nature medicine

Resource

https://doi.org/10.1038/s41591-025-03716-5

A reference model of circulating hematopoietic stem cells across the lifespan with applications to diagnostics

Received: 1 May 2024

Accepted: 11 April 2025

Published online: 27 June 2025

Check for updates

N. Furer[1,17], N. Rappoport[2,3,17], O. Milman[3,17], S. Tavor[4,5], A. Lifshitz[3], A. Bercovich[3], O. Ben-Kiki[3], A. Danin[1], M. Kedmi[6], Z. Shipony[7], D. Lipson[7], E. Meiri[7], G. Yanai[7], S. Shapira[8], N. Arber[8], S. Berdichevsky[9], J. Tyner[10,11], S. Joshi[10,11], D. Landau[12,13,14,15], S. Ganesan[12,13,14,15], N. Dusaj[12,13,14,15], P. Chamely[12,13,14,15], N. Kaushansky[1], N. Chapal-Ilani[1], R. Shamir[2], A. Tanay[3,17] ✉ & L. Shlush[1,5,16,17] ✉

With aging, deviation of human blood counts from their normal range accompanies the transition from health to disease. Hematopoietic stem and progenitor cells (HSPCs) deliver life-long multi-lineage output, but their variation across healthy humans with aging, and their diagnostic utility, haven't been characterized in depth thus far. To address this, we introduced an HSPC reference model using single-cell RNA profiling of circulating $CD34^+$ HSPCs from 148 healthy age- and sex-diverse individuals. We characterized physiological circulating HSPC composition, showed that age-related myeloid bias is predominant in older men and defined age-related transcriptional signatures in lymphoid progenitors. We further demonstrated the potential of this resource to facilitate the diagnosis of myelodysplastic syndrome (MDS) from peripheral blood without bone marrow sampling, defining classes of patients with MDS and abnormal lymphocyte, basophil or granulocyte progenitor frequencies. Our resource provides insights into HSPC reference ranges across the lifespan and has the potential to facilitate the clinical applications of single-cell genomics in hematology.

The basis for understanding and defining human pathophysiological states is a detailed description of interindividual heterogeneity among healthy individuals. Large population studies have identified wide interindividual differences in complete blood counts (CBCs) of healthy individuals[1] and exposed different age-related blood count changes, such as high red blood cell (RBC) distribution width (RDW), macrocytic anemia and a reduction in absolute lymphocyte counts[2]. The establishment of reference values, or population-wide normal ranges for certain blood parameters, has been crucial for patient evaluation, diagnosis and treatment.

Although CBC reference ranges are used in the clinic daily, the equivalent reference range for hematopoietic stem and progenitor cells (HSPCs) has not been established so far. As HSPCs reside mainly in the bone marrow (BM), access to these cells, especially in the healthy population, has been problematic, whereas their general paucity in the circulation made it quite challenging to characterize them efficiently from the blood. This has become feasible given modern technologies such as single-cell RNA sequencing (scRNA-seq).

Individual heterogeneity in the frequency of circulating HSPCs (cHSPCs) has been reported in the past and was linked to age, sex, smoking status, lipid profiles and hereditary factors[3], as well as to different pathological states[4]. Few studies have analyzed HSPC heterogeneity in higher resolution, but their sample size was limited[5,6]. Previous studies, including some based on scRNA-seq analysis[7], demonstrated

A full list of affiliations appears at the end of the paper. ✉e-mail: amos.tanay@weizmann.ac.il; liranshlush3@gmail.com

that most HSPC subpopulations can be identified in the peripheral blood (PB)[8] and functional stem cells were identified in the PB of mice[9] and humans[7].

We have developed a reference model for healthy cHSPC distributions and provided proof-of-concept evidence supporting potential diagnostic applications. We applied scRNA-seq analysis to cHSPCs from 148 healthy age- and sex-diverse individuals, to capture a spectrum of states, from hematopoietic stem cells (HSCs), through early common myeloid and lymphoid progenitor states and more specific progenitor populations. All data can be explored in https://apps.tanaylab.com/MCV/blood_aging. We discovered extensive interindividual heterogeneity in the frequency of cHSPC subtypes and found that these correlate with certain CBC parameters, aging and the presence of clonal hematopoiesis (CH). We then developed tools for projecting new samples on our reference model and analyzed 73 additional samples from patients with cytopenia and myelodysplastic syndrome (MDS), to demonstrate our model's potential applications in MDS diagnosis. The healthy reference model and methodologies used in the present study provide a framework for the deployment of single-cell genomics in hematology, for the diagnosis of MDS, and possibly other stem cell-related blood malignancies, from the PB, reducing the need for BM analysis.

Results

HSPC states observed across humans in PB

To evaluate interpersonal diversity in subtype distribution and regulation of cHSPCs, we combined multiplexed scRNA-seq, bulk DNA genotyping and integrated clinical data (Fig. 1a). Multiplexing was resolved using SNPs identified in the 3′-UTR of cHSPCs' RNA, facilitating precise matching of cells to individuals and improving control for batch effects and doublets. Altogether, we collected cHSPCs from 79 men and 69 women between the ages of 23 years and 91 years (median 61.5 years) (Extended Data Fig. 1a and Supplementary Table 1). We performed deep targeted somatic mutation analysis to identify cases of CH (Supplementary Table 1)[10]. After quality control and filtering, we retained 840,104 single-cell profiles, which were normalized to control for sequencing-platform batch effects and combined to construct and annotate a metacell manifold model[11] (Extended Data Fig. 1b,c). We retained 626,966 $CD34^+$ single cells for downstream analysis (Extended Data Fig. 1d). These formed a rich repertoire of states, associated with cHSPCs and their differentiation trajectories (Fig. 1b and Extended Data Fig. 1e,f). The derived model recapitulated and deepened earlier characterization efforts of HSPC states from the BM. We noted that, although we could not assume that cHSPCs fully reflect BM HSPC dynamics, previous studies, as well as our own BM scRNA-seq comparisons, supported at least partial compatibility between the two[12] (Extended Data Figs. 2 and 3a). One notable characteristic specific to cHSPCs was, however, the repression of cell-cycle gene expression (Extended Data Fig. 3b), previously demonstrated by others[7]. Importantly, we found our cHSPC model to be consistent across individuals. The median number of individuals contributing cells to each metacell was 84 and all metacells included cells from at least 47 individuals. Expression differences between cell states were greater than between individuals, limiting individually specific differential expression when controlling for each sample's cell distribution over the atlas states (Extended Data Fig. 3c,d). Altogether, the data suggest that cHSPCs, although not fully reflecting BM hematopoiesis, can serve as a highly accessible proxy for hematopoietic dynamics.

HLF, *GATA3*, *HOXB5* and *TLE4* as HSC TFs

One of the hallmarks of our cHSPC model is a distinct HSC state that is transcriptionally linked with two major differentiation gradients: the first represents a continuum of common lymphoid progenitors (CLPs; subdivided into early (E) mid (M) and late (L) states). The second, more common branch, represents multipotent progenitor (MPP) states and their differentiation toward granulocyte–monocyte progenitors (GMPs), erythrocyte progenitors (ERYPs) and basophil, eosinophil or mast progenitors (BEMPs). Platelet contamination prevented precise megakaryocyte progenitor modeling (Extended Data Fig. 3e), such that states at the base of the myeloid trajectory were annotated as megakaryocyte, erythrocyte, basophil, eosinophil or mast progenitors (MEBEMPs, subdivided into early (E) and late (L) states).

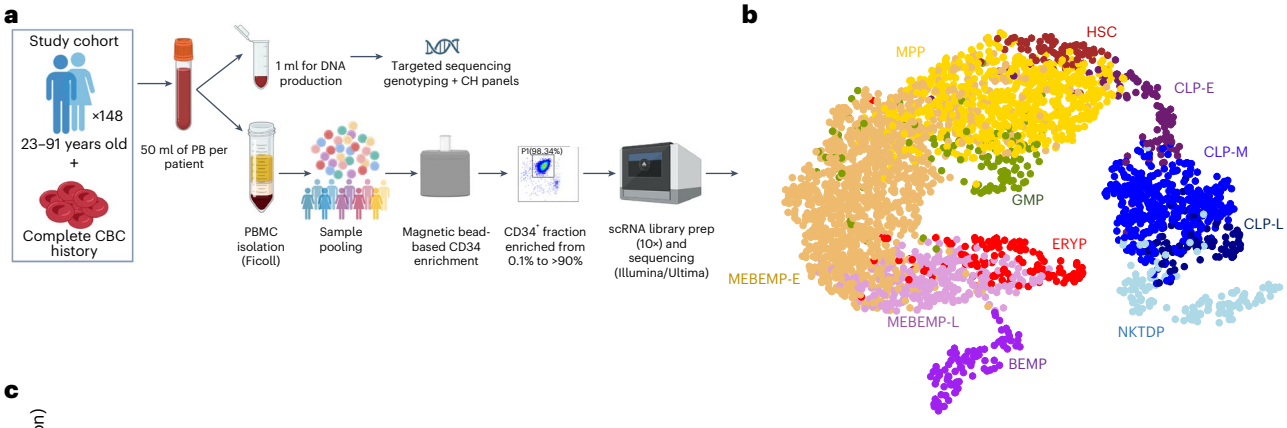
Early HSCs are marked by high *AVP* and *HLF* expression and were previously shown to represent a rare cell population enriched with self-renewal capacity in both the BM and cord blood[13]. Our model included data on 14,440 *HLF* and *AVP* expressing HSCs that could be matched with cells from independent BM atlases[14], suggesting that, under a steady state, HSCs with potential self-renewal capacity are present in the PB (Extended Data Fig. 3f). Further functional studies are needed to establish this finding. Together with *HLF* and *AVP*, we discovered 14 genes expressed at least 1.75-fold higher in HSCs compared with their 2 immediate differentiation branches (Extended Data Fig. 3g and Supplementary Table 2). We identified several transcription factors (TFs) enriched in HSCs, including *HOXB5*, *TLE4* and *GATA3* (Fig. 1c). *GATA3* was previously reported to regulate self-renewal in murine long-term HSCs[15]. Its role in human HSCs has not been studied thus far. We note that, although the HSC state is defined by unique markers that are symmetrically downregulated on exiting to the CLP and MEBEMP trajectories (Fig. 1c), it also expresses several lineage-specific regulators at intermediate levels, which are bifurcating anti-symmetrically on exiting the HSC state to the CLP and MEBEMP trajectories (Fig. 1d and Extended Data Fig. 3g). This may suggest that the multipotent capacity of HSCs is associated with intermediate expression of multiple regulators, which is resolved with differentiation.

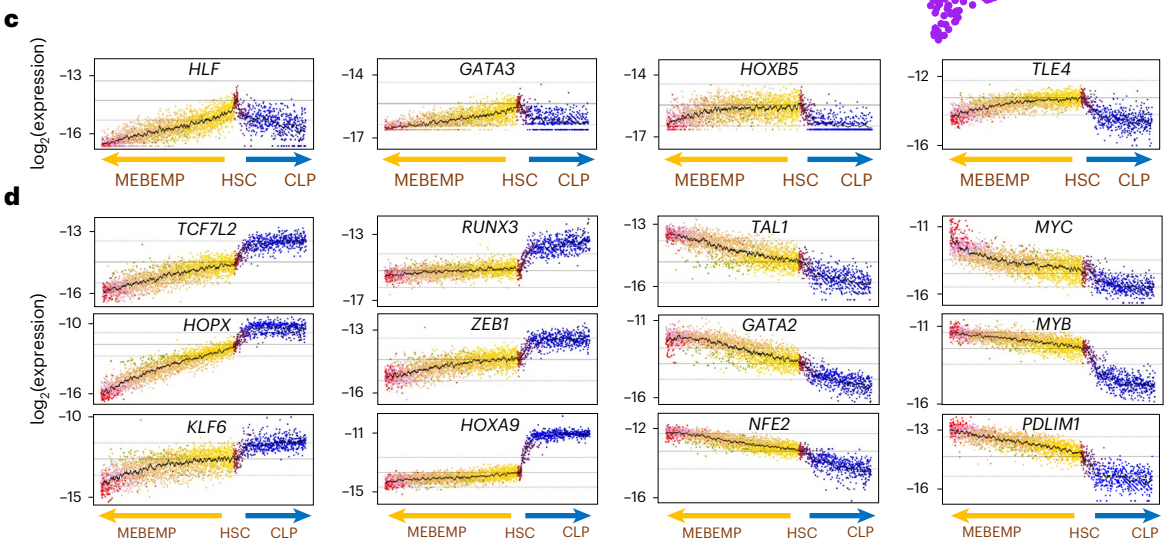
BEMPs and NKTDPs are enriched in cHSPCs

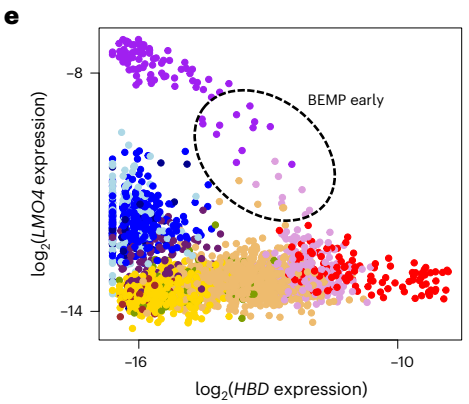
The cHSPC atlas was enriched for BEMPs. Although classic studies linked these cells with a GMP origin, more recent studies suggested that these emerge, at least in part, from erythroid progenitors in both mice and humans[7,16]. Our analysis identified a small population of metacells linking BEMPs with their MEBEMP-L precursors (Fig. 1e). This highlighted TFs (Fig. 1f) and other factors (Extended Data Fig. 4a) positively or negatively regulated in this postulated early stage of BEMP specification. Another interesting cHSPC population included lymphoid states with high *ACY3* expression and intermediate-to-low *DNTT* levels, a combination rarely found in human BM but present in PB (Extended Data Fig. 4b). We observed co-variation of key T cell regulators within this population and anti-correlation of these factors with some Hallmark plasmacytoid dendritic cell (pDC) regulators, as demonstrated

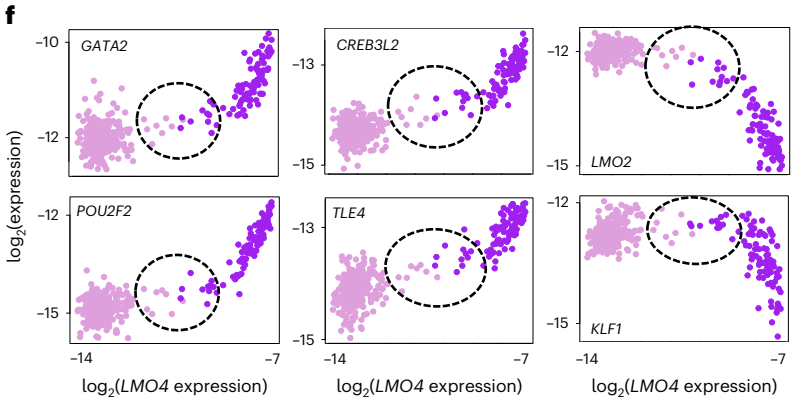
Fig. 1 | Mapping cHSPCs. a, Experimental design. **b**, Annotated two-dimensional Uniform Manifold Approximation and Projection (UMAP) of our metacell manifold after filtration of metacells with low *CD34* expression. For all subsequent panels in Figs. 1–3, metacell color denotes cell state as here. **c**,**d**, Symmetrical (**c**) and asymmetrical (**d**) regulation of specific HSC TFs on bifurcation to the CLP (right) and MEBEMP (left) lineages. Each panel shows the expression of one gene (*y* axis). Metacells in all panels are ordered (left to right) by increasing *AVP* expression in the MEBEMP lineage and decreasing *AVP* expression in the CLP lineage. The *y* axes denote $\log_2$(fractional expression) of each gene. **e**, The metacell population of interest (dashed line) linking BEMPs to their MEBEMP-L precursors. **f**, Positively and negatively regulated TFs involved in early BEMP differentiation. **g**, Gene–gene plot of *IRF8* against *TCF7* expression as Hallmark markers of DC and T cell differentiation, respectively. The high *ACY3* NKTDP metacell population of interest is depicted (dashed line). **h**, This population exhibits high expression of both T cell and DC regulators, forming a gradient consisting of NK or T cell-like progenitors exhibiting a high *TCF7*:*IRF8* expression ratio along with high expression of other T cell Hallmarks such as *CD7*, *MAF*, *IL7R*, *TRBC2* and DC-like progenitors exhibiting a low *TCF7*:*IRF8* expression ratio, along with high expression of other DC Hallmarks, such as the myeloid TF *PU.1* and the MHC-II gene *CD74*. Panel **a** created with BioRender.com.

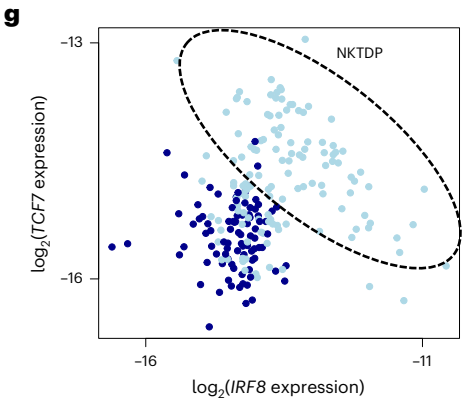
https://doi.org/10.1038/s41591-025-03716-5

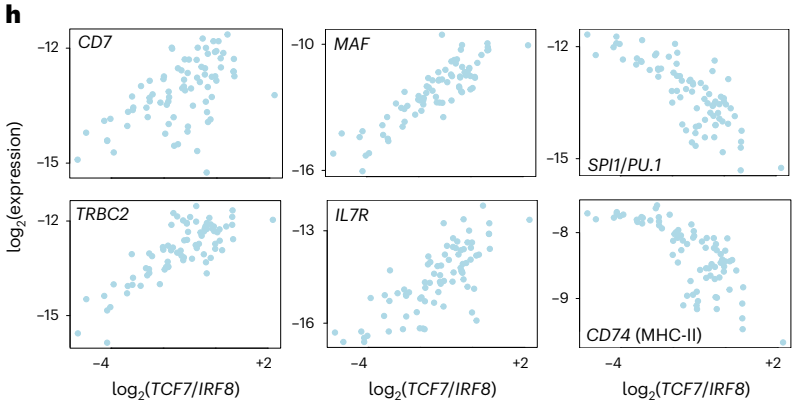












https://doi.org/10.1038/s41591-025-03716-5

by comparison of *TCF7* and *IRF8* expression (Fig. 1g,h and Extended Data Fig. 4c). We therefore termed this population natural killer (NK) cell, T cell and DC progenitors (NKTDPs)[17,18]. To summarize, our map of cHSPCs showed a rich spectrum of progenitor states, which refined previous analyses and a remarkable consistency of these states across individuals. This provided an opportunity for deciphering interindividual hematopoietic variability based on our solid and quantitative definition of cHSPC subtypes.

Interindividual variation in cHSPC state composition

To study interindividual cHSPC variation, we first analyzed cell-state compositions by quantifying cell-state relative frequencies within each individual's single-cell ensemble (Fig. 2a). These frequencies varied extensively between individuals as shown in Fig. 2b. For example, HSCs and CLP-Ms, representing 2.4% and 12.6% of the CD34$^+$ population on average, showed s.d. values of 1.0% and 6.8%, respectively. The abundant MPP and MEBEMP-E states (mean frequencies of 20.7% and 37.6%) showed smaller relative variations (s.d. values of 4.9% and 5.8%, respectively). To analyze the stability of cell-state frequencies across time and sampling instances, we re-sampled 20 individuals 1 year after their original sampling date. Both CLP (CLP-E, CLP-M, CLP-L and NKTDP) and MEBEMP (MEBEMP-E, MEBEMP-L, ERYP and BEMP) frequencies were stable within the same individual across time (Fig. 2c).

To analyze composition in higher resolution, we profiled each individual's enrichment over the CLP and MEBEMP trajectories. Clustering of these enrichment profiles yielded six archetypes of cHSPC composition within the healthy population (classes I–VI, Fig. 2d). These were composed of individuals with relative lymphoid enrichment (classes I and II) or depletion (classes V and VI), further subdivided by a stemness gradient, enriched in classes II, IV and VI and depleted in classes I, III and V. Analysis of technical and biological replicates confirmed this variation to be robust and individual-specific (Extended Data Fig. 5a,b). To summarize, we constructed cHSPC subtype normal reference ranges and showed that, although HSPC cell states are consistent among healthy individuals, their compositions are highly variable.

Circulating HSPC frequencies correlate with CBCs and CH

To extract an initial clinical annotation for the observed interindividual variation in cHSPC state frequencies, we correlated individual compositions with longitudinal CBC data (Methods). We observed a significant positive correlation ($P < 0.01$) between PB mature lymphocyte counts (%) and CLP frequencies (Fig. 2e, left). Given the high variability in female RBC counts and volumes during menstruation, pregnancies and prolonged perimenopausal periods, we analyzed RBC indices (count, hematocrit (HCT), mean corpuscular volume (MCV) and RDW) separately for men and women. We observed a significant negative correlation ($P < 0.02$) between CLP frequencies and HCT (men, Fig. 2e, middle) and a significant positive correlation ($P < 0.01$) between increased RDW and relative CLP depletion (men, Fig. 2e, right). Our previous work[19] and the work of others[20] correlated high RDW with CH and predisposition to acute myeloid leukemia. Our data suggest that reduction in CLP frequencies is associated with CH (Extended Data Fig. 6a). A similar trend was suggested by genotyping of transcriptomes (GoT)[21] performed on one of our *DNMT3A* R882H cases, showing a lower fraction of CLP cells within the mutant clone ($P < 0.005$, Fisher's exact test; Extended Data Fig. 6b). Although this trend was suggested in other GoT data[22], sample size is insufficient to prove it statistically and explore the clonal mechanisms underlying it. To further explore the association between CH and RDW, we studied a cohort of 18,147 healthy individuals for whom we had both longitudinal CBCs and DNA available. We identified 602 individuals with a high RDW (>15%, not meeting minimal criteria for MDS diagnosis) and 602 age- and sex-matched normal RDW controls. We performed deep targeted sequencing to identify leukemia-associated mutations on both high-RDW individuals and controls, and found a significant enrichment of CH$^+$ cases in the high-RDW group (Fisher's exact test, $P < 0.002$; Fig. 2f and Supplementary Tables 3 and 4). Altogether, the data suggest associations across decreased CLP frequencies and elevated RDW and CH. Determination of the existence of a direct (and perhaps three-way) linkage for these variables requires further investigation.

Age-related myeloid bias is predominantly observed in men

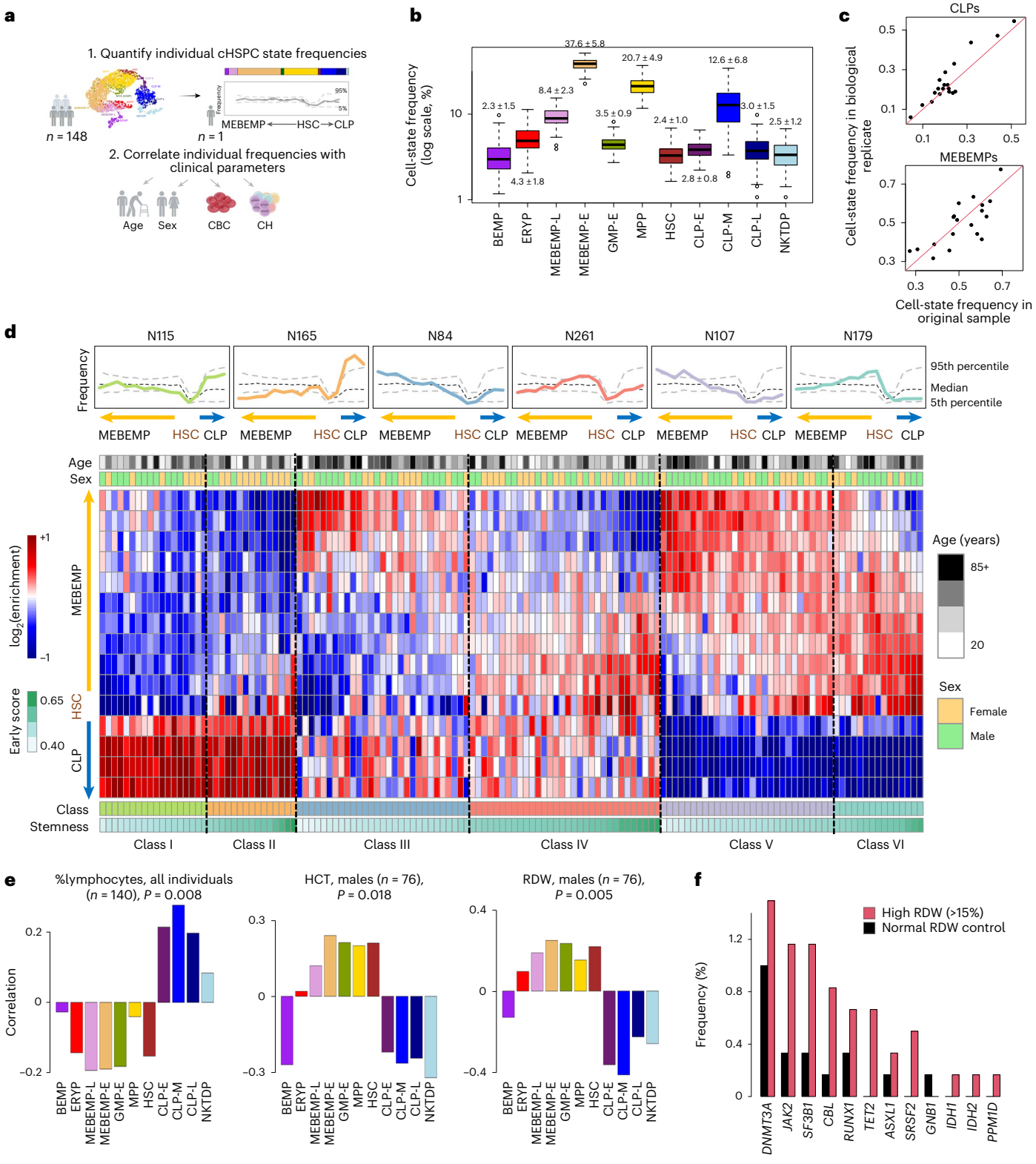
Analysis of age-linked compositional changes in cHSPCs within CH-negative individuals showed a remarkable increase in myeloid (MEBEMP) to lymphoid (CLP) progenitor ratios in men (when comparing <50 to >60-year-old individuals; Fig. 3a and Extended Data Fig. 6c). This effect was insignificant in women. Of note, although both men and women experience a decline in lymphocyte counts with aging, it occurs at an older age in women, as confirmed by recent analyses[2]. Notably, women show a temporary postmenopausal surge in lymphocyte counts, which delays their decline. Within the MEBEMP differentiation trajectory, aging was correlated with over-representation of more differentiated states, once again only in men (Fig. 3b). Of note, the frequency of cHSCs did not significantly change with age (Fig. 3c). Although previous studies suggest that aging is linked with an increase in HSC frequency[23,24], this was not observed with the restrictive definitions employed in the present study. We further identified an age-related decline in CD34$^+$ HSPC frequency in a cohort of 1,000 healthy individuals undergoing peripheral blood mononuclear cell (PBMC) scRNA-seq[25] (Fig. 3d and Extended Data Fig. 6d), which has also been reported by fluorescence-activated cell sorting (FACS) in the past[3]. The sex-specific correlation between age and cHSPC myeloid bias could be related to cell intrinsic properties, such as male-specific leukemia-associated mutations predisposing to myeloid differentiation[26]. This is less likely because canonical CH-positive cases were excluded from this analysis. Alternatively, this predominantly male myeloid bias could be related to cell extrinsic factors such as age-related hormonal and BM microenvironmental changes[27,28].

Fig. 2 | Normal cHSPC composition. a, Characterization of interindividual cHSPC compositional variation and its correlation to clinical parameters (scheme). **b**, Boxplots of cHSPC state frequency distributions across 148 healthy individuals (logarithmic scale). The percentage was calculated from all CD34$^+$ cells within each individual's single-cell ensemble. Boxplot centers, hinges and whiskers represent the median, first and third quartiles and 1.5× the interquartile range, respectively. Outliers are marked by circles. The numbers represent the mean ± s.d. for each distribution. **c**, Comparison of cell-state frequencies between 19 biological replicates and their original samples, for CLP (CLP-E, CLP-M, CLP-L and NKTDP) populations (top) and MEBEMP (MEBEMP-E, MEBEMP-L, ERYP and BEMP) populations (bottom). The diagonal $y = x$ is shown in red. All biological replicates were sampled 1 year after their original sampling date. **d**, Top: cell-state frequency profiles over the HSC-MEBEMP and HSC-CLP differentiation gradients of six sampled individuals (colored lines), each representing one of six archetypes (classes) of cHSPC composition observed in healthy individuals. The dashed lines represent the median (black) and the 5th and 95th percentiles (gray) of the studied population. Bottom: cell-state enrichment map over 15 differentiation bins (rows), for all studied individuals (columns) clustered into six classes (Methods). Classes I and II represent individuals relatively enriched in lymphoid progenitors, whereas classes V and VI represent individuals with relative depletion of lymphoid progenitors. Individuals are sorted by stemness within each class. Age and sex are denoted for each individual. **e**, CBC correlations to cell-state frequencies: %lymphocytes (from white blood cells, calculated for the entire cohort, left), HCT (men only, center) and RDW (men only, right). Missing individuals lacked sufficient cells for analysis. Two-sided permutation test *P* values are displayed for each correlation. See Methods for details on the permutation-based test. **f**, CH frequency (by gene) in age- and sex-matched high (red, $n = 602$) and normal (black, $n = 602$) RDW individuals selected from a cohort of 18,147 individuals. Panel **a** created with BioRender.com.

https://doi.org/10.1038/s41591-025-03716-5

Composition-controlled cHSPC expression correlates with age

As shown above, individual cHSPC compositions provide an initial blueprint of hematopoietic dynamics along the stemness and CLP or MEBEMP axes, with age-dependent changes. Composition-normalized gene expression profiles were further correlated with age, enabling age prediction based on normalized gene expression alone (Fig. 3e and Extended Data Fig. 6e; see Supplementary Tables 5 and 6 for additional screening for age-, CBC-, CH- and sex-associated gene expression). We next looked for gene groups (signatures) that co-variate between individuals, filtering out sex-linked signatures and those showing strong batch effects. The most prominent of these signatures included *Lamin-A (LMNA)* as well as *ANXA1*, *AHNAK*, *MYADM*, *TSPAN2* and *VIM*, among others (Fig. 3f, Extended Data Fig. 6f and Supplementary Table 7). Individual *LMNA* signature expression varied across a range of more than twofold (Extended Data Fig. 6g), exhibiting high expression variability in HSCs and early myeloid and lymphoid cell states, and a homogeneously low expression in late MEBEMPs and CLPs (Extended Data Fig. 6h). Individual *LMNA* signature expression was consistent in myeloid and



https://doi.org/10.1038/s41591-025-03716-5

lymphoid cell states (Fig. 3g) and was stable in our follow-up cohort (Extended Data Fig. 6i). We observed an age-linked increase in *LMNA* signature expression in lymphoid, but not myeloid, cHSPCs (Fig. 3h). Future studies on larger cohorts, enriched with clinical data, could further explore the age-related *LMNA* signature overexpression in CLPs and how it correlates with disease and immune function. Taken together, we show that, in addition to the accumulation of leukemic mutations in HSPCs, aging is linked with changes in the distribution of progenitor cell states within the PB and with notable expression differences in certain gene signatures. The mechanistic basis for this variation and its clinical impact remain unresolved.

Coordination of stemness and myeloid signatures

The differentiation of HSPCs toward MEBEMP and CLP fates involves coordinated activation and repression of specific transcriptional programs which are conserved across individuals. Yet, our screen for interindividual variation in gene signature expression suggested that individuals differed in the way in which they synchronized the opposing effects of these stemness and differentiation programs. To quantify this variation, we compared *AVP* (stemness) and *GATA1* (MEBEMP differentiation) signatures (Supplementary Table 8) on a 20 × 20 bin expression matrix (Fig. 3i). Although most individuals displayed dynamics close to the diagonal line (individuals N16 and N86, for example), following the typical transition from stemness to differentiation, some individuals deviated from the diagonal, indicating skewed synchronization between *AVP* and *GATA1* signatures. We quantified this deviation (that is, off-diagonal frequency) using a synchronization-score (sync-score). This facilitated the identification of individuals with sync-scores as low as 0.12 (N122 and N172, for example, Fig. 3i, top), indicating delayed activation of *GATA1* relative to *AVP* repression. In contrast, individuals exhibiting a high sync-score (N98 and N121, for example, Fig. 3i, bottom) show early activation of *GATA1* expression, which precedes *AVP* repression. Interindividual sync-score variability (Extended Data Fig. 6j) was positively correlated with RBC levels and consistently anti-correlated with MCV in men ($P < 0.01$ (Spearman's) for both RBC and MCV; Fig. 3j). Analysis of the correlation between individual sync-scores and cHSPC compositions in men demonstrated a negative correlation with ERYPs (Fig. 3k). In summary, variation in the coordination of stemness and MEBEMP differentiation programs correlated with RBC counts and volumes. More studies on larger cohorts are needed to explore how this coordination relates to age-related macrocytic anemia.

Circulating HSPC composition abnormality in cytopenia and MDS

Diagnosis of myeloid malignancies requires the identification of clonal markers (mutations or structural variants) and the detection and quantification of blasts and dysplasia, by next-generation sequencing, polymerase chain reaction, cytogenetics, fluorescence in situ hybridization, microscopy and flow cytometry of BM specimens. In Fig. 4a we described a stepwise approach for analysis of myeloid disorders based on sampling of cHSPCs and comparison of their compositions, normalized expression and copy number variations (CNVs) to our normal reference (Extended Data Fig. 7a–c). As proof of concept, we focused on MDS diagnosis. First, we reconstructed the reference model using data from 79 healthy individuals, putting aside some normal samples for classifier training. We then performed additional sequencing to obtain data from 44 patients with MDS and 29 patients with cytopenia (Supplementary Tables 9–11). We developed a streamlined in silico sorting scheme for quantifying the cHSPC composition of a new PB sample given the reference (Extended Data Fig. 7a,b) and used it to identify cases with abnormal compositions (Fig. 4b,c, Extended Data Fig. 8a and Methods). Classification included subpopulations (GMP-L, pre-B, pro-B and MkP) that were rare in the normal reference model and were not shown in Fig. 1. We then marked MDS or cytopenic samples with normal compositions (matching the reference model, group 1) and organized them along the myeloid and lymphoid spectrum. The remaining cases were clustered into distinctive subclasses. Although most cases of MDS showed significantly lower CLP frequencies (groups 3 and 4; Extended Data Fig. 8b), we identified a subclass of MDS and cytopenia with high CLP frequencies (group 2). Other subclasses included high MPP (group 4.2), high BEMP (group 4.1) and high GMP (group 3) frequencies. This sorting scheme partially separated MDS from other, non-MDS-related, cases of cytopenia, with most cytopenia cases exhibiting normal (group 1) compositions. Cases of MDS with abnormal CNVs (Methods) were enriched in groups 2–4 ($P < 0.004$, Fisher's exact test; Fig. 4d and Supplementary Table 12) and patients with high RDW were enriched in group 4 (Extended Data Fig. 8c). In summary, cHSPC compositions reveal molecular features that offer possibilities for identifying MDS subclasses and pathophysiology. Classification of MDS cases with normal cHSPC compositions (group 1) depends on further analysis of genetic and transcriptional states within specific cHSPC subtypes.

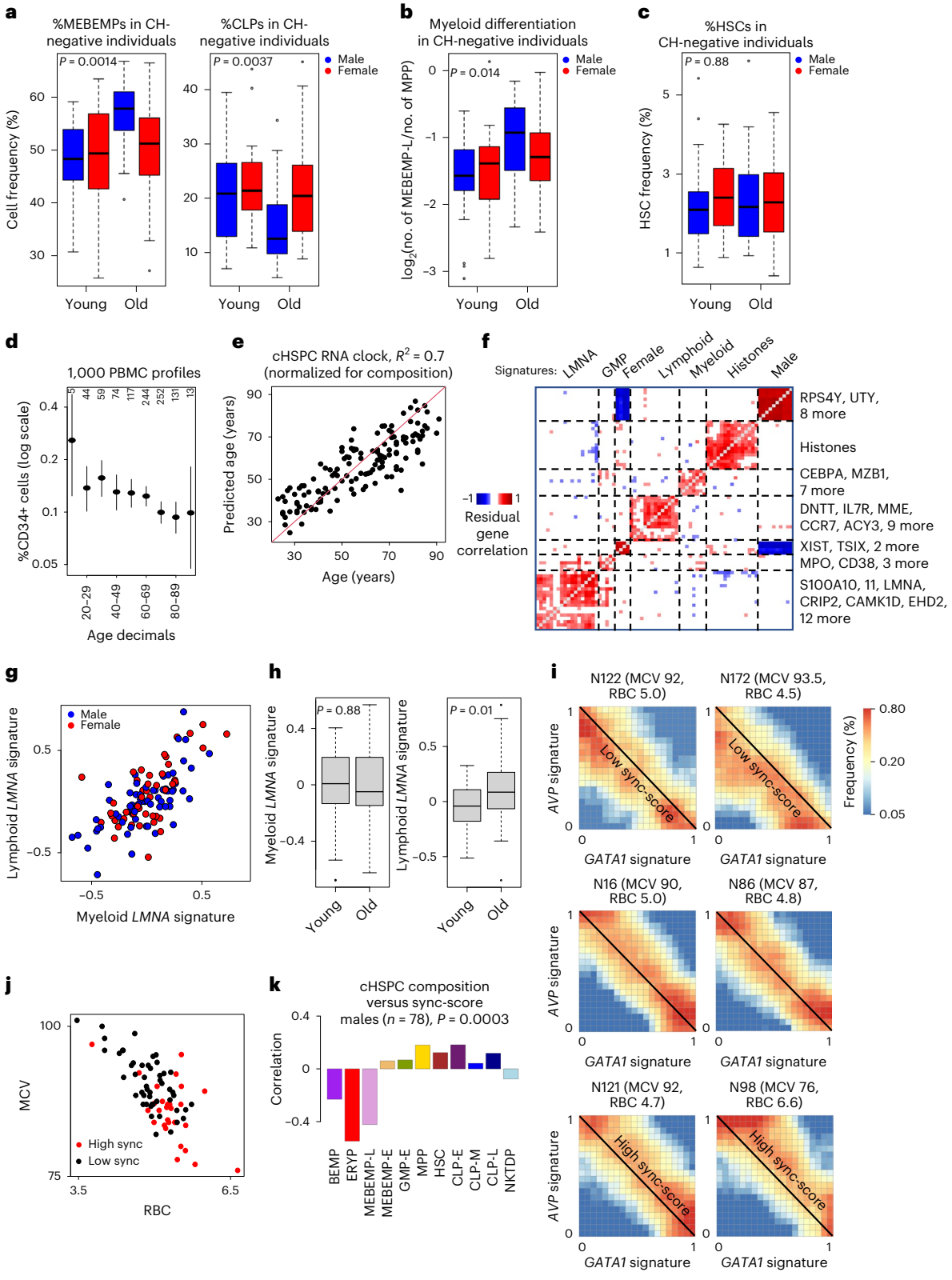
PB-based MDS diagnosis with CBC, mutation and cHSPC RNA data

To improve our diagnostic accuracy, we next derived specific gene signatures showing additional variation within cell types, from the reference model (Supplementary Table 13), and scored these signatures based on their ability to separate patients with MDS and cytopenia from healthy donors. A group of major histocompatibility complex class II (MHC-II) genes in MEBEMP-L, multi-potency genes in BEMP and S-phase genes in MEBEMP-L (Fig. 4e) emerged as top-ranking. These signatures were overall consistent across different samples of the same individuals (Extended Data Fig. 8d). We then combined CBCs, maximum variant allele frequency (VAF), cHSPC compositions and all afore-mentioned expression signatures into a feature set that formed

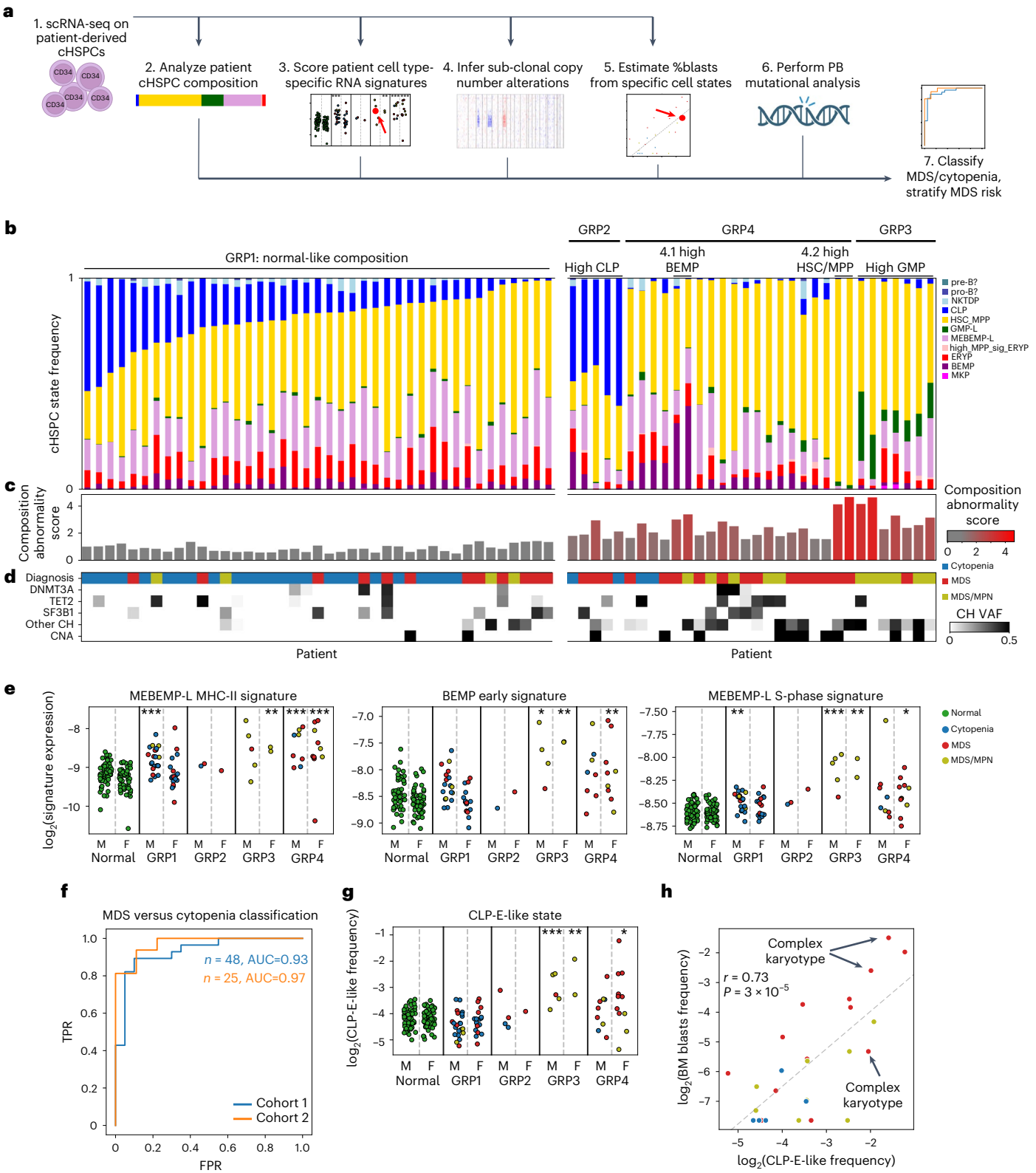
Fig. 3 | Age- and sex-linked changes in cHSPC composition. a, Frequency of MEBEMP (MEBEMP-E, MEBEMP-L, ERYP and BEMP, left) and CLP (CLP-E, CLP-M and CLP-L, right) populations, out of a total CD34$^+$ population, in young (<50 years) versus old (>60 years) individuals without CH, in men (blue) and women (red). The two-sided Kruskal–Wallis *P* values for differences among groups are denoted. The number of individuals per group is (left to right): 31, 15, 24 and 31. **b**, Analysis of age-linked compositional differences within the MEBEMP differentiation trajectory, comparing abundance of more (MEBEMP-L) with less (MPP) differentiated states in young versus old individuals. The two-sided Kruskal–Wallis *P* value for difference among groups is denoted. The number of individuals is as in **a**. **c**, As in **a**, for the HSC population. **d**, cHSPC frequency per age decimal in an scRNA-seq PBMC dataset of 1,000 healthy individuals[25]. For each decade, mean CD34$^+$ cell frequency is shown (Methods). The 95% confidence intervals are indicated as error bars. The number of individuals in each decade is indicated (top). **e**, True age (*x* axis) versus age predicted based on composition-controlled MPP expression (*y* axis). The diagonal $y = x$ is shown in red. **f**, Gene–gene correlation heatmap, calculated over individual-level MPP gene expression controlled for MPP composition. **g**, Individual *LMNA* signatures ($\log_2$(observed:expected ratios)) in lymphoid (CLPs) and early myeloid (MPPs) cell states. **h**, Analysis of age-linked differences in *LMNA* signature expression for CLP (right) and MPP (left) populations in young (<50 years) versus old (>60 years) individuals. The *y* axis denotes $\log_2$(observed:expected expression) normalized for composition. The number of individuals is 66 young, 65 old on the left and 48 young, 53 old on the right. The two-sided Mann–Whitney *U*-test *P* values are indicated. **i**, Individual heatmaps of single-cell counts over 20 bins of stemness (*AVP* signature, *y* axis) and MEBEMP differentiation (*GATA1* signature, *x* axis). Individual identifier, MCV and RBC are denoted at the top. The diagonal is indicated in black for reference. **j**, MCV versus RBC in male donors, with colors indicating high (red) and low (black) sync-scores. **k**, Correlation between individual sync-scores and cell-state compositions in men. The two-sided permutation test *P* value is denoted. All boxplots are as in Fig. 2b.

https://doi.org/10.1038/s41591-025-03716-5

the basis for construction of an MDS diagnostic classifier using standard machine learning tools. We created two cohorts: the first (cohort 1), composed of 28 patients with MDS, 20 patients with cytopenia and 41 healthy individuals, and the second (cohort 2) composed of 16 patients with MDS and 9 patients with cytopenia. We observed classifier training performance (even when aiming to separate MDS from cytopenia cases) was better when including normal cases in the dataset. Analysis of classifier performance showed very high specificity and sensitivity (Fig. 4f; area under the curve (AUC) = 0.93 in leave-one-out cross-validation for cohort 1 separation of MDS from cytopenia). Performance of the cohort 1 model on cohort 2 (which was not used during classifier training) was even higher (AUC = 0.97). Although cohort 2 data was not used in classifier training or feature selection, it was accessible to us during project analysis phases, such that we were cautious not to



https://doi.org/10.1038/s41591-025-03716-5



treat this as formal validation. The most informative feature used by the MDS classifier was the maximum VAF (Extended Data Fig. 9a). Yet, classifier performance was high even when excluding VAF information (Extended Data Fig. 9b,c).

Diagnosis and risk stratification of MDS rely on quantification of BM blast fractions. Our analysis of cohort 1 and cohort 2 samples, with the addition of three cases of MDS exhibiting complex karyotypes, suggests that we can predict this percentage quantitatively from cHSPC data using the fraction of cells showing a mixed HSC and CLP state (Fig. 4g,h and Extended Data Fig. 9d). All in all, this implies that, with further validation and testing, cHSPC profiling has the potential to replace BM analysis for MDS diagnosis and risk stratification, offering substantial benefits, such as noninvasive follow-ups and watchful waiting protocols. We present two case studies supporting this idea in Extended Data Fig. 10a,b. The first is an 82-year-old man showing progressive clonal expansion over a span of 3 years, accompanied by

Fig. 4 | Applying the cHSPC reference model to MDS diagnosis and subclassification. a, Schematics of a diagnostic approach to cytopenia and MDS using scRNA-seq of cHSPCs and a reference model. **b**–**d**, Cytopenia and MDS patient cHSPC compositions and mutations. **b**, Each bar represents a patient's cHSPC composition. Patients exhibiting normal compositions (Methods) are ordered by lymphoid cHSPC frequency (left) and those exhibiting abnormal compositions are ordered by composition hierarchical clustering (right). Cytopenia or MDS subclasses suggested by composition are marked (top). **c**, Patient composition abnormality scores, depicted by both bar height and color, as coded on the right. **d**, Patient diagnosis and CH VAF for three specific mutations and the maximal VAF over all other detected mutations, as color coded on the right. Presence of copy number alterations (CNAs) as detected by scRNA, color coded in black. **e**, Transcriptional signatures across healthy donors and patient groups as in **b**, further subdivided by sex (men, left; women, right). Individuals exhibiting insufficient cell counts for the population of interest were excluded. Color denotes clinical diagnosis. Mann–Whitney Benjamini–Hochberg-adjusted significance of difference from healthy donors of the same sex is indicated by asterisks (*$q < 0.05$, **$q < 0.01$, ***$q < 0.001$). **f**, Receiver operator curve for a classification model predicting MDS status based on cHSPC scRNA-seq, CH VAF and CBC values. FPR, false-positive rate; TPR, true-positive rate. **g**, Frequency of a CLP-E-like cell state across healthy donors and cytopenia and MDS groups, as in **e**. **h**, Comparison of BM blast counts measured by FACS and frequency of a CLP-E-like cHSPC population across individuals. Color denotes clinical diagnosis, as in **e**. Samples excluded from **b**–**g** due to the presence of a complex karyotype (as detected by scRNA-seq) are highlighted (arrows). Linear fit across all individuals ($n = 26$) is shown (dashed line), as well as the corresponding r and (two-sided) P value.

deteriorating anemia. The second is a 65-year-old woman presenting with clonal del5q showing complete cytogenetic remission after lenalidomide treatment. Additional follow-up examples (Extended Data Fig. 10c) suggest small changes in (normal or abnormal) composition across time, further supporting the idea of using cHSPCs for noninvasive assessment of disease progression.

Discussion

The present study characterizes interindividual heterogeneity in cHSPCs across 148 healthy individuals using scRNA-seq analysis of PB $CD34^+$ cells. The magnitude of our cohort, along with the potency and resolution of modern single-cell technologies and the computational methods used in the present study, allowed us to characterize in detail the transcriptional programs of diverse, sometimes rare (NKTDP and BEMP), HSPC subpopulations, refining and augmenting previous findings from smaller cohorts (Fig. 1). We defined a normal reference range for cHSPC subpopulation frequencies within an age- and sex-diverse healthy population and showed that cHSPC subtype compositions were highly variable between individuals, whereas the cell states themselves were remarkably general (Fig. 2). These compositions remained stable over a 1-year follow-up period. Future studies will need to further explore and better define the mechanistic and genetic basis for this compositional heterogeneity. With current sample size, we showed that the known age-related myeloid bias in HSPCs is predominantly male driven and that composition-controlled RNA expression can be used to infer chronological age (Fig. 3).

Our data show that cHSPCs are transcriptionally similar to their BM counterparts (Extended Data Figs. 2 and 3), except for reduced cell-cycle gene expression. Although not a complete model for BM hematopoiesis, cHSPCs serve as a highly accessible proxy for key hematological processes. Interindividual differences in cHSPC compositions and states can thus serve as a tool for capturing key aspects of a patient's hematopoietic state. The relevance and importance of a cHSPC normal reference (Fig. 2b) can perhaps be better understood in view of the normal CBC reference range, developed in the 1930s[1]. The development of a population-wide CBC reference enabled the identification of numerous pathological blood states that characterize distinct clinical entities. In a similar fashion, our cHSPC reference can be used to characterize physiological and pathological states. In Fig. 4 we describe a pipeline for the identification and characterization of blood pathologies based on our normal cHSPC reference and show how this can be applied to MDS diagnostics (including inference of cytogenetics and blast counts from the PB). We present scRNA-seq data on cHSPCs from 73 cases of cytopenia and MDS, greatly extending currently available BM MDS scRNA-seq datasets[5,29–32]. The data described supports MDS diagnosis (over non-MDS-related cytopenia) and suggests the possibility of MDS subclassification based on over-representation of distinct HSPC progenitor populations. The MDS-related gene expression signatures identified in the present study open avenues for research that might contribute to better understanding of MDS pathophysiology and drug design strategies. Importantly, further follow-up, validation in prospective studies and cohort expansion to ethnically diverse populations are needed to prove that the tools introduced here can become a clinical standard. The diagnostic potential of our reference model may be further enhanced upon acquisition and analysis of additional blood subpopulations and disease states in contrast with the reference. Practically, application of scRNA-seq for diagnostics would have to rely on stable and minimally biased cell acquisition and processing technologies that can be deployed across diverse clinical settings and provide consistent and trustworthy results. Development in this domain is promising, but more work must be done to reach clinical standards.

To conclude, our study delves into the basic molecular physiology of cHSPCs at the population level, uncovering age-related phenotypes and proposing a platform for mechanistic and diagnostic insights into blood malignancies. This resource, along with various other tools for profiling genetics and epigenomics in the blood, has the potential to redefine normal versus pathological states in hematology and provide both clinicians and researchers the means for mapping the transition from health to disease.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-03716-5.

References

1. Osgood, E. E. Normal hematologic standards. *Arch. Intern. Med.* **56**, 849–863 (1935).
2. Cohen, N. M. et al. Personalized lab test models to quantify disease potentials in healthy individuals. *Nat. Med.* https://doi.org/10.1038/S41591-021-01468-6 (2021).
3. Cohen, K. S. et al. Circulating $CD34^+$ progenitor cell frequency is associated with clinical and genetic factors. *Blood* **121**, e50–e56 (2013).
4. Mende, N. & Laurenti, E. Hematopoietic stem and progenitor cells outside the bone marrow: where, when, and why. *Exp. Hematol.* **104**, 9–16 (2021).
5. Ainciburu, M. et al. Uncovering perturbations in human hematopoiesis associated with healthy aging and myeloid malignancies at single-cell resolution. *eLife* **12**, e79363 (2023).
6. Quaranta, P. et al. Circulating hematopoietic stem/progenitor cell subsets contribute to human hematopoietic homeostasis. *Blood* **143**, 1937–1952 (2024).
7. Mende, N., Dresden, T. U., Santoro, A. & Lidonnici, M. R. Unique molecular and functional features of extramedullary hematopoietic stem and progenitor cell reservoirs in humans. *Blood* https://doi.org/10.1182/blood.2021013450 (2022).

8. Bender, J. et al. Identification and comparison of CD34-positive cells and their subpopulations from normal peripheral blood and bone marrow using multicolor flow cytometry. *Blood* **77**, 2591–2596 (1991).
9. Goodman, J. W. & Hodgson, G. S. Evidence for stem cells in the peripheral blood of mice. *Blood* **19**, 702–714 (1962).
10. Biezuner, T. et al. An improved molecular inversion probe based targeted sequencing approach for low variant allele frequency. *NAR Genom. Bioinform.* **4**, lqab125 (2022).
11. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biol.* **23**, 100 (2022).
12. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
13. Lehnertz, B. et al. *HLF* expression defines the human hematopoietic stem cell state. *Blood* **138**, 2642–2654 (2021).
14. Regev, A. A single cell immune cell atlas of human hematopoietic system. *Human Cell Atlas Data Explorer* https://data.human-cellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79?catalog=dcp1 (2022).
15. Frelin, C. et al. GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat. Immunol.* **14**, 1037–1044 (2013).
16. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
17. Lavaert, M. et al. Integrated scRNA-seq identifies human postnatal thymus seeding progenitors and regulatory dynamics of differentiating immature thymocytes. *Immunity* **52**, 1088–1104 (2020).
18. Scoville, S. D. et al. A progenitor cell expressing transcription factor RORγt generates all human innate lymphoid cell subsets. *Immunity* **44**, 1140–1150 (2016).
19. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
20. Kar, S. P. et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat. Genet.* **54**, 1155–1166 (2022).
21. Nam, A. S. et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature* **571**, 355–360 (2019).
22. Nam, A. S. et al. Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. *Nat. Genet.* **54**, 1514–1526 (2022).
23. Sudo, K., Ema, H., Morita, Y. & Nakauchi, H. Age-associated characteristics of murine hematopoietic stem cells. *J. Exp. Med.* **192**, 1273–1280 (2000).
24. Pang, W. W. et al. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proc. Natl Acad. Sci. USA* **108**, 20012–20017 (2011).
25. Yazar, S. et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
26. De-Morgan, A., Meggendorfer, M., Haferlach, C. & Shlush, L. Male predominance in AML is associated with specific preleukemic mutations. *Leukemia* **35**, 867–870 (2021).
27. Zioni, N. et al. Inflammatory signals from fatty bone marrow support DNMT3A driven clonal hematopoiesis. *Nat. Commun.* **14**, 2070 (2023).
28. Bacharach, T., Kaushansky, N. & Shlush, L. I. Age-related micro-environmental changes as drivers of clonal hematopoiesis. *Curr. Opin. Hematol.* **31**, 53–57 (2024).
29. Serrano, G. et al. Single-cell transcriptional profile of CD34^{+} hematopoietic progenitor cells from del(5q) myelodysplastic syndromes and impact of lenalidomide. *Nat. Commun.* **15**, 5272 (2024).
30. Wu, Z. et al. Sequencing of RNA in single cells reveals a distinct transcriptome signature of hematopoiesis in GATA2 deficiency. *Blood Adv.* **4**, 2702–2716 (2020).
31. Liu, Y. et al. Single-cell RNA sequencing identifies the properties of myelodysplastic syndrome stem cells. *J. Transl. Med.* https://doi.org/10.1186/s12967-022-03709-9 (2022).
32. Ganan-Gomez, I. et al. Stem cell architecture drives myelodysplastic syndrome progression and predicts response to venetoclax-based therapy. *Nat. Med.* **28**, 557–567 (2022).

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. [3]Department of Computer Science and Applied Mathematics, and Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [4]Hemato-Oncology Department, Assuta Medical Center, Tel Aviv, Israel. [5]Maccabi Healthcare Services, Tel Aviv, Israel. [6]Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel. [7]Ultima Genomics, Fremont, CA, USA. [8]Integrated Cancer Prevention Center, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel. [9]Clalit Health Services, Tel Aviv, Israel. [10]Department of Cell, Developmental and Cancer Biology, Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA. [11]Division of Hematology and Medical Oncology, Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA. [12]New York Genome Center, New York, NY, USA. [13]Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA. [14]Division of Hematology and Medical Oncology, Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [15]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA. [16]Division of Hematology, Rambam Healthcare Campus, Haifa, Israel. [17]These authors contributed equally: N. Furer, N. Rappoport, O. Milman, A. Tanay, L. Shlush. ✉e-mail: amos.tanay@weizmann.ac.il; liranshlush3@gmail.com

Methods

Patient recruitment

All healthy reference model individuals (n = 148, analyzed in Figs. 1–3 and Extended Data Figs. 1–6) volunteered to participate in our study and donated blood at the Weizmann Institute of Science (WIS) between November 2020 and December 2023. They were recruited from the WIS community and primary care clinics and consisted of 79 men and 69 women aged 23–91 (median 61.5) years. Their demographic data and CBCs are included in Supplementary Table 1. Written informed consent allowing access to their demographic, longitudinal CBC and sequencing data (CH and genotyping panels) was obtained from all participants in accordance with the Declaration of Helsinki. All relevant ethical regulations were followed and all protocols were approved by the WIS ethics committee (under Institutional Review Board (IRB) protocol no. 283-1).

For the main reference model (Figs. 1–3), recruitment was intended to allow characterization of the normal variation in cHSPC states. As no such profiling had been previously performed, we could not assume much about the variance in the population a priori. Participants were therefore required to lack any known hematological condition, including hematological malignancy or premalignant state, or any prior evidence of blood clonality. An Illumina-sequenced subset of these 148 individuals (n = 79) was used for constructing the healthy reference model used in Fig. 4 ('Fig. 4 reference model'), filtering out individuals with any blood count abnormality (up to 5 years before sampling) and putting aside 41 healthy samples for classifier training.

Recruitment of the cytopenic cohort (including patients with MDS and non-MDS-related cytopenia, analyzed in Fig. 4 and Extended Data Figs. 7–10) took place between November 2021 and February 2024. These patients were recruited from several outpatient hematological clinics by collaborating physicians to represent the wide clinical spectrum of MDS, from patients with moderate anemia and mild dysplasia as their sole BM abnormality to those with severe cytopenia and excess blasts on BM analysis. Key patient characteristics, including CBCs, BM FACS blasts and mutational data, are included in Supplementary Tables 9 and 11. Median age for the cytopenic cohort was 73 years (range 27–93 years), with men representing 53% of patients. Patient PB samples were either drawn at WIS or transported to WIS within <2 h of blood drawing. All cytopenic samples were processed in an identical fashion to the healthy ones (described below). Longitudinal CBCs, mutational data and most recent BM analyses were collected from patients and analyzed along with their scRNA-seq data.

The cytopenic cohort included a total of 83 individuals, 50 of whom were labeled as cases of 'MDS', based on BM morphology and/or mutational and karyotypic abnormalities (as detected in the clinic or by our CH panel and scRNA CNA analysis). The remaining 33 patients with cytopenia not satisfying MDS criteria were labeled as cases of 'cytopenia'. We note that, consistent with common medical practice in Israel, most of these 33 patients with cytopenia did not undergo BM examination, which may have resulted in missed MDS diagnoses. To address this limitation, we collected the most recent clinical data available for patients with cytopenia, with a median follow-up period exceeding 600 d after cHSPC sampling (Extended Data Fig. 10d and Supplementary Table 9). Importantly, no new diagnoses of myeloid malignancy were recorded in any of the cytopenic cases, except for N193 who was diagnosed with VEXAS syndrome 1 year after cHSPC sampling, exhibiting a *UBA1* mutation c.121A>G:p.Met41Val, but had not been diagnosed with MDS yet. In addition, during this follow-up, median change in RDW% was zero. In contrast, over a similar period (looking at historical records before cHSPC sampling; Extended Data Fig. 10e), median RDW% change in cases of MDS was 0.75, significantly higher than in patients with cytopenia (Extended Data Fig. 10f; P = 0.001, two-sided Mann–Whitney U-test). Overall, these data support the accurate classification of cytopenic cases. Eleven cHSPC samples were acquired from patients under treatment, six of which were included in Fig. 4 (Supplementary Table 9). Of the 83 patients with cytopenia, 17 who presented with asymptomatic, mild cytopenia, were also included in the original healthy cohort of 148 individuals.

Sampling of cHSPCs

We drew 50 ml of PB from each individual into lithium–heparin tubes and put aside 1 ml of blood for DNA production. The remaining volume was used for PBMC isolation via Ficoll using Lymphoprep-filled Sepmate tubes (STEMCELL Technologies), followed by CD34 magnetic bead-based enrichment using the EasySep human $CD34^+$ selection kit II (STEMCELL Technologies). We found this enrichment strategy to be simple and reproducible and chose it for a couple of reasons: (1) RNA-seq data were most reproducible when cells were not sorted, but rather enriched for using beads (lower mitochondrial gene fraction); and (2) CD34 purity could be highly regulated by this method, to achieve anywhere between 50% and 95% enrichment of $CD34^+$ cells, which could later be easily distinguished based on their single-cell expression data.

ScRNA-seq of cHSPCs

ScRNA libraries were generated using the 10x Genomics scRNA-seq platform (Chromium Next Gem single-cell 3′ reagent kit v.3.1). Chip loading was preceded by flow cytometry to verify that enrichment was successful and that enough $CD34^+CD45^{int}$ live cells were gathered. All blood samples were either drawn at WIS or transferred from participating clinics on the morning of each experiment day, and time from blood draw to 10x loading was restricted to 5 h. The motivation for working with fresh samples was based on our previous experience with PB $CD34^+$ cells being vulnerable to freezing–thawing rounds and long manipulation times. The 10x libraries were sequenced on two alternative platforms (Illumina and Ultima Genomics). Twelve libraries were simultaneously sequenced on both platforms for comparison purposes and to demonstrate the scalability of our approach. We observed the Ultima-sequenced data to be highly similar to the Illumina-sequenced data (Extended Data Fig. 5a).

DNA production and sequencing

All healthy and patient DNA was produced from PB at sampling. DNA sequencing was performed on two targeted panels: the first a rich myeloid CH panel (InfiniSeq Myeloid Malignancies Panel, Sequentify, Israel) covering all known pre-leukemic mutations[10] and the second a genotyping panel specifically designed to capture polymorphic sites prevalently expressed by RNA molecules from all cell types in our data. This allowed demultiplexing of individual pools based on individual specific SNP combinations and replaced previous, antibody-based multiplexing methods. Three to six individual samples were pooled on each experiment day after extraction of DNA aliquots, such that CD34 enrichment was performed on the entire pool of PBMCs produced. As with other methods of sample multiplexing, genotype-based multiplexing allows for robust doublet detection during data analysis, which enabled loading of 30,000–40,000 cells on each Chromium Chip lane.

Both our CH and genotyping panels are Molecular Inversion Probe (MIP) panels described in detail previously[10]. For the healthy cohort we used our CH panel v.3, containing 705 probes, covering leukemia-related SNVs and insertions/deletions (indels) in 47 genes, complemented by two amplicon-sequencing reactions to cover GC-rich regions in *SRSF2* and *ASXL1*. For the cytopenic cohort, we used our CH panel v.4 (Supplementary Table 10). For alignment of reads we used Burrow–Wheeler Aligner (BWA)-MEM v.2 (ref. 33). As MIP sequencing is cost-effective yet noisy, we developed an in-house variant calling method to reliably identify low-VAF CH events[10]. For the genotyping panel we used Varscan for variant calling[34]. Each DNA sample was sequenced twice with a minimum depth of 10^6 paired-end reads on an Illumina Novaseq machine. Variant calling was performed as previously reported[10]. Our genotyping panel allows for the simultaneous

detection of >2,000 SNVs. It includes heterozygous sites with at least a 5% minor allele frequency from the 1,000 Genomes project, which were extensively covered in our data (at least 80 unique molecular modifiers (UMIs) across all cells in a test 10x library), excluding sites in repetitive elements and sex chromosomes. Both panels were designed using MIPgen[35] to ensure capture uniformity and specificity.

CH sequencing of high-RDW samples and controls

To compare propensity for CH and high-risk CH mutations in high-RDW cases and normal RDW controls, we performed deep targeted sequencing of DNA samples from 602 high-RDW (>15%) individuals, whose blood count did not meet MDS criteria (11.5 g dl^{-1} ≤ Hg ≤ 15.5 g dl^{-1} (F), 13 g dl^{-1} ≤ Hg ≤ 17 g dl^{-1} (M), 80 fL ≤ MCV ≤ 96 fl, PLT ≥ 100 × 10^9 l^{-1}, Abs Neut ≥ 1.8 × 10^9 l^{-1}) and 602 normal RDW, age- and sex-matched controls. Case–control matching was performed using the R MatchIt package, balanced on age and sex, method = 'nearest', ratio = 1, from a total of 18,147 individuals with longitudinal blood counts and available DNA at the Tel Aviv Sourasky Medical Center (TASMC) Integrated Cancer Prevention Center. All DNA samples were collected after obtaining written informed consent in accordance with the Declaration of Helsinki and were received de-identified from the TASMC. All relevant ethical regulations were followed and all protocols were approved by the TASMC ethics committee (under IRB protocol no. 02-130).

ScRNA-seq processing

We processed fastq files by executing cell-ranger (v.3.1.0) with an hg-38 reference genome. We filtered out cells with at least 20% mitochondrial expression, then removed mitochondrial genes (as well as few other batch-prone genes) and further filtered cells with ≤500 UMIs.

Doublet calling

We performed several steps to assign cells to individuals and to detect doublets. Our pipeline included the following steps:

(1) Demultiplexing cells and calling doublets based on SNPs found in the scRNA-seq data
(2) Building a metacell model using cells from all libraries, including cells previously marked as doublets, identifying and removing metacells made of doublets
(3) Identifying doublet metacells based on expression of marker genes
(4) Building the final metacell model and marking metacells as doublets based on expression markers

In the first step, we identified doublets and assigned cells to individuals using Vireo v.0.3.2 (ref. 36) and Souporcell v.2.4 (ref. 37), which cluster cells based on SNPs found in sequenced RNA molecules. We executed Vireo (preceded by running cellsnp v.0.3.0) and Souporcell on each library separately. Both methods used SNPs from our genotyping panel[10] which were covered by at least 20 UMIs in the library (in Souporcell—at least 10 from the major and minor allele each). We observed high agreement in doublet calling between the two methods.

In the next step, we built a metacell model with cells from all libraries. This model included cells that we already identified as doublets. The model was built with metacell2 (ref. 11), with a target metacell size of 200 cells. We then marked all metacells, where at least 35% of the cells were already marked as doublets, and all metacells that expressed key markers of distinct cell types as doublet metacells. All cells that belonged to a doublet metacell were then marked as doublets. We then built an additional metacell model (see below), without cells that were marked as doublets.

Assignment of cells to individuals

Vireo and Souporcell both cluster cells based on SNPs found in the sequenced RNA, such that cells in the same cluster belong to the same individual. We next assigned clusters of cells to the individual to whom they belonged. To this end, we correlated the genotypes of each cell cluster, as inferred by Vireo, to all genotypes that we measured using the MIP panel (only using sites with sufficient sequencing depth). As a control, this matching was performed against the MIP genotypes of all individuals in the cohort and not only those from the specific library analyzed. We observed clear matchings between Vireo clusters and individuals from the expected libraries in all cases. This method also correctly identified related individuals. The sex of all matched individuals was confirmed by expression of XIST in the RNA data.

Removal of droplets with megakaryocyte signatures

Droplets with complete or partial megakaryocyte expression (at least 5% of UMIs coming from a megakaryocyte gene program including PF4, PPBP and 131 additional genes) were removed from our model as a result of their overall high doublet rate, and a final metacell model was constructed from the retained cells ((1) not marked as doublets, (2) confidently assigned to an individual and (3) not exhibiting megakaryocyte expression).

Correcting for sequencing-platform bias

Our 10x libraries were sequenced on Ultima Genomics and Illumina sequencers. To minimize batch effects related to these sequencing-platform variations, we used libraries that were sequenced on both platforms to calculate an Illumina–Ultima correction factor per gene as the mean $\log_2$(fold-change) in expression of the gene across re-sequenced libraries. We then normalized each Ultima-sequenced library by downsampling genes with at least 0.28 $\log_2$(fold) Ultima overexpression and resampling genes with at least 0.2 Illumina overexpression. The downsampling and resampling were performed for each gene independently, across all cells in each Ultima library. The thresholds for downsampling and resampling were chosen such that the overall number of UMIs per cells remained similar; 87 genes with at least 4-fold-change between Ultima and Illumina were excluded from further processing.

Computing the reference metacell model

Our metacell model was built using metacell2, with a target metacell size of 200 cells, deriving 4,253 metacells. We marked histone, cell-cycle, ribosomal, sex-linked and stress response genes (including *FOS* and *JUN*) as forbidden genes, as well as genes with high technical variation, such as those with high or inconsistent differences between Illumina- and Ultima-sequenced technical replicates. These genes were not used for calculating gene–gene similarities but were included in downstream analyses. We annotated metacells using known markers as illustrated in Extended Data Fig. 1c. We excluded metacells with low *CD34* expression, such as mature monocytes, B cells, T cells, natural killer (NK) cells, dendritic cells (DCs) and endothelial cells, as well as 20 GMP-L metacells, from most downstream analyses. We used UMAP projections of the metacell expression vector over genes with specific enrichment over cell types for visualization of the metacell manifold.

BM comparisons and projections

We used three BM datasets for comparison purposes: a dataset including CD34-enriched cells from two individual BM samples collected by us and processed similarly to PB (Fig. 1a), the Human Cell Atlas (HCA) BM dataset[14] and a $CD34^+CD38^-$ bead-based enriched BM dataset[12]. We previously processed and annotated the HCA dataset in a metacell model. We further constructed a metacell model for the two CD34-enriched BM samples collected by us using metacell2, in a similar fashion to that described previously, and downloaded the Setty et al. sequencing data[12], processed it by running cell-ranger, and created a third BM metacell model from their data. To project our own PB data, our own BM data and the Setty CD34-enriched BM data on the HCA model, we correlated projected metacells from each of these models with HCA metacells over genes showing high variance in the HCA model. We annotated each Setty metacell using the mode of its

five most correlated HCA metacells. We annotated our own BM data using both the mode of the five most correlated HCA metacells and expression of gene markers. We projected metacells from each of these models on the HCA UMAP using the mean *x* and *y* values of the five most correlated HCA metacells. To compare S-phase genes between PB and BM (Extended Data Fig. 3b), we calculated the S-phase signature (mean expression of six cell-cycle genes: *TYMS*, *H2AFZ*, *PCNA*, *MCM4*, *HELLS* and *MKI67*) for each PB and HCA metacell and plotted the distribution of these scores across metacells for each cell type.

HSC differentiation gene programs

To visualize transcriptional dynamics in HSCs, we sorted MEBEMP and CLP metacells based on their *AVP* expression. To calculate differential expression (DE) between HSCs and neighboring cell types (Extended Data Fig. 3g), we took the geometric mean expression of each gene across each of these cell states (within HSC or CLP-M or MEBEMP-E metacells) and calculated the difference of means between HSC and MEBEMP-E and between HSC and CLP-M metacells.

DE between individuals unexplained by the metacell model

We compared each individual's pooled expression profile to a matched expression profile based on the individual's distribution across metacells. We performed this analysis separately for MPPs or MEBEMPs (MPP, GMP-E, MEBEMP-E/-L, ERYP and BEMP) and CLPs (CLP-E/-M/-L and NKTDP). In each of these cell states, we downsampled each cell to have 500 UMIs and summed the UMIs across all cells of each individual, normalized the sum to 1 and calculated the $\log_2$(value) to obtain the observed expression. To compute matched expression, we downsampled each metacell to have 90,000 UMIs and summed all UMIs of the metacell to which each cell belongs for each individual. We normalized this matched expression to sum to 1 and took the $\log_2$(value). For Extended Data Fig. 3c,d, we plotted all genes that were expressed in either the observed or the matched expression in at least one individual ($\log_2$(expression) > $2^{-14.5}$ for MPPs or MEBEMPs, > $2^{-13.5}$ for CLPs), with at least a twofold change between matched and observed in at least one individual. We excluded genes exhibiting strong batch effects.

HSPC compositional analysis

To explore variance in cell-type composition between individuals, we first calculated the distribution of each individual's cells across the CD34$^+$ cell states. We further partitioned cells from the CD34$^+$ cell states into finer-grained bins, using one HSC, four CLPs and ten MPP or MEBEMP bins, for a total of fifteen bins. We assigned HSC cells to bin 0, CLP-E cells to CLP bin 1 and CLP-M/-L cells to CLP bins 2–4 based on an *AVP* expression gradient, such that each of these bins consisted of an equal number of cells. We similarly assigned MPP and MEBEMP-E/-L cells into equal size MPP or MEBEMP bins 1–10 based on decreasing *AVP* expression.

The bottom panel of Fig. 2d shows individual enrichment across bins ($\log_2$(ratio of each individual's cell frequency in each bin to the median cell frequency in that bin across individuals)). We partitioned individuals into three groups based on their mean enrichment across CLP bins 2–4—those with mean enrichment >0.5 are high CLP, those with <−0.5 are low and the rest are intermediate. We next defined the stemness score as the ratio between the number of cells in MPP or MEBEMP bins 1–5 and the total MPP or MEBEMP cells (bins 1–10). Individuals with stemness score >0.5 had enriched stemness. Individuals within each cluster were further sorted based on their stemness score. The combination of CLP enrichment and stemness defines the six classes shown in the figure.

Test for association between cell-state compositions and a numerical label

We used permutation tests to test for an association between cell-state distributions and a label, such as CBC indices or sync-scores. We sorted CD34$^+$ cell states into 11 bins from late MEBEMP differentiation through HSCs to late CLP differentiation (as ordered in Fig. 2b). We correlated each of the 11 cell-state frequency vectors to the numerical label vector. We then looked at triplets of adjacent cell states in this order and calculated the mean correlation for each triplet to obtain nine mean correlation values and took the maximal absolute correlation value as a test statistic. We repeated this process after permuting the label vector 10,000× and used the test statistics from the permutations to derive a *P* value.

CD34$^+$ cell frequency in the OneK1K dataset

We built a metacell model for the cells from Yazar et al.[25]. We labeled all cells in metacells with high *CD34* expression ($\log_2$(fraction of UMIs > −14.3)) as CD34$^+$ cells. We then selected individuals with at least 800 cells in the model and randomly sampled 800 cells from each (Supplementary Table 14). To produce Fig. 3d, we pooled these sampled cells by the decade of their individuals' age and calculated the fraction of CD34$^+$ cells in each decade. The 95% confidence intervals shown in the figure assume a binomial distribution, given the very sparse nature of the data.

Variably expressed gene modules

We detected gene modules with high variance across individuals while controlling for compositional variation. This was performed separately for myeloid and lymphoid states, in the following manner:

(1) For each individual, we calculated the 5th percentile of their number of UMIs across all MPP metacell cells and downsampled all cells to this number. We then pooled all downsampled cells, normalized to sum to 1 and took the $\log_2$(value). This gave us the observed expression profile of each individual.
(2) We then created the expected expression profile for each individual as follows: we partitioned all MPP metacells into 30 equal size bins based on their *AVP* expression, and downsampled metacells to 90,000 UMIs. We then took the average expression of each gene across downsampled metacells in each bin. This defined an expression profile for each of the 30 bins. To obtain an individual's expected expression, we calculated the weighted average expression profile of bins, where the weight of each bin is proportional to the fraction of the individual's cells from that bin, normalized to sum to 1 and took the $\log_2$(value). We then calculated the difference between the observed and expected expression profiles.
(3) Our data showed some batch effect distinguishing samples collected in two calendric periods. As this effect could introduce co-variation between genes across individuals, we applied a correction controlling for it. This was performed using a linear model fitting each gene to the sample collection period. We then subtracted the inferred period factor from the samples that were collected in the second period. We found that this approach substantially reduced emergence of gene clusters linked with sample collection date bias.
(4) We screened for genes with high variance that were unlikely to be affected residually by the main manifold differentiation process. We removed genes with high batch effects, genes with high *AVP* correlation (absolute value Pearson's correlation >0.65), and genes highly correlated (absolute value Pearson's correlation >0.5) with a module of differentially expressed genes between the first and second collection periods. We then calculated each gene's variance in the difference between observed and expected expression across individuals. As some of this variance can be explained by sampling noise, we plotted each gene's variance across individuals against its mean expression across individuals. We sorted genes by this expression value and subtracted from the

variance of each gene a rolling mean of the variances of 100 neighboring genes in that ordering. We chose genes with variance at least 0.08 higher than the rolling mean variance.

(5) We calculated a gene–gene Spearman's correlation matrix for the high variance genes and clustered correlation profiles using hierarchical clustering. We removed genes with low mean correlation (<0.2) to their cluster and then removed gene clusters with low mean correlation between their genes (≤0.25 mean correlation for all gene pairs). We further computed gene–gene correlations using only samples from our first library collection period and required gene clusters to have a high mean correlation (>0.25) between their genes when using only these samples. We removed additional gene modules arising from this analysis resulting from batch effects or traces of MEBEMP differentiation not normalized by this approach. This resulted in Fig. 3f.

We performed a similar analysis for CLPs, with few differences. The analysis included all cells from CLP-M metacells. The cells were partitioned into six equal size bins and partitioning was based on the average of their *DNTT* and *VPREB1* expression. Genes with high absolute correlation to the average *DNTT* and *VPREB1* expression were excluded. This was followed by hierarchical clustering of the gene–gene correlation profiles and removal of genes as described for MPPs.

Age regression

We developed age-regression models for MPP and CLP expression separately. To predict age, we used the difference between an individual's observed and expected gene expression as described above ('Variably expressed gene modules'). We used genes with minimal expression $\geq 2^{-14.5}$ for MPPs and $\geq 2^{-15.5}$ for CLPs across individuals. We trained a LASSO (least absolute shrinkage and selection operator) model using nested leave-one-out cross-validation. For each left-out sample, we performed cross-validation on the remaining samples to select LASSO's λ parameter, trained a model using the selected λ and made a prediction on the left-out sample.

The *LMNA* signature

We used the difference between an individual's observed and expected gene expression and correlated this difference to the difference in *LMNA* expression separately for MPPs and CLPs. We then summed the MPP and CLP correlation values and kept genes with summed correlation >0.7. We further removed genes with high technical variance, retaining 17 genes in the *LMNA* signature. To calculate individual *LMNA* signatures, we took the average value of these 17 genes in the observed–expected matrix of each individual for MPPs and CLPs separately. To plot Extended Data Fig. 6g, we calculated the geometric mean of LMNA signature gene expression for each individual in each one of the ten MPP or MEBEMP bins described earlier in Fig. 2d.

GoT analysis

GoT[21] performed on sample N122 allowed us to mark this individual's cells as wild-type or mutated. As a result of the low VAF of N122's *DNMT3A* mutation, and to increase power, we marked cells with a DNMT3A mutation status that could not be determined by GoT as wild-type cells. For Extended Data Fig. 6b, we examined sample N122's cell distribution across cell states.

Sync-score

We defined the *AVP* signature to include genes with high correlation (>0.6) to *AVP* across HSC, MPP and MEBEMP metacells and the *GATA1* signature to include those with high correlation (>0.7) to *GATA1*. We filtered out genes with mean relative expression $>2^{-10}$ in these metacells, to preclude a small number of genes from dominating the signatures. We then scored all HSC, MPP, MEBEMP-E and MEBEMP-L cells by their fraction of UMIs from the *AVP* and *GATA1* signatures and partitioned them into 20 equal size bins of *AVP* signature expression and 20 equal size bins of *GATA1* signature expression. The sync-score is then defined as the fraction of cells in *GATA1* bins 13 and above (upper two quintiles of *GATA1*) that are in *AVP* bins 9 and above (upper three quintiles of *AVP* expression). To visualize sync-scores (Fig. 3i), we normalized this 20 × 20 bin matrix to sum to 1, smoothed the obtained matrix by averaging cells using a running window of length 3 and took the $\log_2$(value).

Differential gene expression with respect to age and CBC

DE was performed separately for MPP and CLP cells as well as for men and women. The MPP and CLP-M matrices previously used to detect variant gene modules were used here as well. Individual gene expression levels were correlated with age, maximal VAF of CH mutations and 20 CBC indices using Spearman's correlation; the correlation was then tested for significance. The *P* values were false discovery rate (FDR)-corrected (Benjamini–Hochberg) for each label separately. For maximal VAF we additionally performed a Mann–Whitney *U*-test comparing individuals with and without detected mutations. DE between men and women was performed using a Mann–Whitney *U*-test on the same expression matrices.

Reconstruction of MDS classification models using improved cell mapping and filtering

For analyzing MDS classification we re-analyzed sequenced libraries of all disease cases and healthy individuals in two groups, separated by sequencing platform (Illumina and Ultima). Cell filtering was then applied for each of the two datasets using the process described above with the following minor modifications:

- Re-mapping all cells using cell-ranger v.7.0.1
- Both Vireo and Souporcell were limited to 7.4 M SNPs with minor allele frequency >0.05 according to the 1,000 Genome project, rather than SNPs from our genotyping panel
- Refined filter for cell exclusion, excluding 10× particles (cells) with high mitochondrial content (>20%), platelet signature (*PPBP* > 0.2%), neutrophil signature (*LCN2*, *CAMP* and *LTF*) or erythrocyte signature (*HBB*, *HBA1* and *HBA2*), and also excluding cells with low signature of nuclear RNAs (Supplementary Table 9 includes number of excluded cells)
- Adjusting the doublet detection algorithm described above with an additional filter involving clustering cells and removal of cells with UMI count that is higher than 2.5-fold of their computed cluster median (thereby compensating for variable cell sizes across types). In addition, cells with high expression of both monocyte and MEBEMP markers were filtered out. Such extra steps were needed because, in some disease batches, highly specific cell states could contaminate other samples more than in standard reference batches.

In silico sorting for inferring sample cHSPC compositions

To estimate cHSPC state for a given single-cell transcriptional profile, an in silico sorting scheme was developed (Extended Data Fig. 7a). First, our original reference model was used to compile gene signatures. Each signature was based on genes differentially expressed in a given cell type, such that the total number of UMIs for the signature is sufficiently high to allow classification at single-cell resolution (selected gene sets in Supplementary Table 13). Extended Data Fig. 7a shows the gating strategy used for classification using signature scores ($\log_2$(total signature UMI in cell) − $\log_2$(total cell UMI)). Cells with ambiguous gating were defined as unassigned. We confirmed that the gating strategy yields classification that is consistent with the annotation derived by applying metacell analysis to the new Fig. 4 reference model (see below) by projecting inferred classes on the metacell model UMAP projection

(Extended Data Fig. 7b). We noted that sorting may reduce the manifold resolution compared with metacell analysis, but it provides robust results for downstream MDS classification purposes.

Healthy donor reference model for MDS analysis

Samples from 79 individuals who were sequenced on the Illumina platform and showed no evidence for disease were considered as the reference model for MDS classification (Fig. 4). This cohort was used for defining the normal distribution of cHSPC composition. It was also the basis for constructing a reference metacell model used for projecting patient data. This model includes 287,000 cells and 2,090 metacells, constructed using metacell2 with a target metacell size of 140 cells and other parameters similar to the original normal reference model.

Grouping MDS and cytopenia patients by composition

We used the inferred cHSPC compositions of 70 donors in the Fig. 4 reference model to score each composition abnormality of patients with MDS and cytopenia. Composition vectors **p** over cell types were log-transformed first as $\mathbf{lp} = \log_2(\epsilon + \mathbf{p})$ where $\epsilon = 0.02$. The distance between two samples was then defined using a Euclidean distance between their **lp** vectors. The abnormality score of a new sample was defined as the average distance for the four nearest neighbors in the reference model. The 0.98 quantile of the abnormality score of healthy donors (excluding reference donors) was used as a threshold (Extended Data Fig. 8a) for classifying patient compositions (excluding four patients exhibiting complex karyotypes) as normal or abnormal (Fig. 4b; GRP1 or GRP2-4, respectively). Patients with abnormal compositions were further grouped using hierarchical clustering of their **lp** vectors. Patients with normal compositions were ordered along a CLP frequency gradient in Fig. 4b (left), in analogy with Fig. 2d.

Estimating patients' CNAs using scRNA projection on the reference

We constructed a metacell model from the filtered cells of each patient separately. This model was projected over the Fig. 4 reference model, using MCProj[38]. The result was a set of metacells for the patient, such that each metacell m was defined by its observed gene expression $e^{\text{obs}}_{\text{gm}}$ and projected gene expression $e^{\text{proj}}_{\text{gm}}$, as determined by MCProj using best matching reference behavior. Expression values were calculated using the geometric mean. Genes were filtered to remove sex-specific and sequencing-platform-specific genes (559 genes overall; Supplementary Table 15). Further filtering was done for 107 genes showing a consistent difference between observed and projected values over all individuals, as well as 27,317 lowly expressed genes (Supplementary Table 15).

To correct for batch effects leading to small GC content preference per library, we grouped genes into ten equal size bins according to the average GC content of 3′-scRNA-sequenced tags in representative sequenced libraries. For each gene g, we computed total observed and expected UMI counts, given the model's projection on the reference:

$$n^{\text{obs}}_g = \sum_c n_c \times e^{\text{obs}}_{\text{gm}(c)}$$

$$n^{\text{proj}}_g = \sum_c n_c \times e^{\text{proj}}_{\text{gm}(c)}$$

where n_c is the total number of UMIs for the cell c and $m(c)$ is its metacell. The bias per GC bin $\text{bias}_{\text{GCbin}}$ is now approximately defined as the median of $\frac{n^{\text{obs}}_g}{n^{\text{proj}}_g}$ across genes in the GC bin. In practice, we calculated the ratio $\frac{n^{\text{obs}}_g}{n^{\text{proj}}_g}$ after normalizing n^{obs}_g and n^{proj}_g by $\sum_{g'} n^{\text{obs}}_{g'}$ and $\sum_{g'} n^{\text{proj}}_{g'}$, respectively, and adding a regularization term of 10^{-5}. We corrected each observed gene expression value $e^{\text{obs}}_{\text{gm}}$ by dividing by the appropriate $\text{bias}_{\text{GCbin}}$ value, generating $e^{\text{obs}'}_{\text{gm}}$, which is used to obtain a per-gene and metacell observed:expected expression log(ratio) $\delta_{\text{gm}} = \log_2\left(\epsilon + e^{\text{obs}'}_{\text{gm}}\right) - \log_2\left(\epsilon + e^{\text{proj}}_{\text{gm}}\right)$, whereas $\epsilon = 10^{-5}$.

We split each chromosome to contiguous bins encompassing 20 genes (ignoring filtered genes). For each chromosome bin bchrom and each metacell m, the median log ratio was computed: $\delta_{m,\text{bchrom}} = \text{Med}_{g \in \text{bchrom}}(\delta_{\text{gm}})$.

The matrix $\delta_{m,\text{bchrom}}$ of metacells and 20-gene bins describing estimated DE was then normalized by subtracting the median of each row (metacell) and visualized in a heatmap, where metacells were re-ordered using hierarchical clustering. Heatmaps of the derived matrices (see, for example, Extended Data Fig. 10a,b) were examined manually to identify CNAs (Extended Data Fig. 7c and Supplementary Table 12).

Within-state gene signatures

Within-state correlated gene sets were inferred from the reference metacell model by clustering the gene–metacell expression matrix of each annotated cell type, while considering only highly variable genes within the cell type. Clusters were evaluated and selected manually and expanded by adding correlated genes, resulting in the final gene sets (Supplementary Table 13). An MHC-II gene set was added after observation of MDS DE compared with the reference. The signature score per cell was estimated as the log-transformed normalized total UMI count for each gene set.

Signature scores per patient were extracted using the median signature of cells within a respective cell type (following the in silico sorting process described above). In the case of too few such cells, the signature score was considered missing (NA).

After classification of patients with cytopenia and MDS into groups 1–4 (Fig. 4b), we performed, for each within-state gene signature (and each relevant state), a Kruskal–Wallis test comparing signature expression levels between groups 1–4 and healthy donors. The signatures with the lowest P values were the MEBEMP-L MHC-II, S-phase and BEMP early signatures, which were accordingly shown in Fig. 4e and Extended Data Fig. 8d.

Features for MDS classification

The following features were collected to facilitate MDS classification:

- CBC values: we used the values with minimal time gap from the cHSPC sampling date
- Maximal CH VAF across mutations detected in the same blood sample that was used for scRNA-seq
- The cHSPC compositions as inferred through in silico sorting
- Twenty-one signature scores
- Composition abnormality score ('Grouping MDS and cytopenia patients by composition')
- Number of CNAs.

We noted that signature scores might be missing (as a result of insufficient number of cells). In addition, a few individuals were missing CBC values.

All feature values per scRNA sample are included in Supplementary Table 9.

MDS classifier training and testing

XGBoost (xgboost Python package, v.2.0.3) training was performed on cohort 1 including 89 samples (41 normal, 20 cytopenia and 28 MDS). All of the samples in cohort 1 were sequenced on the Ultima platform. MDS (including MDS or myeloproliferative neoplasms) samples were considered positive, whereas cytopenia and normal samples were considered negative.

We applied feature selection separately in each leave-one-out fold, selecting a subset of the features for which the FDR-corrected

Mann–Whitney *P* value for MDS or cytopenia separation was <0.1. We then inferred a classifier using the binary:logistic objective and default hyperparameters, except for tree_method = hist. Model performance was evaluated by pooling accuracy on the left-out samples of their respective folds.

Further evaluation of the approach was done by classifying cohort 2 (16 MDS and 9 cytopenia cases, all sequenced on the Illumina platform) using the classifier trained on cohort 1.

Blasts versus CLP-E analysis

We compared blast fraction from BM samples acquired at most 1 year before or after cHSPC sampling. Comparison was also restricted to cHSPC samples with >500 HSPCs (scRNA samples used in this analysis are specified in Supplementary Table 9). CLP-E-like frequency estimation was as described in Extended Data Fig. 9d.

Reporting summary

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

Data availability

All scRNA-seq data generated in the present study is available in the Gene Expression Omnibus under accession no. GSE285943, in CELLxGENE (https://cellxgene.cziscience.com/collections/5542eeb0-96ef-4ab9-95ea-eb6abc178461) and as metacells at https://apps.tanaylab.com/MCV/blood_aging. Targeted DNA sequencing data of clonal hematopoiesis mutations by MIP is available in the European Nucleotide Archive under accession no. PRJEB85241. All of this data was uploaded in accordance with donor written informed consent.

Code availability

Code to reproduce the figures is available at https://github.com/tanaylab/blood_aging. The Metacell package is available at https://github.com/tanaylab/metacells.

References

33. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
34. Koboldt, D. C. et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
35. Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing. *Bioinformatics* **30**, 2670–2672 (2014).
36. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
37. Heaton, H. et al. Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
38. Ben-Kiki, O. et al. MCProj: metacell projection for interpretable and quantitative use of transcriptional atlases. *Genome Biol.* **24**, 220 (2023).
39. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

Acknowledgements

The present study was generously supported by the Adelis Foundation. Research was also supported by the ISF-IPMP-Israel Precision Medicine Program no. 3165/19. Additional grant support includes LLS and Rising Tide Foundation (grant nos. RTF6005-19, ISF-NSFC 2427/18 and ISF 1123/21), the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine, ERC Advanced grant (Cells2Tissues) and ISF-MAPATZ (grant no. 714459). This research was also supported by the Sagol Institute for Longevity Research, the Barry and Eleanore Reznik Family Cancer Research Fund, the Steven B. Rubenstein Research Fund for Leukemia and Other Blood Disorders, the Applebaum Foundation, the Bolton Hope foundation, the Anthony Beck foundation, the estate of Hartz de Rooij, ICRF-USA-PG, IMOS German Program no. 0004070, and the EU Horizon 2020 funding project MAMLE ID 714731. L.S. is an incumbent of the Ruth and Louis Leland career development chair. The contribution of N.R. is part of a PhD thesis research conducted at Tel Aviv University. N.R. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics, Tel Aviv University and the Planning and Budgeting Committee fellowship for excellent PhD students in Data Sciences. N.R. was also supported by awards from the Herczeg Institute on Aging and the Tel Aviv University Healthy Longevity Research Center. We thank all members of the L.S. and A.T. laboratories for their constructive comments. We thank all the individuals who participated in the present study.

Author contributions

Patient recruitment was performed by N.F., L.S., S.T., and S.B. Blood drawing and sample preparation for scRNA-seq, including CD34 enrichment, and 10x scRNA library preparation and sequencing on the Illumina platform were performed by N.F. ScRNA-seq data processing, including doublet calling, assignment of cells to individuals and metacell model construction and BM projections and comparisons were performed by N.R. and O.M. All analysis toward Figs. 1–3 was performed by N.R. Clinical data curation was performed by N.F. Clinical data association analyses were performed by N.R. MDS analysis in Fig. 4 and related processing were performed by O.M. MCV construction was performed by A.L. CH and genotyping deep targeted sequencing of all study participants and amplicon validation of CH mutations were performed by N.F. Sequencing data analysis and variant calling were performed by N.C.-I., N.F. and L.S. DNA samples for the high-RDW CH analysis were provided by S.S. and N.A. CH sequencing of all high-RDW and control DNA samples was performed by N.F. and analyzed by N.C.-I., N.F. and L.S. Replicate 10x library sequencing on the Ultima Genomics platform was performed by Z.S., D.L., E.M. and G.Y. Biological and technical replicate analysis was performed by N.R. GoT experiments were performed by S.G., analyzed by N.D. and P.C., supervised by D.L. and further analyzed by N.R. GATA3 deep targeted sequencing and Sanger validation were performed by S.S. and supervised by J.T. A.B., A.L. and O.B.-K. contributed to data analysis and interpretation. A.D. provided sample preparation support. M.K. provided 10x guidance and technical support. N.K. contributed to funding applications. L.S. and A.T. designed and supervised all aspects of this study. N.F., N.R., O.M., L.S. and A.T. wrote the paper with help from all authors.

Competing interests

Z.S., D.L., E.M. and G.Y. are all employees and shareholders of Ultima Genomics. L.S. is a shareholder of Sequentify. L.S. and A.T. are shareholders of Cliseq. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-025-03716-5.

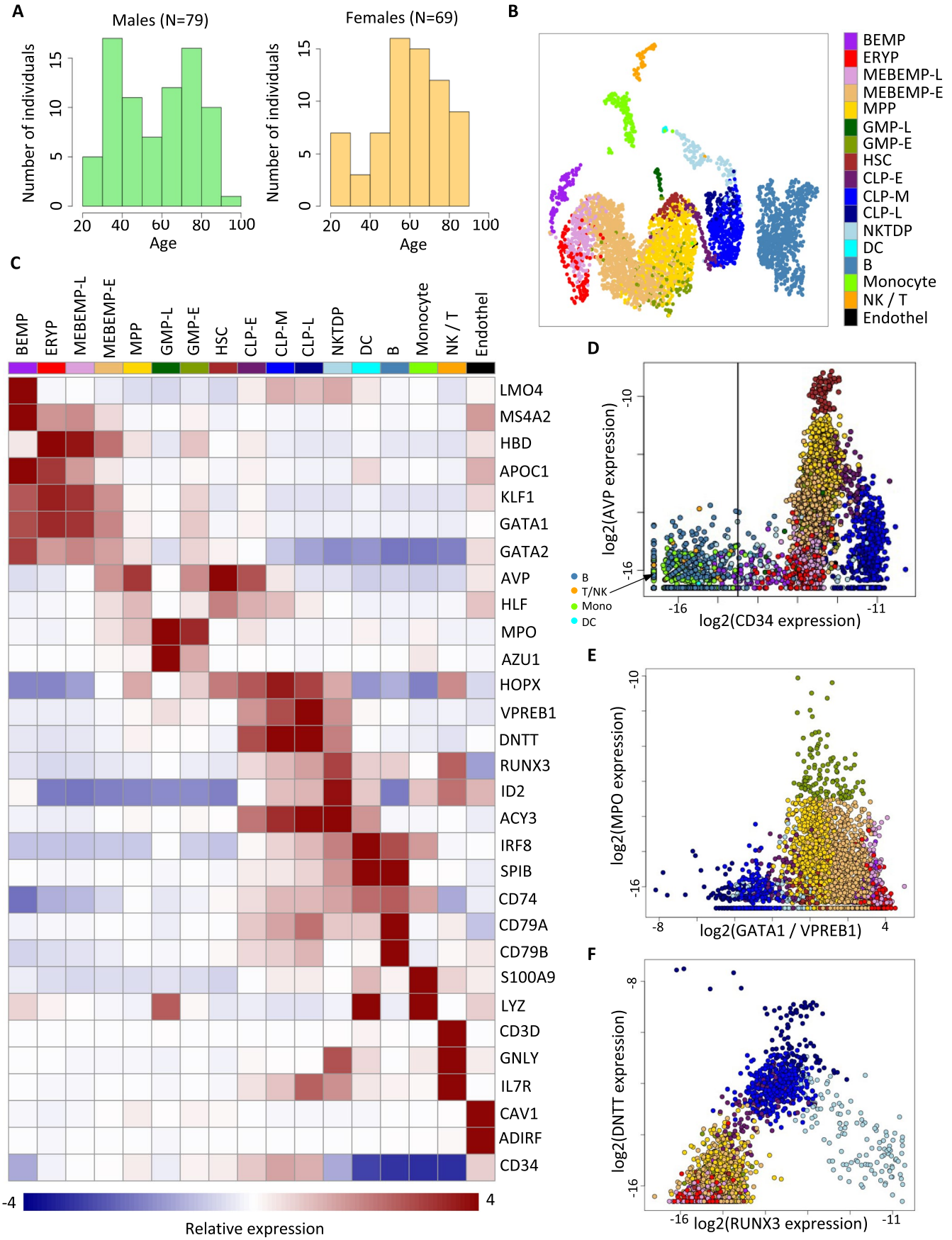
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-03716-5.

https://doi.org/10.1038/s41591-025-03716-5

Correspondence and requests for materials should be addressed to A. Tanay or L. Shlush.

Peer review information *Nature Medicine* thanks Maria Carolina Florian, Peter van Galen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Anna Maria Ranzoni, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-025-03716-5

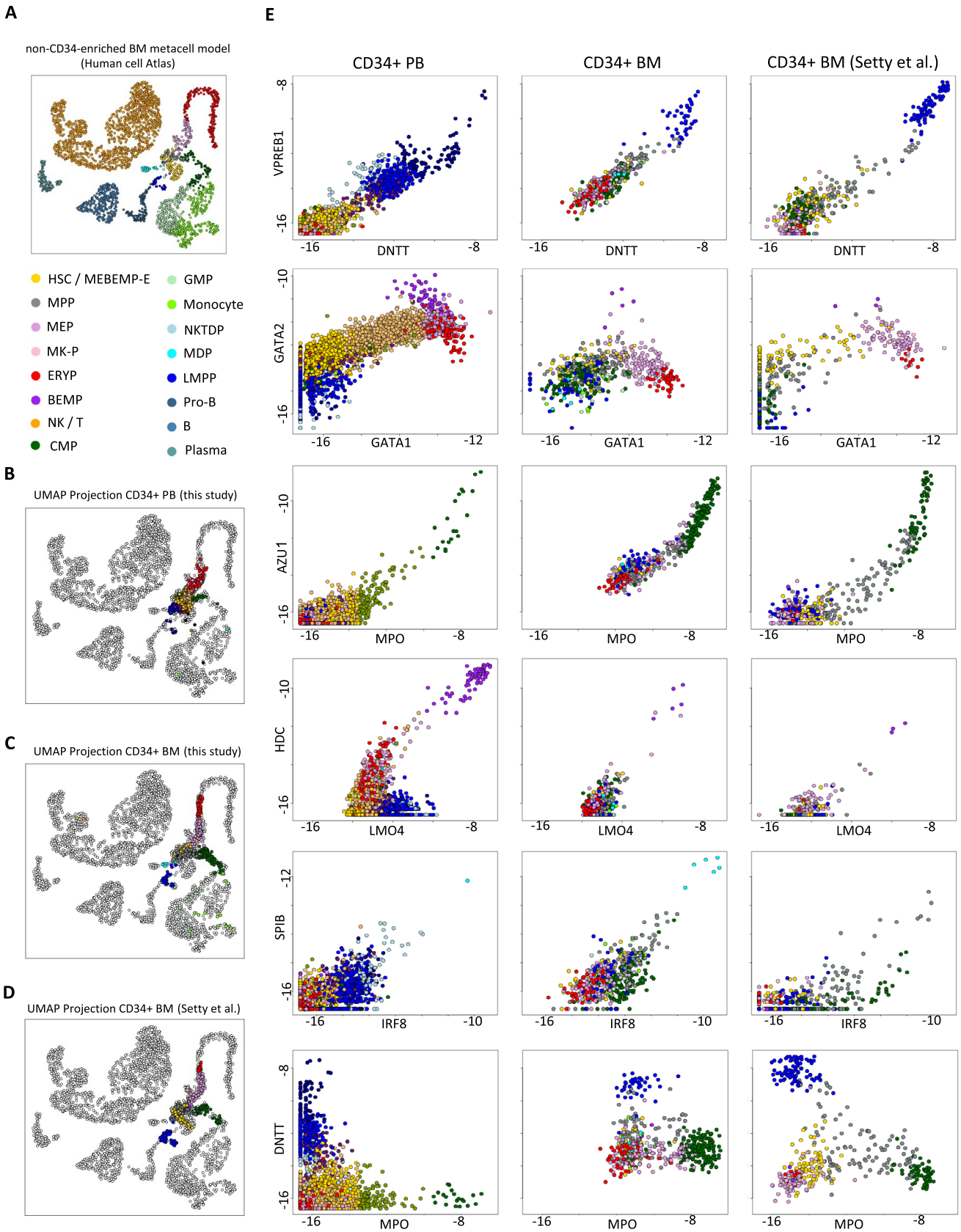


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41591-025-03716-5

Extended Data Fig. 1 | Cell state annotation, major markers and regulators of HSC differentiation and sub-population branching. 1A – age distribution (decimals) of studied population by sex. **1B** – 2D UMAP projection of our metacell model prior to *CD34* metacell filtering. **1C** - relative expression heatmap of cell states (columns) and gene markers used for cell state annotation (rows). **1D** – Metacell filtration on *CD34* expression. **1E** - expression plot of *MPO* and *GATA1/VPREB1* showing all 3 differentiation trajectories (GMPs, CLPs, MEBEMPs) from HSCs. **1F -** gene-gene expression plot of *DNTT* and *RUNX3*, showing early CLP differentiation and their bifurcation into late CLPs and NKTDPs. All gene expression values are obtained by normalizing gene expression to sum to 1 and taking log2.

https://doi.org/10.1038/s41591-025-03716-5

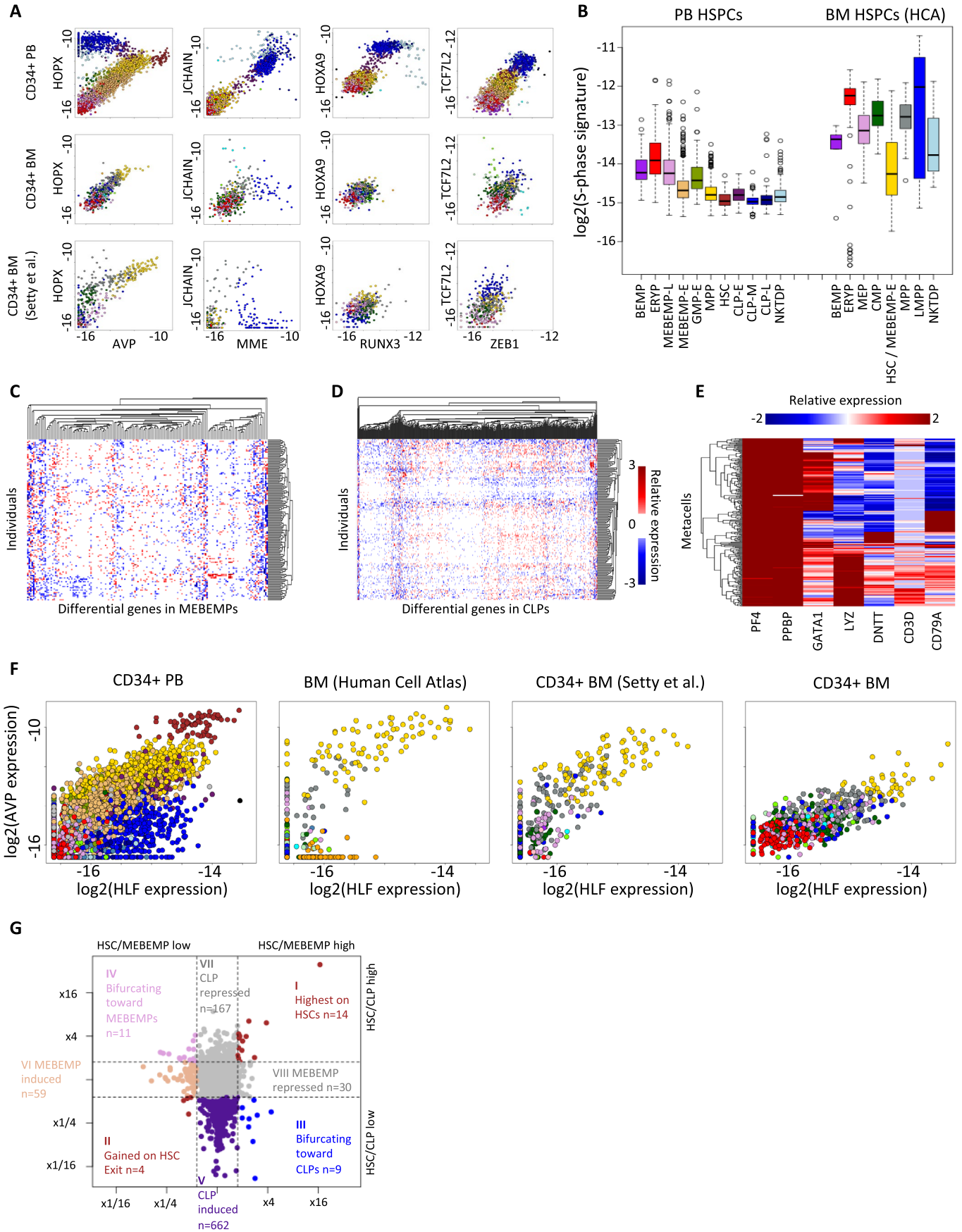


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s41591-025-03716-5

Extended Data Fig. 2 | BM comparisons I. 2A – 2D UMAP projection of a non-CD34-enriched BM metacell model from the Human Cell Atlas[14], colored by a BM-specific cell state annotation. **2B** - projection of our PB CD34$^+$ derived metacells on the non-CD34 enriched BM metacell model. **2C** – projection of our BM CD34$^+$ derived metacells on the non-CD34 enriched BM metacell model. **2D** - projection of BM CD34$^+$ derived metacells from Setty et al.[12] on the non-CD34 enriched BM metacell model. **2E** - gene-gene expression plots comparing PB CD34$^+$ derived metacells with their BM CD34$^+$ counterparts from our study and from Setty et al. for all differentiation trajectories. Panels (top to bottom) represent CLP differentiation, MEBEMP differentiation, GMP differentiation, BEMP differentiation, DC differentiation, and GMP/CLP/MEBEMP trifurcation from HSCs. PB and BM metacells are colored by PB and BM specific annotations.

https://doi.org/10.1038/s41591-025-03716-5

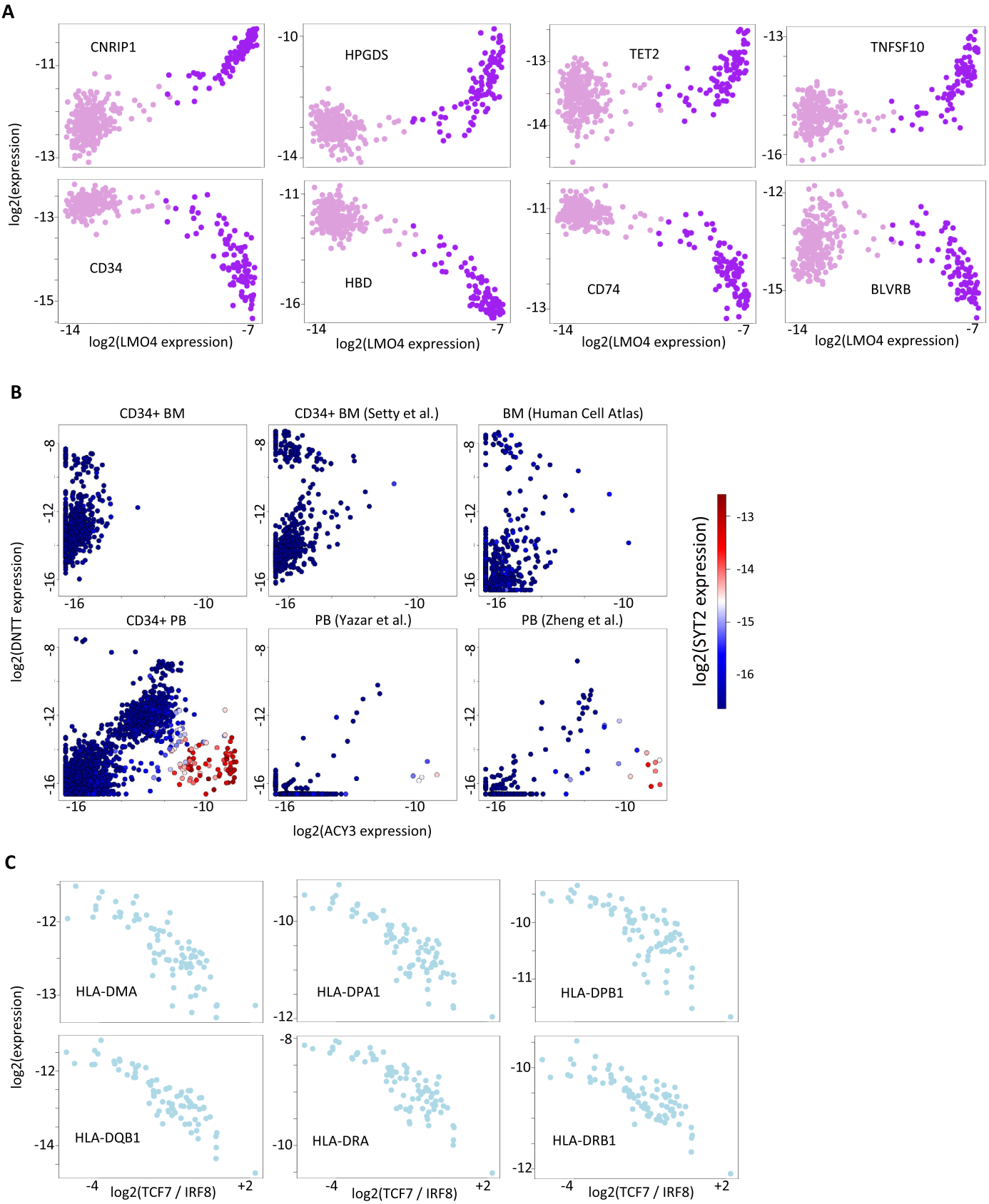


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-025-03716-5

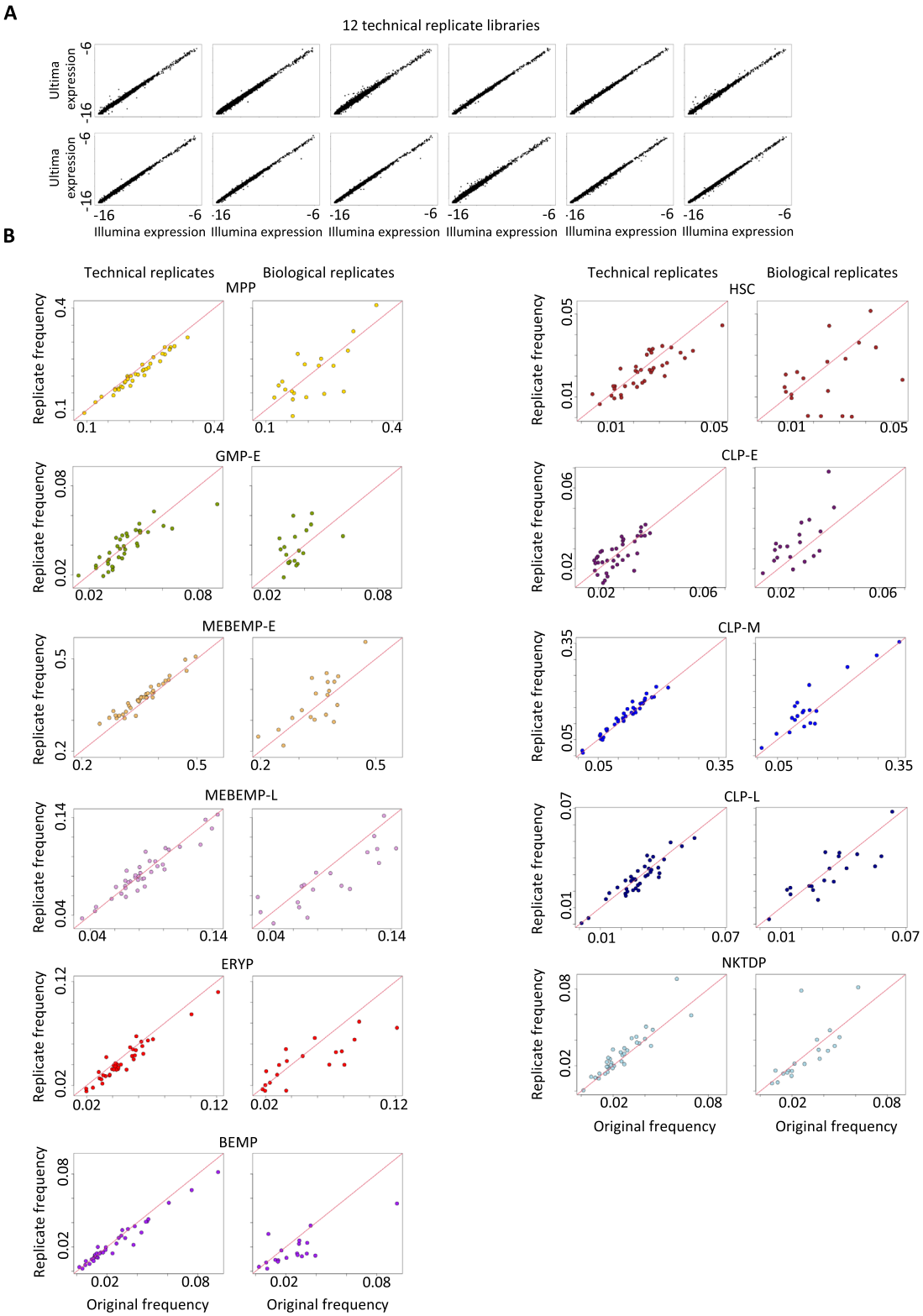
Extended Data Fig. 3 | BM comparisons, Individual-specific state-controlled differential gene expression, megakaryocytic contamination and circulating HSCs. 3A – gene-gene expression plots comparing PB CD34$^+$ derived metacells with their BM CD34$^+$ counterparts from our study and from Setty et al.[12] for markers and regulators of CLP differentiation and bifurcation. **3B** – cell state specific comparison of S-phase signatures in circulating (left) and. BM[14], right) HSPCs. Boxplot centers, hinges and whiskers represent median, first and third quartiles and 1.5× interquartile range, respectively. Outliers are marked by circles. Number of metacells per box (left to right): 89, 124, 261, 1178, 116, 668, 69, 88, 378, 88, 127; 5, 194, 62, 82, 87, 43, 35, 3. **3C,D** – Individual-specific differential gene expression after controlling for distribution across the CD34$^+$ PB manifold in MEBEMPs (**C**) and CLPs (**D**). **3E** - relative expression heatmap of the megakaryocytic markers *PF4* and *PPBP* and cell-state-specific markers, across metacells with high megakaryocytic signature, showing an abnormally high doublet rate involving megakaryocytes. Cells contained in such metacells were accordingly excluded from the final metacell model. **3F** - gene-gene expression plots comparing the PB high *AVP* and *HLF* HSC population (left) with that found in two BM metacell models[12,14] and in our CD34$^+$ BM data. PB and BM metacells are colored by PB and BM specific annotations. **3G** – map of transcriptionally activated genes upon exit from the HSC state and differentiation toward lymphoid (CLP) and non-lymphoid (MEBEMP) fates. Dots represent genes. HSC/CLP and HSC/MEBEMP gene expression ratios are depicted on the y and x axis, respectively. Class I genes are representative of the HSC state; Class II genes exhibit symmetric transcriptional activation upon exit from the HSC state towards CLP and MEBEMP fates, whereas Class III, IV, V, VI exhibit asymmetrical transcriptional activation upon exit from the HSC state towards CLP (class III, V) and MEBEMP (Class IV, VI) fates. n is the number of genes in each class.

https://doi.org/10.1038/s41591-025-03716-5



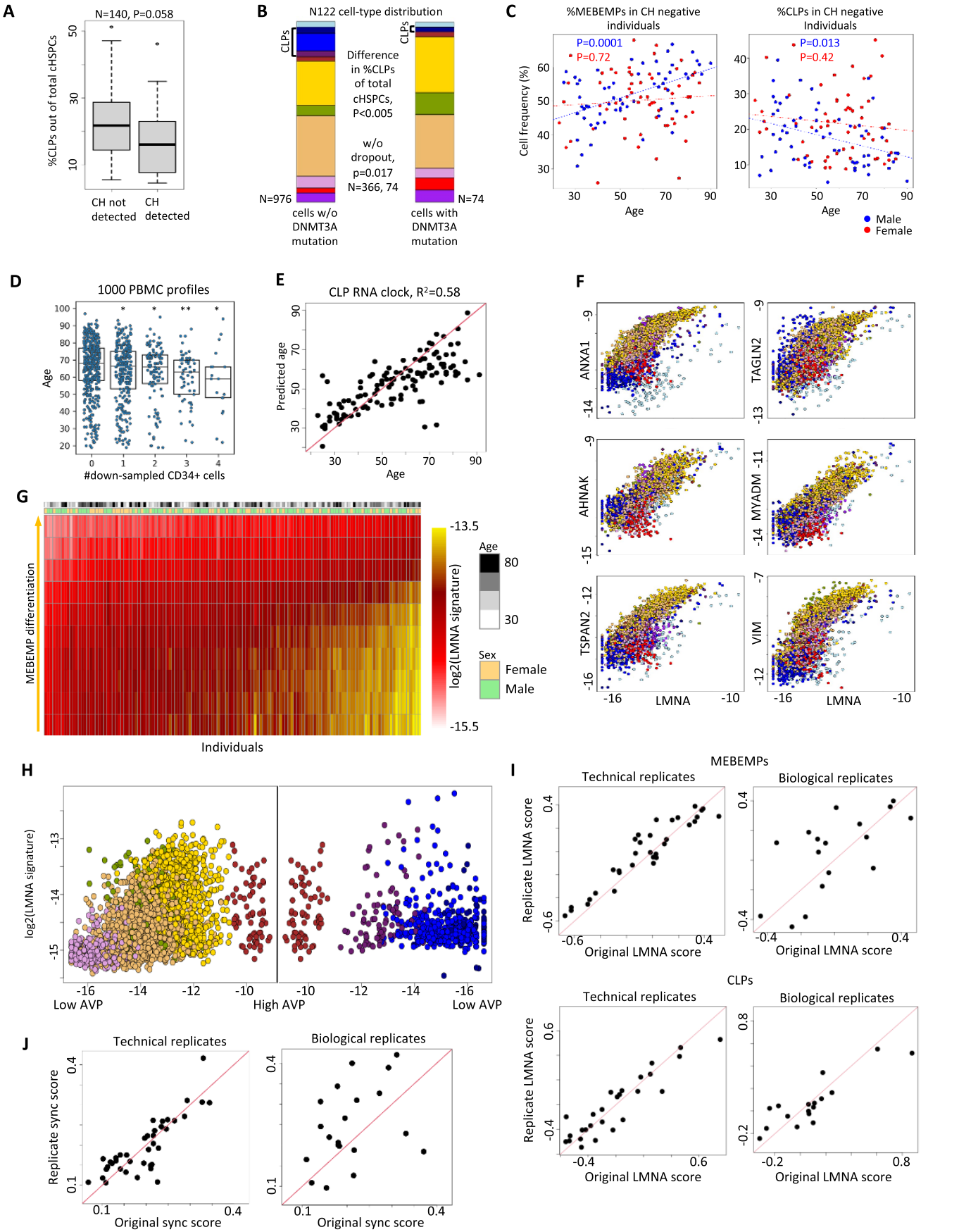
Extended Data Fig. 4 | Factors involved in BEMP and NKTDP differentiation. **4A** - factors positively and negatively regulated in the early stages of BEMP specification. **4B** – gene-gene expression plots of *DNTT* and *ACY3* comparing CD34-enriched[12] and non-enriched[14] BM to our CD34$^+$ BM model (top), as well as non-enriched[25] and partially CD34-enriched PB[39] to our CD34$^+$ PB model (bottom). Metacells are color-coded by *SYT2* expression. The *SYT2* high, *ACY3* high, *DNTT* intermediate population clearly seen in our cHSPC data is completely lacking from the BM datasets. **4C** – anti-correlation of the DC *IRF8-MHC-II* coupled dynamics and the T cell regulator *TCF7*, involved in the bifurcation of the NKTDP state to its sub-populations.

https://doi.org/10.1038/s41591-025-03716-5



Extended Data Fig. 5 | Stability of cell state compositions across technical and biological replicates. 5A – comparison of Illumina and Ultima Genomics sequenced data. Each panel represents one library that was sequenced using both technologies. Points represent genes, and each gene's expression level across all cells in the library as determined by Illumina (x) and Ultima Genomics (y) is shown. **5B** – Cell state frequency comparisons between 39 technical & 19 biological replicates and their original samples. Each pair of panels represents one cell state, denoted on top. Panels on the left of each pair compare the cell state frequency in the original sample, sequenced by Illumina (x), to its technical replicate frequency, sequenced by Ultima Genomics (y). Panels on the right of each pair compare the cell state frequency in the original sample (x) to its frequency in the biological replicate (y). All biological replicates were sampled 1 year following original blood sampling. The diagonal y = x is shown in red.

https://doi.org/10.1038/s41591-025-03716-5

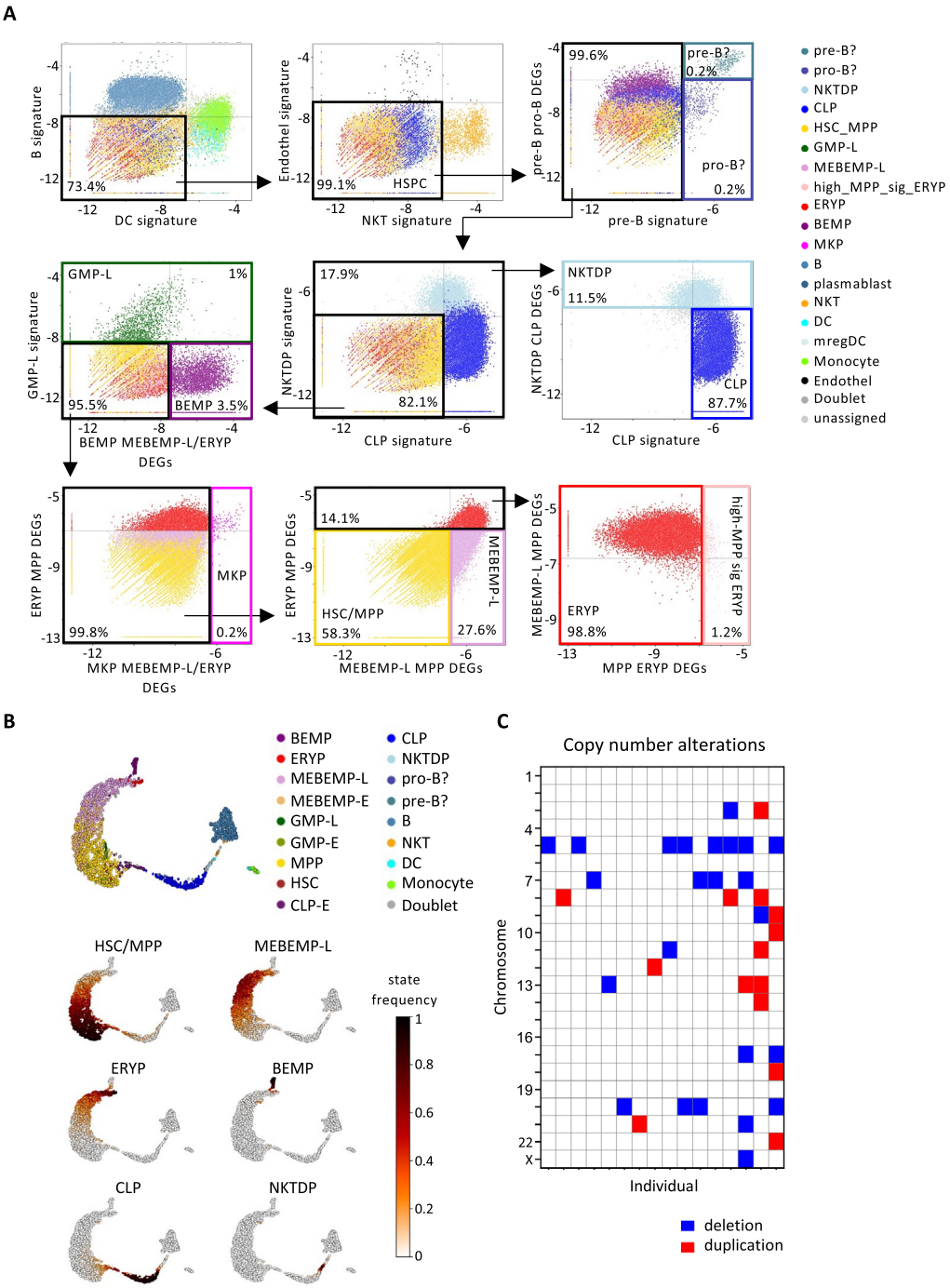


Extended Data Fig. 6 | See next page for caption.

https://doi.org/10.1038/s41591-025-03716-5

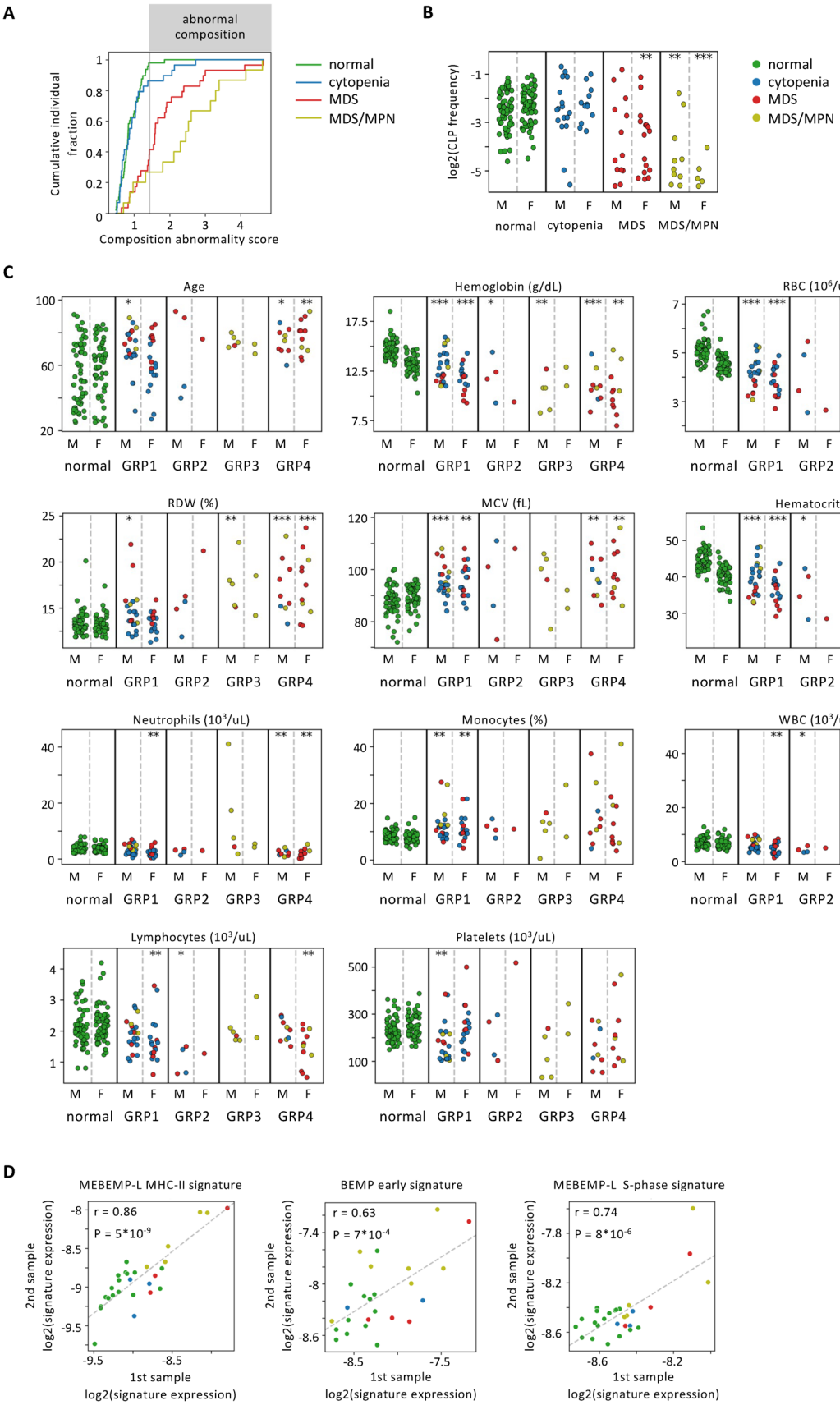
Extended Data Fig. 6 | Composition-controlled transcriptional variation: the *LMNA* signature and sync score. 6A – boxplots showing CLP frequency distributions in individuals with (right) and without (left) clonal hematopoiesis. Boxplot centers, hinges and whiskers represent median, first and third quartiles and 1.5× interquartile range, respectively. Outliers are marked by circles. Two-sided Mann-Whitney U p-value is indicated. **6B** - Relative cell state frequencies in mutant (right) and non-mutant (left) cells following GoT of sample N122 (DNMT3A R882H mutated, VAF = 0.07). P-values were calculated using two-sided Fisher's exact tests. **6C** – Similar data as in Fig. 3a, showing each individual's age and MEBEMP (left) or CLP (right) cell-state frequency. Each dot represents an individual. Males are color-coded in blue and females in red. P-values for Spearman test of independence are indicated for males and females. **6D** – Same data as in Fig. 3d, showing individual age distributions (y axis) stratified by the (down-sampled) number of *CD34*$^+$ single cells. Boxes represent median, first and third quartiles. Stars denote significant difference from the age distribution of individuals with 0 observed down-sampled *CD34*$^+$ cells, as determined by two-sided Mann-Whitney U test. **6E** – True age (x) versus age predicted based on composition-controlled CLP expression (y). The diagonal y = x is shown in red. **6F** – the *LMNA* signature – co-variation of *LMNA* expression with *ANXA1, TAGLN2, AHNAK, MYADM, TSPAN2* and *VIM*. **6G** – heatmap of individual *LMNA* signature expression across the MEBEMP trajectory. Individual age and sex are color-coded on top. **6H** – *LMNA* signature vs *AVP* expression in HSCs (denoted by high *AVP*, center) and throughout MPP / MEBEMP (left) and lymphoid (right) differentiation. **6I** - *LMNA* signature expression correlations between 39 technical & 17 biological replicates and their original samples, calculated across MEBEMPs (top) and CLPs (bottom). The diagonal y = x is shown in red. **6J** - sync score correlations between 39 technical & 20 biological replicates and their original samples. All biological replicates were sampled 1 year following original blood sampling. The diagonal y = x is shown in red.

https://doi.org/10.1038/s41591-025-03716-5



Extended Data Fig. 7 | In-silico sorting scheme and copy number alterations. **7A** – In-silico sorting scheme for a CD34-enriched PB scRNA sample. Each scatter plot demonstrates one or two virtual gates, based on total expression of gene signatures that were compiled using our cHSPC reference model. Representative cells shown belong to the 79 healthy donors comprising our Fig. 4 reference model (see **Methods**). Colors denote sorted cell states. **7B** – Validation of in-silico sorting by annotated metacells. UMAP projection of the metacell model comprised of cells in **A** is shown, colored according to metacell annotation by marker genes (top), as in Extended Data Fig. 1C, and according to in-silico sorted state frequency, for each common cHSPC state (bottom). **7C** – Distribution of copy number alterations (CNAs) identified by abnormal RNA expression over chromosomes (rows) and individuals (columns). Duplications and deletions are shown in red and blue, respectively. None of the individuals exhibited both duplication and deletion in the same chromosome. Individuals w/o CNAs are not shown.

https://doi.org/10.1038/s41591-025-03716-5

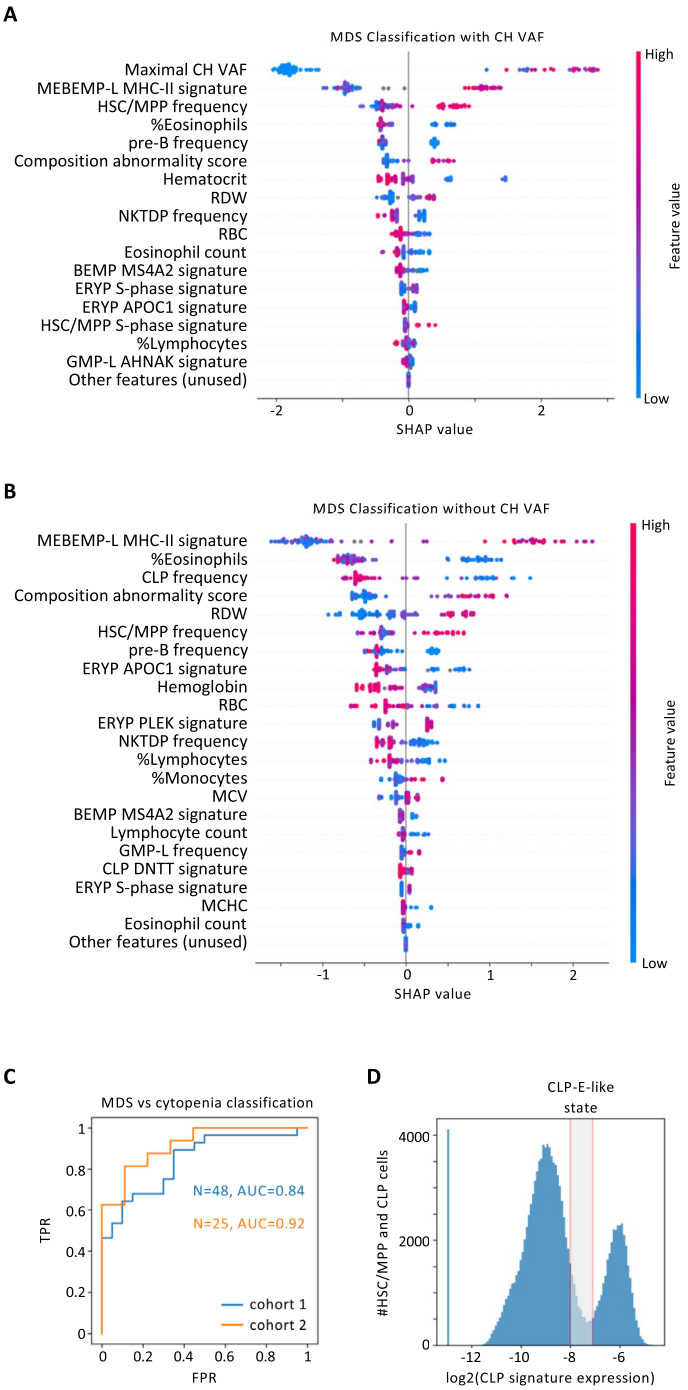


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41591-025-03716-5

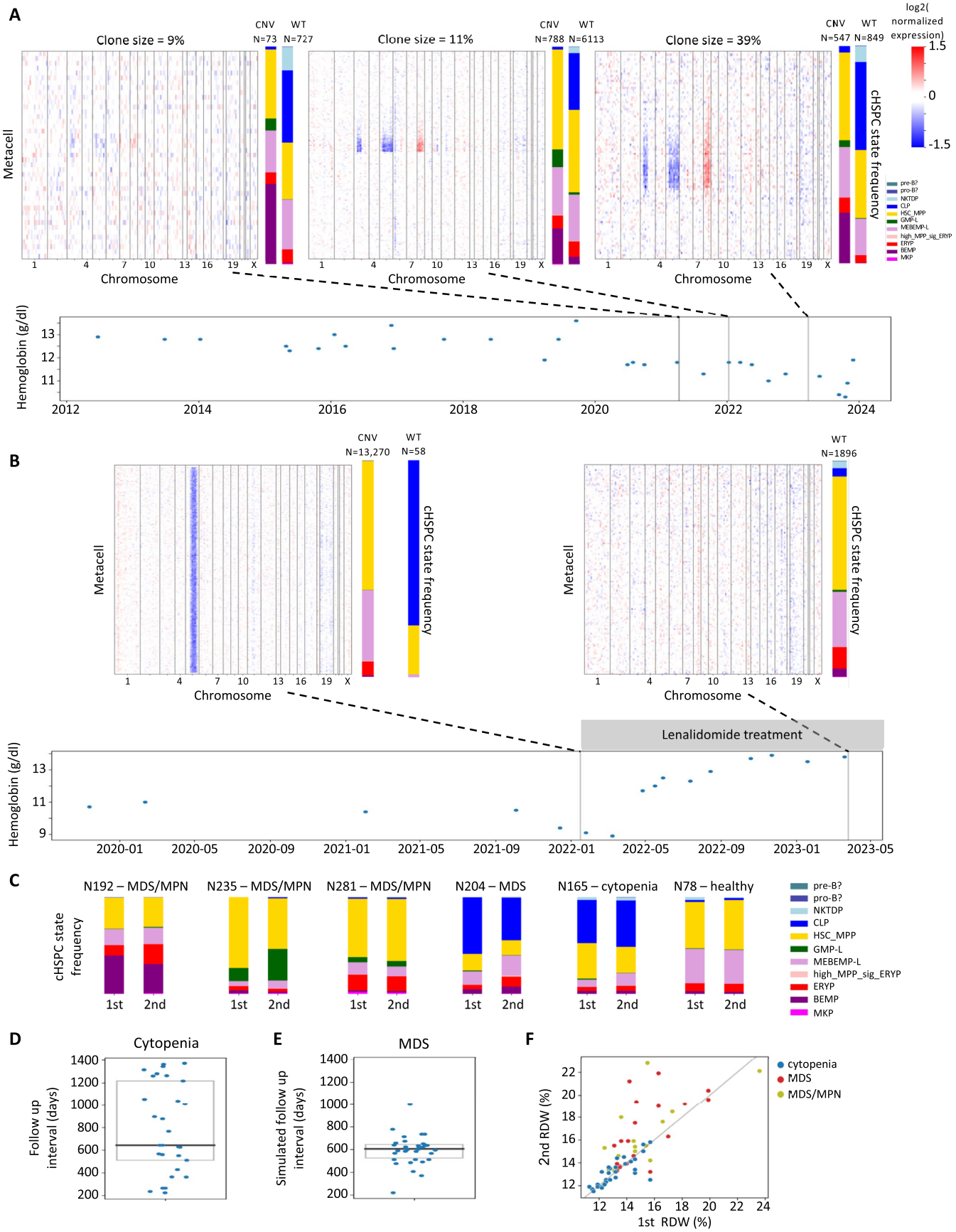
Extended Data Fig. 8 | MDS cHSPC composition groups. 8A – Distribution of composition abnormality scores across clinical diagnoses. The vertical grey line marks the 98th percentile score of healthy donors, defining the normal-like composition (group 1) and abnormal composition (groups 2-4) shown in Fig. 4b. **8B** – In-silico sorted CLP frequencies across clinical diagnoses. Individuals are further separated by sex (male – left, female – right). Colors denote clinical diagnoses. Stars denote BH-adjusted significant difference from healthy donors of the same sex, determined by two-sided Mann-Whitney U test. **8C** – Age and CBC indices values across healthy donors and cytopenia/MDS groups as in Fig. 4e. **8D** – For each signature shown in Fig. 4e, biological replicate data is shown, comparing signature expression between the first and second sample. Individuals with insufficient cell counts for the population of interest were excluded. Linear fit across all individuals (n = 28 – MEBEMP-L MHC-II and S-phase signatures, n = 25 – BEMP early signature) is shown (dashed line), as well as the corresponding r- and (non-adjusted, two-sided) p-values.

https://doi.org/10.1038/s41591-025-03716-5



Extended Data Fig. 9 | MDS classification model, CLP-E-like state. 9A – Features used by the MDS classification model whose performance is shown in Fig. 4f. For each feature used by the model, SHAP analysis shows the estimated impact of the feature on classification for each individual. Colors denote individual feature values. **9B** – same as **A**, but for an MDS classification model that does not use maximal CH VAF as a feature. **9C** – ROC curve as in Fig. 4f, but for an MDS classification model that does not use maximal CH VAF as a feature. **9D** – Definition of the CLP-E-like state. Distribution of the CLP signature is shown across all HSC/MPP and CLP cells (as defined using our in-silico sorting) in the Fig. 4 reference model, along with thresholds (red) defining an intermediate expression level (shaded grey). HSC/MPP and CLP cells exhibiting these intermediate CLP expression levels are considered to be in a CLP-E-like state.

https://doi.org/10.1038/s41591-025-03716-5



Extended Data Fig. 10 | See next page for caption.

https://doi.org/10.1038/s41591-025-03716-5

Extended Data Fig. 10 | Progression and remission case studies, abnormal composition reproducibility and stability of cytopenia patients after cHSPC sampling. 10A – Case study of disease progression. scRNA cHSPC samples were obtained for MDS patient N180 in 2021 (at which time he was diagnosed as cytopenic), 2022 and 2023, and a metacell model was constructed for each. Copy number variation (CNV) analysis is shown for each of these samples (top). Normalized gene expression (by the Fig. 4 reference model) for each metacell (row) and chromosomal region (column) is shown. This analysis revealed a small clone with trisomy 8 (red), and deletions in chromosomes 3 and 5 (del(5q)) (blue). Estimated clone size, as a percentage of total cHSPCs, is specified for each sample. Cell-state frequencies for cells with and without identified CNVs are shown on the right of each CNV analysis. The number of cells in each group is indicated. Longitudinal hemoglobin counts (bottom) are also shown, with grey vertical lines denoting dates of scRNA sampling. **10B** – Case study of remission. Same as **A**, but for MDS del5q patient N211, sampled before and after lenalidomide treatment (bottom, shaded grey). Initial CNV analysis of her scRNA data revealed a very large clone with del(5q), which could not be detected following treatment with lenalidomide. **10C** – Recurring abnormal cell state frequencies. For each of 6 individuals, cHSPC compositions (obtained by in-silico sorting) are shown for two different scRNA samples, demonstrating recurrence of abnormal cHSPC state frequencies (N192 – high BEMP, N235 and N281 – high GMP-L, N204 and N165 – high CLP, N78 – low CLP). **10D** – Distribution of follow-up intervals, between cHSPC sampling and most recent available CBC results, across 29 (out of 33) patients with cytopenia. Boxplot indicates first, second and third quartiles. **10E** – As a positive control for CBC instability in MDS, earlier CBC results were retrieved for MDS patients, targeting an interval of approximately 600 days between 1st and 2nd CBC measurements, to reflect the follow-up duration observed in patients with cytopenia. Shown here is the distribution of these simulated follow-up intervals for patients with MDS (similar to **D**). Treated patients and those with follow-up intervals <200 days were excluded from this analysis. **10F** – Change in RDW values over follow-up intervals shown in **D** and **E**, used as a proxy for disease progression. Colors denote clinical diagnosis at cHSPC sampling.

nature portfolio

Corresponding author(s): Prof. Liran Shlush
Prof. Amos Tanay

Last updated by author(s): Feb 27, 2025

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	we used R MatchIt package for case-control matching in the RDW comparisons (Fig 2F)
Data analysis	scRNA-seq data analysis: vireo (version 0.3.2, with cellSNP version 0.3.0) and souporcell (version 2.4) were used to identify doublets and assign cells to individuals. To produce UMI count matrices, cell-ranger version 3.1.0 was used in the analysis resulting in Figures 1-3 and EDFs 1-6, while cell-ranger version 7.0.1 was used in the analysis resulting in Figure 4 and EDFs 7-10. In both cases, cell-ranger count was used to align reads to hg38. metacell2 was used to construct all cohort and individual metacell models. Targeted DNA-seq (MIP): Varscan was used to call SNPs from the genotyping MIP panel. BWA-MEM (v2) was used to align reads either to hg38, or to hg37 while LiftOver was used to convert hg37 genomic positions to hg38. For predicting MDS diagnosis, xgboost Python package (v2.0.3) was used. Code to reproduce the figures is available at https://github.com/tanaylab/blood_aging. The Metacell package is available at https://github.com/tanaylab/metacells

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

scRNA-seq data generated in this study is available in Gene Expression Omnibus (GEO) under accession GSE285943, in CELLxGENE (https://cellxgene.cziscience.com/collections/5542eeb0-96ef-4ab9-95ea-eb6abc178461), and also as metacells at https://apps.tanaylab.com/MCV/blood_aging/. Targeted DNA sequencing data of clonal hematopoiesis mutations by Molecular Inversion Probes is available in European Nucleotide Archive (ENA) under accession PRJEB85241. All of these data were uploaded in accordance with donor written informed consent.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	No gender-related data was collected. Sex data was self-reported and verified based on XIST and RPS4Y1 expression. Sex data is reported in Supplementary Tables S1 and S9.
Reporting on race, ethnicity, or other socially relevant groupings	No sociopolitical construct (such as race or ethnicity) were considered in this study. We acknowledge that due to practical considerations in recruitment, this study included mostly individuals of Jewish ancestry, and so our results should be validated in a larger and more ancestrally diverse cohort.
Population characteristics	For the healthy cohort (analyzed in Figures 1-3) we collected blood samples from 148 consented healthy individuals, consisting of 79 males and 69 females aged 23-91 (median 61.5). The cytopenic cohort (analyzed in Figure 4) included a total of 83 individuals, 50 of which were labeled as MDS cases based on BM morphology and/or mutational and karyotypic abnormalities (as detected in the clinic, or by our CH panel and scRNA CNA analysis). The remaining 33 cytopenic patients not satisfying MDS criteria were labeled as cytopenia cases. 17 of the 83 cytopenic patients, presenting with asymptomatic mild cytopenia, were also included in the original 148-individual healthy cohort. Median age for the cytopenic cohort was 73 years (range 27-93), with males representing 53% of patients.
Recruitment	Recruitment of the healthy cohort took place between Nov 2020 and Dec 2023. Blood donors were recruited from numerous sources (including the Weizmann Institute of Science and several primary care clinics) as detailed under "Methods". We profiled responsive volunteers with no known hematological malignancy / prior information on blood clonality, balancing sex and seeking a dispersed age distribution biased toward older individuals. We reassessed this strategy following initial sampling, and observed remarkable homogeneity in transcriptional states across individuals. We observed large variation in the composition of cell states across healthy individuals, which suggested that more sampling, without specific preferences, would be appropriate for further characterizing compositional variation. Recruitment of the cytopenic cohort (including MDS and non-MDS-related cytopenic patients) took place between Nov 2021 - Feb 2024. These patients were recruited from several outpatient hematological clinics by collaborating physicians to represent the wide clinical spectrum of MDS, from patients with moderate anemia and mild dysplasia as their sole BM abnormality to those with severe cytopenia and excess blasts on BM analysis. As different physicians use different criteria for BM analysis this can create a selection bias among the cytopenic patients. Furthermore, some of the cytopenic patients did not go through BM analysis as their physicians decided it was not appropriate, which might lead to selection bias in cytopenic cases. A significantly larger study across diverse clinical attributes and locations is needed to overcome these biases.
Ethics oversight	All protocols were approved by the Weizmann Institute of Science ethics committee (under IRB protocol 283-1) or approved by the TASMC ethics committee (under IRB protocol 02-130).

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No calculations of sample size were made. Recruitment of the healthy cohort was intended to allow characterization of the normal variation in cHSPC states. As no such profiling had been previously performed, we could not assume much regarding the variance in the population or define sample size a-priori. We aimed to sample equally from different age groups and both genders.

Data exclusions	In the analysis resulting in Figures 1-3 and EDFs 1-6, We filtered out cells with at least 20% mitochondrial expression, then removed mitochondrial genes (as well as few other batch-prone genes), and further filtered cells with ≤ 500 UMIs (cell barcodes with less than 500 UMIs usually correspond to empty 10x droplets exposed to ambient RNA). In the analysis resulting in Figure 4 and EDFs 7-10, cells were also filtered if they were suspected to be contaminated by platelets, neutrophils or erythrocytes, as described in the Methods section. Few individuals were excluded from certain analyses due to lack of sufficient data. These instances are reported in all relevant figures, their accompanying legends and text.
Replication	We ran technical replicates on 39 individuals and biological replicates on a follow-up cohort of 20 individuals, sampled approximately one year following their original sampling date. These are discussed and analyzed in the text. Specifically, the 20 follow-up replicates exhibited stable CLP and MEBEMP frequencies (Fig 2C).
Randomization	Simultaneously processed human samples were randomly assigned to different pools in order to minimize batch effects and single cells were randomly captured prior to sequencing.
Blinding	Not relevant, as readouts were quantitative and not prone to subjective judgment.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	The EasySep Human CD34 Positive Selection Cocktail uses a class II anti-CD34 antibody clone.
Validation	Available from the manufacturer

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*