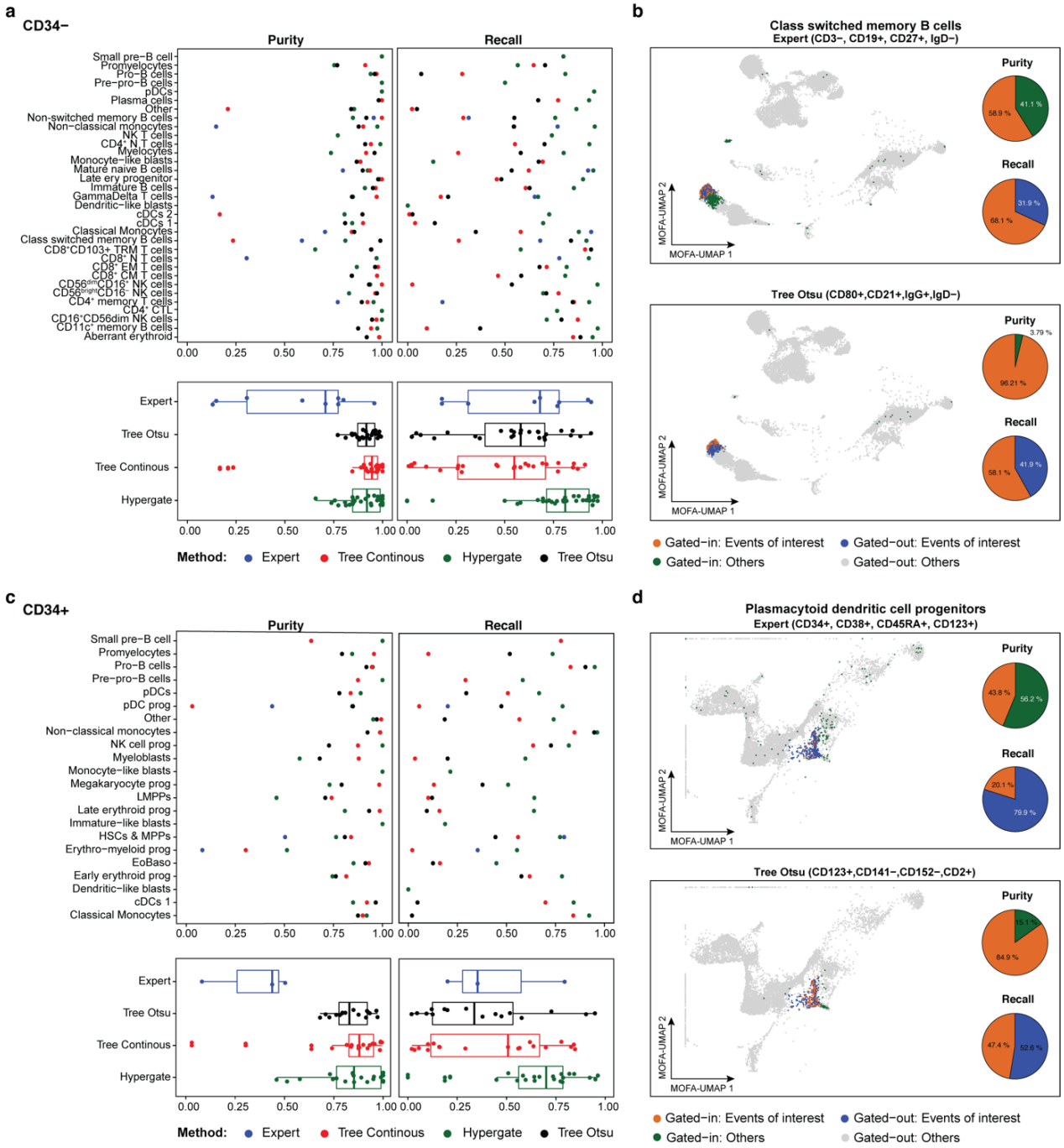


Figure S7. Comparison of data-defined and state-of-the-art (expert-defined) gating schemes. Related to Figure 5. **a.** Performance of different methods used for the definition of gates of CD34- populations. Gates for each cell type were defined from CD34- Abseq data using the following strategies: Blue dots correspond to gates that were manually set by an expert based on the current state of the art for purifying the cell type of interest (Supplementary Table 7). Green dots correspond to that were set using the hypergate algorithm (Becht et al., 2019). Red dots correspond to gates that were set using a decision tree. Black dots correspond to gates that were set using a decision tree with pre-defined thresholds, see Methods, section *Data-driven identification of gating schemes*. For each gating scheme, precision (purity) and recall were calculated. Box plot summarizes precision and recall estimates across all cell types. **b.** Illustration of the calculation of precision (purity) and recall for 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. Green dots correspond to other cells located inside the selected gate (false positives) and grey dots to other cells located outside the gate (true

negatives). Pie charts indicate precision (purity) and recall. Top panel: An expert defined state of the art gating scheme (CD3-CD19+CD27+IgD-) is shown. Bottom panel: A data defined gating scheme (CD80+CD21+IgG+IgD-) is shown. **c.** Like a, except that CD34+ populations are shown. **d.** Like *b*, except that gating schemes to define plasmacytoid dendritic cell progenitors are shown.

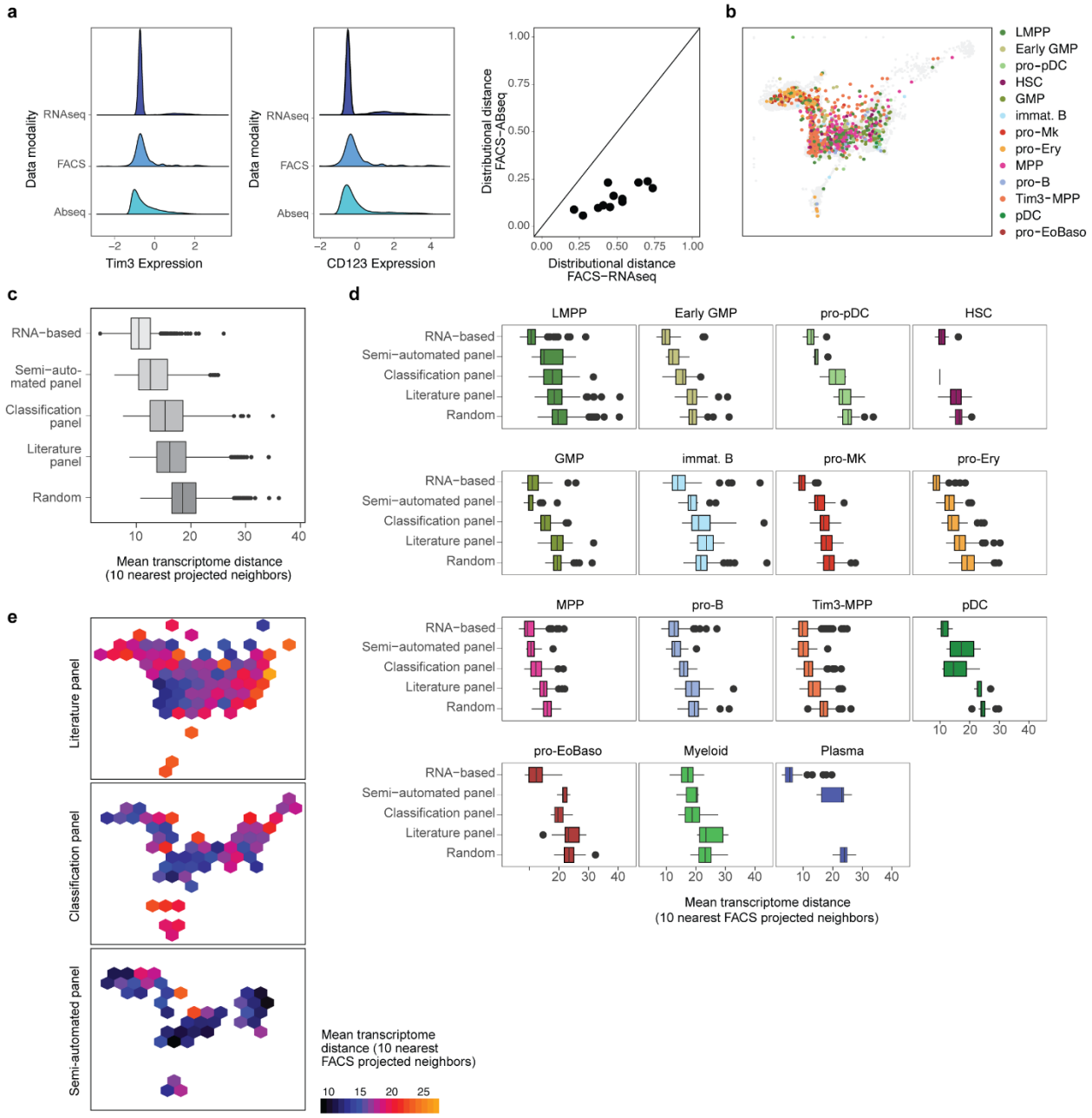


Figure S8. Projection and classification of cytometry data using a single-cell proteo-genomic reference. Related to Figure 7. **a.** Distribution of normalized, scaled expression values of Tim3 (left panel) and CD123 (central panel) measured by scRNA-seq, Abseq, and FACS. Right panel: Scatter plot depicts 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 Figure 6f is shown. **b.** Projection of indexed Smart-seq2 data onto a reference uMAP, like in main Figure 7b. Here, the dataset from Velten et al., 2017, employing a commonly used gating scheme from Doulatov et al. (Doulatov et al., 2010). (CD10, CD34, CD38, CD45RA, CD90, CD135) is used. **c-e.** Quantitative analyses of the similarity of gene expression in cells projecting to similar areas in the uMAPs on main Figure 7b and S8b. For every cell projected on the uMAP, the 10 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. For comparison, data was integrated based on gene expression values using Seurat v3, or 10 nearest neighbors were selected at random. **c.** Boxplot summarizing the distance across data integration strategies. **d.** Boxplots stratified by cell type (from Smart-seq2, see Supplementary Note

7) demonstrate that projection using the semi-automated panel performs close to an RNA-based integration in most cases. **e.** Hexagonal plot summarizing the projection accuracy for different regions of the uMAP.

REFERENCES

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature *404*, 193–197.

Al-Sabah, J., Baccin, C., and Haas, S. (2020). Single-cell and spatial transcriptomics approaches of the bone marrow microenvironment. Curr. Opin. Oncol. *32*, 146–153.

Baccin, C., Al-Sabah, J., Velten, L., Helbling, P.M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L.M., Trumpp, A., and Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. Nat. Cell Biol. *22*, 38–48.

Baron, C.S., Barve, A., Muraro, M.J., van der Linden, R., Dharmadhikari, G., Lyubimova, A., de Koning, E.J.P., and van Oudenaarden, A. (2019). Cell Type Purification by Single-Cell Transcriptome-Trained Sorting. Cell *179*, 527-542.e19.

Becht, E., Simoni, Y., Coustan-Smith, E., Evrard, M., Cheng, Y., Ng, L.G., Campana, D., and Newell, E.W. (2019). Reverse-engineering flow-cytometry gating strategies for phenotypic labelling and high-performance cell sorting. Bioinformatics *35*, 301–308.

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716-729.e27.

Van Dongen, J.J.M., Lhermitte, L., Böttcher, S., Almeida, J., Van Der Velden, V.H.J., Flores-Montero, J., Rawstron, A., Asnafi, V., Lécrevisse, Q., Lucio, P., et al. (2012). EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia *26*, 1908–1975.

Doulatov, S., Notta, F., Eppert, K., Nguyen, L.T., Ohashi, P.S., and Dick, J.E. (2010). Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. Nat. Immunol. *11*, 585–593.

Drissen, R., Thongjuea, S., Theilgaard-Mönch, K., and Nerlov, C. (2019). Identification of two distinct pathways of human myelopoiesis. Sci. Immunol. *4*.

Fagnoni, F.F., Vescovini, R., Mazzola, M., Bologna, G., Nigro, E., Lavagetto, G., Franceschi, C., Passeri, M., and Sansoni, P. (1996). Expansion of cytotoxic CD8+ CD28- T cells in healthy ageing people, including centenarians. Immunology *88*, 501–507.

Frenette, P.S., Pinho, S., Lucas, D., and Scheiermann, C. (2013). Mesenchymal stem cell: Keystone of the hematopoietic stem cell niche and a stepping-stone for regenerative medicine. Annu. Rev. Immunol. *31*, 285–316.

van Galen, P., Hovestadt, V., Wadsworth, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J., et al. (2019). Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell *176*, 1265-1281.e24.

Giladi, A., and Amit, I. (2018). Single-Cell Genomics: A Stepping Stone for Future Immunology

Discoveries. Cell *172*, 14–21.

Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-Wallscheid, N., Dress, R., Ginhoux, F., et al. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. Nat. Cell Biol. *20*, 836–846.

Görgens, A., Ludwig, A.K., Möllmann, M., Krawczyk, A., Dürig, J., Hanenberg, H., Horn, P.A., and Giebel, B. (2014). Multipotent hematopoietic progenitors divide asymmetrically to create progenitors of the lymphomyeloid and erythromyeloid lineages. Stem Cell Reports *3*, 1058–1072.

Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. *37*, 1458–1465.

Haas, S. (2020). Hematopoietic Stem Cells in Health and Disease—Insights from Single-Cell Multi-omic Approaches. Curr. Stem Cell Reports *6*, 67–76.

Haas, S., Trumpp, A., and Milsom, M.D. (2018). Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. Cell Stem Cell *22*, 627–638.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., et al. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. Cell *172*, 1091-1107.e17.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303–309.

Hua, P., Roy, N., de la Fuente, J., Wang, G., Thongjuea, S., Clark, K., Roy, A., Psaila, B., Ashley, N., Harrington, Y., et al. (2019). Single-cell analysis of bone marrow–derived CD34+ cells from children with sickle cell disease and thalassemia. Blood *134*, 2111–2115.

Jacobsen, S.E.W., and Nerlov, C. (2019). Haematopoiesis in the era of advanced single-cell technologies. Nat. Cell Biol. *21*, 2–8.

Kondo, M., Weissman, I.L., and Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell *91*, 661–672.

Laurenti, E., and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. Nature *553*, 418–426.

Loughran, S.J., Haas, S., Wilkinson, A.C., Klein, A.M., and Brand, M. (2020). Lineage commitment of hematopoietic stem cells and progenitors: insights from recent single cell and lineage tracing technologies. Exp. Hematol. *88*, 1–6.

Macaulay, I.C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S.A., and Cvejic, A. (2016). Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. Cell Rep. *14*, 966–977.

Nestorowa, S., Hamey, F.K., Pijuan Sala, B., Diamanti, E., Shepherd, M., Laurenti, E., Wilson, N.K., Kent, D.G., and Göttgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood *128*, e20–e31.

Notta, F., Zandi, S., Takayama, N., Dobson, S., Gan, O.I., Wilson, G., Kaufmann, K.B., McLeod, J., Laurenti, E., Dunant, C.F., et al. (2016). Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. Science (80-.). *351*.

Papalexi, E., and Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. Nat. Rev. Immunol. *18*, 35–45.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell *163*, 1663–1677.

Pei, W., Feyerabend, T.B., Rössler, J., Wang, X., Postrach, D., Busch, K., Rode, I., Klapproth, K., Dietlein, N., Quedenau, C., et al. (2017). Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature *548*, 456–460.

Pellin, D., Loperfido, M., Baricordi, C., Wolock, S.L., Montepeloso, A., Weinberg, O.K., Biffi, A., Klein, A.M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat. Commun. *10*.

Perié, L., Duffy, K.R., Kok, L., De Boer, R.J., and Schumacher, T.N. (2015). The Branching Point in Erythro-Myeloid Differentiation. Cell *163*, 1655–1662.

Peters, M.J., Joehanes, R., Pilling, L.C., Schurmann, C., Conneely, K.N., Powell, J., Reinmaa, E., Sutphin, G.L., Zhernakova, A., Schramm, K., et al. (2015). The transcriptional landscape of age in human peripheral blood. Nat. Commun. *6*.

Rodriguez-Fraticelli, A.E., Wolock, S.L., Weinreb, C.S., Panero, R., Patel, S.H., Jankovic, M., Sun, J., Calogero, R.A., Klein, A.M., and Camargo, F.D. (2018). Clonal analysis of lineage fate in native haematopoiesis. Nature *553*, 212–216.

Schaum, N., Karkanias, J., Neff, N.F., May, A.P., Quake, S.R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M.B., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367–372.

Shahi, P., Kim, S.C., Haliburton, J.R., Gartner, Z.J., and Abate, A.R. (2017). Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci. Rep. *7*, 1–12.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods *14*, 865–868.

Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. *20*, 257–272.

Szabo, P.A., Levitin, H.M., Miron, M., Snyder, M.E., Senda, T., Yuan, J., Cheng, Y.L., Bush, E.C., Dogra, P., Thapa, P., et al. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat. Commun. *10*.

Takeuchi, A., and Saito, T. (2017). CD4 CTL, a cytotoxic subset of CD4+ T cells, their differentiation and function. Front. Immunol. *8*.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331–338.

Tusi, B.K., Wolock, S.L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J.R., Klein, A.M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. Nature *555*, 54–60.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281.

Wilson, A., Laurenti, E., Oser, G., van der Wath, R.C., Blanco-Bose, W., Jaworski, M., Offner, S., Dunant, C.F., Eshkind, L., Bockamp, E., et al. (2008). Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. Cell *135*, 1118–1129.

Zheng, S., Papalexi, E., Butler, A., Stephenson, W., and Satija, R. (2018). Molecular transitions in early progenitors during human cord blood hematopoiesis. Mol. Syst. Biol. *14*.