Check for updates

Science

RESEARCH ARTICLES

Cite as: B. Furtwängler *et al*., *Science* 10.1126/science.adr8785 (2025).

Mapping early human blood cell differentiation using single-cell proteomics and transcriptomics

Benjamin Furtwängler[1,2,5]†, Nil Üresin [1,2,5]†, Sabrina Richter[6,7]†, Mikkel Bruhn Schuster[1,2], Despoina Barmpouri[1,2], Henrietta Holze[5], Anne Wenzel[1,2], Kirsten Grønbæk[2,3,4], Kim Theilgaard-Mönch[1,2,3,4], Fabian J. Theis[6,7]*, Erwin M. Schoof[5]*, Bo T Porse[1,2,3]*

[1]The Finsen Laboratory, Rigshospitalet, Copenhagen University Hospital, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [2]Biotech Research and Innovation Centre, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [3]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark. [4]Department of Hematology, Rigshospitalet, Copenhagen University Hospital, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [5]Department of Biotechnology and Biomedicine, Technical University of Denmark, Lyngby, Denmark. [6]Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany. [7]TUM School of Life Sciences Weihenstephan, Technical University of Munich, Freising, Germany.

***Corresponding author. Email: bo.porse@finsenlab.dk (B.T.P.); erws@dtu.dk (E.S.); fabian.theis@helmholtz-munich.de (F.T.)**

Single-cell transcriptomics (scRNA-seq) has facilitated the characterization of cell state heterogeneity and recapitulation of differentiation trajectories. However, the exclusive use of mRNA measurements comes at the risk of missing important biological information. Here we leveraged recent technological advances in single-cell proteomics by Mass Spectrometry (scp-MS) to generate an scp-MS dataset of an in vivo differentiation hierarchy encompassing over 2500 human CD34+ hematopoietic stem and progenitor cells. Through integration with scRNA-seq, we identified proteins that are important for stem cell function, which were not indicated by their mRNA transcripts. Further, we showed that modeling translation dynamics can infer cell progression during differentiation and explain substantially more protein variation from mRNA than linear correlation. Our work offers a framework for single-cell multi-omics studies across biological systems.

Single-cell transcriptomics (scRNA-seq) of the human hematopoietic stem and progenitor cell (HSPC) compartment has changed our understanding of the hematopoietic system by revealing the gradual transition through transcriptome states during HSPCs differentiation toward functional blood cells (*1*–*4*). Capturing these transitional cell states from single-cell snapshots has enabled the in-depth characterization of cell state heterogeneity and recapitulation of differentiation trajectories (*5*, *6*). Furthermore, modeling RNA velocity from mRNA splicing kinetics using scRNA-seq data has enabled the study of temporal dynamics during differentiation (*7*–*9*), although its application to hematopoiesis has faced challenges due to time dependent kinetic rates (*10*, *11*). However, albeit very successful, these methods have mostly been based on the transcriptome, while the actual workhorse for cellular function, the proteome, has not been widely accessible for single-cell analysis. The importance of proteome characterization has been shown in many multi-omics studies, encompassing mRNA and protein measurements on bulk population level, by revealing that transcript abundance can be a poor proxy for protein abundance (*12*–*16*) and that the proteome can reveal more phenotypical information (*17*, *18*). Furthermore, these bulk multi-omics datasets have enabled the modeling of translation kinetics during cell differentiation (*16*), revealing the temporal dynamics between the two modalities (*19*–*21*). On single-cell level, protein measurements have mostly been limited to targeted antibody-based cell surface epitope assays such as CITE-seq, which were nonetheless instrumental to resolve cell states that were inaccessible with mRNA measurements (*3*, *22*, *23*), and even enabled temporal modeling (*24*). Thus, direct, untargeted and high-throughput measurements of proteins in single cells have the potential to further improve our understanding of the biological systems by providing more phenotypical information about observed cell states and by revealing temporal dynamics in the system.

Recent technological advances have enabled measuring protein abundance in individual cells using single-cell proteomics by Mass Spectrometry (scp-MS) (*25*). Specifically SCoPE-MS (*26*) uses sample multiplexing via isobaric tags enabling measuring 14 to 16 single-cells in a single mass spectrometry (MS) run. The addition of a carrier channel provides more peptide ion copies to the MS to facilitate peptide identification. Based on this method, we have previously introduced our 384-well plate-based scp-MS workflow (*27*, *28*), where single cells are isolated via fluorescence-activated cell sorting (FACS), which enables the incorporation of antibody-based surface markers to isolate or annotate populations of interest.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

While scp-MS has reached a technological level that allows for the quantification of a considerable part of the proteome, reports describing its application to in vivo differentiation hierarchies remain scarce. This poses the questions whether scp-MS can be upscaled for robust data generation of complex biological systems like the human HSPC compartment and whether it captures phenotypical information that is complementary to mRNA-based approaches. This would make scp-MS well-suited for single-cell multi-omics analyses, if additionally, integration of scRNA-seq and scp-MS data are demonstrated.

Here, we present a scp-MS dataset encompassing over 2500 cells that recapitulates the HSPC hierarchy through relative protein abundance changes. We subsequently computationally integrate scp-MS with scRNA-seq, generating a single-cell multi-omics dataset. We discover proteins involved in stem cell function that are not apparent on mRNA level and highlight processes during lineage specification that are better described on protein level. Finally, we quantify temporal dynamics between the modalities and show that these can be modeled via a translation kinetics model that we termed scProtVelo.

Results

A single-cell proteomics dataset of FACS isolated human HSPCs

We isolated CD34+ bone marrow (BM) HSPCs from six healthy donors and stained the cells with a surface marker panel, enabling the isolation of defined subpopulations via FACS (*2*) (Fig. 1A). With the aim of measuring protein abundance in single cells, we applied our scp-MS workflow (*27*, *28*) starting with the deposition of single cells into the wells of a 384-well plate, while recording the FACS parameters for later analysis (Fig. 1B). We applied two different sorting strategies resulting in two kinds of multiplexed sets for scp-MS. First, we used the classical gating strategy to isolate specific subpopulations and distributed them evenly across the 14 single cells per multiplexed set ("Enriched Populations"). Secondly, we sorted any CD34+ cell resulting in the random sampling of the entire HSPC population ("Total HSPC"). For the MS measurement of each multiplexed set, we used our recently described real-time search-assisted acquisition (RETICLE) (*28*). For downstream analysis, we applied our computational workflow SCeptre (*27*), which expands on the functions of Scanpy (*29*) for the processing of scp-MS data, including cell filtering, batch correction, and data normalization (fig. S1, A to D, see methods).

SCeptre processing yielded a dataset with over 2900 proteins quantified across 2500 cells, albeit with variable data completeness (Fig. 1C) and on average 68% missing values per cell, which is comparable or even better than previous SCoPE-MS results (*30*). From the UMAP (*31*) embedding of the protein abundance data, we did not observe any batch effects based on the TMTpro channel, individual donors or plates (fig. S1, B to D), indicating that technical and biological (donor) variations were largely removed. We further validated this using principal component regression (*32*) which revealed that only very low fractions of the variance could be explained by MS run, TMTpro label, donor or age (fig. S2A). Overlaying the cell identities derived from the two different cell-sorting strategies onto the UMAP embedding revealed clustering and branching of the sorted (enriched) populations based on the HSPC differentiation hierarchy (Fig. 1D). Hematopoietic stem cells (HSCs) and multipotent progenitors (MPPs) were primarily located at the top of the hierarchy as a mixed and heterogeneous population, which has also been observed in scRNA-seq before (*33*). Downstream of HSCs/MPPs followed lymphoid-primed multipotent progenitors (LMPPs), which tended to be located upstream of granulocyte-macrophage progenitors (GMPs) and common lymphoid progenitors (CLPs) branches. Some megakaryocyte-erythroid progenitors (MEPs) clustered together with GMPs (further investigated below). Moreover, we observed cells sorted as common myeloid progenitors (CMPs) to be clustered together with both myeloid branches but also forming a separate cluster that we later determined to be basophil, eosinophil, mast cell progenitors (BaEoMa) (see below). This aligns well with findings that immunophenotypical CMPs constitute a heterogeneous population (*34*–*36*) and with functional studies demonstrating BaEoMa potential within the immunophenotypic CMP population (*37*–*39*). Taken together, these results showed that by SCeptre processing of the scp-MS data we successfully removed batch effects from technical covariates, without the need to define them.

Recording the FACS parameters during sorting enabled us to overlay that information onto the UMAP together with the total scp-MS protein signal (Fig. 1E). This demonstrated the alignment of FACS markers and the scp-MS UMAP embedding. Cells differed in their scatter properties, indicative of differences in cell size and granularity. The differences in FSC-A correlated well with the total protein signal (Fig. 1E), indicating a higher protein content for larger cells. Furthermore, our data suggest that Endomucin might serve as an alternative surface marker for enriching LT-HSCs (*40*), since commonly used markers CD90 and CD49f (fig. S2B) exhibited lower or no specificity in our dataset.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

We subsequently used the recorded FACS parameters to perform retrospective computational gating (Fig. 1F). We refined the gating strategy in the CD38+CD10- fraction (lower right panel) by moving the CMP gate down to CD45RA-CD123dim to incorporate a CD45RA-CD123+ gate for BaEoMa (*39*). Furthermore, we moved the GMP gate to CD123-/dim to incorporate a CD45RA+CD123+ population of progenitors for dendritic cells (pDC) (*2*). Overlaying these annotations onto the proteomics-derived embedding revealed well-clustered populations in line with the HSPC differentiation hierarchy (Fig. 1G).

We observed that some MEPs clustered together with the GMP branch (Fig. 1, D and G). MEPs are purified based on the absence of specific markers and this population has been shown to be impure in functional assays (*34*, *41*). To improve the MEP annotation, we generated an additional scp-MS dataset with 922 cells, incorporating CD71 and CD110/BAH-1 FACS markers to further specify the erythroid and megakaryocyte progenitors within the conventional MEP gate (*34*) (Fig. 1H and fig. S3). The combination of CD71 and BAH-1 clearly delineated MEPs, allowing the separation of those MEPs that clustered with GMPs (Fig. 1, H and I). We found proteins associated with MEP differentiation, such as S100A4 and RAP1B, to be highly abundant in the BAH-1 positive MEPs, and proteins associated with granulocytic differentiation, such as ELANE and MPO, to have high levels in some of the putative false MEPs with low BAH-1 levels (Fig. 1J). This observation corroborates the proximity of the BAH-1 negative cells in the MEP gate to the GMP branch.

Overall, our scp-MS data recapitulated the human HSPC hierarchy and highlighted certain limitations of low-dimensional FACS marker combinations in resolving the cellular heterogeneity.

Unsupervised clustering reveals distinct cellular differentiation stages and functional protein covariance

Next, we used unsupervised clustering of the scp-MS data to group cells into 11 clusters (Fig. 2A) that were mainly named after their most prevalent population, based on FACS information and protein abundance. Note the switch from MEP to the more common annotation of erythroid-megakaryocyte progenitors (EMP). Furthermore, we named the first two clusters downstream from HSCs Progenitors 1 and 2, as they should contain the early bifurcation of HSPCs into LMPPs and EMPs, which could not be resolved fully with this clustering approach. Next, we evaluated the quantitative accuracy of the scp-MS data by comparing fold-changes generated from cluster averages (pseudobulks) to an external bulk proteomics dataset derived from human HSPCs (HSC, MEP, CMP, GMP) (*42*) (Fig. 2, B and C). We also included the computationally gated scp-MS populations from Fig. 1G, as they might better resemble the cell composition of the bulk-sorted populations. Pearson correlation values between protein log2 fold-changes of scp-MS-based populations and the four bulk populations revealed that scp-MS populations clustered together by similarity and had the highest correlation coefficients with their corresponding bulk population. The GMP and CMP bulk populations had slightly lower and much lower correlations respectively, likely due to minor differences between the bulk and single-cell FACS gating. Furthermore, we used the bulk data to benchmark cell-wise normalization based on total signal versus median-ratio normalization and found the latter to perform slightly better (fig. S4).

To evaluate the purity and integrity of the scp-MS clusters, we overlaid cluster labels onto the computational gating scheme from Fig. 1F (Fig. 2D and fig. S5). Most clusters were well restricted to their respective FACS gate. Only LMPPs, Progenitors 2 and, to some degree, Progenitors 1 were distributed over multiple FACS gates, indicating that these populations contain a mixture of differentiation stages. This highlights the issues of using a mainly FACS-based approach for cell annotation in a cluster-based context and in the absence of a reference map. Therefore, the UMAP annotation in Fig. 2A should be considered as an intermediate toward the more refined annotation of the scp-MS data presented later. We also note that erythroid-megakaryocyte progenitors (EMP) cluster should give rise to another branch containing megakaryocyte progenitors; however, we did not observe any putative branch, which might be due to their rarity in BM or their depletion through the inclusion of CD61 megakaryocyte marker in the lineage cocktail in FACS.

We next analyzed differential protein abundance to identify marker proteins for the 11 clusters (Fig. 2, E and F). For example, HSCs showed high levels of H1.0 linker histone (H1F0), indicating a compact chromatin structure and quiescent cell state (*43*). Erythroid differentiation was marked by high levels of carbonic anhydrase I (CA1) (*44*). Basophil-eosinophil-mast progenitors (BaEoMa) were marked by CD123 surface expression and high protein levels of Proteoglycan 2 (PRG2) and Charcot-Leyden Crystal Galectin (CLC), components of eosinophil granules (*45*). Granulocyte-monocyte-dendritic cell progenitors (GMDP) displayed high levels of AZU1, PRTN3, ELANE, CTSG and MPO that are found in azurophil granules, the first granules to be

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

formed during neutrophil differentiation (*46*). Progenitors for monocytes and common dendritic cells (MDP) branched off from the GMDP cluster, marked by high levels of LGALS1 and PLD4, which could also be observed in scRNA-seq data (*47*) (fig. S6). Furthermore, MDPs showed increased surface expression of CD123 and high levels of proteins like lysozyme (LYZ) and HLA-DRB1 (*3*). The cluster CLP/pre-pDC that likely contained a mixture of common lymphoid progenitors (CLP) and progenitors for plasmacytoid dendritic cells (pre-pDCs) shared markers with MDPs and progenitors for B-cells (pB). This is in line with the similarity of dendritic cell progenitors derived from myeloid or lymphoid lineages (*48*), however, they could be differentiated from MDPs by the lower levels of LYZ (*3*). Finally, the cluster that contained B cell progenitors (pB) was marked by surface expression of CD10 and high protein levels of topoisomerase IIβ (TOP2B), which is essential for B cell development (*49*). Thus, although HSPC differentiation is a continuous process, protein and FACS data are in good concordance on distinct abundance changes during lineage commitment.

To investigate protein covariation across HSPC differentiation, we performed unsupervised clustering of the most variable protein abundance profiles across clusters and performed gene ontology term enrichment analysis (*50*) on each protein cluster (Fig. 2, G and H). Protein cluster 3 was highly abundant in HSCs and showed enrichment in glycolysis, likely reflecting their reliance on glycolysis instead of oxidative phosphorylation (*51*). Furthermore, HSCs showed high levels of proteins related to the response to oxidative and chemical stress, suggesting that these proteins are used to remove such stressors. The enriched terms found in other protein clusters related to the clusters' high abundance in certain populations (see supplementary text).

Overall, our observations captured a diverse range of biological processes, adding further validity to the biological relevance of our dataset and highlighting the potential to analyze protein covariation at scale using scp-MS data.

Single-cell multi-omics analysis of transcriptomics and proteomics of the human HSPC compartment

With the aim to perform an integrated analysis of mRNA and protein abundance in the HSPC compartment, we generated a CITE-seq (*22*) dataset of human HSPCs (Fig. 3A). We integrated data from four healthy adult humans using totalVI (*52*), resulting in a dataset of 9086 cells. From a UMAP embedding of the mRNA expression data, we did not observe any batch effects based on specific individuals (fig. S7A), indicating that technical and biological variation attributable to any specific donor sample were largely removed. We further validated this using principal component regression (*32*) which revealed that only very low fractions of the variance could be explained by donor or age, especially in the batch corrected scVI-modeled (*53*) mRNA layer (fig. S7B). We annotated cells based on mRNA expression with a reference mapping from Azimuth (*4*) and, similar to the scp-MS data, subsetted GMPs into GMDP and MDP (fig. S7C). The cell annotation in combination with the UMAP embedding, PAGA (*54*) connectivity, and surface marker expression indicated a good recapitulation of the human HSPC hierarchy (Fig. 3A and fig. S7, D and E). Furthermore, while cell cycle phase prediction resulted in an expected phase distribution across differentiation stages (fig. S7F), it also revealed strong cell cycle-dependent clustering of the UMAP embedding, which likely explains the unexpected connectivity between Early Eryth and LMPPs in the same phase. We also observed a very wide range of total mRNA counts and detected genes per cell differentiation stage, with HSCs having on average 3.45 times fewer counts and 2.1 times fewer detected genes compared to Late Eryth, in line with the low transcriptional activity in HSCs (*55*) (fig. S7G). Finally, investigating mRNA expression changes across populations revealed known markers (fig. S8). Since totalVI returns modeled and denoised mRNA expression values, similar to scVI (*53*), we benchmarked the fold-change accuracy using previously published bulk mRNA data (*42*) and found that scVI-modeled and median-ratio normalized mRNA data showed higher correlation with the bulk data compared to the count-based data and similar centering around zero as the bulk data (fig. S9). Nonetheless, we note that the scVI-modeled values should be used with caution as they represent only the mean of a modeled distribution with potentially high uncertainty.

We integrated the scp-MS and CITE-seq datasets with GLUE (*56*) (Fig. 3B), a variational autoencoder (VAE)-based tool designed for unpaired single-cell multi-omics data integration. The resulting joint latent space successfully integrated "mRNA" cells from the CITE-seq dataset and "protein" cells from the scp-MS dataset, with a silhouette score of 0.03. Figure 3C shows the UMAP embedding of the resulting joint latent space with "mRNA" cell labels transferred to the "protein" cells via nearest neighbor matching. Biological variation from the individual modalities was preserved in the integrated embedding as metrics on cell type separation (*32*) were comparable before and after integration for both modalities (Fig. 3D and fig. S10A). Moreover, we found many protein-level markers to agree

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

with mRNA expression (fig. S10B). Overlaying the new labels onto the protein UMAP revealed a more nuanced clustering and annotation compared to the initial annotation (compare Fig. 3E to Fig. 2A). Further, we compared the cell identity of "protein" cells based on the scp-MS clusters to the transferred labels and indeed found that labels were not transferred across lineage branches (Fig. 3F). Furthermore, "protein" cells in scp-MS clusters Progenitors 1 and 2 were assigned EMP and LMPP labels, which suggested that the model learned to separate these lineages at an early stage based on protein abundance. Indeed, in these clusters we found early differences in expression of surface marker CD45RA (FACS) expression and scp-MS-derived proteins like GAPDH, LCP1, and PSME1 that had higher levels in LMPP and its descendants, while the proteins SOD2 and ALDH1A1 had higher levels in EMPs (Fig. 3G and fig. S11). Moreover, the Azimuth label pre-mDC was transferred onto "protein" cells contained in the MDP cluster. Indeed, these cells could be separated from MDPs and pre-pDCs by distinctive protein abundance patterns of, for instance, CD123 (FACS) and scp-MS-derived proteins like S100A11, COTL1, and especially Lysozyme (LYZ) (Fig. 3H). Taken together, these analyses demonstrated the successful integration of mRNA and scp-MS data, and the resulting ability to transfer mRNA-derived labels onto "protein" cells which facilitated the refinement of cell states in scp-MS data.

Trajectory analysis on joint latent space recapitulates HSPC differentiation

We found that the joint latent space represented the biological latent cell states, and we then asked whether we could identify differentiation trajectories and relevant driving factors. We therefore used the integrated latent space as input for a trajectory analysis using cellrank (*5*, *6*). Here, we used the pseudotime kernel (Fig. 4A) that infuses directionality into a k-NN graph to compute a transition matrix. The resulting terminal states and the cell-specific fate probabilities indeed represented the expected differentiation trajectories in our HSPC dataset (Fig. 4A and fig. S12, A to C and supplementary text).

We subsequently investigated whether we could have obtained the same results on either one of the modalities independently. We found that the terminal states identified from the joint latent space were not readily recovered from either the CITE-seq or the scp-MS datasets only (Fig. 4B), even when more macrostates were considered (Fig. 4C). Furthermore, we compared the fate probabilities across the three models, which revealed a better separation of the pre-pDC and pre-mDC trajectories in the joint latent space compared to the individual modalities, specifically correct lineage assignments of CLP, pre-pDC, MDP, and pre-mDC improved from 86% to 91% on the RNA cells and from 65% to 95% on the protein cells (fig. S12D). Taken together, our results indicated that the cellrank model based on the joint latent space outperformed the models based on CITE-seq or scp-MS alone. These findings confirm our hypothesis that the two modalities provide complementary information about hematopoietic differentiation.

Subsequently, we used the trajectory analysis to investigate mRNA and protein abundance changes during differentiation. First, we analyzed HSC quiescence and differentiation by correlating mRNA or protein abundance to pseudotime in an early subset of cells (Fig. 4, D and E). This yielded a correlation-based ranking of mRNA or protein, with positive and negative correlations reflecting increases and decreases during differentiation, respectively. Secondly, we analyzed the process of lineage specification by correlating mRNA or protein abundance to fate probability for a particular lineage (Fig. 4D). This yielded a correlation-based ranking of mRNA or protein, with positive correlations reflecting a coordinated increase of abundance across single-cells during differentiation into that lineage compared to all other lineages. Importantly, the cell feature vectors pseudotime and fate probability were derived from the joint latent space and thus enabled the comparison of mRNA and protein abundance by comparing the correlation vectors between modalities. We envisioned this analysis to be suitable for biological systems characterized by gradual transitions as in HSPCs, in contrast to a cluster-based analysis, where fold-changes between pseudobulk clusters are compared between mRNA and protein level. The correlation-based analysis does not require clustering and is computed on single-cell level, emphasizing features that are changing along the axis of interest (pseudotime or fate probability). Additionally, with the fate probabilities from cellrank, it makes full use of the soft lineage assignments per cell, emphasizing lineage-specific changes.

To show the validity of this approach, we first analyzed only the mRNA expression using the "mRNA" cells. The analysis of HSC quiescence and differentiation on the early subset of cells revealed genes associated with HSCs like the transcription factors *ATF3*, which has been shown to prevent HSC exhaustion (*57*), or *KLF2*, which is important for embryonic stem cell self-renewal (*58*) (Fig. 4, E to G). In contrast, genes associated with stem cell differentiation were *CDK6*, a key regulator of exit from quiescence (*59*), the transcription factor *SOX4* (*60*), and the endonuclease *APEX1* (*61*), which are both important for progenitor formation. Additionally, we modeled these gene expression

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

profiles along pseudotime to enable a visualization (see methods and Fig. 4F). Subsequently, we used the correlation vector to perform a pre-ranked gene set enrichment analysis (GSEA) (*50*, *62*) (Fig. 4H and table S1). This revealed known processes for HSC maintenance such as antigen presentation via MHC II (*63*) (HLA-DRB1, HLA-DPA1, HLA-DRA), prostaglandin metabolism (*64*), nitric oxide metabolism (*65*) and reduction of reactive oxygen species. Processes enriched during HSC differentiation included nucleosome organization and increased expression of the ATP synthase, required for the switch from glycolysis to oxidative phosphorylation (*51*). Furthermore, the analysis of lineage specification revealed lineage-specific gene expression cascades and enrichment of relevant gene sets (figs. S13-17 and table S1). Taken together, these results indicated that the pseudotime represented the main axis of early stem cell differentiation and that cellrank recapitulated correct differentiation trajectories of the HSPC hierarchy and finally, supported the validity of our correlation-based analysis.

Protein level information reveals additional insights into hematopoietic stem cell quiescence and differentiation

Subsequently we continued with the analysis of protein abundance and directly compared the results to mRNA, by comparing correlation vectors between both modalities. For mRNA, we used the above calculated vectors based on counts, but also included correlation vectors based on the modeled and denoised scVI data. For protein, we repeated the analysis on the "protein" cells. The analysis of HSC quiescence and differentiation on the early subset of cells recapitulated the HSC protein markers we observed in Fig. 2F, while proteins associated with HSC differentiation were mainly related to DNA replication. Interestingly, on mRNA level, many of these genes showed only low correlations or even disagreements in the direction of change in abundance (Fig. 5A). Comparison between the full correlation vectors on mRNA and protein level revealed an overall weak correlation below 0.25 between these vectors (Fig. 5B). Moreover, this difference in the correlation vectors also resulted in the enrichment of gene sets that were not enriched on mRNA level (Fig. 5C and table S1). For example, we observed an enrichment of proteins involved in carbohydrate metabolism (glycolysis) in HSCs, which was not enriched on mRNA level, in line with typically low correlations of key metabolic pathways (*14*). Moreover, we found many proteins involved in the protection of HSCs from oxidative stress, including superoxide dismutases SOD1 and SOD2, Peroxiredoxin 1 (PRDX1), and TALDO1, a key enzyme of the pentose phosphate pathway that generates NADPH, which fuels the regeneration of reduced glutathione, a major cellular antioxidant (*66*). Apart from SOD1, a clear decrease of their levels during early HSC differentiation could only be detected via scp-MS and not scRNA-seq (Fig. 5B). We also found many proteins involved in chromatin structure among HSC-correlating factors, including the previously mentioned H1 linker histone H1F0, the H1-like protein HP1BP3, which is required for HSC self-renewal (*67*), the histone macroH2A1 (H2AFY), associated with HSC homeostasis (*68*), and the chromatin regulator HMGA1, important for HSC regenerative capacity (*69*). Again, the decrease of these chromatin regulators during early HSC differentiation was better described via scp-MS. Finally, we observed strong differences between mRNA and protein levels of nuclear lamin B2 (LMNB2). It has been shown that Lamin-A:B stoichiometry regulates nuclear stiffness (*70*), thus, scp-MS could potentially better inform about the nuclear morphology in single cells than scRNA-seq.

Finally, investigation of the proteins and their gene sets related to HSC differentiation revealed processes that were better described via scp-MS. For example, the increase of proteins involved in mRNA processing, RNA export from the nucleus, translation, and DNA replication (Fig. 5C and table S1). Notably, we investigated an early HSC cell-subset of the data (Fig. 5A), where detecting changes can be challenging due to the small amounts of protein and especially mRNA in these cells and due to the small effect sizes. This does not exclude the likely possibility that these terms would be enriched on mRNA level when comparing HSCs to more differentiated cells, where effect sizes would be greater. Furthermore, we focused on proteins that showed absolute correlations above 0.2, as these proteins represented the potential additional information that could be gained from the scp-MS data. It should be noted that the mRNA data also contained distinct findings, which could be relevant to follow-up on.

To validate the functional importance of selected proteins associated with the HSC state that have previously not been interrogated in human, we performed CRISPR/Cas9-mediated knockout of SOD1, TALDO1, and H1F0 in human CD34+ BM cells (Fig. 5D). Subsequently, we conducted a long-term culture-initiating cell assay (LTC-IC) as a surrogate for HSC activity upon knockout (Fig. 5, D to F) as well as colony forming unit (CFU) and 3-week stromal co-culture assays to further assess the functional consequence of deleting these factors. LTC-IC limiting dilution analysis (*71*) revealed a reduction in LTC-IC frequency

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

of TALDO1 and H1F0 deficient cells (1 in 42 and 1 in 80, respectively) compared to the AAVS1 control (1 in 14), suggesting a decreased ability to sustain long-term hematopoiesis upon knockout of these factors (Fig. 5F). Loss of TALDO1 and H1F0 also resulted in mild decrease in the number of CFUs, however the relative proportions of distinct colony types remained unaltered (Fig. 5, G and H), strongly suggesting a specific impact on the most immature cells, namely HSCs. This was further corroborated by stromal cell co-culture assays where loss of these factors only had minor effects on cell proliferation (fig. S18, A and B), as well as on the frequencies of CD34+ HSPCs and different myeloid cell populations (fig. S19). In contrast, SOD1-deficient HSPCs (a cytosolic superoxide dismutase) were largely unable to form colonies (only yielding small CFU-Gs, Fig. 5, G and H). This phenotype was associated with early proliferative cessation (fig. S18, A and B), a block in especially granulocytic differentiation in the stromal co-culture assay (fig. S19), and a dramatic reduction in LTC-IC frequency (1 in 153; Fig. 5F). However, since SOD1 KO HSPCs were nearly unable to form colonies, we could not determine the extent to which the latter is due to a specific HSC phenotype. Finally, we tested the functional consequences of knocking out SOD2 (a distant mitochondrial homolog of SOD1, 25% identify), which nearly replicated the SOD1 KO phenotype (Fig. 5, G and H, and figs. S18 and S19) except for the fact that SOD2-deficient HSPCs did not display a proliferative phenotype. Combined, these findings not only demonstrate the importance of SOD1 (and SOD2) for immature human HSPCs, but also highlight specific differences between mice and human, as Sod1 and Sod2 null mice are not associated with any prominent HSPC phenotypes (*72*, *73*).

In summary, the protein abundance trends during early HSC differentiation revealed additional information compared to scRNA-seq measurements and identified proteins and related pathways important for HSC function.

Proteomics-based trajectory analysis reveals protein level-specific functional covariation

Besides maintenance of the HSC state, we asked which protein-level programs are specific to particular lineages. For this, we analyzed lineage specification on protein level in the integrated dataset and compared it directly to the mRNA level (Fig. 6, A to G, and figs. S20-24). This revealed many of the protein markers from Fig. 2F but also new proteins. For example, the Late Erythroid trajectory was characterized by the retained high abundance of ALDH1A1 (Fig. 6, A and C), as described above.

Furthermore, we detected the transient abundance increase of markers for megakaryocyte progenitors like RAP1B (*4*), and subsequent increase in levels of erythroid related proteins like S100A4, PRDX2, and CASP6 (Fig. 6, A and C). Comparison between the full correlation vectors on mRNA and protein level for this lineage (Fig. 6D) revealed a much higher agreement of up to 0.7 compared to the HSC differentiation (r = 0.19 to 0.24). Nonetheless, differences between mRNA and protein correlation vectors propagated to differences in the GSEA, including processes involving ribosomal proteins (fig. S20A). In the GMDP lineage (Fig. 6, B, C, and E), we found gene sets related to protein glycosylation and endocytosis to be enriched only on protein level (fig. S20B and table S1). The latter set contained Beta-2 microglobulin (B2M), which, in our dataset, was not well correlated between mRNA and protein, since, for example, scp-MS measured an increase of B2M in the GMDP lineage, while the mRNA measurement indicated lower relative expression (Fig. 6, E and F). Nonetheless, the B2M protein abundance profile corresponded well to its complex partners, HLA-A and HLA-B in the major histocompatibility complex (MHC) class I (Fig. 6, F and G).

Based on these results, we hypothesized that proteins with low concordance with their respective mRNA template still covaried with functionally related proteins. To investigate this, we performed a joint mRNA-protein covariation analysis. Since the trajectory analysis yielded six different lineages, for each mRNA and protein molecule, the six correlations with lineage fate probabilities resulted in a vector of six values, representing the relative abundance across lineages. We focused on the 226 most variable proteins and their respective mRNA transcripts and calculated a correlation matrix between those 452 molecules. Then, for each protein, we ranked all molecules by correlation and calculated the rank of the corresponding mRNA. We found that for a large fraction of proteins, the respective mRNA molecule ranked very high, indicating that it was among the best predictors for the protein abundance profile across lineages (Fig. 6H, blue dots). Subsequently, for each protein, we calculated the rank of the highest-ranking protein that was part of the same KEGG pathway. This revealed that many protein abundance profiles could be well explained by a functionally related protein (Fig. 6H, orange dots). Furthermore, looking at the proteins with low mRNA concordance, and thus high mRNA ranks (right side of the plot), we indeed found that many protein abundance profiles could be better explained by functionally related proteins than their respective mRNA (Fig. 6H). Apart from the B2M-HLA-B pair, more examples were ALDOA and GPI, both part of glycolysis, MYL12B and MYH9, both part of the

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

myosin complex, HLA-DRA and HLA-DRB1, both part of the MHC class 2 complex, and lastly, protein pairs involved in protein processing in the endoplasmic reticulum GANAB: DDOST and PDIA6: HSPA5 (Fig. 6H). Finally, to confirm that the observed low concordances between mRNA and protein was not purely due to dataset-specific measurement errors, we compared our results to an external bulk dataset with mRNA and protein measurement of 59 breast cancer cell lines (*74–76*). Calculating the mRNA correlation rank for each protein in this dataset revealed a positive correlation of 0.35 with our ranks (Fig. 6I). These results supported a general trend for some genes to exhibit low correlations between mRNA and protein (*13*, *14*), however also indicating that there might be effects that are specific to the biological system or measurement technique.

scProtVelo models translation dynamics during differentiation

Finally, we hypothesized that having measured both mRNA and protein expression changes during HSPC differentiation could enable the study of temporal dynamics in that system, namely transcription, translation, and degradation of mRNA and protein. Conceptually similar to how the time delay between unspliced and spliced transcript abundance changes enables RNA velocity (*7–9*) to infer splicing kinetics, the temporal delay between mRNA expression changes and the corresponding protein response encodes information about its translation dynamics (Fig. 7A).

Toward this endeavor, we developed scProtVelo (single-cell protein velocity), a latent variable model that explains observed mRNA and protein abundance changes over time through learned gene-specific transcription, translation, mRNA, and protein degradation rates. To validate scProtVelos ability to infer translation dynamics, we applied it to subsets of cells from the erythroid and pre-mDC differentiation trajectories, where the assumption of constant kinetic rates per gene is reasonable, and compared the inferred gene states (activation vs. repression) to ground truth defined by differential expression results between early and late cells. We indeed found the predicted dynamics to closely reflect the observed data (Fig. 7B and fig. S25, A and B) and model predictions of the gene state to be substantially better than random permutations in both differentiation branches (Fig. 7C and fig. S25C), implying that temporally delayed protein abundance changes were detectable. Further, aggregating the inferred protein velocity information of the top 100 genes recapitulated the correct cell progression in both trajectories (Fig. 7D, left, and fig. S25D). In contrast, applying the standard RNA velocity workflow to the scRNA-seq cells of the erythroid trajectory resulted in the previously reported erroneous backflow in velocity vectors from Late to Early Erythroid progenitors (*10*, *11*) (Fig. 7D, right). Also, judging individual proposed cell to cell transitions by whether they coincide with an increase in pseudotime consolidates higher accuracy in trajectory inference based on translation modeling than on RNA velocity (fig. S25, E to I).

To quantify the impact of considering the temporal dynamics between mRNA and protein, we compared scProtVelo to a linear model, which cannot account for temporal delays. We considered 147 genes that were differentially expressed in early and late cells of the erythroid differentiation and incorporated pseudotime annotations as a prior on the inferred time into scProtVelo to ensure correct model fitting. Comparing the variance in protein abundance that can be explained through the corresponding mRNA expression, we found that scProtVelo resulted in more accurate modeling for almost all genes (except anti-correlated mRNA-protein pairs) (Fig. 7, E and F, and fig. S25B). Overall, accounting for translation dynamics when modeling protein abundance changes from mRNA levels led to a 40% increase in explained protein variance as compared to the simple assumption of a linear relationship (median R^2 values of 36% and 50% for the linear model and scProtVelo, respectively (Fig. 7G).

Taken together, our results demonstrate successful modeling of temporal gene expression dynamics from integrated scRNA-seq and scp-MS data, its potential for protein velocity-based trajectory inference, and its contribution to illuminating the often far from ideal mRNA-protein correlations.

Discussion

Here we presented an scp-MS dataset of the primary human HSPC compartment, encompassing over 2500 cells that recapitulated the HSPC hierarchy through relative protein abundance changes. The combination with FACS allowed us to validate the scp-MS-defined cell states and revealed heterogeneity in surface marker-defined populations. Furthermore, integration of scp-MS with CITE-seq resulted in a joint latent space that enabled label transfer between modalities and a trajectory analysis that showed better performance compared to each of the two individual modalities. Vectors derived from the joint latent space such as pseudotime or fate probabilities, enabled us to investigate differences between the modalities without the need for paired data. Here we found that the proteome level better described certain processes during HSC quiescence and differentiation, and we identified proteins which we

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

experimentally validated to be important for the functional behavior of the most immature HSPCs.

In contrast to the HSC state, we found that the concordance between mRNA and protein is much higher during lineage specification, which possibly could be partly explained by the higher effect sizes resulting in increased signal-to-noise measurements. Nonetheless, we were able to show that many of the proteins were not well explained by their mRNA profiles. Further, many of these proteins co-varied with functionally related proteins, highlighting the potential of scp-MS to reveal functional protein units and the importance of considering post-transcriptional regulation of protein abundances. For example, the abundance of Beta-2 microglobulin (B2M) could be regulated on the protein level by protection from degradation via stabilization in the MHC I complex (*77*).

Finally, we investigated the temporal dynamics between the modalities by modeling translation dynamics. We found that this model could better explain the relationship between modalities than a linear regression, highlighting its relevance for shedding light on the often low correlations between the modalities. Further, we successfully used the learned dynamics to infer directionality of the pre-mDC and erythroid differentiation process, which for the latter, has been challenging to achieve using splicing kinetics (*10*, *11*). While the current necessary restrictions on the model complexity limit the simultaneous inference of multiple trajectories, we believe that this analysis serves as a compelling proof-of-concept and lays the groundwork for future, more comprehensive studies of gene expression dynamics. Successfully uncovering the underlying processes of transcription, translation, mRNA, and protein degradation from single-cell data would open up a wide and highly impactful range of applications especially in the context of in vivo cellular differentiation.

Although scp-MS is still associated with relatively low throughput and coverage, technological advances will continue to improve data depth and accuracy. In this context, our work offers a framework for more single-cell mRNA/protein-based multi-omics studies, which have the potential to improve our understanding of cell states and translation dynamics across many biological systems including normal development, stem cell systems, and disease states such as cancer.

Materials and Methods

Human bone marrow processing

Bone marrow samples were collected from healthy donors of European ancestry (young, 20 to 25 years; old (>60 years) following the standard protocol of the Department of Hematology, Rigshospitalet, with prior informed and written consent according to the Helsinki declaration under a protocol approved by the Danish National Ethics committee (approval number 1705391). Briefly, 15 mL of bone marrow was aspirated into a 20 mL syringe containing 5 mL of ice-cold sterile-filtered anticoagulant ACD, and immediately mixed to prevent clumping. The collected BM was then diluted 1:2 with PBS + 2mM EDTA + 0.5% BSA (PBS-EDTA) and filtered through a 70 μM filter. Subsequently, 30 mL of the diluted BM was gently layered over 15 mL of Lymphoprep (ProteoGenix, Cat. 1114544) and centrifuged at 400 *g* for 30 min at 20°C. The resulting supernatant, containing the interphase with the low-density mononuclear cells (LDMNCs), was transferred to a new tube. The LDMNCs were washed at a 1:1 ratio with PBS-EDTA, and the pellet was resuspended in 10 mL of PBS-EDTA. Cell counts were determined using a NucleoCounter, and the cells were frozen in FBS containing 10% DMSO overnight at −80°C before being transferred to liquid nitrogen storage. For thawing, cells were gently warmed in a water bath at 37°C for 2 to 5 min and then transferred to a 50 mL Falcon tube. 20 mL pre-warmed thawing medium (RPMI + 10% FCS + 100 μg/mL DNAse (Roche, Cat. 11284932001) + 5 mM $MgCl_2$) was added dropwise to the cell suspension, followed by centrifugation at 300 *g* for 10 min at room temperature. The supernatant was removed, and the cell pellet was resuspended in 10 mL of pre-warmed thawing medium and incubated for 15 min at 37°C. After the incubation, cells were counted and centrifuged.

FACS staining, analysis, and cell sorting

All FACS experiments included single-color controls using CompBeads (BD Biosciences, Cat. 552844) and Fluorescence Minus One (FMO) controls. Antibodies are shown in table S2.

scp-MS samples

Thawed BM mononuclear cells were CD34+ enriched using CD34 MicroBeads (Miltenyi Biotec, Cat. 130-046-702) according to the manufacturer's instructions. CD34+ HSPCs were subsequently incubated with a cocktail of biotinylated lineage antibodies and the purified, unconjugated Endomucin antibody for 30 min on ice. After incubation, cells were washed with PBS-EDTA and centrifuged at 100 *g* for 10

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

min. Subsequently, cells were incubated for 30 min with Streptavidin-PECy5 and goat anti-rabbit IgG (PE), followed by another wash with PBS-EDTA at 100 *g* for 10 min. Finally, the cells were incubated with the panel of monoclonal antibodies for 30 min. Similar as previously described (*27*, *28*), after three washes with PBS to eliminate any remaining serum, cells were sorted on a BD FACSAria III instrument at a maximum speed of 500 events per second into Eppendorf twin.tec 384 LoBind plates containing 1 µL 20% TFE, 10 mM TCEP, 40 mM CAA, 50 mM TEAB lysis buffer. Live cells were sorted based on the membrane-impermeable viability dye 7-AAD. The 100 µm nozzle was used for all sorts into 384-well plates, and "Single Cell" and "4-way Purity" precision modes were selected for single cell and 500 cells/well carrier sorts, respectively. Immediately after sorting, plates were briefly centrifuged, snap-frozen on dry ice, and then boiled at 95°C in a thermocycler (Applied Biosystems Veriti 384-well) for 5 min. The plates were then briefly centrifuged again and stored at –80°C until further sample preparation.

CITE-seq samples

BM cells were incubated with the cocktail of biotinylated lineage antibodies and the panel of monoclonal antibodies for 30 min on ice. Live cells were sorted based on the membrane-impermeable viability dye 7-AAD. Cells were FACS-sorted as Lin–CD34+ into low-bind Eppendorf tubes using a BD FACSAria III instrument.

scp-MS sample preparation

Plates were thawed on ice and sonicated for 2 min in a water bath sonicator. Protein lysates from the single cells were digested with 2 ng of Trypsin (Promega, Cat. V5280), dissolved in 1 µl of 100 mM TEAB pH 8.5 containing Benzonase (Sigma, Cat. E1014) diluted to a ratio of 1:5000 (vol/vol) to digest any DNA that would interfere with downstream processing. For the carrier plate, containing 500 cells per well, the trypsin amount was increased to 10 ng in each well. All dispensing steps in this procedure were carried out using the Dispendix I-DOT One instrument. After dispensing the digestion buffer, the plates were centrifuged, vortexed, and incubated at 37°C overnight to complete the digestion.

On the next day, peptides were labeled with TMTPro 16plex reagents (ThermoFisher Scientific, Cat. A44520). 6 mM of each label (except 127C) in 1 µL acetonitrile (ACN) was added to the single cell wells, while the 500-cell carrier wells were labeled with 13 mM of TMTPro-126 reagent in each well. The plates were then incubated at room temperature for 1 hour. The labeling reaction was quenched by adding 1 µl of 1.25% hydroxylamine (ThermoFisher Scientific, Cat. 90115) in 100 mM TEAB to single cell wells and 2.5% to carrier wells, followed by 15 min incubation at room temperature. The carrier plate, containing peptides labeled with TMTPro-126, was pooled into one sample and desalted using a SOLAµ Solid Phase Extraction (SPE) column (ThermoFisher Scientific, Cat. 60209-001) following the standard cleanup protocol. Eluted and desalted peptides were concentrated to dryness in an Eppendorf Speedvac, after which they were resuspended in A* (2% ACN, 0.1% TFA). Subsequently, peptides from 14 single cells and 200-cell equivalent of the Lin-CD34+ carrier were pooled and desalted using a SOLAµ SPE 96-well plate, described in the following. Briefly, the SOLAµ plate was prepared by activating with 200 µL methanol, conditioning with 200 µL 80% ACN, 0.1% FA, and equilibrating with 200 µL 0.1% FA. Between each step, the plate was centrifuged at 700 *g* for 1 min. Next, 60 µL of 0.1% FA and 200-cell equivalent of the carrier (in 10 µL volume) were dispensed into each column. The final TMT sets were pooled from 14 single cells using a handheld multichannel pipette onto the prepared SOLAµ columns. After pooling, peptides were bound to the filter by centrifuging at 700 *g* for 1 min. The SOLAµ plate was washed twice with 200 µL 0.1% FA, and the peptides were eluted twice with 20 µL of 40% ACN, 0.1% FA into an Eppendorf Twin.tec 96-well LoBind plate. Finally, the peptides were dried at 60°C in an Eppendorf Speedvac and reconstituted in A* for individual mass spectrometry analysis.

CITE-seq sample preparation

Sorted Lin-CD34+ cells were washed with PBS and stained with Fc block for 10 min and then washed again. After Fc block the cells were stained for 30 min with the CITE-seq antibodies (table S3). After the staining, cells were washed once again, counted to reach the proper concentration, and loaded into the 10x instrument according to the manufacturer's protocol. Single-cell RNA sequencing was performed using the 10x Genomics 3′ v3 platform. Approximately 7000 to 10,000 cells were loaded in each 10x channel, after which the single-cell libraries were prepared according to manufacturer's instructions, except for the added CITE-seq steps. A specific primer was added in the cDNA amplification step and instead of discarding the small fragments in the bead purification, they were kept and used for the CITE-seq/ADT library preparation according to the original publication (*22*). After all libraries were ready, their quality and size were verified in the Bioanalyzer before sequencing on the NextSeq 500 Illumina platform. Two to three samples were pooled and sequenced together in each sequencing run.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

CRISPR/Cas9 targeting in human CD34+ BM cells

For each donor, normal BM was aspirated and the CD34+ cells were isolated as described above. Cells were cultured for 2 days in StemSpan (Stem Cell Technologies, Cat. 09655) supplemented with TPO (100 ng/mL), SCF (100 ng/mL), FLT3 (100 ng/mL), IL6 (100 ng/mL), and UM171 (35 nM) and electroporated with the ribonucleoprotein complexes (RNPs) made with 0.3 μl recombinant Cas9 (10 μg/μl) using sgRNAs harboring 2'-O-Methyl at the first 3 bps and last bp, 3′phosphorothioate bonds between the first 3 bps and last 2 bps as well as an 80-mer SpCas9 scaffold (Synthego). For the control locus AAVS1 we used a single sgRNA (5′-GGGGCCACUAGGGACAGGAU-3′), whereas we used pairs of sgRNAs for targeting H1F0 (5′-GAATTCCACGTCCGCCCCTG-3′ and 5′-UUGGAGGCCUUGGCCCGCUU -3′), SOD1 (5′-UAAUGGACCAGUGAAGGUGU-3′ and 5′-CAUGAACAUGGAAUCCAUGC-3′), SOD2 (5′-GGUUGUUCACGUAGGCCGCG-3′ and 5′-CGUCACCGAGGAGAAGUACC-3′), and TALDO1 (5′-GCCGACACGGGCGACUUCCA-3′ and 5′-UCCAUCCUCUGACGCUUCAC-3′). Electroporation was carried out on an Amaxa D4 electroporator using the program DZ-100.

Edited regions were amplified from gDNA from targeted cells using the following primer pairs (left and right): AAVS1: 5′-AGGGAGTTTTCCACACGGAC-3′ and 5′-CAAAGTACCCCGTCTCCCTG-3′, SOD1: 5′-ACTCACTTTGCTTGCCACTC-3′ and 5′-TTGGGAGGTCGAGGAGGATA-3′, SOD2: 5′-ACTCACTTTGCTTGCCACTC-3′ and 5′-TTGGGAGGTCGAGGAGGATA-3′, TALDO1: 5′-GCTGAGGACCTGGGGAGA-3′ and 5′-ATGAAGGTGCACAGAGGGG-3′, H1F0: 5′-AGGAAAAGCAGCGACTCCT-3′ and 5′-TAGGAGACAGGAGGGAGTGT-3′. Sanger sequencing was carried out using the left (SOD1, SOD2 and TALDO1) or the right (AAVS1 and H1F0) primer, and Variant Allele Frequency (VAF) was assessed against a reference (WT) amplicon using the ICE analysis tool (ice.editco.bio).

Long-term culture initiating cell (LTC-IC) assay

One day after electroporation, CD34+ cells were sorted (10, 30, 100, 300, and 1000 cells; 10 wells for each number) onto Mitomycin-C treated MS5 feeder cells in 96-well plates and cultured for 5 weeks in Gartner's medium (αMEM supplemented with 11% FCS, 12.5% horse serum, 1% Pen/Strep, 1 μM hydrocortisone, 57.2 μM beta-mercaptoethanol, 20 ng/mL TPO, 20 ng/mL IL-3, and 20 ng/mL GCSF) followed by 2 weeks in methylcellulose (MethoCult, Stem Cell Technologies, Cat. H4435). Stem cell frequencies were assessed by counting wells positive for cell growth.

Co-culture assay

One day post-electroporation, CD34+ cells were sorted, and 20,000 cells were seeded in triplicate onto Mitomycin-C treated MS5 feeder cells in 6-well plates and cultured for 3 weeks in Gartner's medium (described in detail above).

FACS analysis of co-culture experiments

After 3 weeks co-culture (see above), adherent cells were trypsinized and stained with the following antibody conjugates: hCD45-BV421 (BD Horizon, Cat.563879, RRID:AB_2744402), CD41a-PECy7 (BD Pharmingen, Cat. 561424, RRID:AB_10642584), CD14-BV786 (BD OptiBuild, Cat. 741034, RRID:AB_2740651), CD11b-BV711 (BD OptiBuild, Cat. 742641, RRID:AB_2740934), CD15-PE (BD Pharmingen, Cat. 582371, RRID:AB_11154049), CD16-BV605 (BD Horizon, Cat. 563172, RRID:AB_2744297), CD66b-APC (BioLegend Cat. 305118, RRID:AB_2566607), and CD34-APC-CY7 (BioLegend, Cat #343514, RRID:AB_1877168). Finally, 7-AAD was added as a viability marker, and the cells were analyzed on a BD LSRFortessa X-20.

Colony Forming Cell (CFC) Assay

One day post-electroporation, 1000 cells were plated in triplicate in 1 mL methylcellulose (MethoCult, Cat. H4435) and cultured for 2 weeks. Subsequently, different classes of colony-forming units (CFUs) were counted under a microscope.

Mass spectrometry data acquisition

Peptides were loaded onto a 2 cm C18 trap column (ThermoFisher 164946), connected in-line to a 15 cm C18 reverse-phase analytical column (ThermoFisher EasySpray ES904), with 100% Buffer A (0.1% Formic acid in water) at 4 μl/min or 750 bar maximum pressure using the ThermoFisher EasyLC 1200, and the column oven operating at 30°C. Peptides were eluted over a linear 140 min gradient at a flowrate of 100 nL/min, using 80% acetonitrile, 0.1% formic acid (Buffer B) going from 10% to 30% B. Spectra were acquired with an Orbitrap Eclipse Tribrid mass spectrometer with FAIMS Pro interface (ThermoFisher Scientific) running Tune 3.4 and Xcalibur 4.3 using the RETICLE acquisition method (*28*). FAIMS switched between compensation voltages of −50 V and −70 V with cycle times of 3 s and 1.5 s, respectively. MS1 spectra were acquired at 120,000 resolution with a scan range from 375 to 1500 m/z, normalized AGC target of 300%, and maximum injection time of 50 ms. Precursors were filtered using monoisotopic peak

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

determination set to peptide, charge state 2 to 6, dynamic exclusion of 120 s with ±10 ppm tolerance excluding isotopes and different charge states, and a precursor fit of 70% in a window of 0.7 m/z. Precursors selected for MS2 analysis were isolated in the quadrupole with a 0.7 m/z window. Ions were collected for a maximum injection time set to 46 ms and normalized AGC target of 300%. Fragmentation was performed with 30 normalized CID collision energy, and MS2 spectra were acquired in the linear ion trap (LIT) at scan rate rapid. LIT MS2 spectra were subjected to real-time search (RTS) using the homo sapiens database obtained from Uniprot (Swiss-Prot with isoforms, 42,297 sequences, downloaded on 30.03.2021), and trypsin set as enzyme (cleavage next to arginine or lysine, but not before proline). Static modifications were TMTpro16plex on Lysine (K) and N-Terminus and carbamidomethyl on cysteine (C). Oxidation of methionine (M) was set as variable modification. Maximum missed cleavages were set to 1 and maximum variable mods to 2. FDR filtering was enabled, maximum search time was set to 40 ms, and the scoring threshold was set to 1.4 XCorr, 0.1 dCn, and 10 ppm precursor tolerance. Use as trigger only was checked, and close-out was enabled with maximum number of peptides per protein set to 4. Precursors identified via RTS were isolated in the quadrupole with a 0.7 m/z window, ions were collected for a maximum injection time of 750 ms and normalized AGC target of 500%, fragmented with 32 normalized HCD collision energy, and MS2 spectra were acquired in the orbitrap at 120,000 resolution with first mass set to 120 m/z.

Mass spectrometry data processing

Raw files were analyzed with Proteome Discoverer 3.1 (ThermoFisher Scientific) with the built-in TMTpro Reporter ion quantification workflows using the standard settings if not further specified. Reporter ion quantification was performed only on FTMS MS2 spectra, and the same spectra were also sent to Sequest for identification, where they were searched with precursor mass tolerance of 10 ppm and fragment mass tolerance of 0.02 Da. The database was the homo sapiens database obtained from Uniprot (Swiss-Prot with isoforms, 42,412 sequences, downloaded on 10.03.2024) with trypsin set as enzyme (cleavage next to arginine or lysine, but not before proline) and maximum number of missed cleavages set to 2. Static modifications were TMTpro16plex on lysine (K) and N terminus, and carbamidomethyl on cysteine (C) was set. Dynamic modifications were set as oxidation (M) and Met-loss on protein N-termini. Results were rescored with Inferys (*78*) and q-values were assigned with percolator and filtered to 1% FDR. For the reporter ion quantification, normalization mode and scaling mode were set to none and average reporter s/n threshold was set to 0. Isotopic error correction was applied.

CITE-seq data processing

Demultiplexed FASTQ files were aligned to the human reference genome (GRCh38, v 2020-A as provided by 10x Genomics, which is based on GENCODE v 32/Ensembl 98) using Cell Ranger software (v 5.0.1) by 10x Genomics. The Cell Ranger "count" standard pipeline was used to obtain the expression matrix of unique molecular identifiers (UMIs) for each individual sample.

scp-MS data analysis

scp-MS data were analyzed using the SCeptre pipeline (*27*) using Python 3.9 with SCeptre version 1.1.0, flowkit 0.9.3 (*79*) and Scanpy 1.9.1 (*29*). FACS .fcs files were processed in flowkit in the following manner. First, compensation and bi-exponential transform was applied and subsequently, data was merged and batch corrected using combat (*80*) with each individual representing a batch. Subsequently, each FACS parameter in the merged data was transformed across single cells using the robust z-score with value clipping from –5 to 5. The resulting data were used to set the gates for the populations to annotate cells with a gated population. FACS data and sort- and label layouts were used to create the metadata for each cell.

Subsequent SCeptre analysis was performed in the following manner. A first pass of the SCeptre workflow was applied to identify clusters of cells that likely contain low quality cells (fig. S1A). Indications for low quality cell clusters were mixing of cells of unrelated FACS populations, low total protein signal and the combination of low FSC-A with high SSC-A. Cells contained in the low-quality clusters were filtered out in the second pass of the SCeptre workflow, which included the following steps. Raw files with less than 2000 peptide-spectrum matches (PSMs) were removed and cells from low quality clusters were removed. To remove proteins whose values likely stem from contamination, a mean single cell to carrier ratio was calculated. For each protein in each cell, the signal ratio to the corresponding carrier channel was calculated. Proteins with a mean ratio of over 0.2 were removed. This also removed proteins where the ratio could not be calculated since there was no single cell signal in any cell. Subsequently, protein abundance data were batch corrected using the SCeptre normalize function. Cells were filtered to have a signal for at least 600 proteins, and a minimum log2 total signal of 13.6. Furthermore, cells were filtered based on a relationship of FSC-A to log2 total signal (Fig. 1E), as these parameters showed high

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

correlation and thus were deemed to be suitable to filter outlier cells. The relationship between the two parameters was modeled using the function sklearn.linear_model.RANSACRegressor and cells with a prediction error of over 0.8 log2 total signal were removed. Finally, proteins were required to have a signal in at least 10 cells. This resulted in a dataset of 2934 proteins across 2506 cells with various data completeness (Fig. 1C) and for the second dataset 2174 proteins across 922 cells.

Subsequently, protein abundance data were median-ratio normalized and log2(x+1) transformed. This represents the "batchcorr_norm_log2" layer that was used for differential expression testing. For the embedding, data were imputed using the k-nearest neighbor algorithm followed by the application of the "regress_out" function of Scanpy using log2 total signal and number of proteins with signal as parameters. To embed the data, principal component analysis (PCA) was applied and used to calculate a neighborhood graph, which was subsequently embedded using UMAP. Unsupervised cell clustering was performed using the Leiden algorithm. Clusters were manually annotated using the abundance of marker genes and FACS information.

The quality of batch effect removal was quantified as the overall variance contribution given a covariate (PC regression value) as implemented in the python single-cell integration benchmarking package scib (*32*). Since the SCeptre batch correction is based on alignments of medians and thus assumes balanced multiplexed sets, we expected good integration of the actively balanced multiplexed sets "Enriched Populations". The observation that the batch correction also worked for the "Total HSPC" sets indicated that random samples of 14 HSPCs were sufficiently balanced.

Differential expressed proteins in each cluster were calculated by comparing protein abundance in each cluster against all other clusters using the layer "batchcorr_norm_log2". For each protein, a two-sided Welch's *t* test was performed, excluding missing values. The test was chosen since we do not assume equal variance and sample size between groups. P-values were corrected via the Benjamini–Hochberg procedure and a cutoff of 5% FDR was applied.

To compare protein log2 fold-changes derived from scp-MS to bulk data, we used proteomics data (*42*) from bulk sorted human bone marrow populations, which were reanalyzed with Spectronaut 18.4 using factory settings. Replicates were aggregated to the mean and the log2 fold-change of each protein in each population was calculated by subtracting the mean of all other populations. scp-MS data were matched to the smallest matching bulk protein group. For each scp-MS cluster and gated population, log2 fold-changes were compared for proteins that had at least 30% coverage in the scp-MS cluster and were detected in the bulk dataset. Each scp-MS cluster and gated population was compared to each bulk population by calculating the Pearson correlation and the resulting matrix was clustered hierarchically.

For unsupervised clustering of proteins, proteins were filtered to have at least one significant log2 fold-change above 0.2 (30% cluster coverage, 1% FDR) in any of the cell clusters. For each cell cluster, values from layer "batchcorr_norm_log2" were aggregated to the mean ignoring missing values and each protein was z-score transformed across cell clusters. Proteins were clustered hierarchically using Euclidean distance as metric and the linkage method ward and the tree was cut to 9 protein clusters. Each protein cluster was tested for enriched GO Biological Process and KEGG terms using gseapy (*50*). The background was set to all proteins in the dataset and terms were filtered 5% FDR and more than one term per gene set. For each protein cluster selected terms were plotted.

CITE-seq data analysis

The CITE-seq expression matrices were analyzed using scanpy 1.9.8 (*29*) and scvi-tools 1.1.2 (*53*). Cells were filtered to contain at least 1800 mRNA counts, at least 800 genes, and a fraction of mitochondrial gene counts between 0.025 and 0.1. Genes were filtered to be detected in at least 99 cells. Cells were annotated via Azimuth (*4*), separately for each individual. We removed cells that were annotated as a rare celltype or celltypes with no evidence in the scp-MS data. Only CITE-seq surface markers that overlapped with scp-MS markers were retained. Genes were mapped to the scp-MS data using an ENSG-Uniprot mapping table. If multiple proteins or mRNAs mapped to the same ENSG, the protein or mRNA with the highest coverage was mapped. Subsequently, we created the "counts_totalnorm_log2" layer by normalizing each cell to total counts and applying log(x+1) transform. Next, genes were subsetted to contain 4000 highly variable genes with flavor "seurat_v3" and batch keys provided, and to contain all genes detected in scp-MS, resulting in 5820 genes. Count data were modeled and batch-corrected using totalVI (*52*). The modeled mRNA expression was median-ratio normalized and log2 transformed, resulting in the "denoised_rna_norm_log2" layer. For the UMAP embedding, the modeled mRNA expression was used. Azimuth GMPs were subsetted into GMDP and MDP using leiden clustering. Cell cycle phase was predicted using a list of marker genes and the

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

"score_genes_cell_cycle" function of Scanpy. To benchmark mRNA log2 fold-changes from either counts or scVI-modeled data, transcriptomics data (*42*) from bulk sorted human bone marrow populations were used. Transcriptomics data were CPM normalized and log(x+1) transformed. Log2 fold-change between two populations was calculated by subtracting the means of the replicates.

Integration of scRNA-seq and scp-MS

Integration was performed using GLUE (*56*). We linked mRNA features from the CITE-seq data to their corresponding proteins in the scp-MS data. The guidance graph was constructed to contain all 2934 proteins of scp-MS and all 5820 mRNAs from scRNA-seq (CITE-seq), resulting in 8754 nodes. 2778 edges with weight 0.9 were inserted between protein and matching mRNA. Self-loop edges with weight 1.0 were inserted for each node. This resulted in 11,532 edges. Further required input for GLUE were an expression matrix and a latent representation thereof per modality. For scp-MS, the imputed data that were not processed with "regress_out" were used as expression matrix and as latent representation, 50 PCs of that were used. The probabilistic generative model used by the decoder was set to "Normal" corresponding to a "LogNormal" distribution considering the prior log-transformation of the scp-MS data. For scRNA-seq, the count data were used as expression matrix and the totalVI latent space was used as representation, which provided additional batch correction. The probabilistic generative model was set to negative binominal ("NB") and the individuals were defined as batch variable. In contrast to the standard GLUE workflow that used a two-stage training procedure to estimate balancing weights, we used a single-stage procedure without balancing weights. The GLUE integration resulted in a joint latent representation of "protein" and "mRNA" cells. After integration, scRNA-seq labels were transferred to scp-MS using the GLUE function "transfer_labels" using the GLUE latent space and 15 nearest neighbors. Modality integration as well as biological signal preservation were measured using scIB metrics (*32*). To quantify the overlap of the distribution of cells from the two modalities in the integrated latent space, we used the average silhouette width score that compares the average distance between cells of the same modality to the average distance of cells from different modalities. The highest possible score for separation of the modalities is 1, while values near 0 indicate well overlapping sets of cells. To assess preservation of biological variance, we compared cell type separation metrics before and after integration per modality. For the "mRNA" cells, we used the Azimuth-derived cell type annotations as ground truth that are to be preserved and the totalVI embedding as baseline. For the "protein" cells, we used the Leiden cluster-based cell types and the PCA embedding that was also used as input to GLUE. We computed the average silhouette width score as just described as well as the normalized mutual information and adjusted rand index scores as implemented in scIB. Both compare the overlap of two clusterings, one being the ground truth cell type annotations and the other one being derived from the embedding structure as data-driven clusterings. A perfect match of the clusterings would lead to a score of 1, uncorrelated or random clusterings to a score of 0.

Cellrank analysis

A cellrank (*5*, *6*) analysis was performed on the integrated dataset, with standard setting applied if not further specified. First, the joint latent space was used to calculate a nearest neighbor graph, following a diffusion map. Then, after setting an immature HSC as root cell, a diffusion pseudotime was calculated, which was used to initialize a cellrank pseudotime kernel. Subsequently, a transition matrix was computed and macrostates were determined via GPCCA. Terminal states were determined by fitting seven macrostates to a Schur decomposition and selecting the top six. Finally, fate probabilities were calculated for the six terminal states.

Expression profiles were modeled using the approach in cellrank where a generalized additive model (GAM) is fitted to the expression values along pseudotime, while using the fate probabilities as weights. Cells were subsetted into the relevant clusters for each trajectory. GAM settings were 12 kn, lam=12 and derivative was set as penalty. For mRNA, the scVI-modeled values were used. For protein, missing values were ignored by setting the weights of cells with missing values for the specific protein to zero. For both modalities, values were not log2 transformed, but used in linear space. Furthermore, values were normalized by dividing them by the global mean of the specific gene.

HSC quiescence and differentiation was investigated using a subset of cells with pseudotime below 0.4. Each protein or mRNA had to be detected in at least 200 cells to be considered. Next, the correlation vector was calculated by calculating the Pearson correlation of each log2 transformed mRNA/protein with pseudotime. For protein, missing values were ignored for this calculation, while for the count-based mRNA, zeros were used. scVI modeled mRNA did not have any missing values.

Lineage specification was investigated using all cells, however using only mRNA/proteins with detections in at least 200 cells. Correlations vectors for each lineage were calculated by calculating the Pearson correlation of each log2 transformed mRNA/protein with fate probability. For

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

protein, missing values were ignored, while for the count-based mRNA, zeros were used. scVI modeled mRNA did not have any missing values.

Gene set enrichment analysis (GSEA) was performed on the correlation vectors using GSEApy (*50*). A pre-ranked GSEA was computed using the gene set GO Biological Process with minimum set size of 3 and maximum of 1000. Resulting terms were filtered to 10% FDR and for the lineage specification additionally to a positive enrichment score. The term list was shortened by removing redundant GO term using REVIGO (*81*).

Gene expression cascades for each trajectory were visualized as heatmaps by filtering the correlation vector to a minimum correlation of 0.25 for mRNA and 0.2 for protein. Then genes were sorted by the pseudotime coordinate of their maximum expression in the modeled abundance profile. Then, this vector was binned into 4 bins and the top 5 mRNAs or top 8 proteins of each bin were selected for visualization.

For the joint mRNA-protein covariation analysis the protein lineage specification correlation vectors were concatenated into a table of protein lineage correlations across the six lineages. In each vector column, if the protein did not pass the 200-cell coverage filter, it was set to missing. The table was subsequently filtered for proteins with Pearson correlation of at least 0.2 in any of the six lineages. For mRNA the count-based correlation vectors were used. Then the intersection of both protein and mRNA tables was concatenated and a pairwise correlation matrix was computed, while requiring at least four values to return a Pearson value. Keeping only columns with proteins yielded a correlation vector of a protein with all other mRNAs and proteins and sorting this vector by correlation resulted in a correlation rank for each mRNA and protein. For each protein, related proteins were defined to be in the same KEGG pathway.

To compare the mRNA correlation ranks for each protein to results from bulk data, we used a dataset containing mRNA and protein measurements of 59 breast cancer cell lines (*74–76*). Log2 transformed tables were normalized by subtracting the row and column medians. Similar to the single-cell data, tables were concatenated, and a correlation matrix was computed and an mRNA correlation rank for each protein was computed. Ranks were then compared between bulk and single-cell data.

Gene expression modeling with scProtVelo

Creating paired single cell data

To create a dataset with paired mRNA and protein expression for all cells, we selected cells of the erythroid and pre-mDC trajectories based on a leiden clustering of the GLUE embedding, so that we obtain the same biological subset of cells from both modalities. For both trajectories, we then removed cells with annotated cell types for which we don't know ground truth transitions. The resulting cell subsets contained 472 cells from the protein dataset and 1677 cells from the scRNA-seq data for the erythroid branch and 262 cells from the protein dataset and 897 cells from the scRNA-seq dataset for the pre-mDC branch. We selected proteins that were detectable in at least 30% of the protein cells within the subsets and mRNAs that had nonzero counts in at least 2% of the scRNA-seq cells. This resulted in 1065 proteins and 5101 mRNAs for the erythroid subset, which gave us a total of 1036 overlapping genes to work with. For the pre-mDC subset, we got 1101 proteins and 5081 mRNAs with an overlap of 1072 genes. mRNA counts and batch corrected proteins were median-ratio normalized and, as also done for RNA velocity, expression values were smoothed per modality over their respective nearest neighborhood graphs. Then, for each cell, values for the not measured modality were interpolated over the 15 nearest neighbor cells of the modality of interest based on the joint GLUE embedding. For all features, 25% and 75% quantiles were computed and used to scale features by $1/\left(75\%quantile-25\%quantile\right)$ to give all features the same weight in the scProtVelo reconstruction loss later on. We want to note here that scaling of expression values only results in scaled transcription and translation rates that can be reverted after model fitting, while a shift cannot be compensated for within the differential equations we define below.

We further define genes to showcase substantial variation during the selected differentiation process, if they are differentially expressed between the upper and lower 30% quantiles of cell with respect to diffusion pseudotime in both modalities. For that we randomly sampled 50 cells from the respective quantiles, performed an unpaired *t* test and selected genes at a significance level of 5% after Benjamini-Hochberg multiple testing correction that had a log-fold-change of at least 0.1 on protein level and 0.5 on mRNA level. This resulted in a selection of 146 genes for the erythroid subset and 46 genes for the pre-mDC branch.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

The translation dynamics model

In a first modeling setup, we followed the concept of RNA velocity (7). We jointly model mRNA $r(t)$ and protein $p(t)$ expression of a gene considering the transcription of a gene to mRNA molecules at rate α, their degradation at rate β as well as translation to proteins at rate κ and their respective degradation at rate δ over time t. We assume fixed (time independent) rates per gene and also model time as gene specific, since orthogonal processes like cell cycle as well as cell differentiation could be the driving factors in different genes. Per gene, this process can be formalised by the following ordinary differential equations:

$$\frac{dr(t)}{dt} = \alpha - \beta r(t)$$

$$\frac{dp(t)}{dt} = \kappa r(t) - \delta p(t)$$

The expression values $r(t)$ and $p(t)$ stand for nearest-neighbor smoothed median-ratio normalised abundances. Both, the time annotations t as well as the kinetic rates α, β, κ and δ are not directly measured and are to be inferred.

Given that we model fixed, time-independent rates, a gene can either be in an induction or repression state, characterised by mRNA expression increasing or decreasing with time. In the RNA velocity formulation, induction is defined to start with unspliced as well as spliced expression being 0. We observe different behavior in mRNA-protein phase portraits. Thus, we assume free initial expression values r_0^{ind} and p_0^{ind} and pose the restriction on the kinetic parameters that the derivatives $\frac{dr(t)}{dt}$ and $\frac{dp(t)}{dt}$ are nonnegative at r_0^{ind} and p_0^{ind}, respectively. For the case of repression, we model the initial expression values r_0^{rep} and p_0^{rep} as a result of an unobserved induction phase that shares all kinetic parameters, but the transcription rate, with the following repression phase and enforce that the derivative of the mRNA expression is non-positive at the inferred starting point (r_0^{rep}, p_0^{rep}). Additionally, similar to the RNA velocity model, we explicitly also allow cells to be in a steady state that we set to be the respective r_0 and p_0 values depending on the general gene state.

On top of this very free model setup, we used a second model variation that is meant to capture one known linear process more accurately. In this setting, we restrict time to be shared across all genes and introduce diffusion pseudotime as a prior on the modeled time.

scProtVelo

We implemented all those considerations into scProtVelo, which is available as a Python package via github.com/theislab/scProtVelo. The scProtVelo model, in short, is a variational autoencoder-like architecture that takes a cell representation as input, learns a latent representation of it that is further used to generate mRNA and protein expression that is close to the original expression of that cell. The crucial part of scProtVelo is that the generation is not performed by a fully connected neural network. Instead, it infers a latent time annotation for a cell as well as the kinetic parameters and combines both to generate the expression values based on the defined expression dynamics. Our python implementation heavily builds on the veloVI framework, and we refer the reader to the original publication (9) for all technical modeling details that are not explained in the following.

The scProtVelo generative model

The differential equations introduced above can be solved analytically and as

$$r(t) = r_0 e^{-\beta t} + \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right)$$

$$p(t) = p_0 e^{-\delta t} + \frac{\alpha\kappa}{\beta\delta}\left(1 - e^{-\delta t}\right) + \frac{\kappa\left(r_0\beta - \alpha\right)}{\beta\left(\delta - \beta\right)}\left(e^{-\beta t} - e^{-\delta t}\right)$$

with initial expression values r_0 and p_0 and as such build the base for the generative process of scProtVelo. We can now formalise all four possible cell-gene states (induction steady state, induction dynamic state, repression steady state, repression dynamic state) described by the gene specific variable $k^{ind/rep}$ and the gene and cell specific variable $k_i^{ss/ds}$ for a cell $i \in I$. Time t as well as kinetic rates α, β, κ and δ and intermediate variables are gene specific, but we omit the subscripts for readability. For the induction steady state, we propose a restriction-free expression pair

$$r\left(\cdot, k^{ind/rep} = 0, k_i^{ss/ds} = 0\right) := r_0^{ind}$$

$$p\left(\cdot,\ k^{ind/rep} = 0,\ k_i^{ss/ds} = 0\right) := p_0^{ind}.$$

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

For the corresponding dynamic case, mRNA and protein expression can be described to follow the analytical solution given above with the initial state being the values defined for the steady state case:

$$r\left(t_i, k^{ind/rep}=0, k_i^{ss/ds}=1\right) := r_0^{ind} e^{-\beta t_i} + \frac{\alpha}{\beta}\left(1-e^{-\beta t_i}\right)$$

$$p\left(t_i,\ k^{ind/rep}=0,\ k_i^{ss/ds}=1\right) := p_0^{ind} e^{-\delta t_i} + \frac{\alpha\kappa}{\beta\delta}\left(1-e^{-\delta t_i}\right) + \frac{\kappa\left(r_0^{ind}\beta-\alpha\right)}{\beta\left(\delta-\beta\right)}\left(e^{-\beta t_i}-e^{-\delta t_i}\right).$$

For the repression modeling, we assume an unobserved induction phase preceding the observed repression phase. We assume that the induction phase shares rates β, κ and δ with the repression phase and that both phases just differ in their transcription rates α_{ind} and α_{rep}. Further assuming an initial state $0,0$ for the induction and a gene-specific time $t_s \in \mathbb{R}_+$ at which the initial expression pair is attained, we propose the repression steady state to be

$$r\left(\cdot, k^{ind/rep}=1, k_i^{ss/ds}=0\right) := \frac{\alpha_{ind}}{\beta}\left(1-e^{-\beta t_s}\right)$$

$$p\left(\cdot,\ k^{ind/rep}=1,\ k_i^{ss/ds}=0\right) := \frac{\alpha_{ind}\kappa}{\beta\delta}\left(1-e^{-\delta t_s}\right) - \frac{\kappa\alpha_{ind}}{\beta\left(\delta-\beta\right)}\left(e^{-\beta t_s}-e^{-\delta t_s}\right).$$

Last, for the dynamic repression state, we obtain

$$r\left(t_i, k^{ind/rep}=1, k_i^{ss/ds}=1\right) := r_0^{rep} e^{-\beta t_i} + \frac{\alpha_{rep}}{\beta}\left(1-e^{-\beta t_i}\right)$$

$$p\left(t_i,\ k^{ind/rep}=1,\ k_i^{ss/ds}=1\right) := p_0^{rep} e^{-\delta t_i} + \frac{\alpha_{rep}\kappa}{\beta\delta}\left(1-e^{-\delta t_i}\right) + \frac{\kappa\left(r_0^{ind}\beta-\alpha_{rep}\right)}{\beta\left(\delta-\beta\right)}\left(e^{-\beta t_i}-e^{-\delta t_i}\right)$$

with r_0^{rep} and p_0^{rep} being the values obtained above for the repression steady state.

For the generative process of an expression pair $\left(r_i, p_i\right)$ of a cell i, we draw a low-dimensional ($d=10$) latent cell representation

$$z_i \sim Normal\left(0,\ I_d\right)$$

as well as a cell-independent distribution over gene states (induction or repression) described by the probability simplex generated by the Dirichlet distribution and then the actual state assignment itself

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

$$\pi^{ind/rep} \sim Dirichlet\left(0.25, 0.25\right)$$

$$k^{ind/rep} \sim Categorical\left(\pi^{ind/rep}\right).$$

Further, we draw a distribution distinguishing between a cell being in steady state or dynamic state as well as the assignment

$$\pi_i^{ss/ds} \sim Dirichlet\left(0.25, 0.25\right)$$

$$k_i^{ss/ds} \sim Categorical\left(\pi_i^{ss/ds}\right).$$

The time annotation t_i, we model as a function of the latent cell representation z_i

$$t_i = h\left(z_i\right)$$

with $h : \mathbb{R}^d \to \left(0,1\right)$ being a fully connected neural network.

Finally, the expression pair $\left(r_i, p_i\right)$ is drawn from two normal distributions as

$$r_i \sim Normal\left(r\left(t_i,\ k^{ind/rep},\ k_i^{ss/ds}\right),\ \sigma^2\right)$$

$$p_i \sim Normal\left(p\left(t_i,\ k^{ind/rep},\ k_i^{ss/ds}\right),\ \sigma^2\right).$$

scProtVelo inference

The goal of the inference is to obtain estimates for all introduced kinetic rates, gene state distribution parameters, neural network weights, initial expression values for the induction cases, switch times for the repression cases and finally posterior distributions over the latent variables $z, \pi^{ind/rep}$ and $\pi^{ss/ds}$. We use variational inference and with $x_{i,glue}$ being a precomputed representation of the mRNA and protein expression of cell $i \in I$, namely its GLUE representation, we propose the following factorization of the posterior distribution

$$q_\phi\left(z,\ \pi^{ind/rep},\ \pi^{ss/ds} \mid x_{glue}\right) := \prod_i^I q_{1,\phi}\left(z_i \mid x_{i,glue}\right) \prod_g^G q_{2,\phi}\left(\pi_g^{ind/rep},\ \pi_{ig}^{ss/ds} \mid z_i\right)$$

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

meaning that a neural network q_1 maps a precomputed cell representation $x_{i,glue}$ to a posterior distribution for the latent cell representation z_i which is itself used as input to a second neural network q_2 that maps it to posterior distributions for the state distributions $\pi_g^{ind/rep}$ and $\pi_{ig}^{ss/ds}$ for all genes $g \in G$. For the likelihoods we integrate over the four defined cell-gene states and obtain:

$$p_\theta\left(r_{ig} \mid z_i,\ \pi^{ind/rep},\ \pi_i^{ss/ds}\right) = \sum_{k_g^{ind/rep} \in 0,1} \sum_{k_{ig}^{ss/ds} \in 0,1} \pi_g^{ind/rep} \pi_{ig}^{ss/ds} Normal\left(r\left(t_{ig},\ k_g^{ind/rep},\ k_{ig}^{ss/ds}\right),\ \sigma^2\right)$$

$$p_\theta\left(p_{ig} \mid z_i,\ \pi^{ind/rep},\ \pi_i^{ss/ds}\right) = \sum_{k_g^{ind/rep} \in \{0,1\}} \sum_{k_{ig}^{ss/ds} \in \{0,1\}} \pi_g^{ind/rep} \pi_{ig}^{ss/ds} Normal\left(p\left(t_{ig},\ k_g^{ind/rep},\ k_{ig}^{ss/ds}\right),\ \sigma^2\right).$$

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Additionally, to minimizing the arising ELBO objective that is composed of the negative log likelihood of the data given the fitted model as well as Kullback-Leibler divergence terms for the latent variables $z, \pi^{ind/rep}$ and $\pi^{ss/ds}$, we include two more loss terms to be optimised. First, we add $L_{param,ind}$ to enforce nonnegative derivatives in mRNA and protein expression at the initial values in the induction case and $L_{param,rep}$ to ensure non-positive derivatives in mRNA expression at the initial values in the repression case

$$L_{param,\,ind}\left(\theta,\ \phi; \{x_{i,\,glue},\ r_i,\ p_i,\ i \in I\}\right) := \frac{1}{|G|}\left(\sum_g^G ReLU\left(r_{g,0}{}^{ind} - \frac{\alpha}{\beta}\right) + \sum_g^G ReLU\left(p_{g,0}{}^{ind} - \frac{\kappa}{\delta} r_{g,0}{}^{ind}\right)\right)$$

$$L_{param,\,rep}\left(\theta,\ \phi; \{x_{i,\,glue},\ r_i,\ p_i,\ i \in I\}\right) := \frac{1}{|G|}\sum_g^G ReLU\left(\alpha_{rep} - \beta r_{g,0}{}^{rep}\right)$$

with loss values set to 0 whenever a denominator is 0 and weighted strongly enough to enforce all wanted conditions.

Secondly, we introduce $L_{init,ind}$ and $L_{init,rep}$. as penalty terms used during a pretraining phase that encourages the modeled steady states to coincide with heuristically observed steady state values $r_g{}^{high}, r_g{}^{low}, p_g{}^{high}$ and $p_g{}^{low}$ that are computed as the observed median expression of the cells that have mRNA expreion within the highest or lowest 1% quantile of that gene

$$L_{init,\,ind}\left(\theta,\ \phi; \{x_{i,\,glue},\ r_i,\ p_i,\ i \in I\}\right) := \frac{1}{|G|}\sum_g^G \left(r_{g,0}{}^{ind} - r_g{}^{low}\right)^2 + \left(p_{g,0}{}^{ind} - p_g{}^{low}\right)^2$$

$$L_{init,\,rep}\left(\theta,\ \phi; \{x_{i,\,glue},\ r_i,\ p_i,\ i \in I\}\right) := \frac{1}{|G|}\sum_g^G \left(r_{g,0}{}^{rep} - r_g{}^{high}\right)^2 + \left(p_{g,0}{}^{rep} - p_g{}^{high}\right)^2.$$

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Trajectory inference using scProtVelo

In the trajectory inference setting, we start model fitting with a pretraining phase with a high focus on the L_{init} loss terms, because we noticed that it helped general model fitting to get initial states approximately correct first. During normal training, this loss term is then turned off.

To quantify the performance in inferring the correct direction of cell progression, we subsetted all genes we fitted the model on to the ones that we accepted as changing along the differentiation trajectory and computed the direction of change as the sign of the difference between mean expression within the late 20% quantile with respect to pseudotime and the mean within the early 20% quantile of the cells. For the predicted state, we chose the state that had the higher value in $\pi_g^{ind/rep}$. To get an impression of the model performance, we sorted the genes by their model likelihood which represents a measure of how well the model captured the expression values of a gene, without knowing whether the direction of change was inferred correctly or not. Finally, we computed accuracy scores on increasing numbers of genes and compared it to the expected accuracy of a random permutation of the model predictions.

We based velocity computations on the 100 genes with the highest likelihood values and followed the same procedure as done for RNA velocity in the veloVI framework.

For the comparison to RNA velocity results, we followed the proposed veloVI workflow on unspliced and spliced mRNA counts of the mRNA cells that were used in the scProtVelo run. To create the .loom files, velocyto 0.17.17 (*7*) was applied on the cellranger count output. To compare the accuracy of the cell-to-cell transition graphs that the stream plot visualizations are based on, we checked whether proposed transitions from one cell to a neighboring cell coincide with an increase in diffusion pseudotime or not. We aggregated these values to accuracies per cell that we inspected visually projected on the respective UMAP embeddings as well as compared their distributions in a histogram.

Gene expression modeling with scProtVelo

To quantify the variance in protein expression that is determined by the corresponding mRNA expression, we figured it made only sense to consider genes that exhibit biological variation and restricted our analysis on the genes we found to be differentially expressed between the start and the end of the chosen differentiation trajectory. To ensure further that we quantify the variance corresponding to cell differentiation, we restricted latent time inference to only one shared time shared across all genes guided by a diffusion pseudotime prior.

To accommodate for those model specification changes, we first included a loss term representing the diffusion pseudotime prior $t_i^{dpt}, i \in I$ as

$$L_{time}\left(\theta,\ \phi; \{x_{i,\,glue},\ r_i,\ p_i, t_i^{dpt},\ i \in I\}\right) := \sum_i^I \left(t_i - t_i^{dpt}\right)^2.$$

Also, the shared time stands in conflict with the prior implementation of the steady states. So instead of modeling a steady state vs dynamic state variable per cell and gene, we learn a switch time per gene at which gene expression for a gene switches from a constant steady state value to dynamic expression. For induction state, this switch time is restricted to be within $(0,1)$ just as the fitted time. For the repression case, we model a switch time in $(-\infty,1]$, where a negative value implies an unobserved earlier initial state and observed cells just cover a more advanced time range. This is essential considering that the initial state for repression is tied to an induction phase. And finally, having access to diffusion pseudotime, we also refine on the heuristically estimated initial states and use the median expression of the cells within the 1% quantile with respect to diffusion pseudotime.

During training we leave out 20% of the cells at random as a test set. To compare protein estimates from scProtVelo against a linear model, we select the protein cells to avoid the computationally interpolated protein values. We then log transform mRNA and protein expression values and fit a linear model (including a bias term) on the train set. Finally, we compute R^2 values for protein predictions from both models in log expression space.

REFERENCES AND NOTES

R1. E. Laurenti, B. Göttgens, From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418–426 (2018). doi:10.1038/nature25022 Medline

2. J. D. Buenrostro, M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang, W. J. Greenleaf, Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535–1548.e1516 (2018). doi:10.1016/j.cell.2018.03.074 Medline

3. S. Triana, D. Vonficht, L. Jopp-Saile, S. Raffel, R. Lutz, D. Leonce, M. Antes, P. Hernández-Malmierca, D. Ordoñez-Rueda, B. Ramasz, T. Boch, J.-C. Jann, D. Nowak, W.-K. Hofmann, C. Müller-Tidow, D. Hübschmann, T. Alexandrov, V. Benes, A. Trumpp, M. Paulsen, L. Velten, S. Haas, Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. *Nat. Immunol.* **22**, 1577–1589 (2021). doi:10.1038/s41590-021-01059-0 Medline

4. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck 3rd, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021). doi:10.1016/j.cell.2021.04.048 Medline
5. M. Lange, V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, D. Pe'er, F. J. Theis, CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022). doi:10.1038/s41592-021-01346-6 Medline
6. P. Weiler, M. Lange, M. Klein, D. Pe'er, F. Theis, CellRank 2: Unified fate mapping in multiview single-cell data. *Nat. Methods* **21**, 1196–1205 (2024). doi:10.1038/s41592-024-02303-9 Medline
7. G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, P. V. Kharchenko, RNA velocity of single cells. *Nature* **560**, 494–498 (2018). doi:10.1038/s41586-018-0414-6 Medline
8. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020). doi:10.1038/s41587-020-0591-3 Medline
9. A. Gayoso, P. Weiler, M. Lotfollahi, D. Klein, J. Hong, A. Streets, F. J. Theis, N. Yosef, Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nat. Methods* **21**, 50–59 (2024). doi:10.1038/s41592-023-01994-w Medline
10. M. Barile, I. Imaz-Rosshandler, I. Inzani, S. Ghazanfar, J. Nichols, J. C. Marioni, C. Guibentif, B. Göttgens, Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biol.* **22**, 197 (2021). doi:10.1186/s13059-021-02414-y Medline
11. V. Bergen, R. A. Soldatov, P. V. Kharchenko, F. J. Theis, RNA velocity-current challenges and future perspectives. *Mol. Syst. Biol.* **17**, e10282 (2021). doi:10.15252/msb.202110282 Medline
12. Y. Liu, A. Beyer, R. Aebersold, On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016). doi:10.1016/j.cell.2016.03.014 Medline
13. G. Kustatscher, M. Hödl, E. Rullmann, P. Grabowski, E. Fiagbedzi, A. Groth, J. Rappsilber, Higher-order modular regulation of the human proteome. *Mol. Syst. Biol.* **19**, e9503 (2023). doi:10.15252/msb.20209503 Medline
14. T. Geiger, Tackling tumor complexity with single-cell proteomics. *Nat. Methods* **20**, 324–326 (2023). doi:10.1038/s41592-023-01784-4 Medline
15. C. Buccitelli, M. Selbach, mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.* **21**, 630–644 (2020). doi:10.1038/s41576-020-0258-4 Medline
16. P. R. van den Berg, N. M. L. P. Bérenger-Currias, B. Budnik, N. Slavov, S. Semrau, Integration of a multi-omics stem cell differentiation dataset using a dynamical model. *PLOS Genet.* **19**, e1010744 (2023). doi:10.1371/journal.pgen.1010744 Medline
17. P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, X. Wang, J. W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J. T. Lei, M. L. Gatza, M. Wilkerson, C. M. Perou, V. Yellapantula, K. L. Huang, C. Lin, M. D. McLellan, P. Yan, S. R. Davies, R. R. Townsend, S. J. Skates, J. Wang, B. Zhang, C. R. Kinsinger, M. Mesri, H. Rodriguez, L. Ding, A. G. Paulovich, D. Fenyö, M. J. Ellis, S. A. Carr; NCI CPTAC, Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016). doi:10.1038/nature18003 Medline
18. B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. C. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. C. Slebos, D. C. Liebler; NCI CPTAC, Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382–387 (2014). doi:10.1038/nature13438 Medline
19. L. Peshkin, M. Wühr, E. Pearl, W. Haas, R. M. Freeman Jr., J. C. Gerhart, A. M. Klein, M. Horb, S. P. Gygi, M. W. Kirschner, On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Dev. Cell* **35**, 383–394 (2015). doi:10.1016/j.devcel.2015.10.010 Medline
20. M. Jovanovic, M. S. Rooney, P. Mertins, D. Przybylski, N. Chevrier, R. Satija, E. H. Rodriguez, A. P. Fields, S. Schwartz, R. Raychowdhury, M. R. Mumbach, T. Eisenhaure, M. Rabani, D. Gennert, D. Lu, T. Delorey, J. S. Weissman, S. A. Carr, N. Hacohen, A. Regev, Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347**, 1259038 (2015). doi:10.1126/science.1259038 Medline
21. K. Becker, A. Bluhm, N. Casas-Vila, N. Dinges, M. Dejung, S. Sayols, C. Kreutz, J.-Y. Roignant, F. Butter, S. Legewie, Quantifying post-transcriptional regulation in the development of Drosophila melanogaster. *Nat. Commun.* **9**, 4970 (2018). doi:10.1038/s41467-018-07455-9 Medline
22. M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017). doi:10.1038/nmeth.4380 Medline
23. A. Konturek-Ciesla, P. Dhapola, Q. Zhang, P. Säwén, H. Wan, G. Karlsson, D. Bryder, Temporal multimodal single-cell profiling of native hematopoiesis illuminates altered differentiation trajectories with age. *Cell Rep.* **42**, 112304 (2023). doi:10.1016/j.celrep.2023.112304 Medline
24. G. Gorin, V. Svensson, L. Pachter, Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biol.* **21**, 39 (2020). doi:10.1186/s13059-020-1945-3 Medline
25. V. Petrosius, E. M. Schoof, Recent advances in the field of single-cell proteomics. *Transl. Oncol.* **27**, 101556 (2023). doi:10.1016/j.tranon.2022.101556 Medline
26. B. Budnik, E. Levy, G. Harmange, N. Slavov, SCoPE-MS: Mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018). doi:10.1186/s13059-018-1547-5 Medline
27. E. M. Schoof, B. Furtwängler, N. Üresin, N. Rapin, S. Savickas, C. Gentil, E. Lechman, U. A. D. Keller, J. E. Dick, B. T. Porse, Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nat. Commun.* **12**, 3341 (2021). doi:10.1038/s41467-021-23667-y Medline
28. B. Furtwängler, N. Üresin, K. Motamedchaboki, R. Huguet, D. Lopez-Ferrer, V. Zabrouskov, B. T. Porse, E. M. Schoof, Real-Time Search-Assisted Acquisition on a Tribrid Mass Spectrometer Improves Coverage in Multiplexed Single-Cell Proteomics. *Mol. Cell. Proteomics* **21**, 100219 (2022). doi:10.1016/j.mcpro.2022.100219 Medline
29. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018). doi:10.1186/s13059-017-1382-0 Medline
30. C. Vanderaa, L. Gatto, Replication of single-cell proteomics data reveals important computational challenges. *Expert Rev. Proteomics* **18**, 835–843 (2021). doi:10.1080/14789450.2021.1988571 Medline
31. E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018). doi:10.1038/nbt.4314 Medline
32. M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, F. J. Theis, Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022). doi:10.1038/s41592-021-01336-8 Medline
33. N. A. Jakobsen, S. Turkalj, A. G. X. Zeng, B. Stoilova, M. Metzner, S. Rahmig, M. S. Nagree, S. Shah, R. Moore, B. Usukhbayar, M. Angulo Salazar, G.-A. Gafencu, A. Kennedy, S. Newman, B. J. L. Kendrick, A. H. Taylor, R. Afinowi-Luitz, R. Gundle, B. Watkins, K. Wheway, D. Beazley, A. Murison, A. G. Aguilar-Navarro, E. Flores-Figueroa, S. G. Dakin, A. J. Carr, C. Nerlov, J. E. Dick, S. Z. Xie, P. Vyas, Selective advantage of mutant stem cells in human clonal hematopoiesis is associated with attenuated response to inflammation and aging. *Cell Stem Cell* **31**, 1127–1144.e17 (2024). doi:10.1016/j.stem.2024.05.010 Medline
34. F. Notta, S. Zandi, N. Takayama, S. Dobson, O. I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, J. D. McPherson, L. D. Stein, Y. Dror, J. E. Dick, Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016). doi:10.1126/science.aab2116 Medline
35. F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, E. David, N. Cohen, F. K. B. Lauridsen, S. Haas, A. Schlitzer, A. Mildner, F. Ginhoux, S. Jung, A. Trumpp, B. T. Porse, A. Tanay, I. Amit, Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015). doi:10.1016/j.cell.2015.11.013 Medline
36. L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, T. N. Schumacher, The Branching Point in Erythro-Myeloid Differentiation. *Cell* **163**, 1655–1662 (2015). doi:10.1016/j.cell.2015.11.059 Medline
37. Y. Mori, H. Iwasaki, K. Kohno, G. Yoshimoto, Y. Kikushige, A. Okeda, N. Uike, H.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Niiro, K. Takenaka, K. Nagafuji, T. Miyamoto, M. Harada, K. Takatsu, K. Akashi, Identification of the human eosinophil lineage-committed progenitor: Revision of phenotypic definition of the human common myeloid progenitor. *J. Exp. Med.* **206**, 183–193 (2009). doi:10.1084/jem.20081756 Medline
38. R. Drissen, S. Thongjuea, K. Theilgaard-Mönch, C. Nerlov, Identification of two distinct pathways of human myelopoiesis. *Sci. Immunol.* **4**, eaau7148 (2019). doi:10.1126/sciimmunol.aau7148 Medline
39. J. Grootens, J. S. Ungerstedt, M. Ekoff, E. Rönnberg, M. Klimkowska, R.-M. Amini, M. Arock, S. Söderlund, M. Mattsson, G. Nilsson, J. S. Dahlin, Single-cell analysis reveals the KIT D816V mutation in haematopoietic stem and progenitor cells in systemic mastocytosis. *EBioMedicine* **43**, 150–158 (2019). doi:10.1016/j.ebiom.2019.03.089 Medline
40. K. Reckzeh, H. Kizilkaya, A. S. Helbo, M. E. Alrich, A. G. Deslauriers, A. Grover, N. Rapin, F. Asmar, K. Grønbæk, B. Porse, N. Borregaard, D. Vestweber, C. Nerlov, K. Theilgaard-Mönch, Human adult HSCs can be discriminated from lineage-committed HPCs by the expression of endomucin. *Blood Adv.* **2**, 1628–1632 (2018). doi:10.1182/bloodadvances.2018015743 Medline
41. C. Sanada, J. Xavier-Ferrucio, Y.-C. Lu, E. Min, P.-X. Zhang, S. Zou, E. Kang, M. Zhang, G. Zerafati, P. G. Gallagher, D. S. Krause, Adult human megakaryocyte-erythroid progenitors are in the CD34+CD38mid fraction. *Blood* **128**, 923–933 (2016). doi:10.1182/blood-2016-01-693705 Medline
42. S. Amon, F. Meier-Abt, L. C. Gillet, S. Dimitrieva, A. P. A. Theocharides, M. G. Manz, R. Aebersold, Sensitive Quantitative Proteomics of Human Hematopoietic Stem and Progenitor Cells by Data-independent Acquisition Mass Spectrometry. *Mol. Cell. Proteomics* **18**, 1454–1467 (2019). doi:10.1074/mcp.TIR119.001431 Medline
43. O. Valiron, C. Gorka, Histone H1(0) expression is restricted to progenitor cells during human hematopoiesis. *Eur. J. Cell Biol.* **72**, 39–45 (1997). Medline
44. J. L. Villeval, U. Testa, G. Vinci, H. Tonthat, A. Bettaieb, M. Titeux, P. Cramer, L. Edelman, H. Rochant, J. Breton-Gorius, Carbonic anhydrase I is an early specific marker of normal human erythroid differentiation. *Blood* **66**, 1162–1170 (1985). doi:10.1182/blood.V66.5.1162.1162 Medline
45. D. D. Metcalfe, R. Pawankar, S. J. Ackerman, C. Akin, F. Clayton, F. H. Falcone, G. J. Gleich, A.-M. Irani, M. W. Johansson, A. D. Klion, K. M. Leiferman, F. Levi-Schaffer, G. Nilsson, Y. Okayama, C. Prussin, J. T. Schroeder, L. B. Schwartz, H.-U. Simon, A. F. Walls, M. Triggiani, Biomarkers of the involvement of mast cells, basophils and eosinophils in asthma and allergic diseases. *World Allergy Organ. J.* **9**, 7 (2016). doi:10.1186/s40413-016-0094-3 Medline
46. S. Rørvig, O. Østergaard, N. H. Heegaard, N. Borregaard, Proteome profiling of human neutrophil granule subsets, secretory vesicles, and cell membrane: Correlation with transcriptome profiling of neutrophil precursors. *J. Leukoc. Biol.* **94**, 711–721 (2013). doi:10.1189/jlb.1212619 Medline
47. Tabula Muris Consortium; Overall coordination; Logistical coordination; Organ collection and processing; Library preparation and sequencing; Computational data analysis; Cell type annotation; Writing group; Supplemental text writing group; Principal investigators, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018). doi:10.1038/s41586-018-0590-4 Medline
48. D. A. Anderson 3rd, C. A. Dutertre, F. Ginhoux, K. M. Murphy, Genetic models of human and mouse dendritic cell development and function. *Nat. Rev. Immunol.* **21**, 101–115 (2021). doi:10.1038/s41577-020-00413-x Medline
49. L. Broderick, S. Yost, D. Li, M. D. McGeough, L. M. Booshehri, M. Guaderrama, S. D. Brydges, K. Kucharova, N. C. Patel, M. Harr, H. Hakonarson, E. Zackai, I. G. Cowell, C. A. Austin, B. Hügle, C. Gebauer, J. Zhang, X. Xu, J. Wang, B. A. Croker, K. A. Frazer, C. D. Putnam, H. M. Hoffman, Mutations in topoisomerase IIβ result in a B cell immunodeficiency. *Nat. Commun.* **10**, 3644 (2019). doi:10.1038/s41467-019-11570-6 Medline
50. Z. Fang, X. Liu, G. Peltz, GSEApy: A comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, btac757 (2023). doi:10.1093/bioinformatics/btac757 Medline
51. T. Simsek, F. Kocabas, J. Zheng, R. J. Deberardinis, A. I. Mahmoud, E. N. Olson, J. W. Schneider, C. C. Zhang, H. A. Sadek, The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. *Cell Stem Cell* **7**, 380–390 (2010). doi:10.1016/j.stem.2010.07.011 Medline
52. A. Gayoso, Z. Steier, R. Lopez, J. Regier, K. L. Nazor, A. Streets, N. Yosef, Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021). doi:10.1038/s41592-020-01050-x Medline
53. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018). doi:10.1038/s41592-018-0229-2 Medline
54. F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019). doi:10.1186/s13059-019-1663-x Medline
55. N. Cabezas-Wallscheid, D. Klimmeck, J. Hansson, D. B. Lipka, A. Reyes, Q. Wang, D. Weichenhan, A. Lier, L. von Paleske, S. Renders, P. Wünsche, P. Zeisberger, D. Brocks, L. Gu, C. Herrmann, S. Haas, M. A. G. Essers, B. Brors, R. Eils, W. Huber, M. D. Milsom, C. Plass, J. Krijgsveld, A. Trumpp, Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell* **15**, 507–522 (2014). doi:10.1016/j.stem.2014.07.005 Medline
56. Z. J. Cao, G. Gao, Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022). doi:10.1038/s41587-022-01284-4 Medline
57. Y. Liu, Y. Chen, X. Deng, J. Zhou, ATF3 Prevents Stress-Induced Hematopoietic Stem Cell Exhaustion. *Front. Cell Dev. Biol.* **8**, 585771 (2020). doi:10.3389/fcell.2020.585771 Medline
58. J. Jiang, Y.-S. Chan, Y.-H. Loh, J. Cai, G.-Q. Tong, C.-A. Lim, P. Robson, S. Zhong, H.-H. Ng, A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.* **10**, 353–360 (2008). doi:10.1038/ncb1698 Medline
59. E. Laurenti, C. Frelin, S. Xie, R. Ferrari, C. F. Dunant, S. Zandi, A. Neumann, I. Plumb, S. Doulatov, J. Chen, C. April, J.-B. Fan, N. Iscove, J. E. Dick, CDK6 levels regulate quiescence exit in human hematopoietic stem cells. *Cell Stem Cell* **16**, 302–313 (2015). doi:10.1016/j.stem.2015.01.017 Medline
60. H. Zhang, M. Ye, R. S. Welner, D. G. Tenen, Sox4 Is Required for the Formation and Maintenance of Multipotent Progenitors. *Blood* **124**, 1577 (2014). doi:10.1182/blood.V124.21.1577.1577
61. S. Zaunz, L. Lauwereins, M. Bajaj, B. Guapo Neves, F. Cadacio, L. Cleuren, C. Notredame, J. Marteijn, K. De Keersmaecker, C. M. Verfaillie, Loss of the Base Excision Repair Gene Leads to Dysfunctional Adult Hematopoietic Stem and Progenitor Cells. *Blood* **138**, 3267 (2021). doi:10.1182/blood-2021-146369
62. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005). doi:10.1073/pnas.0506580102 Medline
63. P. Hernández-Malmierca, D. Vonficht, A. Schnell, H. J. Uckelmann, A. Bollhagen, M. A. A. Mahmoud, S.-L. Landua, E. van der Salm, C. L. Trautmann, S. Raffel, F. Grünschläger, R. Lutz, M. Ghosh, S. Renders, N. Correia, E. Donato, K. O. Dixon, C. Hirche, C. Andresen, C. Robens, P. S. Werner, T. Boch, D. Eisel, W. Osen, F. Pilz, A. Przybylla, C. Klein, F. Buchholz, M. D. Milsom, M. A. G. Essers, S. B. Eichmüller, W.-K. Hofmann, D. Nowak, D. Hübschmann, M. Hundemer, C. Thiede, L. Bullinger, C. Müller-Tidow, S. A. Armstrong, A. Trumpp, V. K. Kuchroo, S. Haas, Antigen presentation safeguards the integrity of the hematopoietic stem cell pool. *Cell Stem Cell* **29**, 760–775.e10 (2022). doi:10.1016/j.stem.2022.04.007 Medline
64. J. Hoggatt, P. Singh, J. Sampath, L. M. Pelus, Prostaglandin E2 enhances hematopoietic stem cell homing, survival, and proliferation. *Blood* **113**, 5444–5455 (2009). doi:10.1182/blood-2009-01-201335 Medline
65. T. Michurina, P. Krasnov, A. Balazs, N. Nakaya, T. Vasilieva, B. Kuzin, N. Khrushchov, R. C. Mulligan, G. Enikolopov, Nitric oxide is a regulator of hematopoietic stem cell activity. *Mol. Ther.* **10**, 241–248 (2004). doi:10.1016/j.ymthe.2004.05.030 Medline
66. K. Banki, E. Hutter, E. Colombo, N. J. Gonchoroff, A. Perl, Glutathione levels and sensitivity to apoptosis are regulated by changes in transaldolase expression. *J. Biol. Chem.* **271**, 32994–33001 (1996). doi:10.1074/jbc.271.51.32994 Medline
67. A. Almars, "Heterochromatin protein HP1BP3 modulates haematopoiesis and leukemogenesis", thesis, University of Nottingham, (2021).
68. O. Bereshchenko, O. Lo Re, F. Nikulenkov, S. Flamini, J. Kotaskova, T. Mazza, M.-M. Le Pannérer, M. Buschbeck, C. Giallongo, G. Palumbo, G. Li Volti, V. Pazienza, L. Cervinek, C. Riccardi, L. Krejci, S. Pospisilova, A. F. Stewart, M. Vinciguerra, Deficiency and haploinsufficiency of histone macroH2A1.1 in mice recapitulate

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

hematopoietic defects of human myelodysplastic syndrome. *Clin. Epigenetics* **11**, 121 (2019). doi:10.1186/s13148-019-0724-z Medline
69. Z. Thompson, L. Z. Luo, B. Wang, J.-H. Kim, H. Woo, L. Cope, L. Resar, High Mobility Group A1 Chromatin Regulators Drive Regenerative Capacity during Stress Hematopoiesis through Modulation of TNF-α and Aging Transcriptional Networks. *Blood* **142**, 4050 (2023). doi:10.1182/blood-2023-189964
70. J. W. Shin, K. R. Spinler, J. Swift, J. A. Chasis, N. Mohandas, D. E. Discher, Lamins regulate cell trafficking and lineage maturation of adult human hematopoietic cells. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18892–18897 (2013). doi:10.1073/pnas.1304996110 Medline
71. Y. Hu, G. K. Smyth, ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J. Immunol. Methods* **347**, 70–78 (2009). doi:10.1016/j.jim.2009.06.008 Medline
72. S. Hadjur, K. Ung, L. Wadsworth, J. Dimmick, E. Rajcan-Separovic, R. W. Scott, M. Buchwald, F. R. Jirik, Defective hematopoiesis and hepatic steatosis in mice with combined deficiencies of the genes encoding Fancc and Cu/Zn superoxide dismutase. *Blood* **98**, 1003–1011 (2001). doi:10.1182/blood.V98.4.1003 Medline
73. J. S. Friedman, V. I. Rebel, R. Derby, K. Bell, T.-T. Huang, F. A. Kuypers, C. J. Epstein, S. J. Burakoff, Absence of mitochondrial superoxide dismutase results in a murine hemolytic anemia responsive to therapy with a catalytic antioxidant. *J. Exp. Med.* **193**, 925–934 (2001). doi:10.1084/jem.193.8.925 Medline
74. G. Kustatscher, P. Grabowski, J. Rappsilber, Pervasive coexpression of spatially proximal genes is buffered at the protein level. *Mol. Syst. Biol.* **13**, 937 (2017). doi:10.15252/msb.20177548 Medline
75. J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, J. K. Pritchard, Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010). doi:10.1038/nature08872 Medline
76. A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, Y. Gilad, Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015). doi:10.1126/science.1260793 Medline
77. E. McShane, C. Sin, H. Zauber, J. N. Wells, N. Donnelly, X. Wang, J. Hou, W. Chen, Z. Storchova, J. A. Marsh, A. Valleriani, M. Selbach, Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**, 803–815.e21 (2016). doi:10.1016/j.cell.2016.09.015 Medline
78. D. P. Zolg, S. Gessulat, C. Paschke, M. Graber, M. Rathke-Kuhnert, F. Seefried, K. Fitzemeier, F. Berg, D. Lopez-Ferrer, D. Horn, C. Henrich, A. Huhmer, B. Delanghe, M. Frejno, INFERYS rescoring: Boosting peptide identifications and scoring confidence of database search results. *Rapid Commun. Mass Spectrom.* **39**, 9128 (2021). doi:10.1002/rcm.9128 Medline
79. S. White, J. Quinn, J. Enzor, J. Staats, S. M. Mosier, J. Almarode, T. N. Denny, K. J. Weinhold, G. Ferrari, C. Chan, FlowKit: A Python Toolkit for Integrated Manual and Automated Cytometry Analysis Workflows. *Front. Immunol.* **12**, 768541 (2021). doi:10.3389/fimmu.2021.768541 Medline
80. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007). doi:10.1093/biostatistics/kxj037 Medline
81. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6**, e21800 (2011). doi:10.1371/journal.pone.0021800 Medline
82. B. Furtwängler, N. Üresin, S. Richter, F. Theis, E. M. Schoof, B. T. Porse, Mapping early human blood cell differentiation using single-cell proteomics and transcriptomics, Version 1, Zenodo (2025); https://zenodo.org/records/15554000.
83. H. Sasaki, D. Kurotaki, T. Tamura, Regulation of basophil and mast cell development by transcription factors. *Allergol. Int.* **65**, 127–134 (2016). doi:10.1016/j.alit.2016.01.006 Medline
84. E. I. Zuniga, D. B. McGavern, J. L. Pruneda-Paz, C. Teng, M. B. Oldstone, Bone marrow plasmacytoid dendritic cells can differentiate into myeloid dendritic cells upon virus infection. *Nat. Immunol.* **5**, 1227–1234 (2004). doi:10.1038/ni1136 Medline
85. G. Barbarani, C. Fugazza, J. Strouboulis, A. E. Ronchi, The Pleiotropic Effects of GATA1 and KLF1 in Physiological Erythropoiesis and in Dyserythropoietic Disorders. *Front. Physiol.* **10**, 91 (2019). doi:10.3389/fphys.2019.00091 Medline
86. A. Fagnan, R. Drissen, C. Di Genua, Y. R. Meng, C. Nerlov, Myelo-Erythroid Lineage Segregation Is Regulated By the GATA-2 Interactome. *Blood* **142**, 2687 (2023). doi:10.1182/blood-2023-190285
87. U. Samulowitz, A. Kuhn, G. Brachtendorf, R. Nawroth, A. Braun, A. Bankfalvi, W. Böcker, D. Vestweber, Human endomucin: Distribution pattern, expression on high endothelial venules, and decoration with the MECA-79 epitope. *Am. J. Pathol.* **160**, 1669–1681 (2002). doi:10.1016/S0002-9440(10)61114-5 Medline

ACKNOWLEDGMENTS

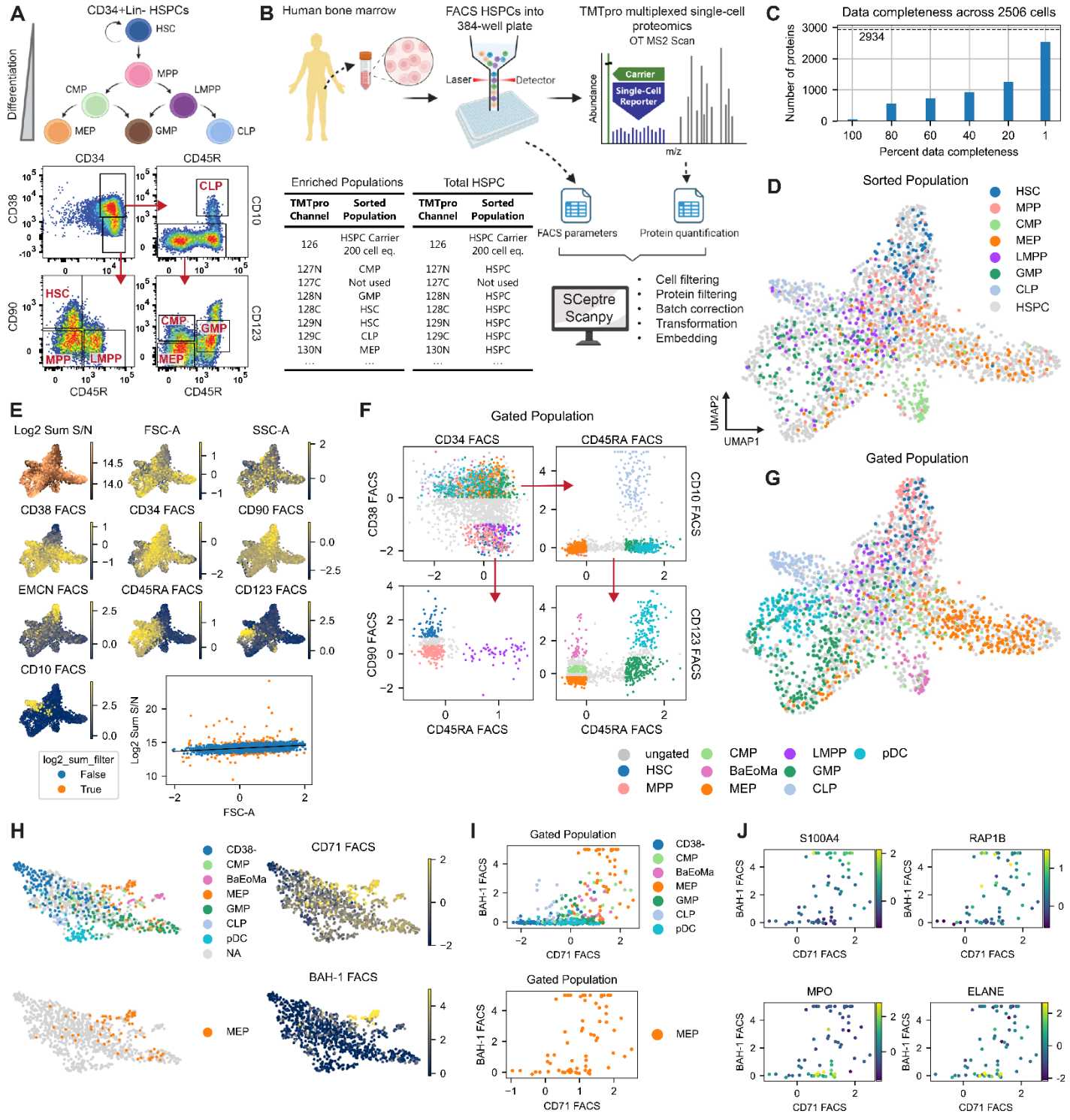
We thank Philipp Weiler for the setup of the first version of scProtVelo and the FACS core facility at BRIC for assistance with cell sorting. We thank Anne-Katrine Frank for coordinating access to human samples. **Funding:** B.F. is the recipient of a Copenhagen Bioscience Ph.D. stipend by the Novo Nordisk Foundation, NNF19SA0035442. S. R. is supported by the Helmholtz Association under the joint research school Munich School for Data Science. Work in the Porse lab was supported by grants from the Svend Andersen Foundation, the Candys Foundation, the Danish Cancer Society, the Eva and Henry Frænkel Memorial Foundation, the Independent Research Fund Denmark and through a center grant from the Novo Nordisk Foundation (Novo Nordisk Foundation Center for Stem Cell Biology, DanStem; Grant Number NNF17CC0027852). This work has been performed in the context of the Danish Research Center for Precision Medicine in Blood Cancers funded by the Danish Cancer Society (R223-A13071) and Greater Copenhagen Health Science Partners. Work in the Schoof lab was supported by grants from the Independent Research Fund Denmark (2067-00053B), Danish Cancer Society (R324-A17978), Novo Nordisk Foundation (NNF21OC0071016), and the Lundbeck Foundation (R413-2022-869). **Author contributions:** NÜ performed scp-MS sample collection and preparation. BF performed scp-MS measurements. DB performed CITE-seq sample preparation and measurement. AW performed CITE-seq raw data analysis. MBS performed the CRISPR knockout experiments. SR and FT created scProtVelo modeling. HH contributed to data interpretation. BF and SR analyzed and interpreted the data. KG and KTM aspirated human bone marrow samples, provided methodology and contributed to data interpretation. FT, ES, BTP Supervision. ES, BTP conceived the study. BF, NÜ, SR Writing - Original Draft. FT, ES, BTP Writing - Review & Editing. All authors approved the final version of the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The raw scp-MS and scRNA-seq data generated in this study have been deposited in European Genome-phenome Archive (EGA) under accession number EGAS00001007930. Code and data, including experimental data and scProtVelo, are available on Zenodo (*82*). The scProtVelo package is also available on github via github.com/theislab/scProtVelo. The external HSPC proteomics data were accessed via the PRIDE repository identifier PXD009255, and the accompanying RNA sequencing data were accessed via the Expression Omnibus with identifier GSE113182. **License information:** Copyright © 2025 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/content/page/science-licenses-journal-article-reuse

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adr8785
Supplementary Text
Figures S1 to S25
Tables S1 to S3
References (*83*–*87*)
MDAR Reproducibility Checklist

19 July 2024; resubmitted 15 February 2025
Accepted 14 July 2025
Published online 21 August 2025
10.1126/science.adr8785

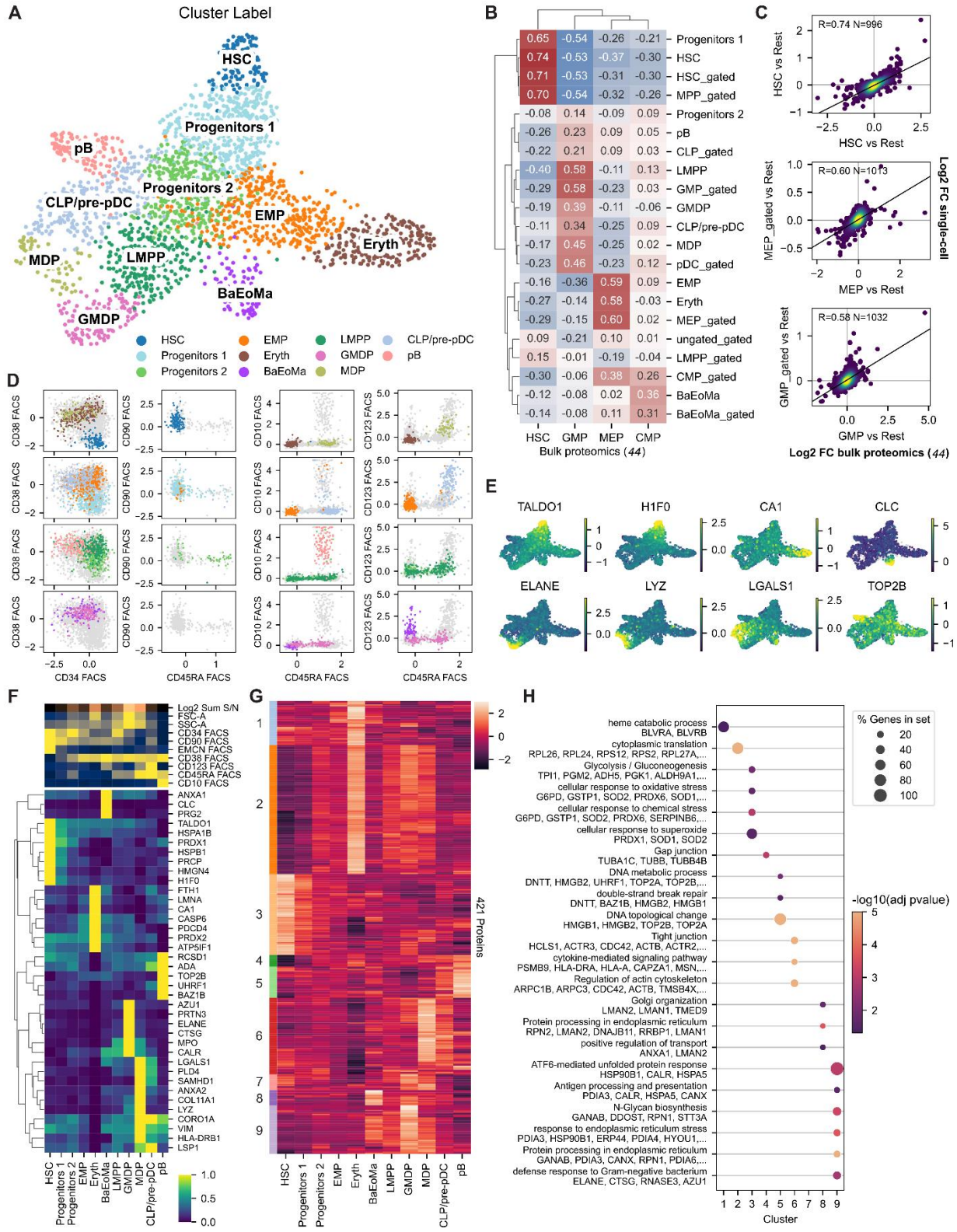
Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025



Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Fig. 1. A single-cell proteomics dataset of FACS-isolated human HSPCs. (**A**) The HSPC differentiation hierarchy (top) and sorting strategy (bottom). (**B**) Experimental workflow. Two different kinds of multiplexed sets were generated by sorting cells from specific gates ("Enriched Populations") or any Lin-CD34+ cell (Total HSPC). (**C**) Data completeness of scp-MS dataset. (**D**) UMAP of cells from scp-MS data, annotated with their respective sorting gate. (**E**) Overlay of total MS signal and FACS parameters onto the UMAP. Scatterplot shows the relationship between FACS forward scatter area and scp-MS total signal. Cells were removed if the absolute deviation of the Log2 Sum S/N from the regression line was above 0.8. (**F**) Computational gating strategy for retrospective annotation of cells from both enriched populations and total HSPC sorts. (**G**) UMAP annotated with gated populations from F. (**H**) UMAP of second scp-MS dataset with FACS markers CD71 and BAH-1 enabling positive selection of MEPs. (**I**) Scatterplot showing the putative true BAH-1+CD71+ MEP subset. Populations were annotated using the strategy in (F). Additional FACS markers reveal true MEPs. (**J**) Scatterplots from I with overlaid abundance levels of MEP (top) and GMP (bottom) marker proteins.

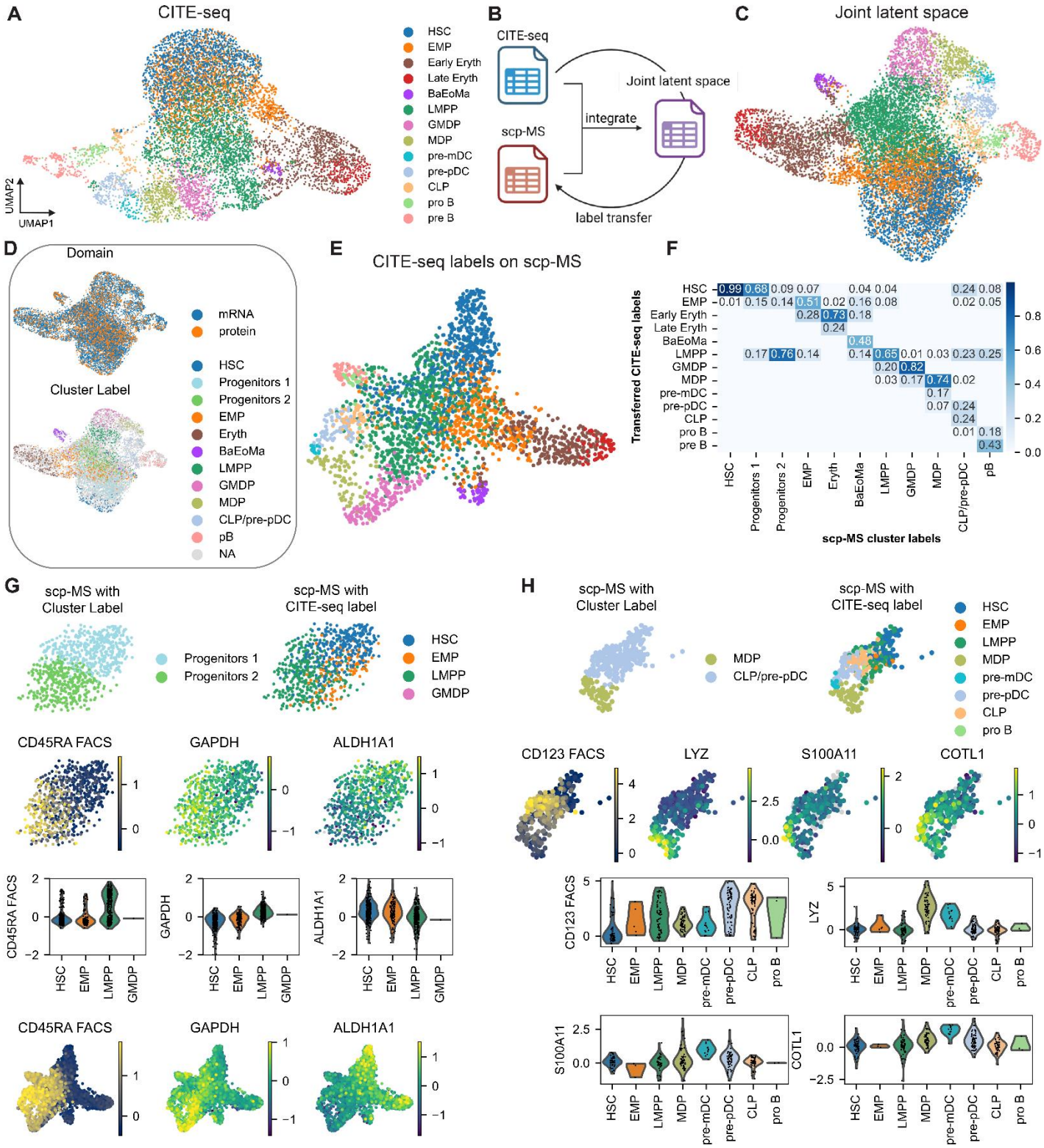
Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025



Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Fig. 2. **Unsupervised clustering reveals distinct cellular differentiation stages and functional protein covariance.** (**A**) UMAP of cells annotated with Leiden clusters. (**B**) Clustered heatmap showing the correlation of abundance changes between scp-MS measurements and bulk proteomics data (*42*), computed as Pearson correlation values of mean protein log2 fold-changes of a population vs all other populations. (**C**) Scatterplots of selected comparisons from (B). (**D**) Cell clusters are visualized on FACS gating scheme from Fig. 1F. Clusters were distributed over 4 panels for visibility. (**E**) UMAP annotated with the abundance of selected proteins. (**F**) Clustered heatmap of selected cluster-specific proteins. Values were mean aggregated and normalized between 0 and 1. Top shows aggregated FACS parameters and total signal. (**G**) Clustered heatmap of protein abundance across Leiden clusters. Values were mean aggregated. (**H**) Dotplot of top enriched terms per protein cluster. -Log10(adj. pvalue) was truncated to max. 5.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025



Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Fig. 3. Single-cell multi-omics analysis of transcriptomics and proteomics of the human HSPC compartment. (**A**) UMAP embedding of annotated CITE-seq dataset. (**B**) Schematic of GLUE integration approach. (**C**) UMAP embedding of integrated dataset of CITE-seq and scp-MS, overlaid with CITE-seq cell annotation that was transferred to the "protein" cells. (**D**) Distribution of "mRNA" cells and "protein" cells (top). Overlaid cluster labels from scp-MS (bottom). (**E**) scp-MS UMAP embedding with transferred CITE-seq cell annotation. (**F**) Heatmap related to the label transfer from "mRNA" to "protein" cells. Rows are the cell annotation from CITE-seq, columns the cluster labels from scp-MS. The values represent the fraction of cells in the respective scp-MS cluster that was assigned the respective CITE-seq label. Thus, columns add up to one. (**G**) Validation of cell label transfer of the CITE-seq labels EMP and LMPP onto scp-MS labels Progenitors 1 and 2. In the first three rows, cells were subsetted to Progenitors 1 and 2. Surface marker and protein abundance profiles align with CITE-seq label, revealing the early bifurcation into EMP and LMPP. (**H**) Validation of cell label transfer of the CITE-seq label pre-mDC onto scp-MS label MDP. Cells were subsetted to MDP and CLP/pre-pDC. Surface marker and protein abundance profiles align with the placement of the pre-mDC CITE-seq label, revealing an additional population missed by the scp-MS clustering.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

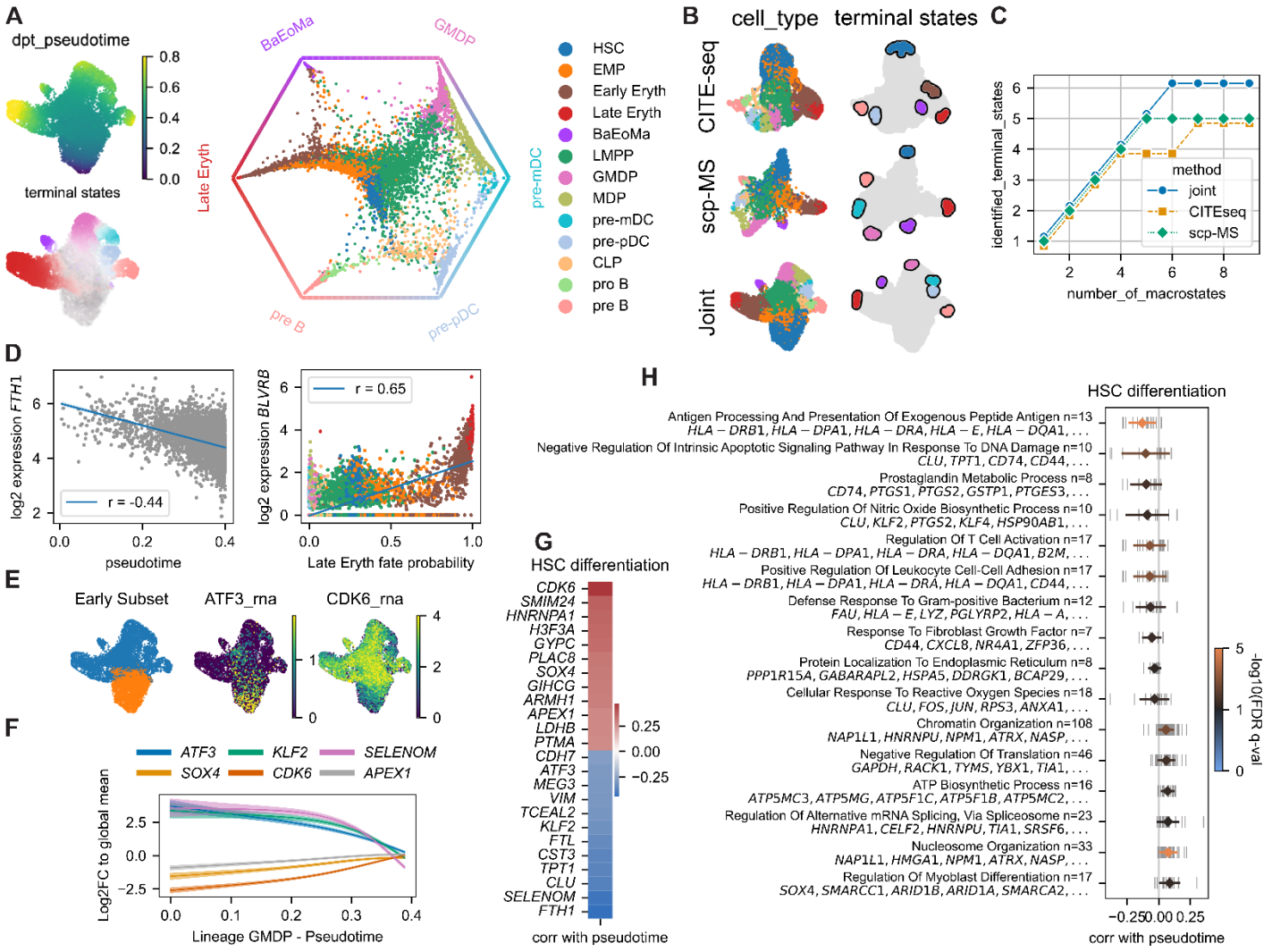
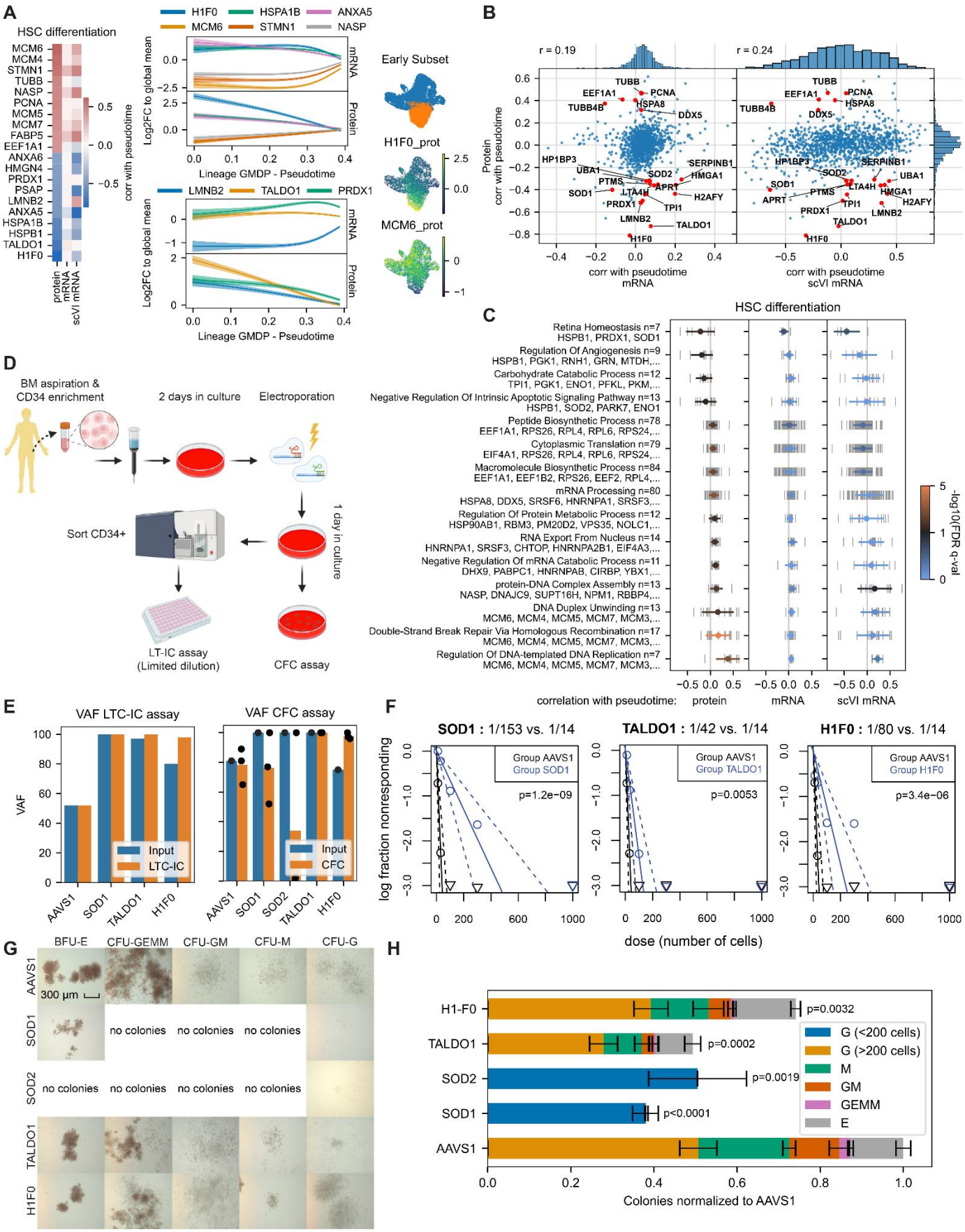


Fig. 4. Trajectory analysis on joint latent space recapitulates HSPC differentiation. (A) UMAP of joint dataset with overlaid pseudotime and terminal states determined by cellrank (left). Cellrank circular projection showing cells ordered by their relative fate probability for each lineage (right). (B and C) Comparison of cellrank terminal state detection between CITE-seq, scp-MS and the joint dataset. B shows the top six detected terminal states. C shows the number of correct terminal states (y-axis) at different numbers of macrostates used for terminal state determination (x-axis). (D) Plots exemplify the calculation of correlation of gene expression with pseudotime (left) or fate probability (right). (E to G) mRNA expression analysis in an early subset of cells to investigate HSC differentiation. (E) Cells that were selected as early subset and mRNA expression of selected genes with high correlations. (F) Modeled gene expression profiles along pseudotime. The early part of the GMDP lineage was used for visualization. (G) Genes with the highest correlation or anticorrelation with pseudotime. (H) GSEA of HSC differentiation. X-axis shows correlation of expression with pseudotime. Vertical gray lines represent individual genes in the set. The horizontal bar represents the mean and standard deviation.

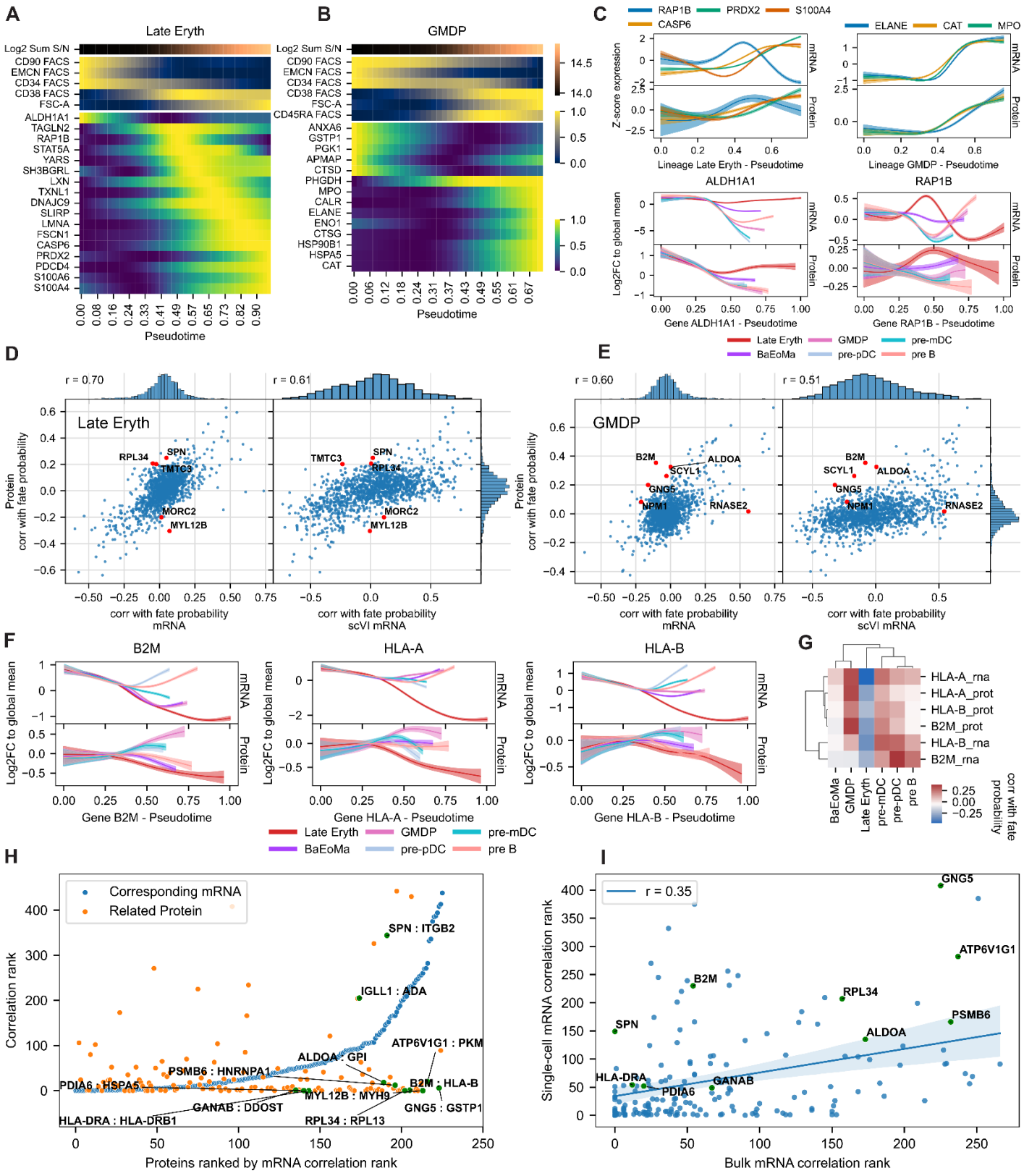
Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025



Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Fig. 5. Hematopoietic stem cell quiescence and differentiation are described differently on protein level. (**A**) Protein abundance change analysis in an early subset of cells to investigate HSC differentiation. Left: Proteins with the highest correlation and anticorrelation with pseudotime, including mRNA level information. Middle top: Modeled protein and mRNA abundance changes of selected genes. GMDP lineage was used as an example. Middle bottom: Modeled protein and mRNA abundance changes of selected genes that are not HSC-specific on mRNA level. Right: Cells that were selected as early subset and abundances of selected proteins. (**B**) Scatterplot comparing the correlation vectors between protein and mRNA (count-based and scVI modeled). Selected genes were annotated. (**C**) GSEA analysis of correlation vectors of HSC differentiation. Vertical gray lines represent individual genes in the set. The horizontal bar represents the mean and standard deviation. (**D**) Experimental workflow for the functional assessment of selected proteins. (**E**) Variant allele frequency (VAF) of CRISPR knockout. (**F**) Quantification of LTC-IC limiting dilution assay from one donor. Each data point was derived from seeding a given dilution in 10 wells. Statistical significance for the limited dilution experiments was assessed as described in (*71*). (**G**) Representative images of colonies of CFC assay. BFU=Burst Forming Unit, CFU=Colony Forming Unit, E=Erythroid, GEMM=Granulocyte/Erythroid/Megakaryocyte/Macrophage, M=Granulocyte/Macrophage, M=Macrophage, G=Granulocyte. (**H**) Quantification of CFCs. Total colony numbers were normalized to AAVS1, and statistical significance (total CFC numbers; unpaired *t* test) for the individual knockouts (3 technical replicates) was determined by comparison to their respective AAVS1 controls. Error bars represent SDs. Experiments in (E) to (H) were performed twice.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025



Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

Fig. 6. Protein level trajectory analysis reveals protein level-specific functional covariation. (**A** and **B**) Heatmap showing abundance profiles of selected proteins in the Late Eryth (A) and GMDP (B) trajectories. (**C**) Modeled abundance profiles for protein and mRNA level of selected genes. (**D** and **E**) Scatterplot comparing the correlation vectors between mRNA and protein level in the Late Eryth (D) and GMDP (E) trajectories. (**F**) Modeled abundance profiles of MHC I complex members B2M, HLA-A, and HLA-B. (**G**) Clustered heatmap of abundance profiles of MHC complex members on both mRNA and protein level. (**H**) Scatterplot of proteins (x-axis) and the correlation ranks of both mRNA and protein molecules (y-axis). Proteins (x-axis) were sorted by the correlation rank of the corresponding mRNA (blue dots). Orange dots depict the related protein (same KEGG pathway) with the lowest correlation rank. (**I**) Comparison of mRNA correlation rank between bulk and scp-MS data.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025

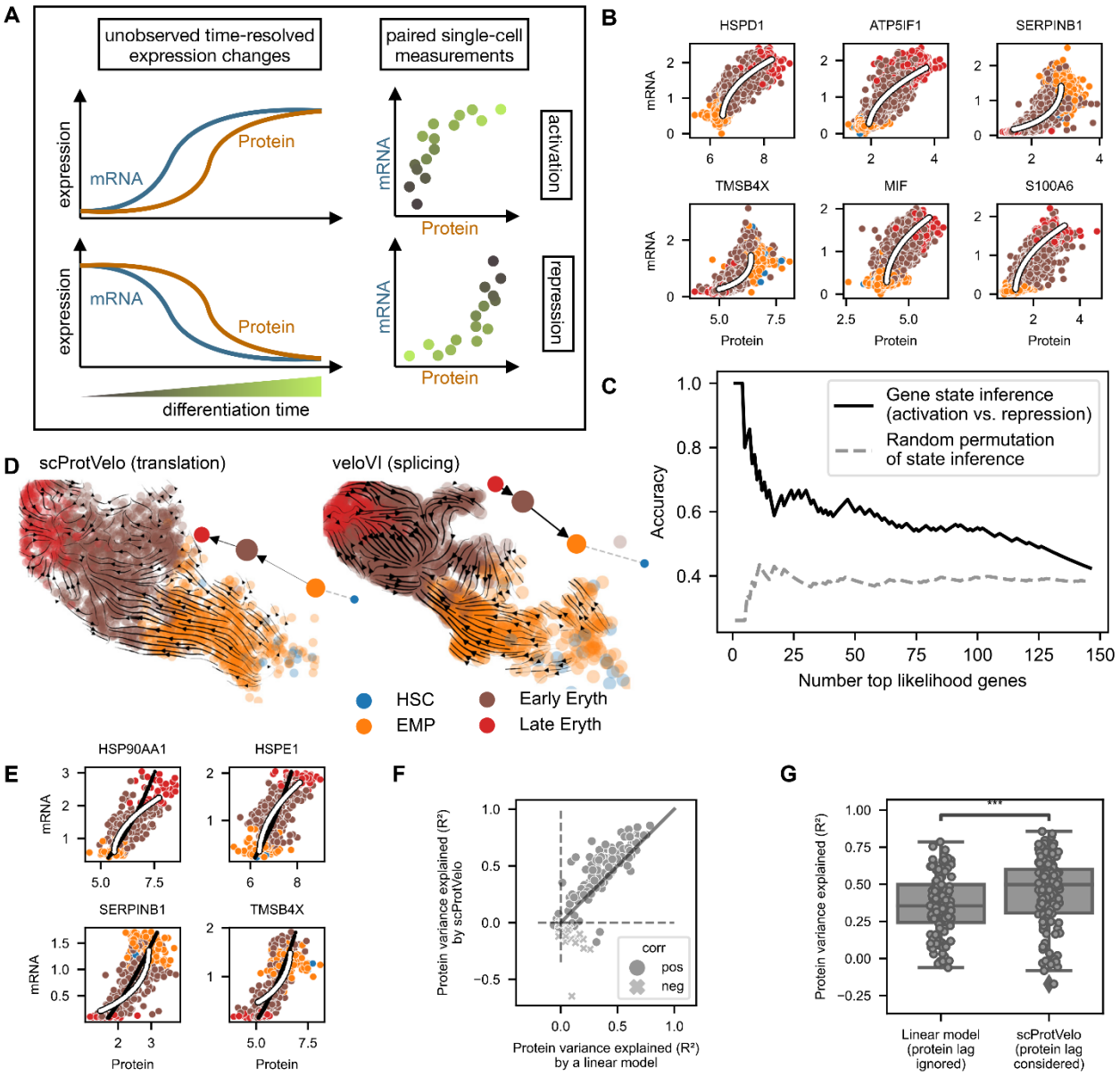


Fig. 7. scProtVelo models translation dynamics in cellular differentiation. (A) mRNA and protein abundance changes, time resolved vs. observed in paired single-cell data. The time delay between the modalities allows the distinction of up- and down-regulation in the data. (B) Exemplary model fits. (C) Performance of gene state predictions compared to uninformed predictions on genes that substantially change along the erythroid differentiation trajectory. (D) Translation-based (left) and splicing-based (right) velocities derived from the top 100 likelihood genes projected onto a UMAP embedding as well as the corresponding PAGA representations. (E) Exemplary comparison between fits from scProtVelo (white) and a linear model fitted in log space (black). (F) Gene-wise comparison of protein abundance variance explained by scProtVelo and a linear model, differentiated by the sign of the linear regression slope. (G) Global comparison of the protein abundance variance explained by scProtVelo and a linear model.

Downloaded from https://www.science.org at Helmholtz Zentrum Mnchen - Zentralbibliothek on October 13, 2025