# Natural and age-related variation in circulating human hematopoietic stem cells

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

Affiliations

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. Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel.
5. Ultima Genomics, 7979 Gateway Blvd, Newark, CA 94560.
6. Integrated cancer prevention center, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel.
7. Clalit Health Services.
8. Hemato-Oncology Department, Assuta Medical Center, Tel Aviv, Israel.
9. Maccabi Healthcare Services.
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.

*Contributed equally to this work

#corresponding authors

**Abstract**

Hematopoietic stem and progenitor cells (HSPCs) are intended to deliver life-long, consistent output. However, with age, we observe changes in blood counts and clonal disorders. Better understanding of inter-individual variation in HSPC behavior is needed to understand the transition from health to age-related hematological disorders. Here we study 360K single circulating HSPCs (CD34+) from 99 healthy individuals together with clinical information and clonal hematopoiesis (CH) profiles to characterize population variability in hematopoiesis. Individuals with CH were linked with reduced frequencies of lymphocyte progenitors and higher RDW. We describe a Lamin-A transcriptional signature across the HSPC spectrum and show it is reduced in CH individuals. We define and estimate HSPC composition bias and an age-related increased S-phase gene signature and show how they form a heterogeneous and multifactorial aging trend in the blood. The new comprehensive model of normal HSPC variation will allow the study of stem cell-related disorders. As a proof of concept, we present methodologies for analyzing myeloid malignancies in comparison to our reference atlas. Together, our data and methodologies shed light on age-related changes in blood counts, CH and can be used to study stem cell-related disorders in the future.

**Introduction**

The basis for understanding and defining human pathophysiological states is a detailed description of inter-individual heterogeneity among healthy individuals. Variability between healthy humans is multifactorial and determined by the interaction between germline/somatic mutations and the environment. The identification of inter-individual changes in complete blood counts (CBC) in large cohorts of healthy individuals exposed different age-related deviations from the reference. Such studies uncovered age-related macrocytic anemia with increased RDW and a reduction in absolute lymphocyte counts[1]. The mechanisms responsible for both phenomena remain enigmatic. Another aspect of heterogeneity in the blood is the appearance of somatic mutations in hematopoietic stem and progenitor cells (HSPCs). All HSPCs acquire somatic mutations[2], however, certain mutations in leukemia-related genes, namely pre-leukemic mutations - pLMs[3,4], can lead to clonal expansion of HSPCs, a phenomenon termed clonal hematopoiesis (CH)[5,6]. While CH is quite common among the elderly[7], it remains poorly

understood why pLMs lead to clonal expansion, and how CH and other age-related blood phenomena are related to each other.

One of the major gaps for understanding these age-related phenomena in the blood is our insufficient knowledge of HSPC variability across healthy, age-diverse individuals. While the various HSPC subpopulations and their functions have been extensively studied, it remains poorly understood how these differ between individuals. Inter-individual heterogeneity in the frequency of CD34+ peripheral blood (PB) HSPCs has been reported in the past, and was linked to age, smoking, sex, and hereditary factors[8], as well as different pathological states[9]. Some studies analyzed HSPC heterogeneity in higher resolution, but their sample size was limited[10]. No study specifically determined the inter-individual heterogeneity in HSPC transcriptional programs in a large cohort of healthy individuals, and how these correlated with CBC, CH and age.

Such a reference map has not yet been described, as the tools to characterize transcriptional programs in HSPCs with minimal bias, and at single cell resolution, have just been recently developed. In addition, as most HSPCs reside within the bone marrow (BM), access to these cells, in particular from healthy donors, has been problematic. However, previous studies have demonstrated that most HSPC populations can be identified in the PB[11], including some based on scRNAseq analysis[12], and functional stem cells were identified in the PB of mice[13] and humans[12]. As the PB connects the BM to other extramedullary stem cell sites, it can be enriched in unique stem cell populations[9]. All this suggests that PB HSPCs can be a good surrogate for studying inter-individual HSPC transcriptional heterogeneity.

In the current study, we analyzed 99 healthy individuals across age (25-91 years), sex and somatic mutations by highly reproducible scRNAseq, and describe transcriptional programs of 360,000 cells and how they correlate with clinical attributes (all data can be observed in https://tanaylab.weizmann.ac.il/MCV/blood_aging/). We discovered rare circulating HLF/AVP positive hematopoietic stem cells (HSCs) known to have extensive self-renewal capacity and previously reported in the BM. We identified a T and dendritic cell progenitor population which does not decline with age. Inter-individual heterogeneity in the frequency of specific HSPCs and in their transcriptional programs were highly correlated with blood indices. Specifically, both a gene signature that includes Lamin-A (*LMNA)* and the frequency of lymphoid progenitors were correlated with CH. We discovered a complex set of interacting factors in blood aging. Finally, as proof of concept, we introduce novel methodologies for the analysis of Myelodysplastic Syndrome (MDS) and Acute Myeloid Leukemia (AML) cases in comparison to the normal reference

map provided. This study portrays the map of circulating human HSPCs and lays the foundations for understanding HSPC aging and related disorders.

## Results

### Universal stem and progenitor states observed across humans in CD34+ peripheral blood

To evaluate interpersonal diversity in the distributions and regulation of HSPCs from healthy humans, we combined multiplexed scRNAseq with genotyping, and integrated clinical data. We used multiplexing to reduce costs and batch bias, relying on common SNPs we identified in the 3' UTR of HSPC RNA and their targeted genotyping, for precise matching of cells to individuals (**Fig 1A**). This design was also instrumental in reducing doublet effects. Altogether, we collected HSPCs from 47 males and 52 females between the ages of 25 and 91 years (median 66), sequencing single cells through a standardized pipeline using 10X and Illumina sequencing (**EDF 1A**, **Supplementary Table S1**). We ran technical replicates on 11 individuals, and biological replicates on a follow-up cohort of 10 individuals, sampled one year following their original sampling date. Replicates were sequenced on an alternative platform (Ultima Genomics) to demonstrate the scalability of our approach. We collected longitudinal CBCs from all individuals up to 5 years prior to sampling, and performed deep targeted somatic mutation analysis on DNA produced from all blood samples, to identify cases of CH (**Supplementary Tables S2, S3**)[14]. Following quality control and filtering, we retained 360,000 single cell profiles with which we constructed a metacell manifold model[15], annotated using known markers (**EDF 1B,C**). From the 1435 metacells we derived, we filtered 251 as showing low *CD34* expression and a strong association with known features of B, NK, T, Monocyte and Dendritic cells (**EDF 1D**). The remaining metacells were visualized in 2D (**Fig 1B)**, showing a rich repertoire of states associated with circulating HSCs and their differentiation trajectories (**EDF 1C-F**). The derived model recapitulated and deepened previous observations from BM (**EDF 2**) and small samples of circulating HSPCs[12]. The model defines a distinct HSC state that is transcriptionally linked with two major differentiation gradients. The first one represents a continuum of common lymphoid progenitor (CLP) programs. The second, and more common differentiation branch, represents multipotent progenitor (MPP) states and their differentiation toward granulocyte-monocyte progenitors (GMP), erythrocyte progenitors (ERYP) and basophil/eosinophil/mast progenitors (BEMP).

Technical limitations of cell disassociation in scRNAseq prevented precise megakaryocyte program modeling (**EDF 2F**). We therefore annotated states at the base of this trajectory as megakaryocyte/erythrocyte/basophil/eosinophil/mast progenitors (MEBEM-P) as these are also presumed to be the cells of origin of megakaryocytes[16]. The depth of our HSPC sample allowed for detailed characterization of rare progenitor populations that were previously difficult to acquire and profile.

**High resolution circulating HSC map shows *HLF*, *GATA3*, *HOXB5* and *TLE4* as distinct HSC TFs**

Early HSCs are marked by high *AVP* and *HLF* expression and were shown by others to represent a rare cell population with self-renewal capacity in BM and cord blood[17]. Our model included data on ~4700 *HLF*/*AVP* HSCs that could be matched with cells from independent BM atlases[18], suggesting that under steady-state, HSCs with the highest self-renewal capacity constantly leave the BM (**EDF 3A**). Together with HLF and AVP, we discovered 26 genes expressed at least 1.75-fold higher in HSCs compared to their two immediate differentiation branches (**EDF 3B, Supplementary Table S4**). We specifically identified several transcription factors (TFs) enriched in HSCs, including the genes *HOXB5*, *TLE4* and, importantly, *GATA3* (**Fig 1C**). *GATA3* was previously reported to regulate self-renewal in mice long-term HSCs[19], yet its role in human HSCs has not been studied in depth thus far. We hypothesized that if *GATA3* is indeed an important HSC TF, it could be mutated in AML. We therefore screened for *GATA3* mutations in exome sequencing datasets of AML[20], and discovered a mutation hotspot at position R353K, which is part of the DNA binding domain, in ~1% of AML patients (**EDF 3C**).

We note that while the HSC state is defined by unique markers that are down-regulated in both the CLP and MEBEMP trajectories (symmetrically) upon exit from the HSC state (**Fig 1C**), it is also expressing a number of lineage-specific regulators at intermediate levels which are bifurcating anti-symmetrically to the CLP and MEBEMP lineages (**Fig 1D, EDF 3B**). These remarkable dynamics may suggest that the multipotent capacity of HSCs is correlated with intermediate expression of multiple regulators that is resolved with differentiation.

**NK-T-dendritic and basophil-eosinophil-mast progenitors are enriched in circulating HSPCs**

The circulating CD34+ atlas was enriched for basophil-eosinophil-mast progenitors (BEMP) that were mapped as one possible terminus of the HSC differentiation trajectories. While classical studies linked these cells with a granulocyte/monocyte progenitor (GMP) origin, more

recent studies suggested these to be emerging, at least in part, from erythroid progenitors in mice and humans[12,21]. Our analysis allowed us to zoom in on a small population of metacells linking BEMPs with their MEBEMP-L precursors (**Fig 1E**). This highlighted TFs (**Fig 1F**) and other factors (**EDF 4A**) that are regulated positively or negatively in this postulated early stage of BEMP specification. Another rare HSPC population we could zoom in on included lymphoid states with high *ACY3* expression and intermediate-to-low *DNTT* levels, a combination that could only be rarely found in the human BM, but is present in peripheral blood (**EDF 4B**). Interestingly, we observed co-variation of key T cell regulators within this population, but also anti-correlation of these factors with some hallmarks of a dendritic cell (DC) program. This can be demonstrated by comparison of *TCF7* and *IRF8* expression (**Fig 1G**), and the matching *TCF7*-coupled dynamics of *CD7*, *MAF*, and *IL7R*, or *IRF8*-coupled dynamics of the myeloid TF *SPI1* (PU.1) and multiple MHC-II genes (**Fig 1H and EDF 4C**). We therefore termed this subpopulation NK/T/DC progenitors (NKTDP). To summarize, our map of circulating HSPCs showed a rich spectrum of differentiation trajectories and progenitor states that refined previous analyses, and provided an opportunity for deciphering inter-individual hematopoietic variability.

**Inter-individual variation in HSPC stemness and in lymphoid/myeloid differentiation bias**

We found our circulating HSPC model to be consistent among individuals. The median number of individuals contributing cells to each metacell was 73, and all metacells included cells from at least 14 individuals. Individual-specific differential expression was limited after controlling for each sample's cell distribution over the atlas states (**EDF 5**). To study inter-individual HSPC variation we combined characterization of compositional state variation, with quantification of within-state differential expression. The compositional analysis is approached by computing the relative frequencies of cell states in the single-cell ensemble acquired for each individual **(Fig 2A)**. These frequencies are observed to vary extensively (**Fig 2B**). For example, HSCs are represented at 1.8% (SD 1.1%) of the CD34+ population, and CLP-Ms at 7.9% (SD 5.2%). The abundant MPP and MEBEMP states (mean frequency of 21.6% and 38%, respectively) showed smaller relative variation (SD 4.7% and 6.5%, respectively). Inter-individual correlation of cell state frequencies (**Fig 2C**) showed co-variation of lymphoid frequencies (CLP-M, CLP-L, NKTDP), and of advanced MEBEMP states (MEBEMP-L, ERYP, BEMP). Interestingly, the HSC state representation was positively correlated with the representation of the related (but already bifurcated) progenitor

states MPP and CLP-E, suggesting that for some individuals, the most potent HSPC states are over-represented compared to the average.

To analyze composition in higher resolution, we profiled each individual's enrichment over the entire MEBEMP and CLP differentiation gradients divided into 15 bins, clustering the resultant profiles over all individuals to derive six archetypes of HSPC composition across normal individuals (denoted classes I – VI) (**Fig 2D**). This showed groups of individuals with relative lymphoid enrichment (class I-II) or depletion (class V-VI) and within them a gradient of stemness enrichment (classes II, IV and VI) or depletion (class I, III and V). We observed the Ultima-sequenced data to be highly similar to the Illumina-sequenced data in our technical replicates (**EDF 6A**), and used it to validate the stability of cell type compositions in our follow-up cohort (**EDF 6B**). The discovery of systematic variation in the distribution of HSPC populations among healthy individuals, laid the grounds to study the impact of this variation on diverse clinical outcomes.

**Circulating HSPC frequencies correlate with CBCs and CH**

Analysis of CBC correlations with our single-cell map reinforced our previous finding of inter-individual HSPC composition variation. We observed a correlation between PB mature lymphocyte percentages and CLP frequencies (**Fig 2E**), consistent with a possible contribution of CLP production to the level of B-cells in healthy individuals. Higher PB monocyte percentages were similarly associated with lower CLP levels (**EDF 6C**). We detected a significant correlation between HSPC cell type distribution and HCT and RDW among males (**Fig 2E**). Specifically, CLP frequencies were negatively correlated with RDW, such that high RDW individuals demonstrated lower CLP frequencies. Female CBC parameters did not show a significant association with HSPC composition, most likely due to perimenopause effects. All CBC correlation analyses were performed using median values for each blood count parameter over 5 years preceding scRNAseq. The mean and median number of blood counts per individual during this 5 year period were 8.8, and 7 respectively.

Our previous work[22] and the work of others[23] correlated increased RDW values with high risk for CH and predisposition to AML. We demonstrate that low CLP frequencies are associated with CH (two-sided Mann-Whitney test; **Fig. 2F, EDF 6D**), and enhance our observation by performing Genotyping of Transcriptomes on one of our DNMT3A R882 cases[24], identifying a

lower fraction of CLP cells in the mutant clone (Fisher's exact test, **Fig. 2G**). To further explore this association, we studied a cohort of 18147 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) and 602 age and sex-matched normal RDW controls. We performed deep targeted sequencing to identify pLMs 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-value < 0.002, **Fig. 2H**, **Supplementary Tables S5, S6**). Altogether, the data demonstrate a 3-way linkage between decreased CLP frequencies, a high RDW, and CH.

**Inter-individual variation in HSPC Lamin-A signature is linked with CH.**

As shown above, an individual's HSPC composition provides an initial blueprint of hematopoietic dynamics along the stemness and CLP/MEBEMP axes. Further analysis of transcriptional variation can now be performed while fully controlling for such compositional effects, aiming to characterize additional individualized gene expression signatures and associate them with clinical parameters (**Fig 3A**). We systematically screened for such signatures by testing the inter-individual correlation of normalized gene expressions over the HSC-MEBEMP (**Fig 3B**) and the HSC-CLP gradients (**EDF 7A**). The most prominent of these signatures were sex related signatures, an S-phase signature (discussed later) and a *Lamin-A (LMNA)* signature, which included *ANXA1*, *AHNAK*, *MYADM, TSPAN2,* and *VIM,* among others (**EDF 7B**, **Supplementary Table S7**). While exhibiting a highly variable expression in HSCs and early myeloid and lymphoid cell states, the *LMNA* signature showed a more homogeneously low expression in late MEBEMPs and CLPs (**Fig 3C**). Individual MEBEMP *LMNA* signature expression varied across a range of more than 4-fold (**Fig 3D**, **EDF 7C**), and was stable in the follow-up cohort (**EDF 7D**). Independent quantification of *LMNA* signatures in CLPs and MEBEMPs showed a strong correlation (**Fig 3E**). Interestingly, high average *LMNA* signatures in MEBEMPs correlated with a skewed MEBEMP/CLP composition (**Fig 3F**). Moreover, individuals with CH showed low MEBEMP *LMNA* signatures (two-sided Mann-Whitney test, P-value < 0.05, **Fig 3G**). The association between CH and low *LMNA* signatures was also demonstrated within the single cell sample of individual #122, where *DNMT3A*-mutated cells (GoT[24]-based, n=78 out of 1031) showed lower *LMNA* signatures (two-sided Mann-Whitney test, **Fig 3H**). The weak anti-correlation of *LMNA* signatures and CLP frequencies (**Fig 3F**), standing in contrast to the negative association of both factors with CH, highlights the complexity of the CH phenotype. Taken together, using the defined inter-individual

HSPC compositional variation as background, we quantified an individualized *LMNA* gene signature in HSPCs, whose expression was low in individuals with CH.

**Rapid repression of stemness signatures in MEBEMPs is linked with lower red cell counts and higher red cell volumes**

The differentiation of HSPCs toward MEBEMP and CLP fates involves coordinated activation of specific transcriptional programs that were generally universal among individuals. Yet, our screen for individual-specific gene signatures suggested that some individuals up- or down-regulated these differentiation programs, even when controlling for compositional differences. This variation in balancing stemness and differentiation signatures could thus characterize individuals. We developed a novel synchronization score based on comparison of AVP-correlated genes (stemness) and *GATA1*-correlated genes (MEBEMP differentiation, **Supplementary Table S8**). We classified each MPP/MEBEMP cell according to how highly it expresses these two signatures, using 20 bins for each score. As expected, these signatures were anti-correlated. However, different individuals synchronized this anti-correlation differently (**Fig 3I**). While most individuals displayed dynamics close to the diagonal line (individuals #16, #86), some individuals deviated from it, indicating skewed synchronization between the *AVP* and *GATA1* signatures. To quantify the level of synchronization we examined cells with high *GATA1* signature, and computed the fraction of these cells that still express the *AVP* signature to a moderate degree, a quantity we termed the synchronization-score (sync-score). We observed individuals with sync-scores as low as 0.12 (e.g., #122 and #172, **Fig. 3I**, left), indicating a delayed rise in *GATA1* signature expression. Namely, while these individuals rapidly reduce their *AVP* expression, their increase in *GATA1* and *GATA1*-related genes is delayed. In contrast, other individuals exhibited a high sync-score (e.g., #98 and #121, **Fig. 3I**, right), suggesting a rapid rise in *GATA1* expression that precedes the decrease in *AVP* expression. We detected significant stability of the sync-score in our follow-up cohort (**EDF 7E**). Inter-individual sync score variability (**Fig 3J**) was positively correlated with RBC levels, and consistently anti-correlated with MCV in males (P-value for Spearman's rho equality to zero < 0.01 for both RBC and MCV; **Fig 3K**). Analysis of the correlation between individual sync-scores and HSPC compositions demonstrated a positive correlation with HSC frequencies and a negative correlation with ERYPs and BEMPs (**Fig 3L**).

To summarize, we demonstrated variation in the coordination of stemness and MEBEMP differentiation programs that is correlated with red blood cell counts and volumes. The possible impact of this signature on the regulation of efficient erythropoiesis should be further explored.

**Age-related perturbation of HSPC composition and transcriptional signatures**

Aging in the blood represents a complex and multi-factorial process that is likely driven by intrinsic hematopoietic effects (e.g., pre-malignant mutations) and extrinsic physiological effects (e.g., hormonal changes). We therefore anticipated multiple properties to define a multi-layered age-HSPC correlation. We first tested the association between HSPC compositions and age and did not observe an apparent directional increase or decrease in HSPC sub-types with aging (**EDF 8A**). We did demonstrate an increase in the variance of cell state frequencies, with a significantly higher variance above the age of 65 ($p < 0.01$). To quantify each individual's deviation from expected cell state frequencies, we computed an HSPC composition bias score, which significantly increased with age (**Fig 4A**, $p < 0.02$, test for Spearman's rho). This supported the notion of multiple age-related processes that perturb the highly homogeneous and robust HSPC landscape seen in young adults.

We used several HSPC signatures to further study inter-individual variation in aged hematopoiesis, including the *LMNA* and sync signatures described above, as well as an S-phase signature, quantifying expression of S-phase related cell-cycle genes (**Supplementary Table S9**), previously shown to have high inter-individual composition-normalized gene expression correlation (**Fig 3B**). The S-phase signature was robust in the follow-up cohort (**EDF 8B**), supporting its role in characterizing an individual quality rather than a transient effect. Circulating HSPCs did not generally express S-phase transcriptional signatures, in contrast to their bone-marrow counterparts (**Fig 4B**). However, weak, but significant, expression of DNA replication genes was observed in the late MEBEMP trajectory of some individuals, with a strong positive association with age (**Fig 4C,** $p < 0.04$, test for Spearman's rho). Comparison of S-phase signatures to HSPC composition bias scores suggested the two increased independently with age (**Fig 4D**). In contrast, increased HSPC bias scores could be associated with lower LMNA signatures (**Fig 4E**), strengthening the association between CH and low *LMNA* expression. Sync scores were not directly correlated with age (**Fig 4F**), despite their associations with RBC and MCV as described above. All individual scores and signatures, including *LMNA* signatures, S-

phase signatures, sync-scores, composition-bias scores, and CD34+ cell type distributions are summarized in **Supplementary Table S10**.

Case studies of individuals with highly abnormal HSPC distributions, and integration of these with clinical markers and mutation profiling illustrate the multi-modal nature of hematopoietic aging. Individual #151, an 80yo MDS-diagnosed male, defined by a *TET2*/*DNMT3A*/*CBL* clone with high variant allele frequency (VAF; *TET2* VAF=70%) and exhibiting high RDW anemia, shows extreme HSPC bias, a low LMNA signature and a high S-phase signature (**Fig 4G**). Individual #98, a 69yo male, represented another distinct behavior, with polycythemia, a high sync signature and high RDW. Taken together, the analysis of HSPC composition and transcriptional signatures (see additional screening for age-, CBC- and sex- associated gene expression in **EDF 8C-E**, **Supplementary Tables S11, S12**) provided insights to the various mechanisms that drive hematopoietic aging. In particular, our analysis separates the spectrum of effects associated with CH, from those associated with changes in HSPC regulation and differentiation. High resolution characterization of these effects enables the analysis of patients with blood malignancies at high molecular depth.

**Using the HSPC atlas for mapping, dissecting and annotating myeloid malignancies**

The current approach for diagnosing myeloid malignancies involves identifying clonal markers, such as mutations or structural variants, and characterizing blasts through microscopy and flow cytometry. We propose an alternative framework for analyzing leukemia cases using the normal reference HSPC atlas presented herein. In **Fig 5A** we describe a stepwise approach for leukemia analysis applied to two MDS and two AML cases. The first MDS case (#N249) carried an *SF3B1* Y623H mutation with 25.7% VAF. The second MDS case (#N48.1) was sampled twice during our study, initially showing an *SRSF2* P95L mutation with 13% VAF and no cytopenia, and later presenting with deteriorating blood counts and several additional mutations, including a frameshift mutation in *TET2* L1340 VAF=36.7%, *IDH2* R140Q VAF=7.3%, and four other truncating mutations in *TET2* with ~3% VAF (**Supplementary Table S3**). His *SRSF2* mutation was quite stable at 8.4% VAF. scRNAseq karyotyping did not identify any major copy number variations (CNVs) nor any population substructure for these two MDS cases (**EDF 9A**). Analysis of each individual MDS sample's transcriptional states through construction of a metacell model and projection onto the healthy reference atlas (**Fig 5B**, middle) showed overall similarity to the normal atlas states (**Fig 5C**). Projection of MDS cells to our 15 MEBEMP-CLP trajectory bins allowed us to identify

deviations from the normal differentiation route (**Fig 5D**). Both MDS-1 (#249) and MDS-2 (#48.1) belonged to the low-CLP high-stemness archetype.

We next studied two secondary (post-MDS) AML cases with no somatic mutations based on targeted sequencing. Clinical cytogenetics was uninformative for both cases. Projection of AML cells onto the healthy reference atlas showed high transcriptional differences (**Fig 5B**, right), but suggested that the tumor cells were most similar to cells in the HSC-MPP-CLP area (**Fig 5C,** right). scRNAseq-based karyotyping (**EDF 9B**) identified two clones in AML-1: a smaller clone (AML-1-1) with normal karyotype and a larger clone (AML-1-2) with +9,+10,+22 and del20 (**EDF 9C**). scRNAseq-based karyotyping of AML-2 (**EDF 9D**) identified +8,+11,+13,+14 in all metacells, with no population substructure (**EDF 9E**). We used normal gene signatures to identify subpopulations of AML cells with CLP, HSC and MEBEMP characteristics (**Fig 5E**). AML-2 was characterized by an early CLP signature, with a subset of the cells showing MPP/MEBEMP characteristics. In contrast, AML-1-1's transcriptional states were more balanced between cell types, including a subpopulation with a high HSC signature (**Fig 5E, EDF 10A**). AML-1-2's cells did not highly express any of the healthy signatures we tested, though few of his cells expressed MEBEMP or CLP signatures. While the AML cells showed variance in their expression of the atlas gene signatures, they differed greatly from healthy cells even in the expression of these genes (**EDF 10B-C**), including major differentiation regulators. The malignant state was characterized by multiple additional gene signatures described by de-novo identification of gene clusters over the AML-1 and AML-2 metacell models (**Fig 5F-G**). As an example, this analysis revealed overexpression of *BCL2* in the AML-1-2 clone compared to the AML-1-1 clone, suggesting a potential differential response to BCL2 inhibitor therapy. To conclude, the atlas of normal HSPC states presented herein enables characterization of AML cases, their subclonal structure and potential transcriptional dynamics, over skewed states that in some cases retain characteristics of normal HSPC differentiation programs.

## Discussion

We used scRNAseq of samples from 99 healthy individuals to facilitate high resolution analysis of interindividual heterogeneity in circulating HSPCs. This led to characterization of widespread variation in the composition of HSPCs over a spectrum spanning myeloid and

lymphocyte differentiation gradients. This data is forming a new basis for understanding how age-related hematological deterioration is transforming normal variation in HSPCs toward disease. It sheds light on three major questions underlying aging in the human blood system, including: a) the mechanistic basis for age-related macrocytic anemia and high RDW, b) the processes leading to decrease in lymphocyte counts with age, and c) the etiology of CH and its functional consequences on HSPC behavior.

We discovered that low hematocrit levels correlated with lower HSC/MPP frequency (**Fig 2**). Furthermore, MCV was correlated with synchronization in closing stem cell programs and opening erythroid programs (**Fig 3I-L**). Individuals which closed their stem cell program early and delayed erythroid programs (quantified using a novel transcriptional sync-score) had higher MCV and lower RBC. This suggests a role in age-related anemia. These findings should be further explored to better distinguish between inter-individual variation caused by genetic, epigenetic and external aging effects.

With regard to age related decline in lymphocytes, we found that it correlated with the frequency of HSPCs differentiating toward CLPs (**Fig 2E**). We could not observe a reciprocal age-related decline in HSPCs that further differentiate toward fates linked with T cells or DC-like programs (denoted NKTDP), as would be expected based on the paradigm of thymus involution. Our results stress the lymphoid-myeloid characteristics of NKTDP (**Fig 1G,H**), suggesting that they might be involved in dynamics of lymphocytic seeding in other tissues.

We discovered that low CLP levels (**Fig 2 F,G**), low *LMNA* transcriptional signature (**Fig 3 G,H**) and high RDW (**Fig 2H**) were all correlated with CH. Altogether the data exposed a novel 3-way linkage between decreased CLP frequency, high RDW, and CH among healthy individuals. It remains unclear whether this correlation is due to cell-intrinsic effects of preleukemic mutations on differentiation, or cell-extrinsic factors which induce all three conditions. Importantly, this phenomenon is unrelated to age-related anemia, and these two age-related phenotypes can occur independently, stressing yet again the complexity of the aging phenotype. The mechanisms leading to variance in the *LMNA* gene signature expression within and between individuals, and how such variance may be involved in regulating HSPC functions, are still poorly understood. It is clear that levels of *LMNA* control nuclear shape and rigidity[25]. For example, reduction in *LMNA* is needed for nuclear segmentation of polymorphonuclear cells[26]. Changes in the nuclear lamina and its connections with chromatin have been described in aged HSCs from rodents[27]. It remains

unclear whether different pLMs all lead to changes in the *LMNA* gene signature, or that the *LMNA* gene signature is related to preleukemic stem cell expansion indirectly. Future studies will resolve this correlation.

So far, our data exposed correlations with functional age (macrocytic anemia, lymphopenia and CH). We next searched for correlations with chronological age. We hypothesized that the ability to describe aging with the different attributes we discovered will better dissect aging to its various etiologies and quantify each one of them in each individual. The prominent age-related correlations we discovered were the cell frequency composition heterogeneity which increased with age and the appearance of S-phase gene signature in MEBEMP-L (**Fig4 A, C**). When we considered these two attributes, together with the functional age-related signatures (*LMNA*, Sync-score and RDW), we were able to better dissect each individual's aging etiologies and to expose different factors involved among different individuals (**Fig 4 D-G**).

Due to their ease of access, circulating CD34+ cells could facilitate disease diagnosis and monitoring. However, the connection between circulating HSPCs and HSPCs residing in the bone marrow and secondary immune organs is still largely unknown. The associations we find between circulating CD34+ cells and clinical labels such as CBCs suggest that they provide an at least partial picture of BM HSPCs. But additional factors such as dynamics of BM ingress, egress, cell proliferation, cell death, and the interaction of these factors with different progenitor cell types, can all influence the snapshot provided by circulating CD34+. These factors should be studied to understand the advantages and limitations of PBMC profiling.

As discussed above, the description of normal HSPC variation promotes the understanding of the mechanisms leading to age-related hematological deterioration. However, this new model can also be applied immediately as a novel platform for the analysis of PB CD34+ cells from patients with uncharacterized disease (**Fig 5A**). We give an example of how quantitative scRNAseq analysis of leukemia samples in reference to our normal map can facilitate and enhance various steps of leukemia diagnosis and classification. We suggest that this can be readily extended to a large cohort of stem cell-related diseases such as myeloid malignancies, but possibly also to other clinical scenarios. We hypothesize that this can become the method of choice for diagnosis and classification of MDS and AML, and that similar maps of other tissues from large cohorts will facilitate the next generation of molecular medicine.

## Methods

### Sample procurement and handling

During the period from Dec. 2020 to Apr. 2021, we collected fresh peripheral blood samples from 99 healthy individuals (47 males, 52 females) aged 25-91. Their demographics and molecular data are presented in **Supplementary Tables S1 and S3** [Basic participant demographics and CH status]. All sample donors were considered healthy, their CBCs were within normal range, and they were not known to have any CH defining mutations prior to sequencing. Written informed consent allowing access to longitudinal CBCs 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 Weizmann Institute of Science ethics committee (under IRB protocol 283-1).

50 ml of PB were drawn from each individual into lithium-heparin tubes. 1 ml of blood was used for DNA production, and 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 positive selection kit II (StemCell technologies). We found this enrichment strategy to be simple and reproducible and chose it for several reasons: 1) RNAseq data was most reproducible when cells were not sorted, but rather enriched-for using beads (lower mitochondrial gene fraction). 2) CD34 purity could be highly regulated by this method, to achieve anywhere between 50-95% enrichment of CD34-positive cells, which could later be easily distinguished based on their single cell expression data. In terms of cell numbers - 50 ml of blood would yield anywhere between 50 to 100 million PBMCs following Ficoll, 1/1000 of which are expected to be CD34+, such that we increased this population's representation from 0.1% in the periphery to at least 50% of cells loaded for analysis.

### ScRNAseq of CD34+ PBMCs

Single cell RNA libraries were generated using the 10x genomics scRNAseq platform (Chromium Next Gem single cell 3' reagent kit V3.1). 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 freshly drawn at the Weizmann Institute of Science on the morning of each experiment day, and time from blood draw to 10x loading was restricted to 5 hours. 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.

All 10x libraries were pooled and sequenced on the NovaSeq6000 platform using a single S2100 kit, and all data was analyzed using the Metacell2 R package [15].

**Genotype-based demultiplexing**

All cells were traced back to their sample of origin using genotype-based de-multiplexing. This method allowed pooling of blood samples immediately following extraction of the DNA aliquot, such that CD34-enrichment was performed on the entire pool of PBMCs produced. The use of SNP-based multiplexing has several advantages to alternative antibody-based cell hashing methods: 1) it is extremely cost effective, such that the cost of sequencing a single individual on a 2000 SNP Molecular Inversion Probe (MIP) panel at a depth of 1000X per SNP (adequate for de-multiplexing purposes) is several folds cheaper than antibody staining, 2) genotyping eliminates the need to keep samples separated prior to loading, it entails shorter handling times and less cell manipulation, as it does not require antibody incubation periods and multiple wash centrifuges. This was very evident in cell viability prior to chip loading. As with other methods of sample multiplexing, genotype-based multiplexing allows for robust doublet detection during data analysis, which enabled loading of 30-40K cells from between 4-6 individuals on each Chromium Chip lane, yielding 15-25k cells per library.

**Molecular Inversion Probe (MIP) panels**

Both our CH and genotyping panels are Molecular inversion probes (MIP)-based panels described in detail previously[14]. Our CH panel contains 705 probes, covering pre-leukemic SNVs and Indels in 47 genes, and is complemented by 2 amplicon sequencing reactions to cover GC rich regions in *SRSF2* and *ASXL1*. Our genotyping panel allows for the simultaneous detection of >2000 common genetic variants, all of which are extensively covered in all cell types in our data. It includes heterozygous sites with at least 5% minor allele frequency from the 1K genomes project, which were highly covered by RNA molecules in our data (at least 80 UMIs across all cells in a test 10x library), excluding sites in repetitive elements and in sex chromosomes. Both panels were designed using MIPgen[28] to ensure capture uniformity and specificity.

**Variant calling and identification of ARCH mutations**

As MIP sequencing is cost-effective yet noisy, we developed an in-house variant-calling method to identify low VAF CH events [14].

**ARCH sequencing of high RDW samples and controls**

In order to compare propensity for CH and high risk CH mutations[22] in high RDW cases and normal RDW controls, we performed deep targeted sequencing of DNA samples from 602 high RDW (>15%) individuals, who did not show signs of anemia and whose blood count did not meet MDS criteria (11.5g/dL≤Hg≤15.5g/dL [F], 13g/dL ≤Hg≤17g/dL [M], 80fL≤MCV≤96fL, PLT≥100×$10^9$/L, Abs Neut≥1.8×$10^9$/L), and 602 normal RDW (11.5g/dL ≤Hg≤15.5g/dL [F], 13g/dL≤Hg≤17g/dL[M], 80fL≤MCV≤96fL, PLT≥100×$10^9$/L, Abs Neut≥1.8×$10^9$/L), age and gender-matched controls. Case-Control matching was performed using the R MatchIt package, balanced on age and gender, method = "nearest", ratio = 1, from a total of 18,147 individuals with longitudinal blood counts and available DNA. All DNA samples and corresponding blood counts were received de-identified from the Tel Aviv Sourasky Medical Center (TASMC) Integrative Cancer Prevention Clinic. All DNA samples were collected after obtaining written informed consent and in accordance with the Declaration of Helsinki. All relevant ethical regulations were followed, and all protocols were approved by the TASMC ethics committee (under IRB protocol 02-130). CBCs and sequencing results of all cases and controls are presented in **Supplementary Tables S5, S6.**

**scRNAseq processing**

We processed fastq files by executing cellranger-3.1.0 with an hg-38 reference genome. We filtered cells with at least 20% mitochondrial expression and ≤ 500 UMIs from unfiltered genes.

**Doublet calling**

We performed several steps to assign cells to their individuals and to detect doublets. The pipeline is made of several steps:

1. Demultiplexing cells and calling doublets based on SNPs found in the scRNAseq data
2. Detecting doublets based on cell expression profiles
3. Building a metacell model using cells from all the libraries, including cells previously marked as doublets, and identifying metacells made of doublet cells.

In the first step, we identify doublets and assign cells to individuals using Vireo and Souporcell, which cluster cells based on SNPs found in sequenced RNA molecules. We executed Vireo[29]

(preceded by running cellsnp) and Souporcell[30] on each library separately. Both methods used SNPs from our genotyping panel[14] 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.

We next identified doublet cells based on gene expression. We executed Scrublet[31] and DoubletFinder[32] on each library separately. Both of these methods require a threshold on their output scores for doublet calling, and we set different thresholds for different libraries. We considered the Vireo doublet calls as ground truth, and set the doublet thresholds, as well as the need to be called as doublet by both Scrublet[31] and DoubletFinder[32], to achieve high precision and recall in doublet calling for each library.

In the next step, we built a metacell model with cells from all libraries, including those identified as doublets by either their SNPs or expression. The model was built with metacell2, with a target metacell size of 500K UMIs. We then marked all metacells where at least 40% of the cells were already marked as doublets, and all metacells that expressed unique markers of distinct cell types, as doublet metacells. All cells that belonged to a doublet metacell were then marked as doublets.

**Assignment of cells to individuals**

Vireo[29] and Souporcell[30] both cluster cells based on SNPs found in the sequenced RNA, such that cells in the same cluster belong to the same individual. We observed very high agreement between the two methods in their assignment of cells into individuals. In two 10x libraries where the two methods did not agree (due to individuals with a very small number of cells), we reran the methods on a subset of the cells and a smaller target number of clusters. In all libraries we took Vireo's clustering, except for one library where we took Souporcell's, because of better matching to the genotype data (described below). We marked cells that were not clustered by Vireo as 'unassigned', even if they were assigned by Souporcell.

In the next step we assigned clusters of cells to the individual they originated from. To this end, we correlated the genotypes of each cell cluster, as inferred by Vireo, to all genotypes we measured using the MIP panel (using sites with sufficient sequencing depth in the MIP panel). As a control, we performed matching against the MIP genotypes of all individuals in the cohort, and not just individuals from one library. We observed in all cases a clear matching to a single

individual from the expected library. The assignment also correctly identified related individuals, and the sex of the matched individual was confirmed by expression of XIST in the RNA data.

**Metacell model**

We next built a second metacell model with the cells that were not marked as doublets, excluding droplets with complete or partial megakaryocyte expression (those in a metacell with PF4 expression > 2^-11.5 in the previous model) due to their overall high doublet rate. The model was built with metacell2, with a target metacell size of 500K UMIs. We marked forbidden genes such as histone genes, cell cycle related genes, ribosomal genes, stress response genes (e.g. FOS, JUN) and other genes that we found to have high technical variation. These genes were not used by metacell2 when calculating gene-gene similarities, but were included in downstream analysis. We annotated the metacells using known markers as illustrated in EDF 1C. We excluded from downstream analyses metacells from cell types with low CD34 expression (monocytes, B cells, T cells, NK cells, DCs), and one metacell expressing endothelial markers.

**BM projections**

We used two BM datasets: the Human Cell Atlas (HCA) dataset[18] and a CD34+ bead-enriched BM datasets from[33]. We have previously processed and annotated the HCA datasets in metacell. We downloaded the Setty et al. sequencing data and processed it by running cellranger and creating a metacell model. To project both our PB data and the Setty dataset on the HCA dataset, we correlated between the HCA metacells and the Setty and PB metacells over genes showing high variance over the HCA metacell model. We annotated each Setty metacell using the mode of the 5 most correlated HCA metacells. To plot Setty and PB metacells on the HCA's UMAP projection, we located each metacell on the mean x and y values of its 5 most correlated HCA metacells. To compare S-phase genes between the PB and BM (Fig 4B), we calculated for each PB metacell its S-phase signature (described in a separate section), and the mean S-phase signature for the 5 HCA metacells most correlated to it.

**HSC differentiation gene programs**

To visualize transcriptional dynamics in HSC cells, we sorted MEBEMP and CLP metacells based on their AVP expression. To calculate differential expression between HSC and neighboring cell types (EDF 3B), we calculated the geometric mean of each gene across HSCs, CLP-E and MPP metacells, and took the difference between HSC and MPP, and between HSC and CLP-E.

**Differential expression between individuals unexplained by the metacell model**

To create EDF 5, we compared each individual's pooled expression profile to a matched expression profile based on the individual's distribution across metacells. We performed the analysis separately for MPP / MEBEMPs (BEMP, EP, MEBEMP-E/L, GMP-E and MPP) and CLPs (CLP-E/M/L, NKTDP). In each group of cell types, we summed all the UMIs of each individual, normalized the sum to 1 and calculated log2, to obtain the observed expression. To compute matched expression, instead of summing over an individual's cells' expression profile, we summed all UMIs of the metacell each cell belongs to, and divided by the number of cells in that metacell. This way the UMIs in each metacell were equally divided between all the cells that belonged to that metacell. We normalized this matched expression to sum to 1, and took log2. For EDF 5 we plotted all genes that were expressed in either the observed or matched expression in any individual (log2 expression > 2^-14.5), that had at least 2-fold change between observed and matched in at least one individual, and that were not exhibiting strong batch effects (Kruskal-Wallis p-value < 1e-4, where individuals are grouped by their 10x library).

**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 types. To perform compositional analysis at higher resolution than cell types, we partitioned cells from CD34+ cell types into finer grained bins. We used one HSC bin, four CLP bins, and ten MEBEMP / MPP bins, for a total of 15 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 decreasing AVP expression of their metacells, such that bins 2-4 had the same number of cells. We similarly assigned MPP and MEBEMP-E/L cells into 10 bins based on AVP such that these bins had an equal number of cells.

For Figure 2D, we calculated the enrichment of each individual's cells in each bin (log2 of the ratio compared to the median across individuals). We partitioned individuals into three group with different CLP numbers based on each individual's 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 / MEBEMP bins 1-5 and the total MPP / MEBEMP number (cells in bins 1-10). Individuals with stemness score > 0.5 had enriched stemness. The combinations of CLP enrichment and

stemness define the six classes shown in the figure. For visualization we further sorted inidividuals within each cluster based on their stemness score.

**Test for association between cell type distribution and a numerical label**

We used permutation tests to test the relation between cell type distribution and a label (age, CBC, sync-score or *LMNA* signature). We sorted 11 CD34+ cell types from late MEBEMP differentiation through HSC and to late CLP differentiation (cell types are displayed by this order in Fig 2B). We looked at triplets of adjacent cell types in this ordering, and calculated for each triplet the total frequency each individual has from these cell types, obtaining a vector of length 9 per individual. We then correlated each of these 9 sums to the label, and took the maximal absolute value from all these correlation values as a test statistic. We repeated this process after permuting the label 10000 times, and used the test statistics from the permutations to derive a p-value.

**Variant gene modules**

We detected genes modules with high variance across individuals in MPP / MEBEMP and CLP cells separately. For MPP / MEBEMP, we performed the following steps:

A. we pooled all cells for each individual from the HSCs, MPP and MEBEMP-E/L metacells, normalized to sum to 1 and took log2. This gave us the observed expression profile of each individual across the MEBEMPs.

B. We created an expected expression profile for each individual as follows. We partitioned the MEBEMP metacells into 30 bins based on their AVP expression, and calculated for all genes the geometric mean expression across all metacells in each bin. This defined an expression profile for each of the 30 bins. To obtain an individual's expected expression, we calculated a weighted mean of the 30 bins' expression profiles, where the weight of each bin is proportional to the fraction of the individual's cells from that bin. We then calculated the difference between the observed and expected expression profiles.

C. We screened for genes with high variance. We removed genes with high batch effects (Kruskal-wallis p-value < 1e-3 when using an individual's 10x batch as a covariate), and genes with high AVP correlation (absolute value Pearson correlation > 0.65). We then calculated each gene's variance in the difference between the observed and expected expression across

individuals. As some of the variance can be explained due to sampling noise, we plotted each gene's variance across individuals compared to its mean geometric expression across all metacells from which the individual's cells were taken. 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.

D. We calculated a gene-gene Spearman correlation matrix for high variance genes, and clustered the correlation profiles using hierarchical clustering. We removed gene clusters with low mean correlation between their genes (< 0.2 mean correlation of all gene pairs), and genes with low mean correlation (< 0.2) to their cluster's genes. We additionally removed one gene module involving PCDH9 and CHRM3, as it represented residual MEBEMP transcription program that could not be fully normalized by our binning and pooling approach. This resulted in Fig 3B.

We performed a similar analysis for CLPs, with a few differences. The analysis included cells from CLP-E/M/L metacells. The cells were partitioned into 6 bins, and the partitioning was based on the average of their DNTT and VPREB1 expression. Genes with high absolute correlation to the average of DNTT and VPREB1 were excluded. After clustering the gene-gene correlation profiles, gene clusters with mean correlation < 0.3 were removed, and gene clusters with remaining correlation to CLP differentiation were removed.

**LMNA and S-phase signatures**

We partitioned the MPP / MEBEMP cells into 10 bins based on the AVP expression of their metacells as described previously. We then pooled for each individual the UMIs from all its cells in each of the 10 bin and obtained a gene by individual matrix per bin. We normalized the sum of UMIs from each individual to 1, took log2, and calculated the mean of the following genes in each bin: LMNA, AHNAK, MYADM, TSPAN2, ANXA1 and ANXA2. This gave a LMNA signature per individual per bin, as visualized in EDF 7C. An individual's LMNA signature is the mean of the individual's signature across the 10 bins. The CLP LMNA score (Fig 3E) was calculated in the same manner, but using CLP-M cells and only one bin.

We similarly defined the S-phase signature. We used the following genes: CLSPN, PCLAF, TYMS, H2AFZ, PCNA, TUBA1B, MCM4, HELLS to calculate an S-phase signature per individual in each

bin, and took the individual S-phase score to be the mean score across bins 6-10 (later stages of MEBEMP differentiation).

**GoT Analysis**

GoT[24] performed on sample #122 allowed us to mark this individual's cells as wild-type or mutated. Due to the low VAF of #122's DNMT3A mutation, and in order to increase power, we marked cells whose DNMT3A mutation status could not be determined by GoT as wild-type cells. For Figure 2G, we examined #122's cells' distribution across cell types.
We compared the LMNA signature between mutant and wild-type cells, while normalizing for the distribution across MEBEMP differentiation stages as follows. We sorted MPP and MEBEMP-E/L metacells based on their AVP expression, and reduced from each gene in each metacell the gene's rolling mean expression across the 30 nearest metacells in the ordering. These calculations were performed in log2 scale. The mean of the LMNA signature genes was then defined as the metacell's LMNA signature, and a cell's signature is the signature of the metacell it belongs to.

**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 correlation > 0.7 to GATA1. We filtered genes with mean relative expression > 2 ^ -10 in these metacells, to preclude a small number of genes from dominating the signatures. We then scored each cell by the fraction of its UMIs from the AVP and GATA1 signatures, and partitioned all cells into 20 bins of AVP signature expression and 20 bins of GATA1 signature expression, such that all AVP bins and all GATA1 bins had the same number of cells. The sync-score is then defined as the fraction of cells in GATA1 bin 13 and above (upper two quintiles of GATA1) that are in AVP bin 9 and above (upper three quintiles of AVP expression).
To visualize the sync scores, we normalized the 20 bins x 20 bins matrix to sum to 1, smoothed the obtained matrix by averaging cells using a running window of length 3, and taking log2.

**Ultima data processing for technical and biological replication**

We processed the ultima data using cellranger as we previously described for the Illumina sequenced data. We used the technical replicates to assess the gene expression technical variation, and found minor differences (**EDF 6A**). We marked a total of 210 genes with at least 2-

fold difference between Illumina and Ultima, or whose log2(Ultima / Illumina) expression had a range higher than 0.5 across the six technical replicates, as technology-dependent. We note that 98% of the genes showing high variance in the PBMC model were consistent between the sequencing platforms.

We assigned cells to individuals and detected doublets as described previously, but detected expression-based doublets only by building a metacell model and finding doublet metacells. We then built an integrated model with cells from both Illumina and Ultima libraries. The integrated model contained only cells from individuals for which both Illumina and Ultima data was available, and included both technical and biological replicates. When building the integrated model, we did not include technology-dependent genes as features, in addition to the genes we excluded previously while building the 360K cells' model.

We validated that in the integrated model, metacells included cells from both sequencing technologies. We then annotated each metacell using our reference 360K metacell model, by annotating each metacell with the annotation of its most highly correlated reference metacell, where the correlation is across the metacell's model highly variant genes. We used the annotations to calculate the cell type frequencies for all individuals in the integrated model, and binned cells from the integrated model into 15 bins as described previously for the 360K cells' model. We then calculated each individual's LMNA and S-phase signatures as described for the 360K metacell model.

The sync-score, unlike other scores, is based on calculation at the single cell level and without cell pooling. It is therefore more difficult to correct for technological variance. We calculated the sync-score as described previously for the 360K cells model, but with several modifications. First, we excluded technology-dependent genes from the AVP and GATA1 gene signatures. Second, we partitioned Illumina and Ultima cells separately into 20 bins based on the AVP and GATA1 signatures. Third, for the cells sequenced by Ultima, before we summed the UMIs from genes in the AVP and GATA1 signatures, we first multiplied each gene by a technology correction factor we derived from the technical replicate 10x library.

**Cell type variance and composition bias**

To test for increased cell types variance in aging, we downsampled the number of cells from CD34+ cell types per individual to 500, and calculated each individual's distribution across cell

types. We then transformed the values into z-scores by subtracting the mean frequency of each cell type and dividing by the frequency's standard deviation. The obtained z-score matrix of individuals by cell types was then given as input to a permutation test. Individuals were partitioned to those at age 65 and above, or below 65. In each age group, the mean z-score per cell type was subtracted from the z-score vector of each individual. These values were squared, summed across all cell type in an individual, averaged across individuals, and the root of the average was taken. The difference between the root in the old and young groups was taken as a test statistic, and was used to derive a p-value across 10000 permutations of the ages of the individuals.

The composition bias of an individual was defined as the sum of the absolute values of the individual's z-scores across all CD34+ cell types.

**Differential gene expression with respect to age and CBC**

Differential expression was performed separately for MPP / MEBEMP cells, and for CLPs. The MEBEMP and CLP matrices that were normalized for the differentiation distribution, and which were used to detect variant gene modules, were here used for differential expression. Differential expression was performed separately for males and females. Each gene's expression value was correlated with age and CBC using Spearman correlation, and the correlation was tested for significance. p-values were FDR-corrected (Benjamini-Hochberg) for each label separately. Differential expression between sexes was done using Wilcoxon test on the same expression matrices.

**MDS and AML scRNA-seq initial processing**

We processed additional 10x libraries, some of which were sequenced by Illumina and some by Ultima, using cellranger as described previously. We detected doublets using only Vireo and Souporcell[30], and assigned cells into individuals as we described above. We then created a metacell model for each of 6 individuals separately: 2 healthy individuals, 2 MDS patients and 2 AML patients. As before, we excluded cells with less than 500 UMIs, with more than 20% expression of mitochondrial genes, or with high expression of megakaryocyte genes. We ignored ribosomal genes and genes that are high in megakaryocytes while building the metacell model, and in two individuals removed megakaryocyte genes altogether from the expression matrix due to high ambience levels of these genes. We used a target metacell size of 75K UMIs.

**Projection of disease data on the HSPC model**

To project each individual's metacells on the healthy reference, we correlated the query metacells with the reference's metacells, using 366 highly variable genes in the CD34+ metacells of the reference, and after excluding genes upregulated in MK cells and technology-dependent genes. The correlation was performed in log2 scale, and when projecting an ultima dataset, a normalizing factor was added to each gene based on its differential expression in the technical replicates. Metacells that mapped to CD34- reference metacells were then discarded for the following analyses.

Fig 5B shows the distribution of the number of differentially expressed genes between query metacells and their most highly correlated reference metacell. The genes included in the count are only those with expression at least 2 ^ -13 in either the query or atlas metacell, and with at least 2-fold difference. We further ignore genes high in MK cells, ribosomal genes, sex-related genes, and genes that we found to have high batch effects between 10x libraries in the reference metacell model.

Fig 5C measures the mean correlation between each query metacell and its 5 most correlated reference metacells, where the correlation is across the genes used for the projection. For Fig 5D we projected single cells, rather than metacells, from the query. We projected each cell to its most correlated reference metacell, where the correlation used the raw UMI counts (and was not in log2 scale), and used genes with high variance in the reference. Each query cell was classified to the bin that was most common among cells in the metacell to which it mapped.

**Karyotype analysis**

To perform karyotype analysis, we calculated the log2 total expression (sum of UMIs) from each autosomal chromosome in each query metacell, and subtracted from it the log2 of the geometric mean of the total expression from each chromosome across the 5 most correlated reference metacells. The total expression didn't include expression UMIs from genes high in MKs, sex-related genes, genes with high batch effects in the reference, ribosomal genes and technology-dependent genes. A similar calculation was performed for the EDF 9 right side figures, but the expression of each gene was measured across all query metacells, and all the

reference metacells to which they were projected. Only genes that were expressed in either the query or reference metacells the query was mapped to (expression > 2 ^ -15.5) were plotted.

**Profiling signatures in disease cases**

We separated AML-1 metacells into AML-1-1 and AML-1-2 by their expression of BCL2 and ROCK1, which were both higher in AML-1-2. To search for variance in the AML samples in gene programs from the healthy reference, we created gene lists as follows:

- NKTDP program – genes with least 1.5 higher expression (in log2 scale) in NKTDP metacells compared to both CLP-M and CLP-L.
- CLP program – genes with at least 1.5 higher expression (in log2 scale) in CLP-M metacells, compared to all the following populations: NKTDP, HSC, MPP, MEBEMP-E/L, BEMP and ERYP.
- HSC program – genes with at least 1 higher expression (in log2 scale, that is 2-fold difference) in HSCs compared to: NKTDP, CLP-M, MEBEMP-E/L, BEMP and ERYP.
- MEBEMP program – genes highly correlated to GATA1 (the same genes that were used in the sync-score calculation).

For the gene list selection, the expression of a gene in a cell type is the geometric mean of its expression in all metacells that belong to that type. We scored each AML metacell by the geometric mean of all genes in each gene list. We set thresholds for a metacell to express a particular gene program as the 25$^{th}$ percentile across reference metacells in the relevant cell population (e.g. NKTDP metacells for the NKTDP gene program, see dashed lines in Fig 5E).

For the heatmap in EDF 10B, we selected genes from the above gene programs, as well as genes high in AML-1, high in AML-2, and high in AML-1-2 compared to AML-1-1.

To select genes high in AML-1, we looked at the annotation each AML-1 metacell received from its projection on the healthy reference. We calculated the mean expression of each gene across all metacells that were projected to the same cell type, and the mean expression using the reference metacells that the AML metacells were projected onto. We then selected genes

higher in AML-1 (at least 1.5 higher in log2) than in the reference across all cell types. A similar gene selection was performed for AML2.

To select AML-1-2 specific genes, we compared gene expression between metacells from AML-1-1 and AML-1-2 that were mapped to the same reference cell type. We selected only genes with higher AML-1-2 expression compared to AML-1-1 expression (1.5 in log2) in all of the following three cell types: CLP-E, MPP and MEBEMP-E.

To discover de-novo gene programs in the AML samples, we selected genes that the metacell algorithm identified as having high variance in the AML metacell models, calculated their correlation across metacells, and clustered their correlation profiles.

**Data Availability**

All data can be explored in: https://tanaylab.weizmann.ac.il/MCV/blood_aging/.

**Code Availability**

Detailed code for the figures will be provided prior to publication.

The Metacell R package is available at https://github.com/tanaylab/metacell

**Author Information**

These authors contributed equally to this work: Nili Furer, Nimrod Rappoport

These authors jointly supervised this work: Liran Shlush, Amos Tanay

Authors and Affiliations

**Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel**

Nili Furer, Nimrod Rappoport, Adi Danin, Nathali Kaushansky, Noa Chapal-Ilani, Amos Tanay
Liran Shlush

**Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel**
Nimrod Rappoport, Ron Shamir

**Department of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot, Israel**
Nimrod Rappoport, Aviezer Lifshitz, Akhiad Bercovich, Oren Ben-Kiki, Amos Tanay

**Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel**
Merav Kedmi

**Ultima Genomics, 7979 Gateway Blvd, Newark, CA 94560**
Zohar Shipony, Doron Lipson, Eti Meiri, Gila Yanai

**Integrated Cancer Prevention Center, Tel Aviv Sourasky Medical Center, Tel Aviv, Israel**
Shiran Shapira, Nadir Arber

**Clalit Health Services**
Berdichevsky Svetlana

**Hemato-Oncology Department, Assuta Medical Center, Tel Aviv, Israel**
Sigal Tavor

**Maccabi Healthcare Services**
Sigal Tavor, Liran Shlush

**Department of Cell, Developmental and Cancer Biology, Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA**
Jeffrey Tyner, Sunil Joshi

**Division of Hematology and Medical Oncology, Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA**

Jeffrey Tyner, Sunil Joshi

**New York Genome Center, New York, NY, USA**

Dan Landau, Neville Dusaj, Saravanan Ganesan, Paulina Chamely

**Sandra and Edward Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA**

Dan Landau, Neville Dusaj, Saravanan Ganesan, Paulina Chamely

**Division of Hematology and Medical Oncology, Department of Medicine, Weill Cornell Medicine, New York, NY, USA**

Dan Landau, Neville Dusaj,Saravanan Ganesan, Paulina Chamely

**Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY, USA**

Dan Landau, Neville Dusaj, Saravanan Ganesan, Paulina Chamely

**Division of Hematology, Rambam Healthcare Campus, Haifa, Israel**

Liran Shlush

**Contributions**

Patient recruitment was performed by NF, LS, SB, and ST. Blood drawing and sample preparation for scRNAseq, including CD34 enrichment, were performed by NF. 10x scRNA library preparation and sequencing on the Illumina platform were performed by NF. scRNAseq data processing, including doublet calling, assignment of cells to individuals and metacell model construction were performed by NR. BM projections and comparisons were performed by NR. All compositional, differential gene expression and variant gene module analyses were performed by NR. Clinical data curation was performed by NF. Clinical data association analyses were performed by NR. MDS and AML patient –derived metacell projections and analyses were performed by NR. MCV construction was performed by AL. ARCH and genotyping deep targeted

sequencing of all study participants were performed by NF. Amplicon validation of ARCH mutations was performed by NF. Sequencing data analysis and variant calling were performed by NC, NF and LS. DNA samples for the high-RDW-CH analysis were provided by SS and NA. ARCH sequencing of all high-RDW and control DNA samples was performed by NF and analyzed by NC, NF and LS. Replicate 10x library sequencing on the Ultima genomics platform was performed by ZS, DL, EM and GY. Biological and technical replicate analysis was performed by NR. GoT experiments were performed by SG, analyzed by ND and PC, supervised by DL and analyzed by NR. Gata3 deep targeted sequencing and Sanger validation were performed by SS and supervised by JT. AB, AL, and OBK contributed to data analysis and interpretation. AD provided sample preparation support. MK provided 10x guidance and technical support. NK contributed to funding applications. NF and NR wrote the manuscript. LS and AT designed and supervised all aspects of the study and wrote the manuscript.

**Corresponding authors**

Correspondence to Liran Shlush and Amos Tanay

**Ethics declarations**

Competing interests

ZS, DL, EM and GY are all employees and shareholders of Ultima Genomics.

LS is a shareholder of Sequentify.

The remaining authors declare no competing interests.

**Acknowledgements**

This work was supported by the following grants: LLS and Rising Tide Foundation Grant ID: RTF6005-19, ISF-NSFC 2427/18, ISF-IPMP-Israel Precision Medicine Program 3165/19, ISF 1123/21,and the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine. LS is an incumbent of the Ruth and Louis Leland career development chair. 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 Rising Tide Foundation and the Applebaum Foundation. The contribution of NR is part of a Ph.D. thesis research conducted at Tel Aviv University. NR was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics, Tel Aviv University, and by the Planning and Budgeting Committee (PBC) fellowship for excellent PhD students in Data Sciences. NR was also supported by awards from the Herczeg Institute on Aging, and from the Tel Aviv University Healthy Longevity Research Center.

**Supplementary Information**

**Supplementary Table S1 – Basic participant demographics**. This table contains data on sex and age of all studied participants at sampling time, along with their allocated study IDs.

**Supplementary Table S2 – Longitudinal CBC's**. This table contains median values for each CBC parameter for 5 years prior to sampling.

**Supplementary Table S3 – ARCH sequencing**. This table contains variant calling data on all ARCH positive individuals, including chromosome number, position, reference and alternate nucleotides, the gene involved, whether or not this alternation is considered a leukemic hotspot, and the reported average VAF from single (Amplicon) or duplicate (MIP) sequencing instances.

**Supplementary Table S4 – HSC expression differences**. This table contains each gene's mean expression across metacells in HSCs, MPPs, CLP-Es, and the differences in expression between HSCs and the latter two cell types.

**Supplementary Table S5 – high RDW case-control longitudinal CBC's**. This table contains all blood count instances of high RDW individuals and controls performed at the TASMC Integrative Cancer Prevention Center.

**Supplementary Table S6 – high RDW case-control ARCH sequencing**. This table contains variant calling data on all ARCH positive high-RDW cases and controls, including chromosome number, position, reference and alternate nucleotides, the gene involved, whether or not this alternation

is considered a leukemic hotspot, and the reported average VAF from single (Amplicon) or duplicate (MIP) sequencing instances.

**Supplementary Table S7 – LMNA gene correlations**. This table contains the Spearman correlation of each gene to the LMNA signature across metacells (using MPP metacells), and across individuals (using pools of cells in MEBEMP bins 1-3).

**Supplementary Table S8 – sync genes**. This table contains data for how the genes used to calculate the sync-scores were selected. It lists each gene's correlation to AVP and GATA1 across the HSC-MPP-MEBEMP trajectory, each gene's mean expression across metacells in these cell types, and whether the gene is part of the AVP and GATA1 gene signatures.

**Supplementary Table S9 – s-phase gene correlations**. This table contains the Spearman correlation of each gene to the S-phase signature across metacells (using MEBEMP-L metacells), and across individuals (using pools of cells in MEBEMP bins 8-10).

**Supplementary Table S10 – MEBEMP differential expression screen**. This table contains correlations between gene expression and different clinical labels across individuals. Correlations are between expression values normalized for distribution across the MPP-MEBEMP population, and age, maximal mutant VAF, and CBCs. For each gene and label, the Spearman correlation is reported, as well as the p-value and BH corrected q-value for the correlation's equality to 0 (two-sided test). Additionally, each gene's association with CH (treating it as a binary trait) and sex is listed. Association is tested using a Mann-Whitney two-sided test, and the test's p-value and BH corrected q-value are listed. Only genes whose expression levels exceeded a minimal threshold, and who displayed small technical variation, are listed.

**Supplementary Table S11 – CLP differential expression screen**. Similar to Table 8, using the CLP cell types.

**Supplementary Table S12 – Individual scores and signatures**. Key scores and signatures are listed for each individual: *LMNA* signature, S-phase signature, sync-score, composition-bias score, as well as CD34+ cell type distribution.

## References


1. Cohen, N. M. *et al.* Personalized lab test models to quantify disease potentials in healthy individuals. *Nat Med* (2021) doi:10.1038/S41591-021-01468-6.

2. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

3. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).

4. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc Natl Acad Sci U S A* **111**, 2548–2553 (2014).

5. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine* **371**, 2488–2498 (2014).

6. Busque, L., Buscarlet, M., Mollica, L. & Levine, R. L. Concise Review: Age-Related Clonal Hematopoiesis: Stem Cells Tempting the Devil. *Stem Cells* vol. 36 1287–1294 Preprint at https://doi.org/10.1002/stem.2845 (2018).

7. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep* **10**, 1239–1245 (2015).

8. Cohen, K. S. *et al.* Circulating CD34+ progenitor cell frequency is associated with clinical and genetic factors. *Blood* **121**, e50–e56 (2013).

9. Mende, N. & Laurenti, E. Hematopoietic stem and progenitor cells outside the bone marrow: where, when, and why. *Exp Hematol* **104**, 9–16 (2021).

10. Ainciburu, M. *et al.* Uncovering perturbations in human hematopoiesis associated with healthy aging and myeloid malignancies at single-cell resolution. *Elife* **12**, (2023).

11. 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).

12. 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* (2022) doi:10.1101/2020.01.26.919753.

13. GOODMAN, J. W. & HODGSON, G. S. Evidence for stem cells in the peripheral blood of mice. *Blood* **19**, 702–14 (1962).

14. Biezuner, T. *et al.* An improved molecular inversion probe based targeted sequencing approach for low variant allele frequency. *NAR Genom Bioinform* **4**, (2022).

15. 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).

16. Mende, N. *et al.* Unique molecular and functional features of extramedullary hematopoietic stem and progenitor cell reservoirs in humans. *Blood* **139**, 3387–3401 (2022).

17. Lehnertz, B. *et al. HLF* expression defines the human hematopoietic stem cell state. *Blood* **138**, 2642–2654 (2021).

18. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp Hematol* **68**, 51–61 (2018).

19. Frelin, C. *et al.* GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat Immunol* **14**, 1037–1044 (2013).

20. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

21. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (1979)* **367**, (2020).

22. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, (2018).

23. 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).

24. Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature* **571**, 355–360 (2019).

25. Ramdas, N. M. & Shivashankar, G. V. Cytoskeletal Control of Nuclear Morphology and Chromatin Organization. *J Mol Biol* **427**, 695–706 (2015).

26. Manley, H. R., Keightley, M. C. & Lieschke, G. J. The Neutrophil Nucleus: An Important Influence on Neutrophil Migration and Function. *Front Immunol* **9**, (2018).

27. Grigoryan, A. *et al.* LaminA/C regulates epigenetic and chromatin architecture changes upon aging of hematopoietic stem cells. doi:10.1186/s13059-018-1557-3.

28. 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).

29. 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).

30. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat Methods* **17**, 615–620 (2020).

31. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291.e9 (2019).

32. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337.e4 (2019).

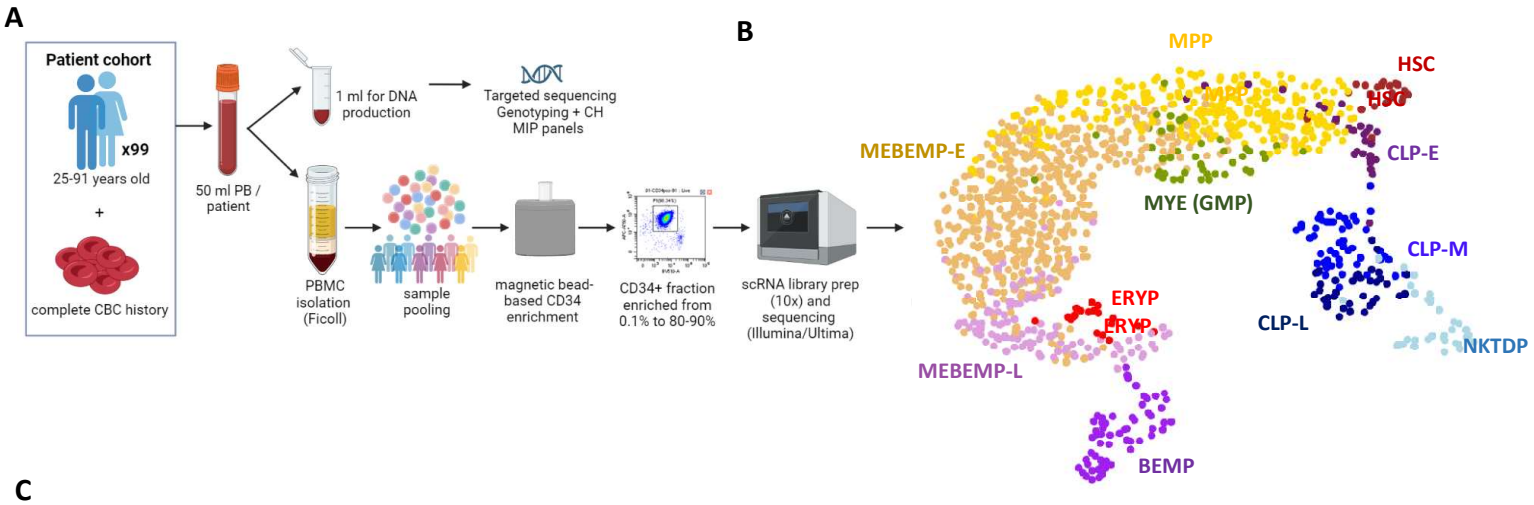
33. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451–460 (2019).

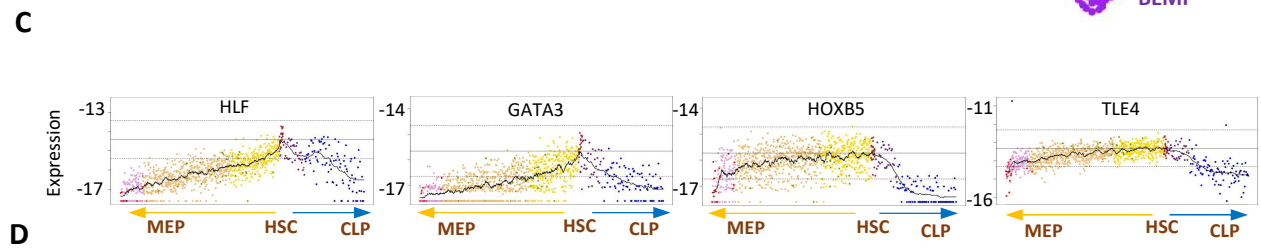
34. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science (1979)* **376**, (2022).

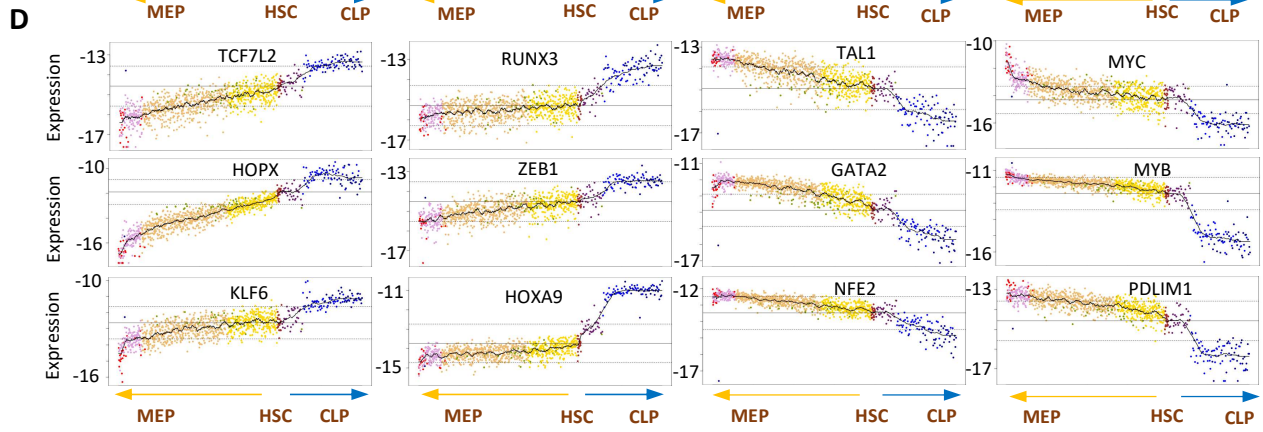
35. Zhang, P. *et al.* Single-cell RNA sequencing to track novel perspectives in HSC heterogeneity. *Stem Cell Res Ther* **13**, 39 (2022).

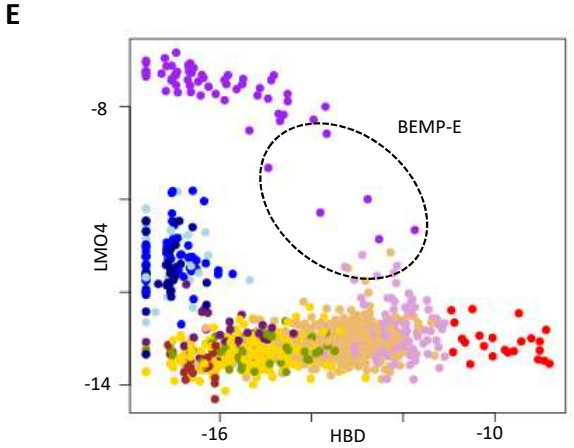
# Figure 1

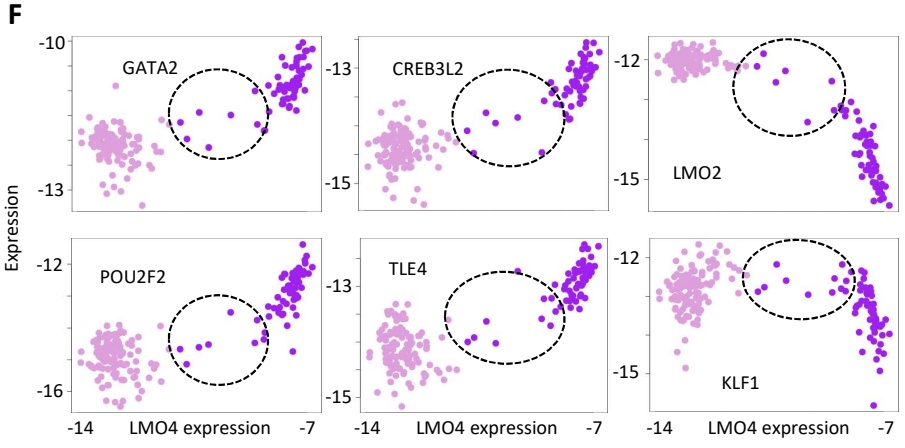


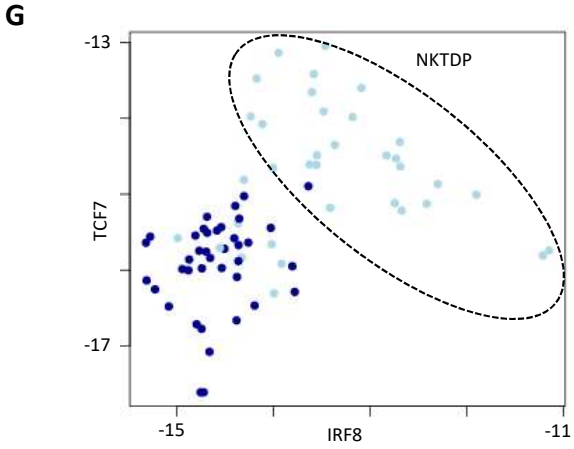


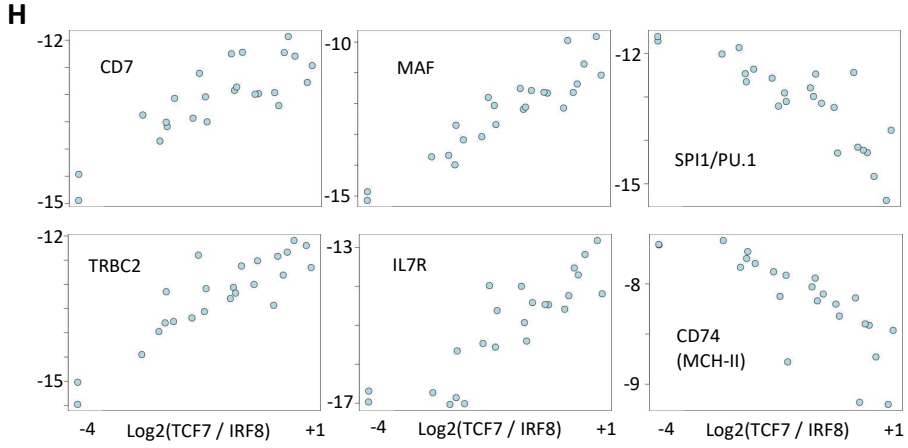



**Figure 1a** – experimental design, **1b** – annotated 2D UMAP projection of our metacell manifold following filtration of metacells with low CD34 expression. Symmetric (**1c**) and asymmetric (**1d**) regulation of specific HSC markers upon bifurcation to the CLP (right) and MEBEMP (left) lineages. 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. Units for gene expression in all the figure panels are log2 of each gene's fractional expression. **1e** – the BEMP-E metacell population of interest (dotted line) linking BEMPs to their MEBEMP-L precursors. **1f** – positively and negatively regulated TFs involved in early BEMP differentiation. **1g** – 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 (dotted line). This population exhibits high expression of both T and dendritic cell regulators, forming a **gradient** consisting of **NK/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 class II gene *CD74* (**1h**).

# Figure 2



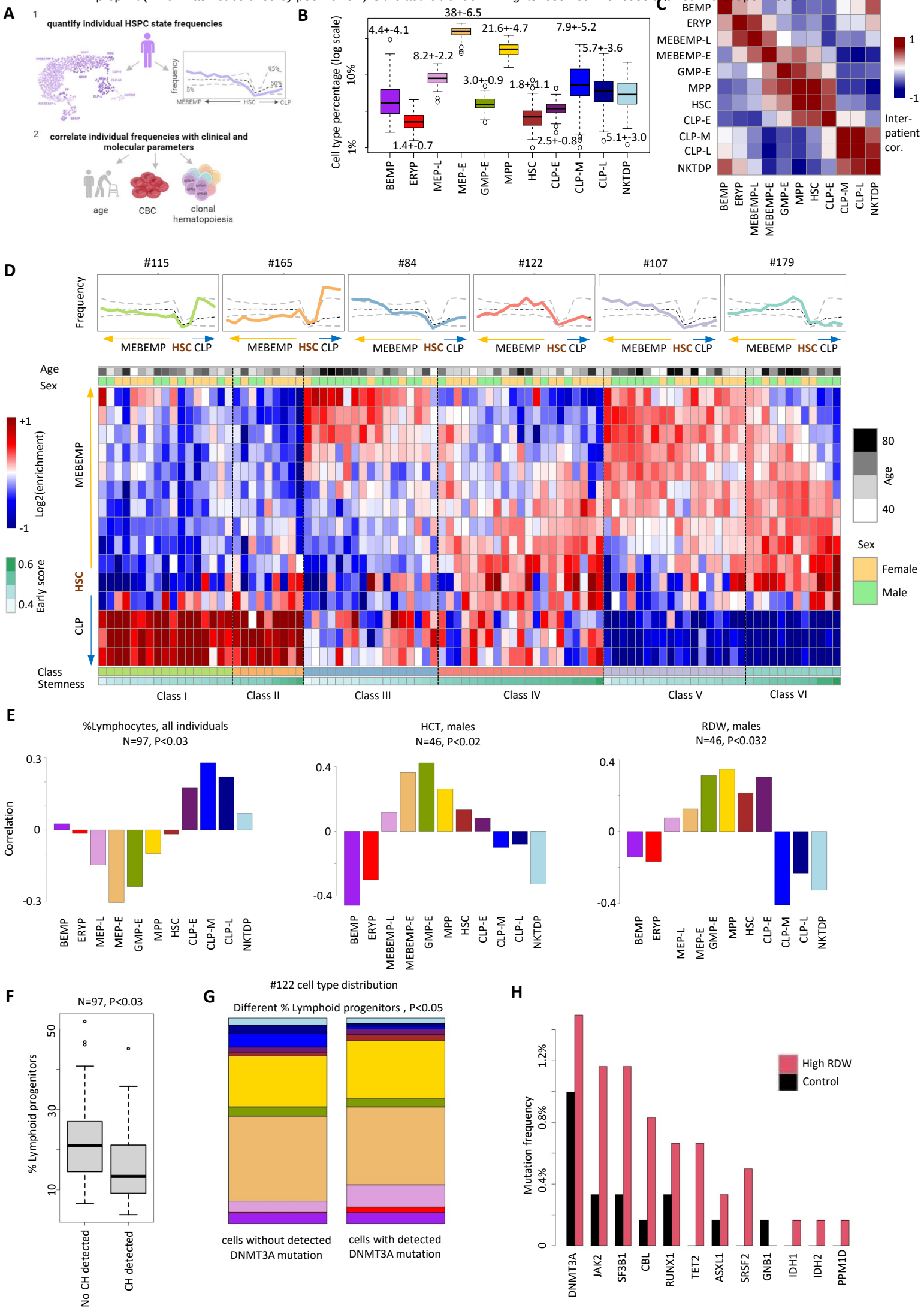


**Figure 2a** - characterization of inter-individual HSPC compositional state variation (scheme). **2b** – boxplots of cell state frequency distributions across individuals (logarithmic scale). Percents calculated out of *CD34*+ population. Boxplot centers, hinges and whiskers represent median, first and third quartiles and 1.5× interquartile range, respectively. Numbers represent mean +/- SD for each distribution. **2c -** correlation of cell state frequencies across individuals. **2d** (**top**) - individual cell state frequency profiles over the HSC-MEBEMP and HSC-CLP differentiation gradients of 6 subjects (colored lines), each representing one of six archetypes (classes) of HSPC composition in healthy individuals. Dashed lines represent the median (black) and $5^{th}$ and $95^{th}$ percentiles (grey) of the studied population. **2d** (**bottom**) cell state enrichment map over 15 differentiation bins (rows), for all studied individuals (columns) clustered into 6 classes. Classes I & II represent individuals relatively enriched in lymphoid progenitors, whereas classes V & VI represent individuals with relative depletion of lymphoid progenitors. Individuals are sorted by stemness in each class. Age and sex bins are denoted for each individual (top). **2e** – CBC correlations to cell type frequencies: %Lym (from WBC, calculated for entire cohort, left), HCT (males, center), RDW (males, right). Missing individuals lacked sufficient cells for analysis. Permutation test p values are displayed for each correlation. **2f** – boxplots of CLP frequency distributions in individuals with (right) and without (left) clonal hematopoiesis. **2g** – Relative cell state frequencies in mutant (right) and non-mutant (left) cells following GoT of sample #122 (DNMT3A mutated, VAF = 0.07). **2h** – CH frequency (by gene) in age- and sex-matched high (red) and low (black) RDW individuals.

# Figure 3

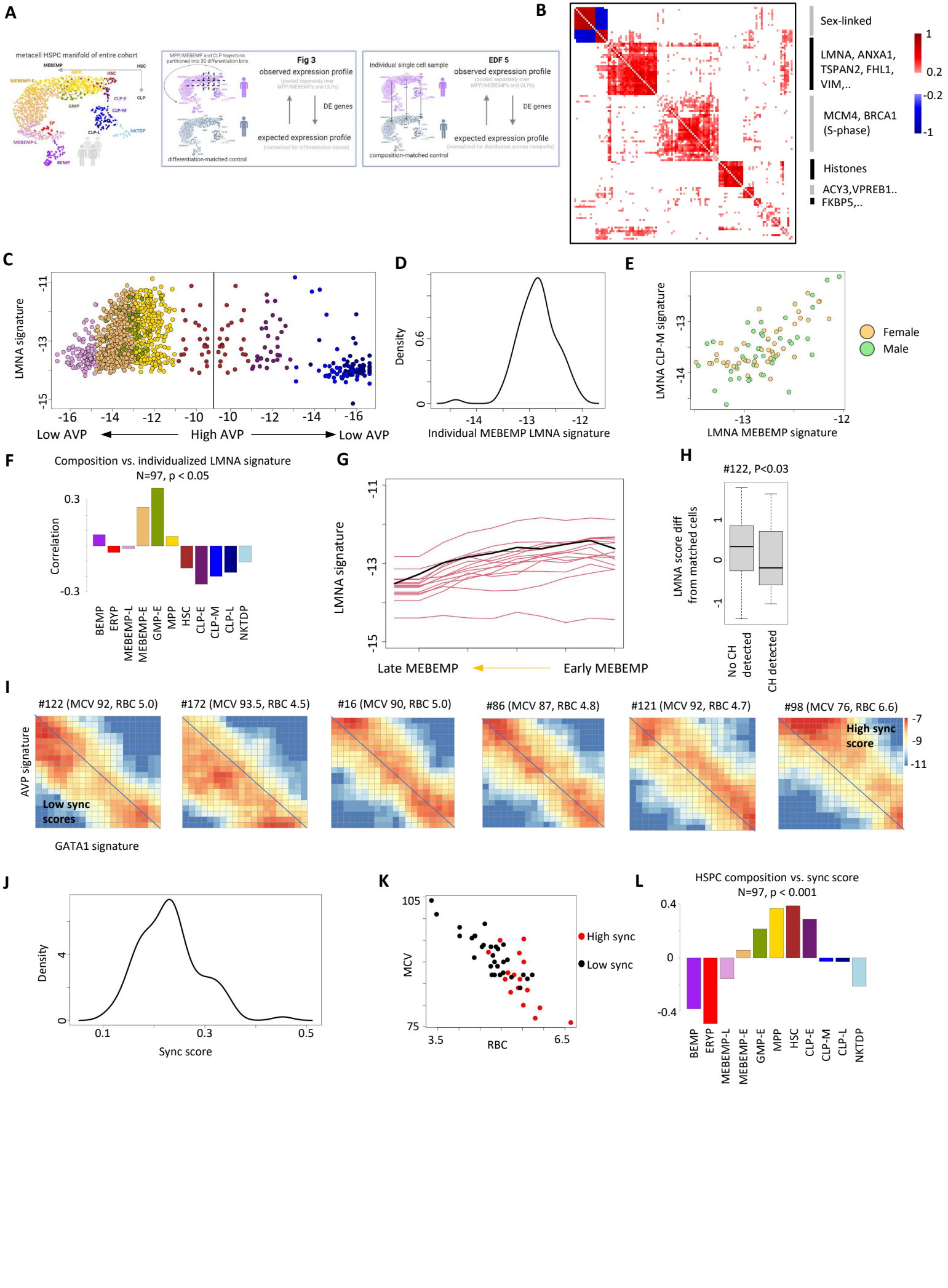
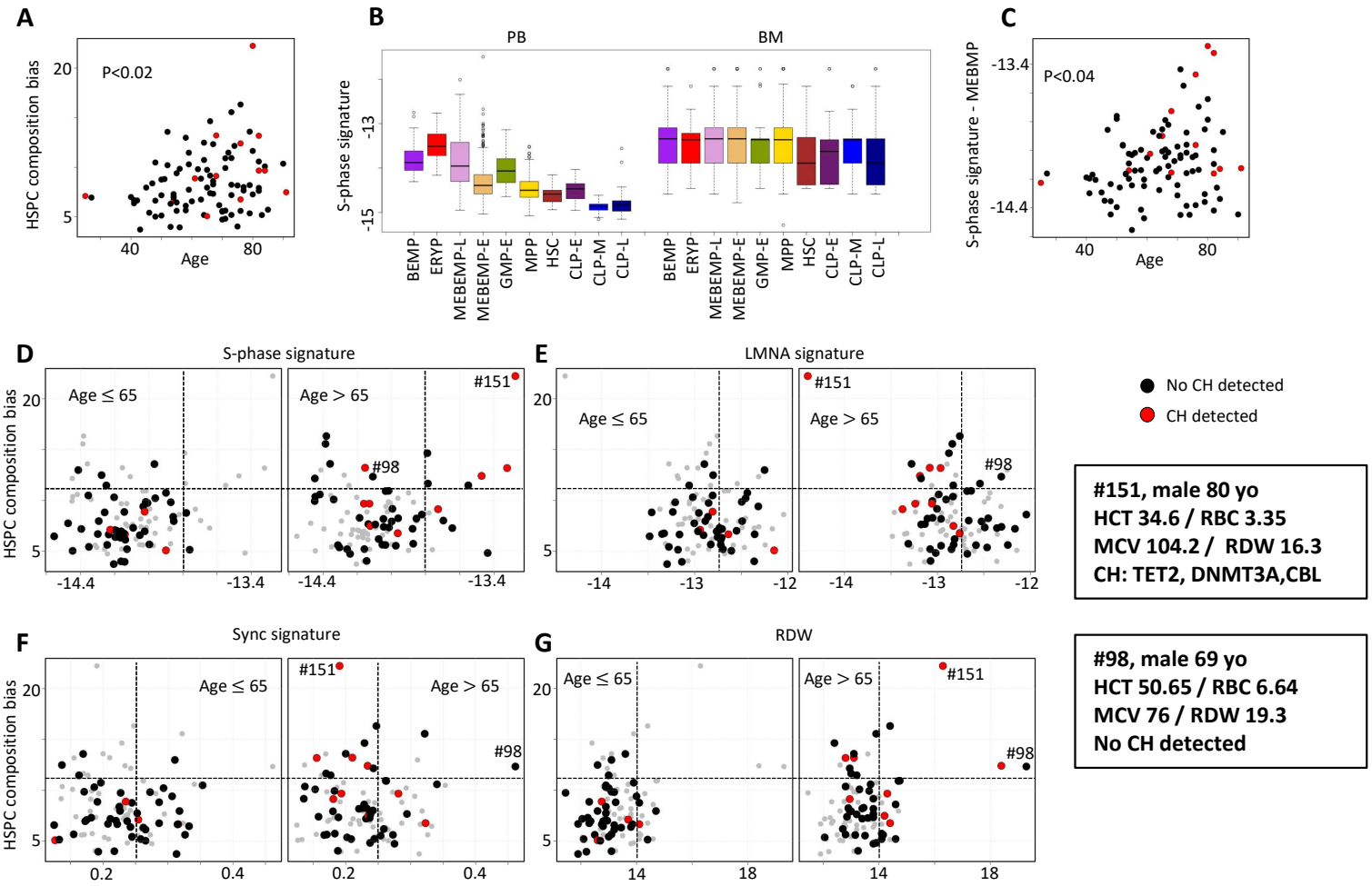


**Figure 3a** - Compositional-controlled characterization of differentially expressed gene signatures and their association with clinical parameters (scheme). **3b** – gene-gene correlation heatmap, calculated over individual-level HSC-MEBEMP gene expression normalized for HSC-MEBEMP composition. **3c** – *LMNA* signature in HSCs (denoted by high *AVP*) and throughout MPP / MEBEMP (left) and lymphoid (right) differentiation. **3d** – density curve of individual MEBEMP *LMNA* signatures. **3e** - intra-individual correlation of *LMNA* signatures in CLPs and MEBEMPs. Male samples are in green, female samples in orange. **3f** – correlation between an individual's average MEBEMP *LMNA* signature and his/her HSPC composition. Permutation test p value denoted on top. **3g** – *LMNA* signatures of CH+ individuals across MEBEMP differentiation. Each red line denotes an individual, black line denotes median *LMNA* signature across the CH- sampled population. **3h** - boxplots comparing *LMNA* signatures between WT and mutated cells within the single cell sample of individual #122 (*DNMT3A* mutated, VAF = 0.07). Y axis measures *LMNA* signature compared to matched cells from the MEBEMP trajectory. **3i** – individual heatmaps of single cell counts over 20 bins of stemness (*AVP* signature, y axis) and MEBEMP differentiation (*GATA1* signature, x axis). Individual identifier, RBC, and MCV are denoted on top. **3j** - density curve of individual sync scores. **3k** – comparison between individual sync scores and clinical parameters (RBC/MCV) across males. High and low sync scores define clinically distinct populations. **3l** – correlation between individual sync scores and cell type composition. Permutation test p value denoted on top.

Figure 4




**#151, male 80 yo**
**HCT 34.6 / RBC 3.35**
**MCV 104.2 / RDW 16.3**
**CH: TET2, DNMT3A,CBL**

**#98, male 69 yo**
**HCT 50.65 / RBC 6.64**
**MCV 76 / RDW 19.3**
**No CH detected**

**Figure 4a** – composition bias score variation with age. **4b** - cell type-specific comparison of S-phase signatures in circulating (left) vs. BM (right) HSPCs. **4c** - S-phase signature variation with age in the late MEBEMP trajectory. **4d** – corresponding individual S-phase signatures (X axis) and composition bias scores (Y) for individuals younger (left) and older (right) than 65 years. **4e-g** – like 4d, but showing the *LMNA* signature, sync scores, and RDW instead of S-phase, respectively.

Figure 5



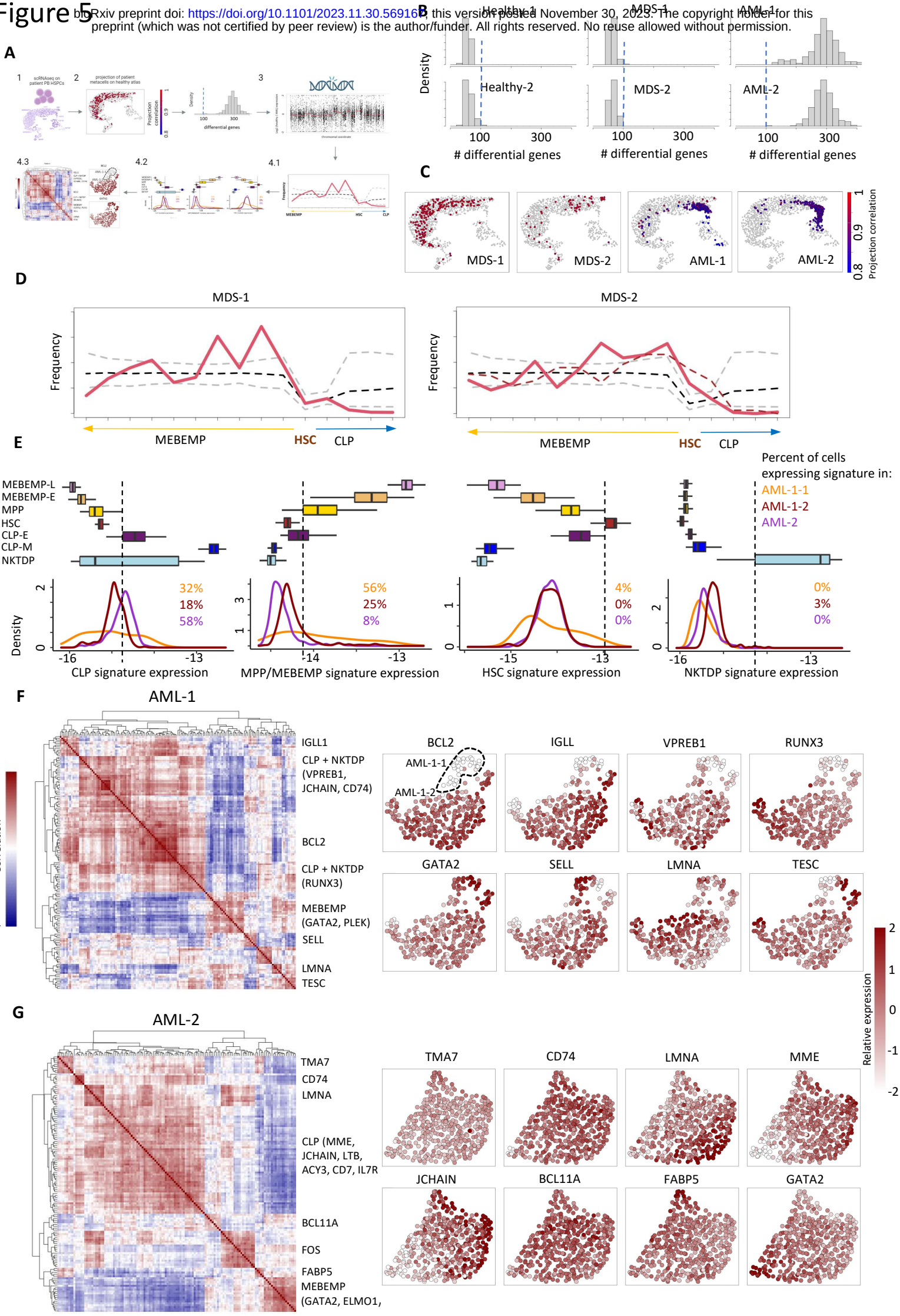
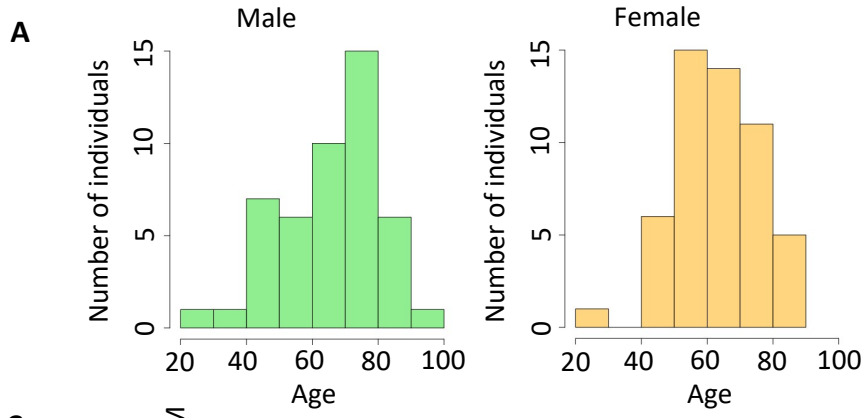
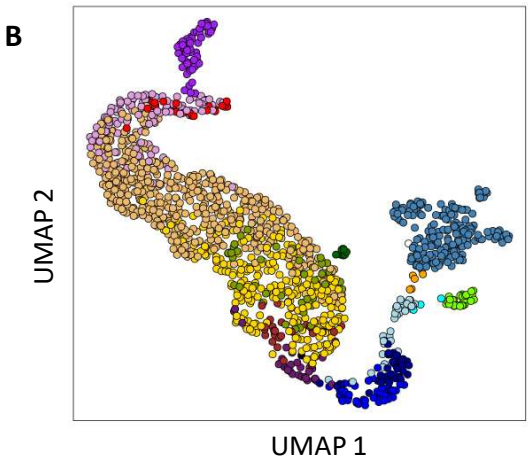
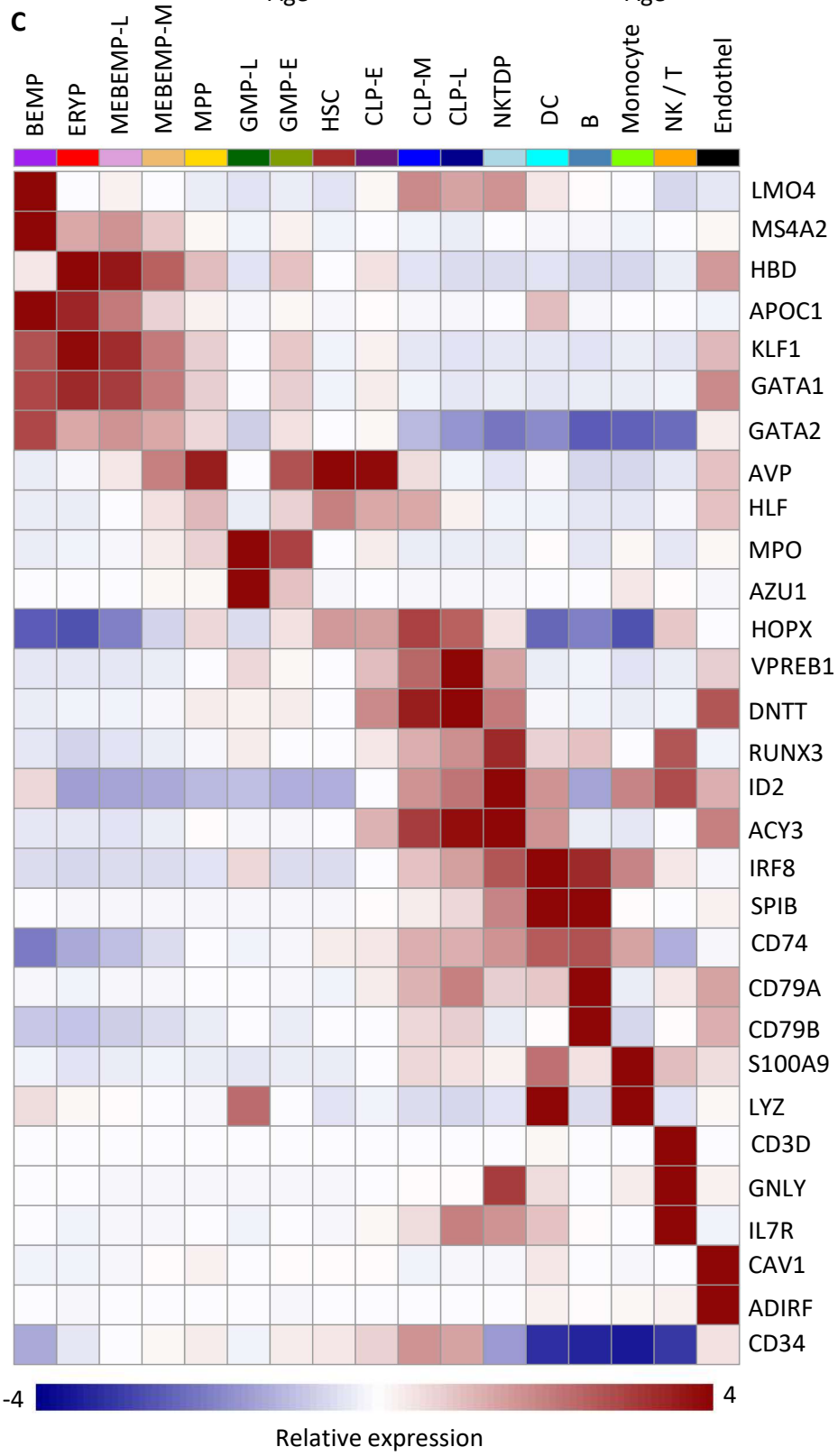


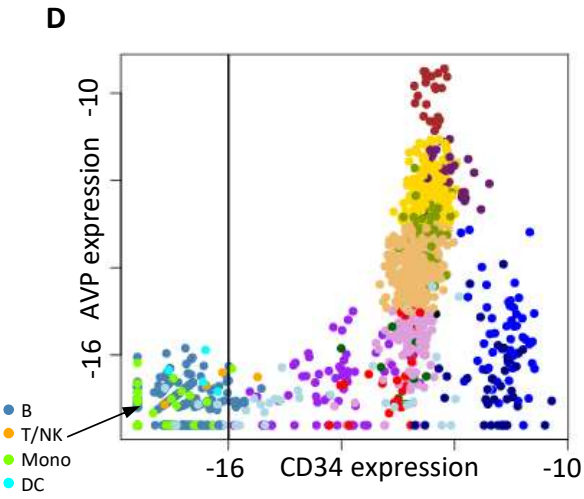
**Figure 5a** - diagnostic approach to leukemia analysis using our HSPC reference atlas (scheme): 1. scRNAseq on CD34-enriched PB and construction of a patient-specific metacell model, 2. Projection of patient derived metacells on the healthy reference atlas - compositional variance and differential gene expression analysis, 3. Mutational and CNV analysis using targeted DNA sequencing and RNA-based karyotyping, 4. RNA-based clonal hierarchy and population substructure analysis using: 4.1 individual cell state frequency profiles over the HSC-MEBEMP and HSC-CLP differentiation gradients, 4.2 sub-population identification of AML cells with CLP, HSC and MEBEMP characteristics 4.3 de-novo identification of clonal specific gene clusters and signatures. **5b** – density plot of the number of differentially-expressed genes (≥2-fold) per metacell as compared to its projection counterpart on our healthy HSPC atlas, for 2 healthy (left), 2 MDS (middle), and 2 AML (right) patients. **5c** – projection of metacells derived from 2 MDS (left) and 2 AML (right) patients on our healthy HSPC reference metacell model. **5d** – individual cell state frequency profiles over the HSC-MEBEMP and HSC-CLP differentiation gradients for 2 MDS cases (red lines). Dashed lines represent the median (black) and $5^{th}$ and $95^{th}$ percentiles (grey) of the healthy population, and MDS-2's initial profile (red, right panel, 8 month prior to current profiling). **5e** – each of the 4 panels refers to a different cell state gene signature as noted on the x-axis. Top - boxplots of gene module expression distributions for different cell states in our reference atlas. Bottom - Gene signature expression density plots for each of the AML subclones. Reference gene signature distributions (top) were used to identify subpopulations of AML cells with CLP, HSC and MEBEMP characteristics (bottom). Dashed lines represent the threshold for expressing a gene signature, and the fraction of cells expressing a signature per AML clone is listed. **5f** - left – correlation heatmap of differentially expressed gene signatures for AML-1. The malignant state is characterized by multiple novel gene expression signatures in addition to aberrant expression of "healthy" differentiation-related modules, right – UMAP projection of the metacell model of AML-1, colored by relative expression of differentially expressed genes. Overexpression of BCL2 in AML-1-2 compared to AML-1-1 can be seen on the top left panel. **5g** – same as 5f for AML-2.

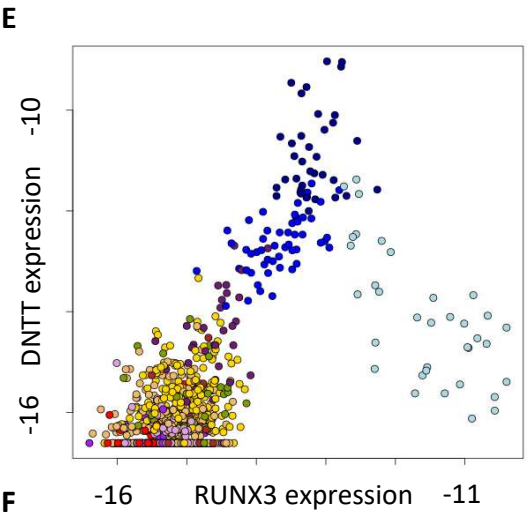
# EDF1

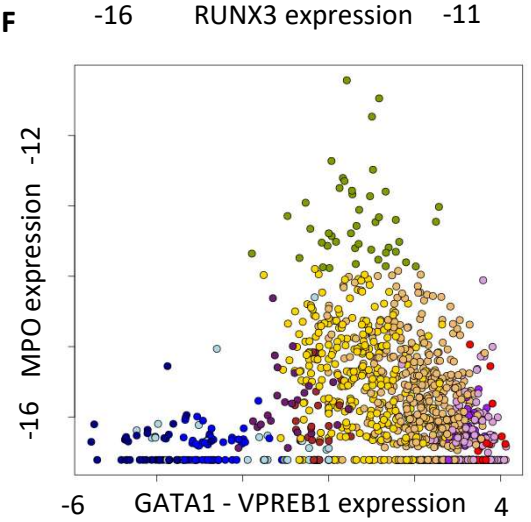


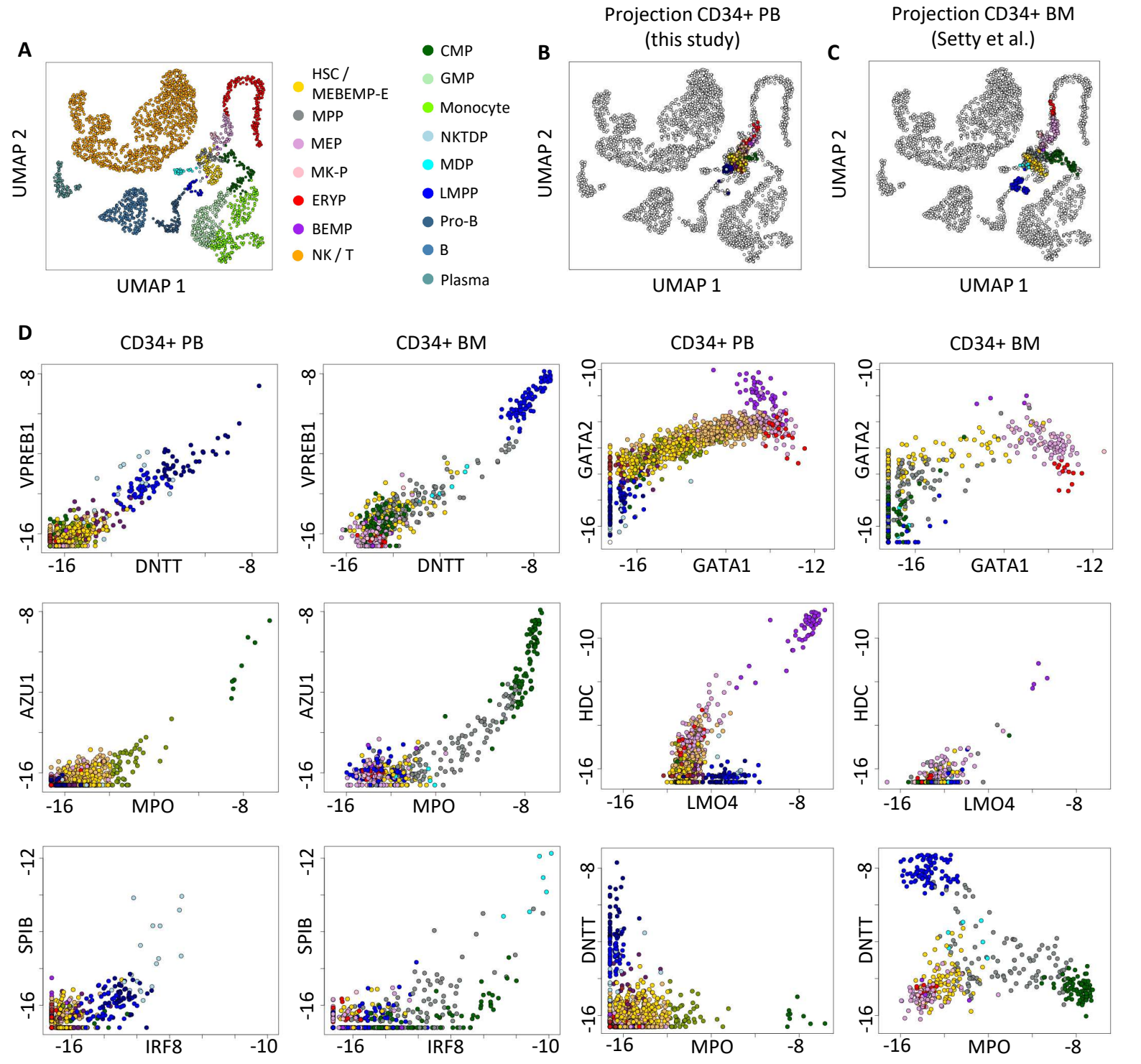



# EDF 1 – Cell state annotation, major markers and regulators of HSC differentiation and sub-population branching

**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 markers used for cell state annotation (rows). **1d** – Filtering metacells with low *CD34* expression. **1e** - gene-gene expression plot of *DNTT* and *RUNX3*, showing early CLP differentiation and their bifurcation into late CLPs and NKTDPs. **1f** –expression plot of *MPO* and *GATA1/VPREB1* expression showing all 3 differentiation routes (GMPs, CLPs, MEBEMPs) from HSCs, and highlighting the GMP's bifurcation from MPP / MEBEMP. All gene expression values are obtained by normalizing gene expression to sum to 1 and taking log2.

# EDF2

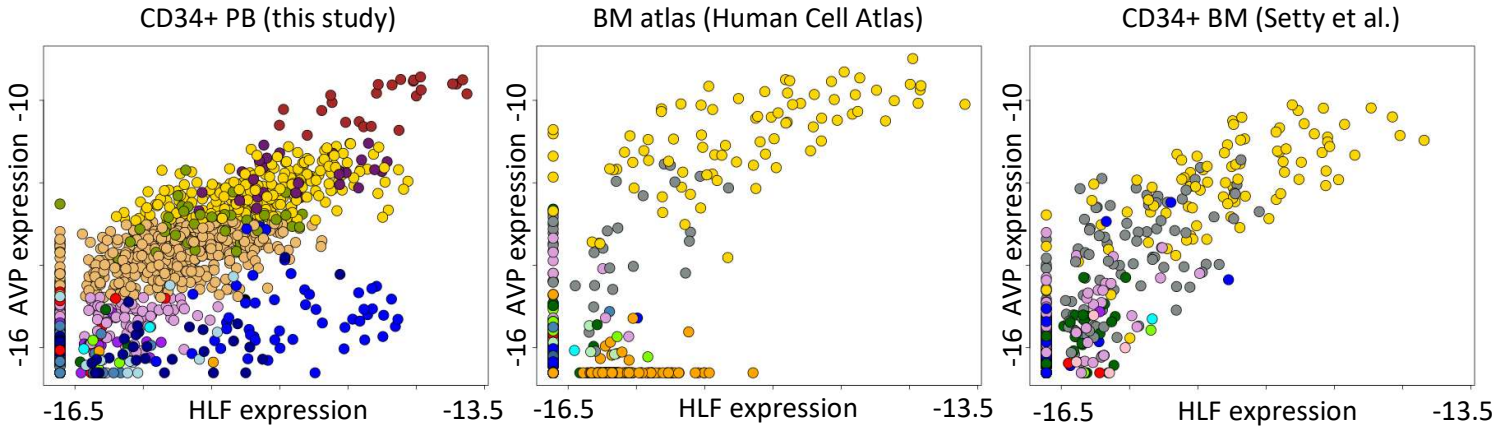


## EDF 2 – BM comparisons

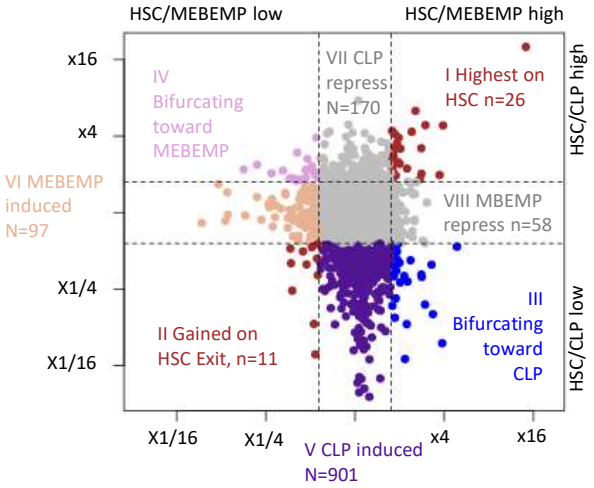
**2a** – 2D UMAP projection of a non-*CD34*-enriched BM metacell model from the Human Cell Atlas[18], colored by a BM-specific cell type annotation. **2b** - projection of our PB *CD34*+ derived metacells on the non-*CD34* enriched BM metacell model. **2c** – projection of BM *CD34*+ derived metacells [Setty et al.] on the non-*CD34* enriched BM metacell model. **2d** - gene-gene expression plots comparing PB *CD34*+ derived metacells with their BM *CD34*+ counterparts for all differentiation trajectories. Panels (left to right, top to bottom) represent CLP differentiation, MEBEMP differentiation, GMP differentiation, BEMP differentiation, DC differentiation, and MPO/CLP/MEBEMP trifurcation from HSCs. The first 4 differentiation panels represent similar PB and BM behaviors, while the last 2 show dissimilarities between them. PB / BM metacells are colored by PB / BM annotations, respectively. **2e** – gene-gene expression plots comparing PB *CD34*+ derived metacells with their BM *CD34*+ counterparts for markers and regulators of CLP differentiation and bifurcation. **2f** – relative expression heatmap of the megakaryocytic markers *PF4* and *PPBP* and cell type specific markers, across metacells with high megakaryocytic signature. The figure shows an abnormally high doublet rate involving megakaryocytes.

# EDF3

**A**



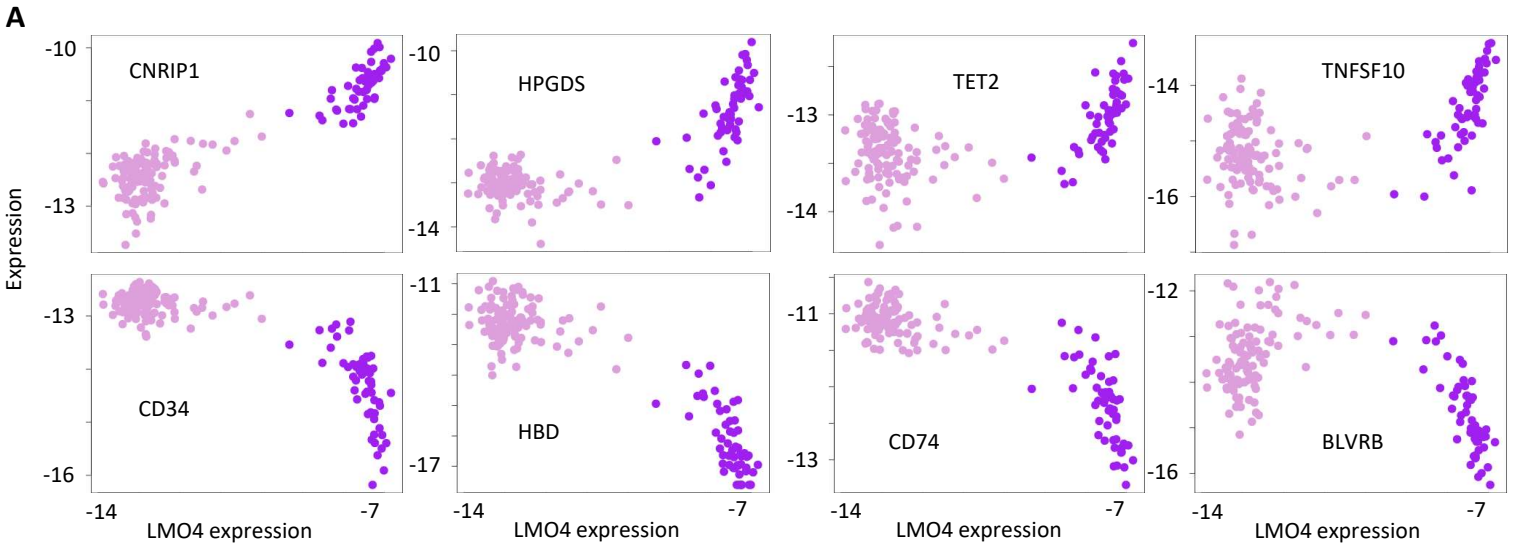

**B**

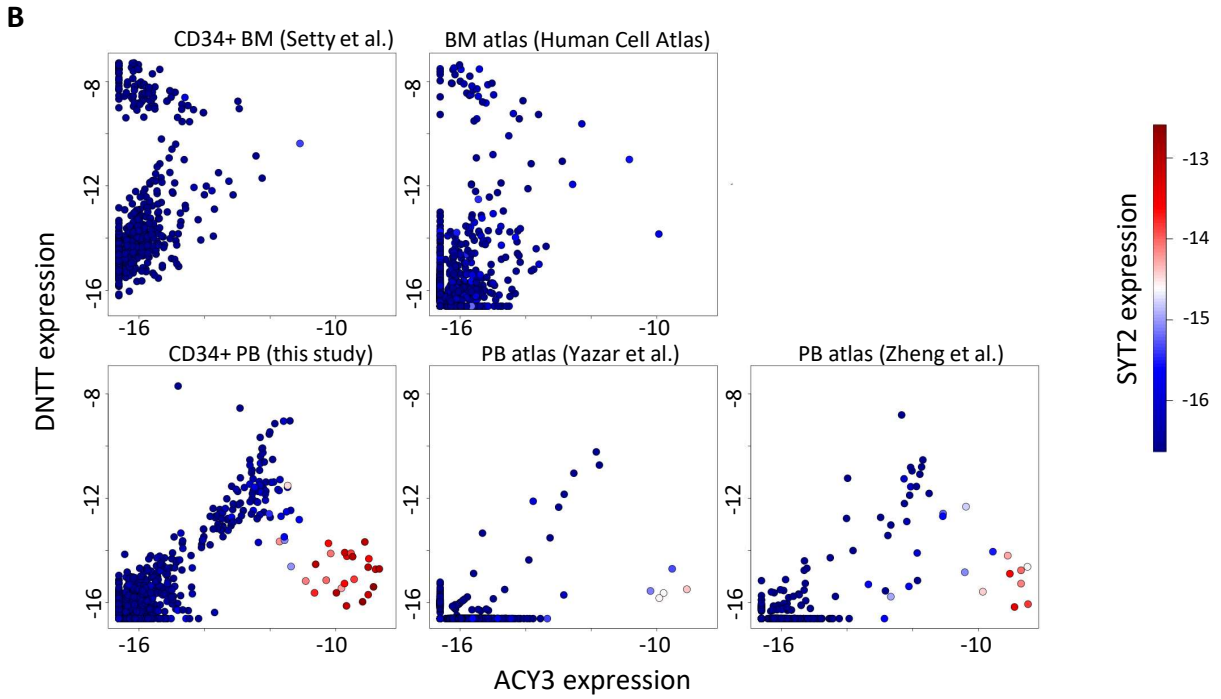


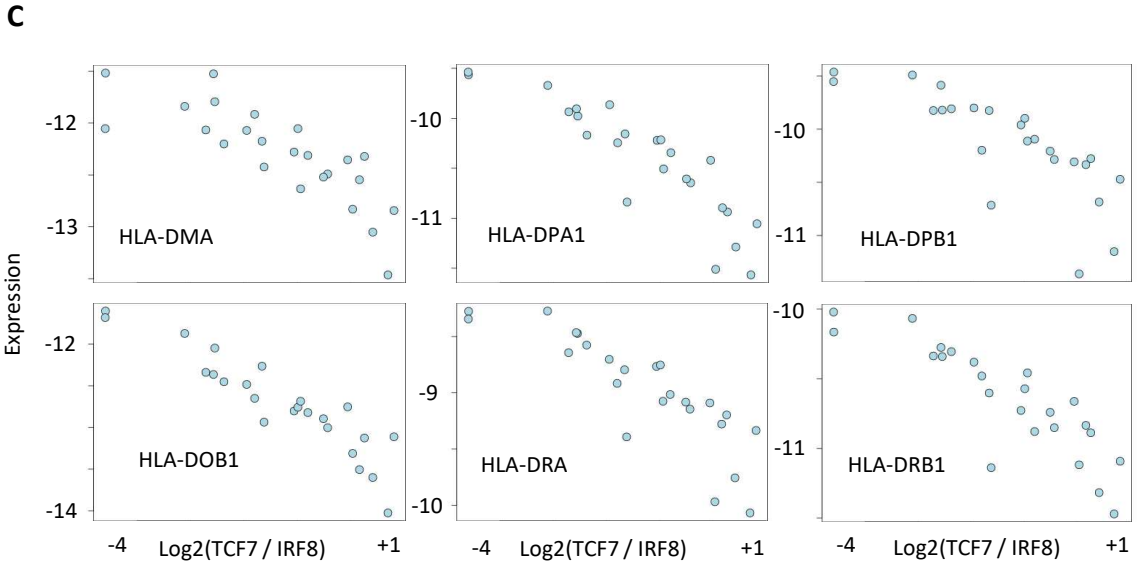
## **EDF 3** – **High HLF/AVP HSCs**

**3a** - gene-gene expression plots comparing our high *AVP*/*HLF* HSC population (left) with that found in two BM metacell models (right[18,33]). PB and BM metacells are colored by PB and BM annotations, respectively. **3b** – 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, **3c** – *GATA3* mutation screening on the Beat-AML WES data detected 5 positive cases for a hotspot in *GATA3* R353K, out of 826 sequenced cases (~1%). These had co-occurring mutations in *DNMT3A*, *TET2*, and *SRSF2*.

# EDF4



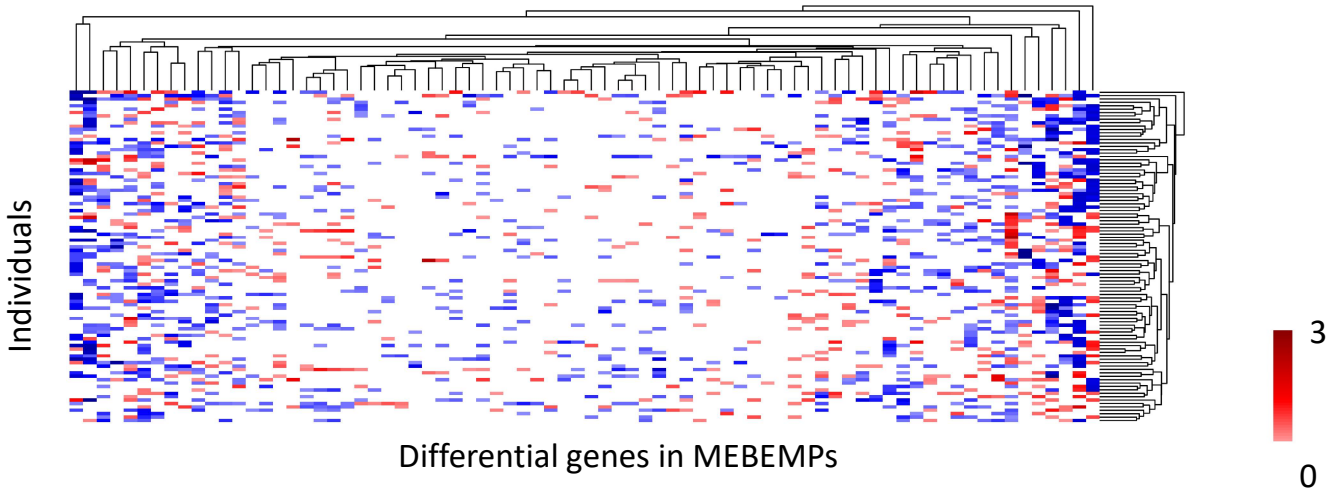



## EDF 4 – Factors involved in BEMP and NKTDP differentiation

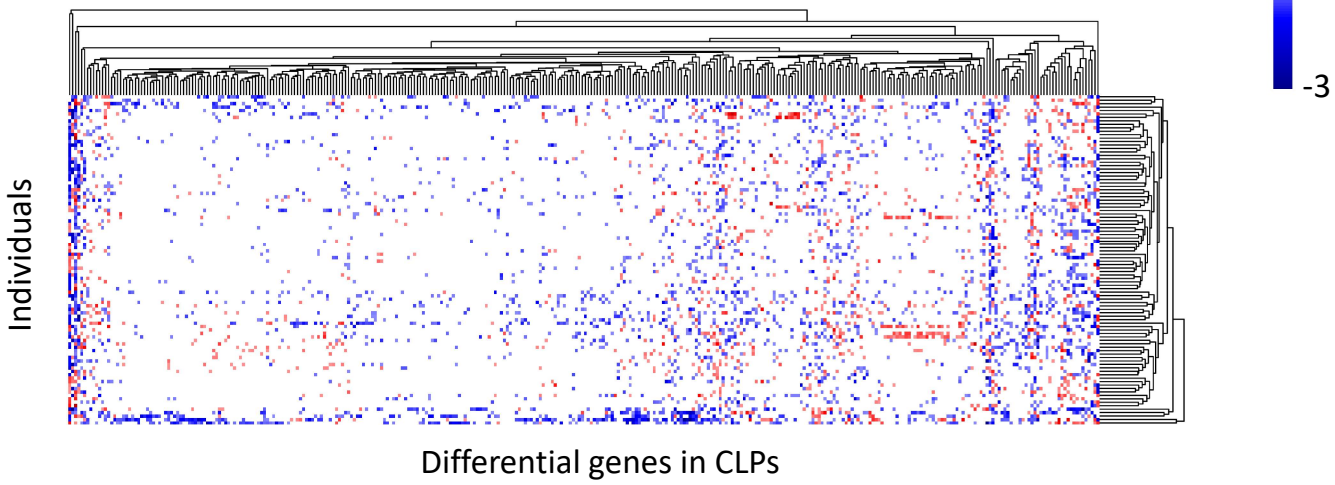
**4a** - factors positively and negatively regulated in the early stages of BEMP specification. **4b** – gene-gene expression plots of *DNTT* and *ACY3* comparing *CD34*-enriched and non-enriched BM (top[18,33]), as well as non-enriched and partially enriched PB (bottom[3435]] to our CD34+ PB model. Metacells are color-coded by *SYT2* expression (log2 transformed). The *SYT2* high, *ACY3* high, *DNTT* intermediate population clearly seen in our 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 cell state to its sub-populations.

# EDF5

**A**



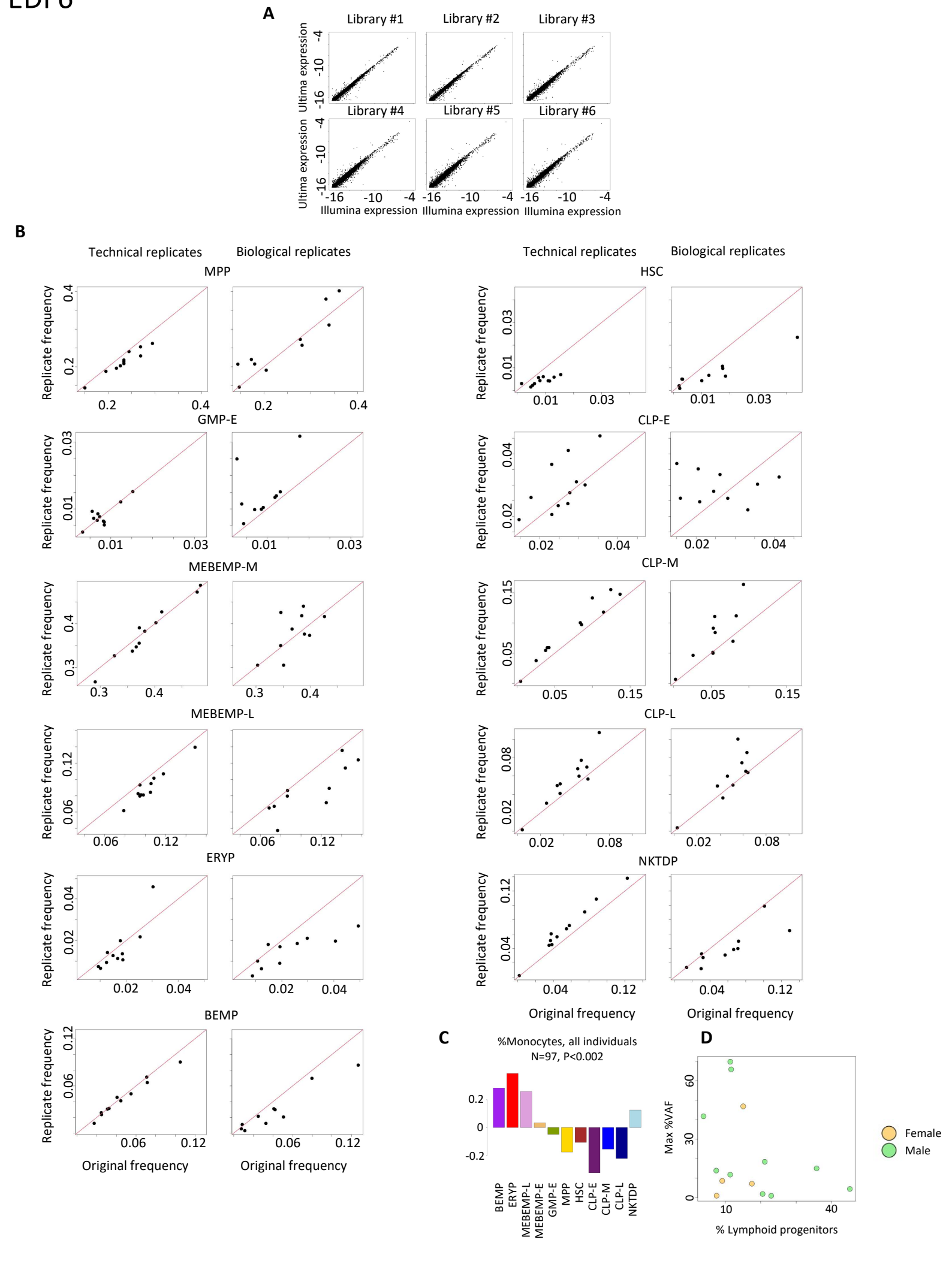

**B**



# EDF 5 – Individual-specific composition-controlled differential gene expression

**5a,b –** Individual-specific differential gene expression after controlling for each individual's distribution across the *CD34*+ PB manifold in MEBEMPs (top) and CLPs (bottom).

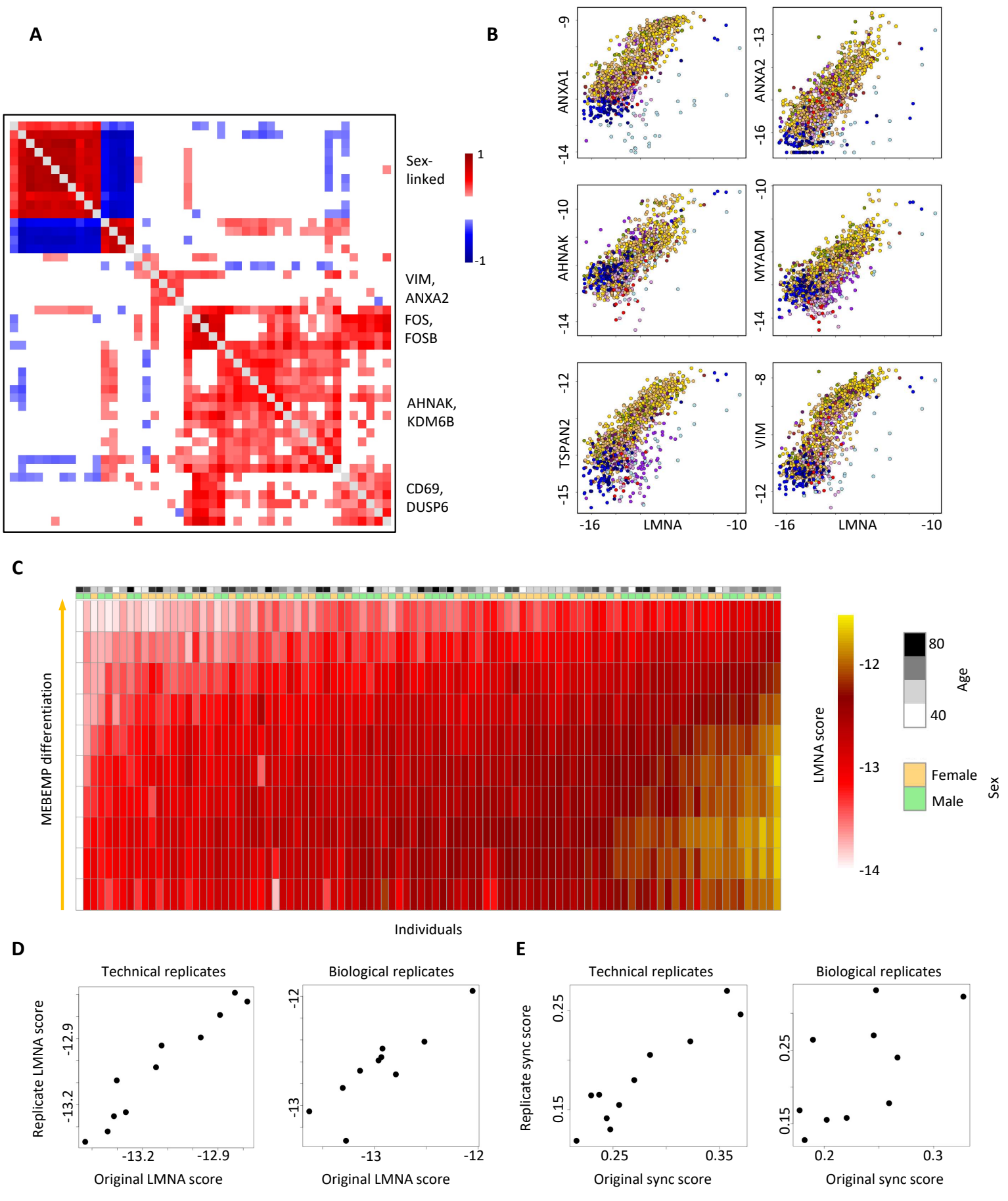
# EDF6



# EDF 6 – Stability of cell state frequencies across technical and biological replicates

**6a** – comparison of Illumina-sequenced and Ultima-sequenced data. Each panel represents one library that was sequenced by both technologies. Points represent genes, and each gene's expression level across all cells in the library as determined by Illumina (X) and by Ultima (Y) is shown. **6b** – cell state frequency correlations between 11 technical & 10 biological replicates and their original samples. Specific cell states are denoted on top of each panel. All biological replicates were sampled ~1 year following original blood draw. **6c** –correlations between CBC %Mono (from WBC) and cell type frequencies. Permutation test p value denoted on top. Missing individuals did not have sufficient cells for analysis. **6d –** CLP frequency against maximal VAF for individuals with CH.

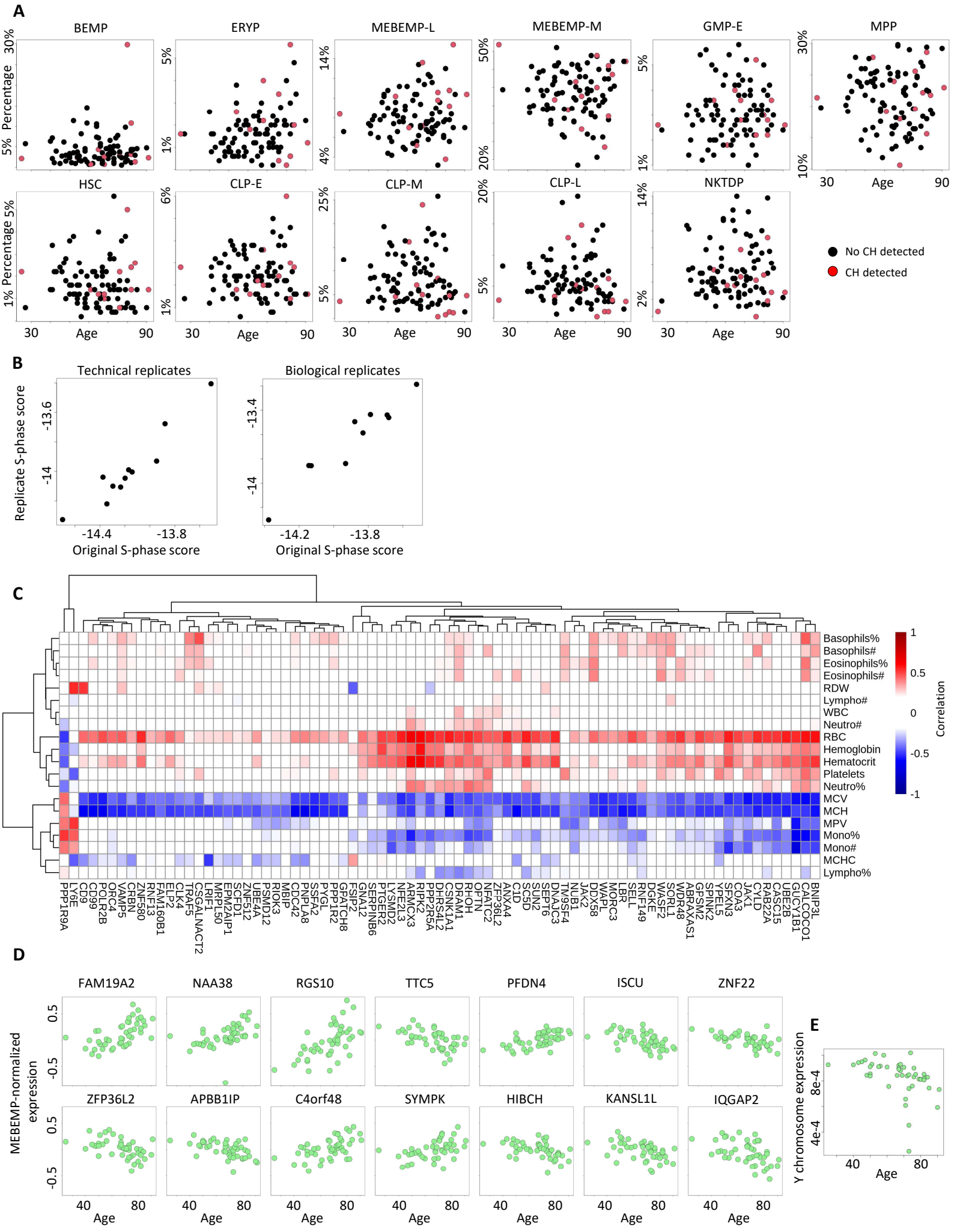
# EDF7



## EDF 7 – Composition-controlled transcriptional variation: the *LMNA* signature and sync score

**7a** - gene-gene correlation heatmap, calculated over individual-level CLP gene expression normalized for CLP composition. **7b** – the *LMNA* signature – co-variation of *LMNA* expression with *ANXA1/2*, *AHNAK*, *MYADM, TSPAN2* and *VIM*. **7c** – heatmap of individual *LMNA* signatures across the MEBEMP trajectory. Individual age and sex are color-coded on top. **7d** – *LMNA* signature correlations between 11 technical & 10 biological replicates and their original samples. **7e** - sync score correlations between replicates and their original samples. All biological replicates were sampled ~1 year following original blood draw.

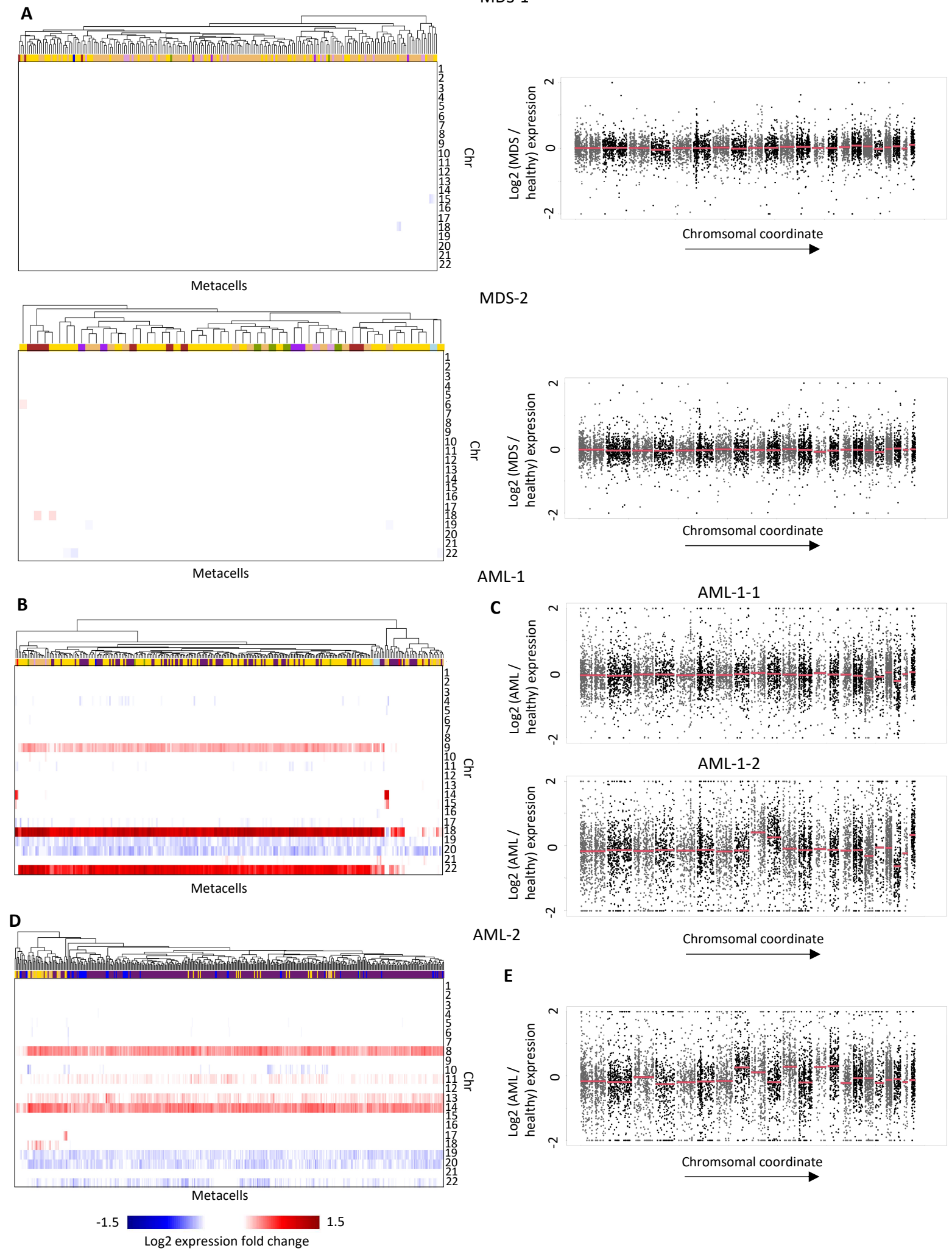
# EDF8



## EDF 8 - Age related perturbation of HSPC compositions and transcriptional signatures

**8a –** Age compared to HSPC cell type frequency. Each panel represents a different cell state. Each dot represents an individual. Individuals are colored by their CH status. **8b** – S-phase signature correlations between 11 technical & 10 biological replicates and their original samples. All biological replicates were sampled ~1 year following original blood draw. **8c** – genes with CBC correlated and anti-correlated expression across males (using Spearman correlations). All genes displayed had an FDR-corrected p-value < 0.1 (two-sided test for Spearman's rho) for at least 1 CBC parameter. **8d** – Age correlated and anti-correlated genes (Spearman), FDR < 0.1. Each dot represents an individual. Correlation applies only to males. **8e** – anti-correlation of total Y chromosome expression with age.

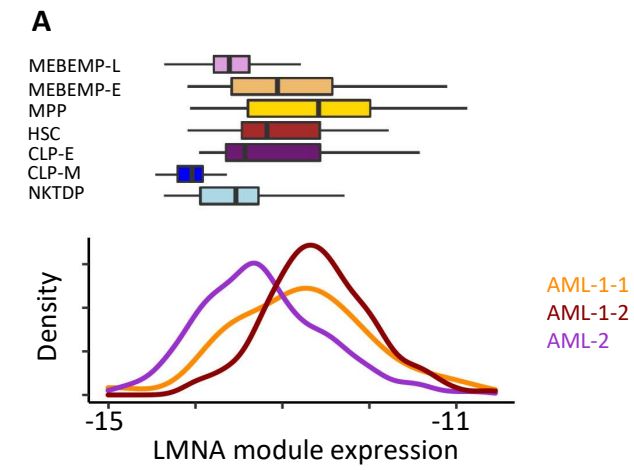
# EDF9

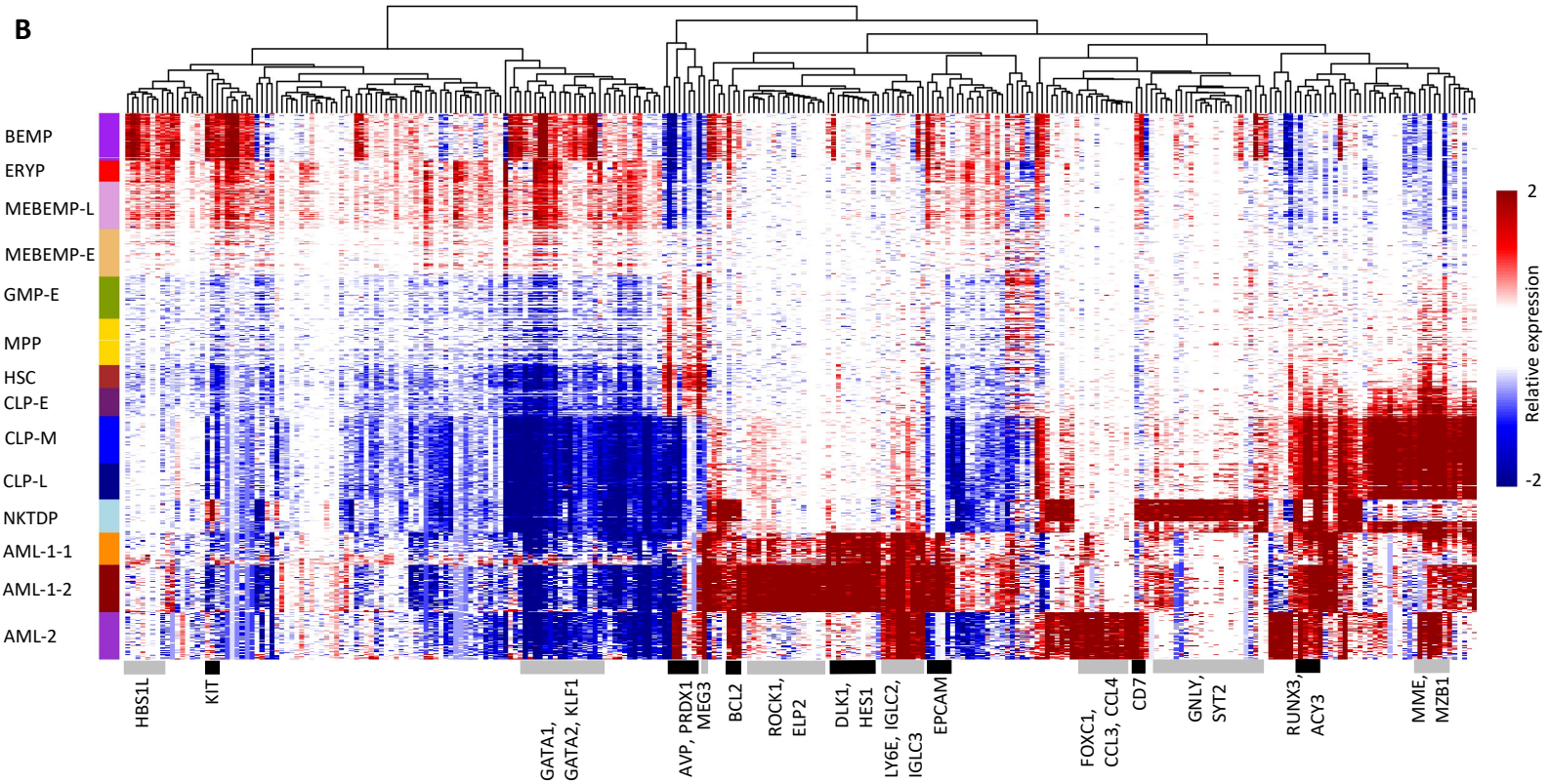


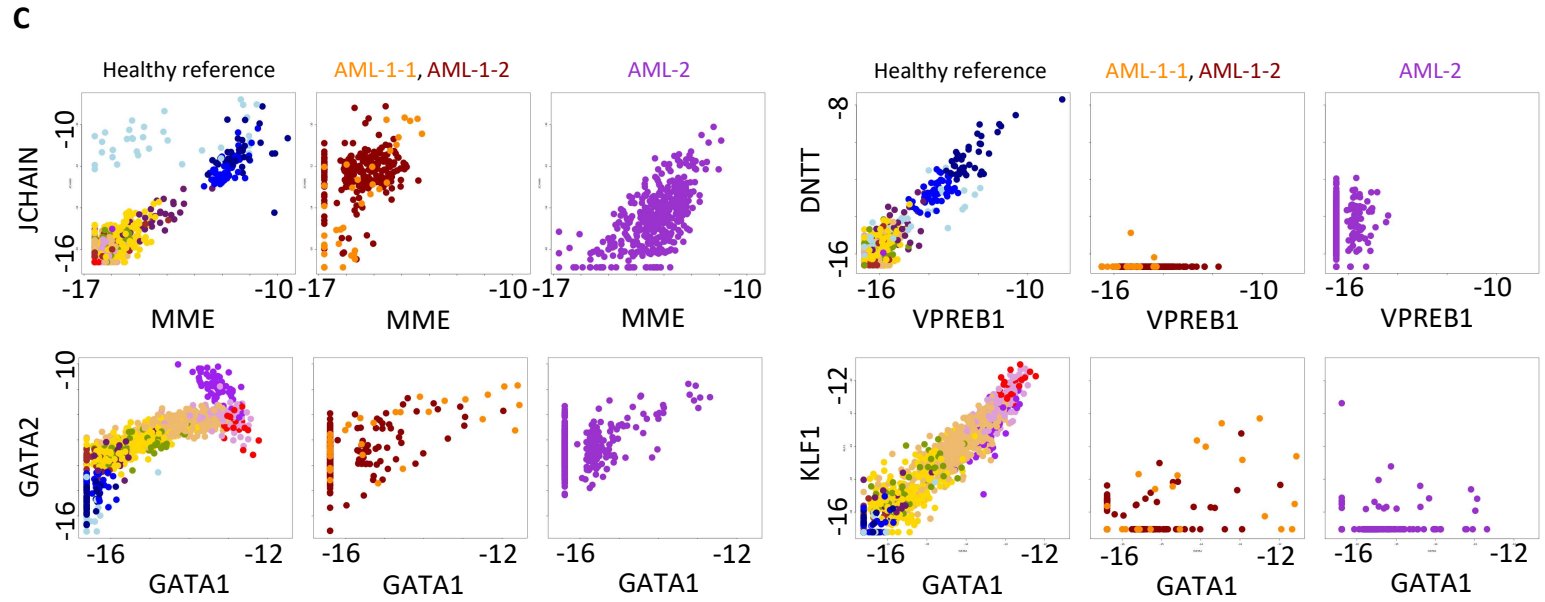
## EDF 9 – scRNAseq karyotyping

**9a** – scRNAseq karyotyping for MDS-1 (top) and MDS-2 (bottom). Metacell models were created for each MDS/AML patient and projected over our healthy reference map. Coupled reference and projected (patient) metacells were then used for calculating expression ratios over all expressed genes in all chromosomes. Left – heatmap of log2 expression fold-change (patient/reference) per metacell over all genes expressed in each chromosome. Expression is first summed over all genes expressed in a certain chromosome by either metacell and ratios (patient sum/reference sum) are then calculated for each metacell couple. Right – log2 fold-change expression (healthy/patient) for all genes expressed by either metacell across all chromosomes. Red lines represent the median of each chromosomal fold-change distribution. **9b,c** - same as 9a for AML-1 cases. scRNAseq karyotyping identified two clones: a smaller clone (AML-1-1, c top) with normal karyotype, and a larger clone (AML-1-2, c bottom) with +9,+10,+22 and Del20. **9d,e** - same as 9a for AML-2. scRNAseq karyotyping identified a single clone with +8,+11,+13,+14 in all metacells (no population substructure).

# EDF10



# EDF 10 – The transcriptional landscape of AML: aberrant expression of reference genes and multiple novel gene signatures

**10a** – same as Fig 5e for *LMNA* expression across AML-1 and AML-2 clones. Top - boxplots of *LMNA* signature distributions for different cell states in our reference atlas. Bottom – *LMNA* signature density plots for each of the AML clones showing high variability in *LMNA* expression. **10b** – Expression heatmap of several gene signatures across reference cell states and AML subclones. The malignant state differs greatly from the healthy state both in the expression of reference genes and by multiple additional gene expression signatures. **10c** – panels compare gene expression of major differentiation regulators associated with lymphoid (top) and MPP / MEBEMP (bottom) differentiation, in reference (healthy) (left), AML-1 (middle, color coded by subclones) and AML-2 (right) metacells.