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

Furer N. [*1], Rappoport N. [*2,3], Milman O. [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) deliver life-long multi-lineage output. However, with aging, we exhibit certain characteristic blood count changes and accumulation of clonal disorders. Better understanding of inter-individual variation in HSPC behavior is needed to understand these age-related phenomena and the transition from health to chronic and acute hematological malignancies. Here we study 627K single circulating CD34+ HSPCs (cHSPCs) from 148 healthy individuals, along with their clinical information and clonal hematopoiesis (CH) profiles, to characterize population-wide and age-related hematopoietic variability. Individuals with CH were linked with reduced frequencies of lymphocyte progenitors and higher RDW. An age-related decrease in lymphoid progenitors was observed, predominantly in males. Inter-individual transcriptional variation in expression of a Lamin-A signature and stemness gene programs were linked with aging and presence of macrocytic anemia. Based on our model for healthy cHSPC variation we construct the normal reference for cHSPC subtype frequencies. We show how compositional and expression deviations from this normal reference can robustly identify myeloid malignancies and pre-malignant states. Together, our data and methodologies present a novel resource, shedding light on various age-related hematopoietic processes, and a comprehensive normal cHSPC reference, which can serve as a tool for diagnosing and characterizing hematological disorders.

50 **Introduction**

51

52 The basis for understanding and defining human pathophysiological states is a detailed
53 description of inter-individual heterogeneity among healthy individuals. Large population
54 studies have identified wide inter-individual differences in complete blood counts (CBCs) of
55 healthy individuals[1] and exposed different age-related blood count changes, such as high RDW,
56 macrocytic anemia, and a reduction in absolute lymphocyte counts[2]. The establishment of
57 reference values, or population-wide normal ranges for certain blood parameters, has been
58 crucial for patient evaluation, diagnosis and treatment. This is especially true in the field of
59 hematology, where multiple parameters are measured simultaneously, and diagnosis is usually
60 dependent upon the coordinated interpretation of more than one test[3].

61 While CBC reference ranges are used in the clinic daily, the equivalent reference range
62 for hematopoietic stem and progenitor cells (HSPCs) has not been established so far. As HSPCs
63 mainly reside in the bone marrow (BM), access to these cells, especially in the healthy
64 population has been problematic, whereas their general paucity in the circulation made it quite
65 challenging to characterize them efficiently from the blood. This has become feasible given
66 modern technologies such as single cell RNA sequencing (scRNAseq), allowing high resolution
67 expression-based quantification and characterization of these rare cellular states.

68 Individual heterogeneity in the frequency of circulating HSPCs (cHSPCs) has been
69 reported in the past, and was linked to age, smoking, sex, and hereditary factors[8], as well as to
70 different pathological states[9]. Few studies analyzed HSPC heterogeneity in higher resolution, but
71 their sample size was limited[10]. Previous studies have demonstrated that most HSPC
72 subpopulations can be identified in the PB[11], including some based on scRNAseq analysis[12].
73 Functional stem cells were identified in the PB of mice[13] and humans[12], but, as the PB connects
74 the BM to other extramedullary sites, it could also be enriched in unique stem cell populations[9].

75 Here we apply scRNAseq analysis to cHSPCs from 148 healthy age- and gender-diverse
76 individuals, to capture key aspects of interindividual heterogeneity. We describe transcriptional
77 programs of 626,966 single CD34+ cells and how these correlate with various clinical attributes.
78 This gave rise to a high-resolution model of cHSPCs, including a spectrum of states, from
79 HLF/AVP positive hematopoietic stem cells (HSCs), through early common myeloid and
80 lymphoid progenitor states, and more specified erythrocyte, basophil-eosinophil-mast, GMP, B,
81 NK/T and DC progenitor populations. All data can be observed in

82 https://apps.tanaylab.com/MCV/blood_aging/. Notably, we discovered extensive inter-
83 individual heterogeneity in the frequency of specific cHSPC subtypes and found these to
84 correlate with certain blood count parameters, presence of CH and aging. We further developed
85 various tools for evaluating patient blood through compositional analysis (equivalent to the
86 classical CBC, but at the HSPC level), followed by in-depth analysis of patient gene expression
87 signatures, and karyotype analysis. These tools greatly enhance and streamline the analysis of
88 hematological disorders, as we show in several case studies.

Results

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

To evaluate interpersonal diversity in subtype distribution and regulation of cHSPCs in healthy humans, we combined multiplexed scRNAseq with genotyping, and integrated clinical data. Multiplexing was resolved using SNPs we identified in the 3' UTR of cHSPC RNA, facilitating precise matching of cells to individuals, and improving control for batch effects and doublets (**Fig 1A**). Altogether, we collected HSPCs from 79 males and 69 females between the ages of 23 and 91 years (median 61.5) (**EDF 1A**, **Supplementary Table S1**). We ran technical replicates on 39 individuals, and biological replicates on a follow-up cohort of 20 individuals (1 year following their original sampling date). We collected longitudinal CBCs up to 5 years prior to scRNAseq along with various other clinical parameters, and performed deep targeted somatic mutation analysis on DNA produced from their blood at sampling, to identify cases of CH (**Supplementary Tables S2, S3**)[14]. Following quality control and filtering, we retained 846,762 single cell profiles, which were normalized to control for sequencing-platform batch effects and combined to construct and annotate a metacell manifold model[15] (**EDF 1B,C**). We retained 672,000 CD34+ single cells for downstream analysis (**EDF 1D)**. These formed a rich repertoire of states, associated with cHSCs and their differentiation trajectories (**Fig 1B**, **EDF 1E,F**). The derived model recapitulated and deepened earlier characterization efforts of HSPC states from the BM, and while not fully reflecting BM dynamics, was compatible with our own BM scRNAseq data and that of others[16,17] (**EDF 2,3A**), suggesting that cHSPCs can serve as a highly accessible proxy for hematopoietic dynamics, both within and between individuals. One notable characteristic specific to cHSPCs, however, was the repression of cell cycle gene expression (**EDF 3B**). Importantly, we found our cHSPC model to be consistent among individuals. The median number of individuals contributing cells to each metacell was 84, and all metacells included cells from at least 47 individuals. Individual-specific differential expression was limited after controlling for each sample's cell distribution over the atlas states (**EDF 4A,B**).

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

One of the hallmarks of our cHSPC model is a distinct HSC state that is transcriptionally linked with two major differentiation gradients: The first representing a continuum of common

119 lymphoid progenitor (CLP) programs; The second, and more common branch, representing
120 multipotent progenitor (MPP) states and their differentiation toward granulocyte-monocyte
121 progenitors (GMP), erythrocyte progenitors (ERYP) and basophil/eosinophil/mast progenitors
122 (BEMP). Technical limitations of cell disassociation in scRNAseq prevented precise
123 megakaryocyte program modeling (**EDF 4C**). We therefore annotated states at the base of this
124 trajectory as megakaryocyte/erythrocyte/basophil/eosinophil/mast progenitors (MEBEMP) as
125 these are also presumed to be the cells of origin of megakaryocytes[12].

126 Early HSCs are marked by high *AVP* and *HLF* expression, and were previously shown to
127 represent a rare cell population enriched with self-renewal capacity in both BM and cord
128 blood[18]. Our model included data on ~14,440 *HLF*/*AVP* HSCs that could be matched with cells
129 from independent BM atlases[16,17], suggesting that under steady-state, HSCs with potential self-
130 renewal capacity are present in the PB (**EDF 5A**). Together with HLF and AVP, we discovered 14
131 genes expressed at least 1.75-fold higher in HSCs compared to their two immediate
132 differentiation branches (**EDF 5B, Supplementary Table S4**). We specifically identified several
133 transcription factors (TFs) enriched in HSCs, including the genes *HOXB5*, *TLE4,* and *GATA3* (**Fig**
134 **1C**). *GATA3* was previously reported to regulate self-renewal in mice long-term HSCs[19], yet its
135 role in human HSCs has not been studied thus far. We note that while the HSC state is defined
136 by unique markers that are symmetrically down-regulated upon exit to the CLP and MEBEMP
137 trajectories (**Fig 1C**), it also expresses several lineage-specific regulators at intermediate levels,
138 which are bifurcating anti-symmetrically upon exit from the HSC state to the CLP and MEBEMP
139 trajectories (**Fig 1D, EDF 5B**). This may suggest that the multipotent capacity of HSCs is
140 associated with intermediate expression of multiple regulators which is resolved with
141 differentiation.

142 **NK-T-dendritic and basophil-eosinophil-mast progenitors are enriched in cHSPCs**

143 The cHSPC atlas was enriched for basophil-eosinophil-mast progenitors (BEMP), mapped
144 as one possible terminus of HSC differentiation. While classical studies linked these cells with a
145 granulocyte/monocyte progenitor (GMP) origin, more recent studies suggested these emerge,
146 at least in part, from erythroid progenitors in both mice and humans[12,20]. Our analysis allowed us
147 to focus on a small population of metacells linking BEMPs with their MEBEMP-L precursors (**Fig**
148 **1E**). This highlighted TFs (**Fig 1F**) and other factors (**EDF 6A**) positively or negatively regulated in
149 this postulated early stage of BEMP specification. Another rare HSPC population we could now

150 focus on included lymphoid states with high *ACY3* expression and intermediate-to-low *DNTT*
151 levels, a combination rarely found in human BM but present in PB (**EDF 6B**). Interestingly, we
152 observed co-variation of key T cell regulators within this population, but also anti-correlation of
153 these factors with some hallmarks of a dendritic cell (DC) program. This can be demonstrated by
154 comparison of *TCF7* and *IRF8* expression (**Fig 1G**), and the matching *TCF7*-coupled dynamics of
155 *CD7*, *MAF*, and *IL7R*, or *IRF8*-coupled dynamics of the myeloid TF *SPI1* (PU.1) and multiple MHC-
156 II genes (**Fig 1H, EDF 6C**). We therefore termed this population NK/T/DC progenitors
157 (NKTDP)[21,22]. To summarize, our map of cHSPCs showed a rich spectrum of differentiation
158 trajectories and progenitor states that refined previous analyses, and a remarkable universality
159 of states, which provided an opportunity for deciphering inter-individual hematopoietic
160 variability.

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

162 To study inter-individual cHSPC variation, we first looked at individual-specific cell state
163 compositions. This was performed by quantifying cell state relative frequencies within each
164 individual's single-cell ensemble **(Fig 2A)**. These frequencies varied extensively as shown in **Fig**
165 **2B**. For example, HSCs and CLP-Ms, representing 2.4% and 12.6% of the CD34+ population on
166 average, respectively, showed standard deviations of 1.0% and 6.8%, respectively. The abundant
167 MPP and MEBEMP-E states (mean frequency of 20.7% and 37.6%, respectively) showed smaller
168 relative variations (SD 4.9% and 5.8%, respectively). To analyze the stability of cell state
169 frequencies over time and sampling instances, we re-sampled 20 individuals one year following
170 their original sampling date. Both lymphoid progenitor frequencies (CLP-M, CLP-L, NKTDP), and
171 MEBEMP (MEBEMP-E, MEBEMP-L, ERYP, BEMP) frequencies were stable within the same
172 individual across time (**Fig 2C**).

173 To analyze composition in higher resolution, we profiled each individual's enrichment
174 over the CLP and MEBEMP trajectories. Clustering of these enrichment profiles yielded six
175 archetypes of cHSPC composition within the healthy population (class I-VI, **Fig 2D**). These were
176 composed of individuals with relative lymphoid enrichment (class I-II) or depletion (class V-VI)
177 further sub-divided by a stemness gradient, enriched in classes II, IV and VI, and depleted in
178 classes I, III and V. Analysis of technical and biological replicates confirmed this variation to be
179 robust and patient specific (**EDF 7A,B**). To summarize, we provide the first cHSPC subpopulation
180 normal reference range (**Fig 2B**), characterized by extensive variation among healthy individuals,

181 and show these compositional differences are a true individual characteristic, with potential
182 clinical implications.

183 **cHSPC frequencies correlate with CBCs and CH**

184 Analysis of CBC correlations with our single-cell atlas enhanced our previous findings on
185 the inter-individual variation in cHSPC compositions. All CBC correlation analyses were
186 performed using median values for each blood count parameter over 5 years preceding
187 scRNAseq. The mean and median number of blood counts per individual during this 5-year
188 period were 8, and 6 respectively. We observed a significant positive correlation ($P<0.01$)
189 between PB mature lymphocyte percentages and CLP frequencies (**Fig 2E**, left). Given the very
190 high variation in female red blood cell counts and sizes during young adulthood (with menarche
191 and pregnancy effects) as well as during prolonged periods of perimenopause, we analyzed RBC
192 indices, including RBC, HCT, MCV, and RDW, separately in males and females. We observed a
193 significant negative correlation ($P<0.02$) between CLP frequencies and Hematocrit (HCT, males,
194 **Fig 2E -** middle), as well as a significant positive correlation ($P<0.01$) between increased RDW - a
195 cHSPC myeloid bias - and a relative CLP depletion (males, **Fig 2E -** right). Our previous work[24] and
196 the work of others[25] correlated increased RDW with high risk for CH and predisposition to AML.
197 We demonstrate that low CLP frequencies are associated with CH (two-sided Mann-Whitney
198 test; **Fig 2F, EDF 8A**), and further enhance this observation by performing Genotyping of
199 Transcriptomes[26] on one of our DNMT3A R882 cases, identifying a lower fraction of CLP cells in
200 the mutant clone ($P<0.005$, Fisher's exact test, **Fig 2G**). To further explore this association, we
201 studied a cohort of 18,147 healthy individuals for whom we had both longitudinal CBCs and DNA
202 available. We identified 602 individuals with a high RDW (>15%, not meeting minimal criteria for
203 MDS) and 602 age and sex matched normal RDW controls. We performed deep targeted
204 sequencing to identify pLMs on both high-RDW individuals and controls and found a significant
205 enrichment of CH+ cases in the high RDW group (Fisher's exact test $P<0.002$, **Fig 2H**,
206 **Supplementary Tables S5,S6**). Altogether, the data demonstrate a 3-way linkage between
207 decreased CLP frequencies, a high RDW, and CH.

208 **Age-related myeloid bias is predominantly observed in males**

209 Blood aging is a complex and multi-factorial process, likely driven by intrinsic factors
210 such as pre-leukemic mutations, and extrinsic effects, such as cytokine and hormonal changes.

211 In order to decouple these factors as much as possible, we studied age-related changes in cHSPC
212 populations in individuals without CH mutations. Analysis of age-linked compositional changes
213 in cHSPCs within this group showed a remarkable increase in myeloid (MEBEMP) to lymphoid
214 (CLP) ratios in males (when comparing <50 to >60-year-old individuals, **Fig 3A**). This effect was
215 not significant in females. In this regard it is important to note that a decline in lymphocyte
216 counts can be observed in both elderly males and females, however it appears in females at an
217 older age[2]. Interestingly, females exhibit a surge in lymphocyte counts immediately following
218 menopause, contributing to this delay in lymphocyte decline. Within the MEBEMP
219 differentiation trajectory, aging was correlated with over-representation of more differentiated
220 states, once again only in males (**Fig 3B**). Of note, the frequencies of cHSCs did not significantly
221 change with age (**Fig 3C**). While previous studies suggested aging is linked with an increase in
222 HSC frequencies[27], such increase was not observed with the restrictive definitions employed
223 here, as well as when determined from CD34+ PB HSPC frequencies in a recent cohort of 1000
224 healthy individuals undergoing PBMC scRNAseq[28] (**Fig 3D**). The sex-specific correlation between
225 age and cHSPC myeloid bias highlights the role of such non-intrinsic effects on this classical
226 hallmark of blood aging.

227 **Composition-controlled HSPC expression correlates with age**

228 As shown above, an individual's cHSPC composition provides an initial blueprint of
229 hematopoietic dynamics along the stemness and CLP/MEBEMP axes. Further analysis of
230 transcriptional variation could now be carried out, while controlling for the dominant effect of
231 cHSPC composition, in order to characterize additional gene expression signatures that could
232 distinguish between individuals. Composition-controlled individual expression profiles showed
233 high information content when correlated with age, enabling age prediction based on
234 normalized expression alone (**Fig 3E**, **EDF 8B; see Supplementary Tables S7,S8 for additional**
235 **screening for age-, CBC-, CH- and sex-associated gene expression**). We next looked for gene
236 groups (signatures) that co-variate between individuals, filtering out sex-linked signatures and
237 those showing strong batch effects. The most prominent of these signatures included *Lamin-A*
238 *(LMNA)* as well as ANXA*1*, *AHNAK*, *MYADM, TSPAN2,* and *VIM,* among others (**Fig 3F**, **EDF 8C,D**,
239 **Supplementary Table S9)**. Individual LMNA signature expression varied across a range of more
240 than 2-fold (**EDF 8E**), exhibiting high expression variability in HSCs and early myeloid and
241 lymphoid cell states, and a homogeneously low expression in late MEBEMPs and CLPs (**EDF 8F**).

242 Individual LMNA signature expression was consistent in myeloid and lymphoid cell states (**Fig**
243 **3G**) and was stable in our follow-up cohort (**EDF 8G**). Interestingly, we observed an age-linked
244 increase in LMNA signature expression in lymphoid, but not myeloid, cHSPCs (**Fig 3H**). Taken
245 together, we show that in addition to the accumulation of pLMs in HSPCs, aging is strongly
246 linked with changes in the distribution of progenitor cell states in the PB, and with significant
247 differences in the expression of certain gene signatures. The mechanistic basis for this variation
248 and its clinical impact remain unresolved.

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

251 The differentiation of HSPCs toward MEBEMP and CLP fates involves coordinated
252 activation and repression of specific transcriptional programs that were generally universal
253 among individuals. Yet, our screen for inter-individual variation in gene signatures suggested
254 that individuals differed in the way they synchronized the opposing effects of these stemness
255 and differentiation programs. To quantify this variation, we compared *AVP* (stemness) and
256 *GATA1* (MEBEMP differentiation) signatures (**Supplementary Table S10**) on a 20x20 bin
257 expression matrix (**Fig 3I**). While most individuals displayed dynamics close to the diagonal line
258 (individuals #16, #86, for example), following the typical transition from stemness to
259 differentiation, some individuals deviated from the diagonal, indicating skewed synchronization
260 between the *AVP* and *GATA1* signatures. We quantified this deviation (i.e. off-diagonal
261 frequency) using a new synchronization-score. This facilitated the identification of individuals
262 with sync-scores as low as 0.12 (#122 and #172, for example, **Fig 3I**, top), indicating delayed
263 activation of *GATA1* relative to *AVP* repression. Namely, while these individuals rapidly reduce
264 their *AVP* expression, their increase in *GATA1* and *GATA1*-related genes is delayed. In contrast,
265 individuals exhibiting a high sync-score (#98 and #121, for example, **Fig 3I**, bottom), show early
266 activation of *GATA1* expression, which precedes *AVP* repression. We detected stability of the
267 sync-score in our follow-up cohort (**EDF 8H**). Inter-individual sync score variability was positively
268 correlated with RBC levels, and consistently anti-correlated with MCV in males ($P<0.01$
269 (spearman) for both RBC and MCV, **Fig 3J**). Analysis of the correlation between individual sync-
270 scores and cHSPC compositions in males demonstrated a negative correlation with ERYPs and
271 BEMPs (**Fig 3K**). To summarize, we demonstrated variation in the coordination of stemness and
272 MEBEMP differentiation programs that is correlated with red blood cell counts and volumes.

273 Hypotheses on the mechanisms underlying this heterogeneity and its greater clinical
274 consequences should be further explored. Individual scores and signatures (LMNA signature in
275 MEBEMPs and CLPs, sync-scores, and HSPC cell state distributions) are summarized in
276 **Supplementary Table S11**.

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

278 Diagnosis of myeloid malignancies requires the identification of clonal markers
279 (mutations / structural variants) and the detection and quantification of blasts by microscopy
280 and flow cytometry. In **Fig 4A** we describe a new stepwise approach for analysis of myeloid
281 disorders based on sampling of cHSPCs and comparison of their compositions, normalized
282 expression and copy number variations (karyotyping) to our new normal reference. As proof of
283 concept, we analyzed sampled cells from 2 new healthy individuals, 3 CMML patients, 2 MDS
284 patients, 1 MDS/MPN overlap patient and 1 myelofibrosis patient. We further sampled and
285 analyzed 2 AML cases to demonstrate how acute disease is manifested when projected over the
286 cHSPC model (**EDF 9-11**). Projection of patient metacells showed a high gene expression
287 correlation between metacell pairs (**Fig 4B**, color coded dots), for all pathological cases except
288 for the AMLs. Cell state compositions were, however, skewed in all pathological cases (**Fig 4C,D**),
289 but not for the healthy controls. All patient samples showed a remarkable reduction in CLP
290 populations (**Fig 4C**,**D**). Two of the CMML patients (N192, N235) demonstrated highly abnormal
291 enrichment of specific (basophil and myeloid) cell states, while the rest showed a relatively
292 balanced distribution over the MEBEMP differentiation spectrum (with slight enrichment of
293 stem states). Compositional-controlled gene expression comparisons of patient samples to the
294 normal model identified specific genes that were recurrently induced or repressed in disease
295 (**Fig 4E**). While both healthy individuals and the treated MDS case showed a minimal number of
296 DE genes, all leukemic cases exhibited a substantial increase in the number of DE genes when
297 compared to the healthy population model (**Fig 4E**).

298 Detection of karyotypic abnormalities based on gene expression dosage effects, previously
299 suggested and implemented in several tools[29], can be readily implemented on cHSPCs, as shown
300 in **Fig 4F** and **EDF 9A**,**B**, where clear chromosomal aberrations are observed in the two AML
301 cases analyzed. We further detected deletion of the long arm of chromosome 5, del(5q) in one
302 of our MDS samples with complete cytogenic remission following lenalidomide treatment
303 (N211A & N211B, **Fig 4F**, **EDF 9A**). To conclude, the atlas of normal cHSPC states presented

304 herein enables high resolution characterization of myeloid disorders, based on their cell states,
305 compositional-controlled transcriptional variance, and abnormal karyotypes.

306 **Discussion**

307 This study characterizes interindividual heterogeneity in circulating HSPCs across 148
308 healthy individuals, analyzing 627K PB CD34+ cells via scRNAseq. The magnitude of our cohort,
309 along with the potency and resolution of modern single cell technologies and the computational
310 methods used here, allowed us to characterize in detail the transcriptional programs of diverse,
311 sometimes rare (NKTDP, BEMP), HSPC sub-populations, refining and augmenting previous
312 findings from much smaller cohorts (**Fig 1**). We define a normal reference range for cHSPC
313 subpopulation frequencies within a large age- and sex-diverse healthy population and show that
314 cHSPC subtype compositions are highly variable between individuals, while the cell states
315 themselves are remarkably universal (**Fig 2**). These compositions remained stable over a one-
316 year follow-up period, placing them as a strong individual characteristic. Future studies will need
317 to further explore and better define the mechanistic and genetic basis for this compositional
318 heterogeneity. At the population level we discovered a significant correlation between low CLP
319 frequencies, CH and increased RDW (**Fig 2**). We further show that the known age-related
320 myeloid bias in HSPCs is significantly male driven (**Fig 3**). Analysis of composition-controlled
321 transcriptional variance identified an RNA expression clock which correlated with age (**Fig 3**),
322 and a novel gene module (*LMNA*) whose expression increased in CLPs with age. We discovered
323 that healthy individuals regulate the transition from stemness to myeloid differentiation
324 differently and found that premature down-regulation of stemness genes correlates with high
325 MCV anemia (**Fig 3**). We conclude by demonstrating how this resource can be used to effectively
326 identify and diagnose pathological cases based on their abnormal cHSPC subpopulation
327 frequencies, differential gene signatures, and chromosomal aberrations (**Fig 4**), all from PB
328 samples.

329 The nature and magnitude of our cohort allowed us to produce **Fig 2B**, which in our opinion,
330 presents the key finding of this study - the first circulating HSPC subtype normal reference range
331 in a healthy population, which is both age and sex diversified. Though the relation between
332 circulating HSPCs and those found in the BM has not been completely characterized yet, our PB-
333 BM comparisons provide evidence for the consistency of our model. Our data demonstrates that
334 cHSPCs resemble their BM counterparts transcriptionally, with the exception of lower cell cycle
335 gene expression (**EDF 2,3**). Importantly, we are *not* arguing that cHSPCs are (or should be) a
336 complete model for hematopoiesis in the BM, but that they are nonetheless useful proxies for

337 major processes driving hematological physiology and pathology – and in this context - their
338 accessibility and robustness are of major utility. Inter-individual differences in cHSPC
339 compositions and states can serve as an indispensable mean for capturing key aspects of a
340 patient's hematopoietic state. The relevance and importance of this normal reference can
341 perhaps be better understood in view of the normal CBC reference range, as only once a
342 population-wide reference was set for CBCs, could deviations from it be considered
343 pathological. In **Fig 4** we provide a detailed pipeline for the identification and characterization of
344 such pathologies based on our normal reference. We fully acknowledge that our reference
345 model is not universal, that it should be explored in diverse populations and in a wider range of
346 pathologies, and that as a resource it surely can and will be further improved.

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

353 **Methods**

354 **Sample procurement and handling**

355 We collected fresh peripheral blood samples from 148 healthy individuals (79 males, 69
356 females) aged 23-91. Their demographics and molecular data are presented in **Supplementary**
357 **Tables S1 and S3** (Basic participant demographics and CH status). All sample donors were
358 considered healthy, their CBCs were within normal range, and they were not known to have any
359 CH defining mutations prior to sequencing. Written informed consent allowing access to
360 longitudinal CBCs and sequencing data (CH and genotyping panels) was obtained from all
361 participants in accordance with the Declaration of Helsinki. All relevant ethical regulations were
362 followed, and all protocols were approved by the Weizmann Institute of Science ethics
363 committee (under IRB protocol 283-1).

364 Recruitment was intended to allow characterization of the normal variation in circulating HSPC
365 states. As no such profiling had been previously performed, we could not assume much
366 regarding the variance in the population a priori. We therefore aimed to profile responsive
367 volunteers with normal blood counts, balancing sex and seeking a dispersed age distribution
368 biased toward older individuals. We reassessed this strategy following initial sampling, and
369 observed remarkable homogeneity in transcriptional states across individuals sampled from our
370 immediate community as well as from HMO out-patient clinics, emergency medical centers,
371 hospital wards etc. This was critically important, as it confirmed the universality of our model
372 across individuals. We observed large variation in the composition of states across healthy
373 individuals, which suggested that more sampling, without specific preferences, would be
374 appropriate for further characterizing compositional variation. Our results support two
375 characteristics of human cHSPCs that we believe to be nontrivial and robust for alteration in
376 sampling strategies: the universality of molecular cell states among individuals (allowing one to
377 use the model as a reference for any patient), and the high degree of variation in composition
378 between healthy individuals (allowing one to look at interesting clinical populations vis-à-vis our
379 model for normal variation).

380 50 ml of PB were drawn from each individual into lithium-heparin tubes. 1 ml of blood was used
381 for DNA production, and the remaining volume was used for PBMC isolation via Ficoll, using
382 Lymphoprep filled Sepmate tubes (StemCell technologies), followed by CD34 magnetic bead-

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

393 **ScRNAseq of CD34+ PBMCs**

394 Single cell RNA libraries were generated using the 10x genomics scRNAseq platform (Chromium
395 Next Gem single cell 3' reagent kit V3.1). Chip loading was preceded by flow-cytometry to verify
396 that enrichment was successful, and that enough CD34+CD45^{int} live cells were gathered. All
397 blood samples were freshly drawn at the Weizmann Institute of Science on the morning of each
398 experiment day, and time from blood draw to 10x loading was restricted to 5 hours. The
399 motivation for working with fresh samples was based on our previous experience with PB CD34+
400 cells being vulnerable to freezing/thawing rounds and long manipulation times.

401 10x libraries were sequenced on two alternative platforms (Illumina / Ultima Genomics). 12
402 libraries were simultaneously sequenced on both platforms for comparison purposes and in
403 order to demonstrate the scalability of our approach. We observed the Ultima-sequenced data
404 to be highly similar to the Illumina-sequenced data (**EDF 7A**).

405 **Genotype-based demultiplexing**

406 All cells were traced back to their sample of origin using genotype-based de-multiplexing. This
407 method allowed pooling of blood samples immediately following extraction of the DNA aliquot,
408 such that CD34-enrichment was performed on the entire pool of PBMCs produced. The use of
409 SNP-based multiplexing has several advantages to alternative antibody-based cell hashing
410 methods: 1) it is extremely cost effective, such that the cost of sequencing a single individual on
411 a 2000 SNP Molecular Inversion Probe (MIP) panel at a depth of 1000X per SNP (adequate for
412 de-multiplexing purposes) is several folds cheaper than antibody staining, 2) genotyping

413 eliminates the need to keep samples separated prior to loading, it entails shorter handling times
414 and less cell manipulation, as it does not require antibody incubation periods and multiple wash
415 centrifuges. This was very evident in cell viability prior to chip loading. As with other methods of
416 sample multiplexing, genotype-based multiplexing allows for robust doublet detection during
417 data analysis, which enabled loading of 30-40K cells from between 4-6 individuals on each
418 Chromium Chip lane, yielding 15-25k cells per library.

419 **Molecular Inversion Probe (MIP) panels**

420 Both our CH and genotyping panels are Molecular inversion probes (MIP)-based panels
421 described in detail previously[14]. Our CH panel contains 705 probes, covering pre-leukemic SNVs
422 and Indels in 47 genes, and is complemented by 2 amplicon sequencing reactions to cover GC
423 rich regions in *SRSF2* and *ASXL1*. Our genotyping panel allows for the simultaneous detection of
424 >2000 common genetic variants, all of which are extensively covered in all cell types in our data.
425 It includes heterozygous sites with at least 5% minor allele frequency from the 1K genomes
426 project, which were highly covered by RNA molecules in our data (at least 80 UMIs across all
427 cells in a test 10x library), excluding sites in repetitive elements and in sex chromosomes. Both
428 panels were designed using MIPgen[30] to ensure capture uniformity and specificity.

429 **Variant calling and identification of ARCH mutations**

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

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

433 In order to compare propensity for CH and high risk CH mutations[24] in high RDW cases and
434 normal RDW controls, we performed deep targeted sequencing of DNA samples from 602 high
435 RDW (>15%) individuals, who did not show signs of anemia and whose blood count did not meet
436 MDS criteria (11.5g/dL≤Hg≤15.5g/dL [F], 13g/dL ≤Hg≤17g/dL [M], 80fL≤MCV≤96fL,
437 PLT≥100×10^9/L, Abs Neut≥1.8×10^9/L), and 602 normal RDW (11.5g/dL ≤Hg≤15.5g/dL [F],
438 13g/dL≤Hg≤17g/dL[M], 80fL≤MCV≤96fL, PLT≥100×10^9/L, Abs Neut≥1.8×10^9/L), age and gender-
439 matched controls. Case-Control matching was performed using the R MatchIt package, balanced
440 on age and gender, method = "nearest", ratio = 1, from a total of 18,147 individuals with
441 longitudinal blood counts and available DNA. All DNA samples and corresponding blood counts
442 were received de-identified from the Tel Aviv Sourasky Medical Center (TASMC) Integrative

443 Cancer Prevention Clinic. All DNA samples were collected after obtaining written informed
444 consent and in accordance with the Declaration of Helsinki. All relevant ethical regulations were
445 followed, and all protocols were approved by the TASMC ethics committee (under IRB protocol
446 02-130). CBCs and sequencing results of all cases and controls are presented in **Supplementary**
447 **Tables S5, S6.**

448 **scRNAseq processing**

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

451 **Doublet calling**

452 We performed several steps to assign cells to their individuals and to detect doublets. The
453 pipeline is made of the following steps:
454 1. Demultiplexing cells and calling doublets based on SNPs found in the scRNAseq data
455 2. Building a metacell model using cells from all the libraries, including cells previously marked
456 as doublets, and identifying metacells made of doublet cells.
457 3. Identifying metacells with doublet cells based on expression of marker genes.
458 4. Building the final metacell model and marking metacells as doublets based on expression
459 markers.

460 In the first step, we identify doublets and assign cells to individuals using Vireo and Souporcell,
461 which cluster cells based on SNPs found in sequenced RNA molecules. We executed Vireo[26]
462 (preceded by running cellsnp) and Souporcell[27] on each library separately. Both methods used
463 SNPs from our genotyping panel[15] which were covered by at least 20 UMIs in the library (in
464 Souporcell – at least 10 from the major and minor allele each). We observed high agreement in
465 doublet calling between the two methods.

466 In the next step, we built a metacell model with cells from all libraries, including all biological
467 and technical replicates. This model included cells that we identified as doublets, and also
468 included cells from individuals that are not part of this study, but were multiplexed with our
469 libraries. The model was built with metacell2, with a target metacell size of 200 cells. We then
470 marked all metacells where at least 35% of the cells were already marked as doublets, and all
471 metacells that expressed unique markers of distinct cell types, as doublet metacells. All cells
472 that belonged to a doublet metacell were then marked as doublets.

473 We then built an additional metacell model (see below), without the cells that were marked as
474 doublets, and only with cells of individuals belonging to this study. In this model we further
475 filtered a small number of metacells that expressed key marker of distinct cell types as doublets.

476 **Assignment of cells to individuals**

477 Vireo[26] and Souporcell[27] both cluster cells based on SNPs found in the sequenced RNA, such that
478 cells in the same cluster belong to the same individual. We observed very high agreement
479 between the two methods in their assignment of cells into individuals. In two 10x libraries
480 where the two methods did not agree (due to individuals with a very small number of cells), we
481 reran the methods on a subset of the cells and a smaller target number of clusters. We mostly
482 used Vireo's clustering, but in some libraries we also assigned cells that were unassigned by
483 Vireo but were assigned to an individual by souporcell, if they matched well to the genotype
484 data (see below). For one library we took Souporcell's assignment, because of better matching
485 to the genotype data.

486 In the next step we assigned clusters of cells to the individual they originated from. To this end,
487 we correlated the genotypes of each cell cluster, as inferred by Vireo, to all genotypes we
488 measured using the MIP panel (using sites with sufficient sequencing depth in the MIP panel). As
489 a control, we performed matching against the MIP genotypes of all individuals in the cohort, and
490 not just individuals from one library. We observed in all cases a clear matching to a single
491 individual from the expected library. The assignment also correctly identified related individuals,
492 and the sex of the matched individual was confirmed by expression of XIST in the RNA data.

493 **Removal of droplets with Megakaryocyte signatures**

494 We identified droplets with complete or partial megakaryocyte expression (at least 5% of the
495 UMIs coming from a megakaryocyte gene program including PF4, PPBP and 131 additional
496 genes) due to their overall high doublet rate. We then built a metacell model from the retained
497 cells (not marked as doublets, confidently assigned to an individual, did not show
498 megakaryocyte expression).

499 **Correcting for sequencing platform bias**

500 Some of our 10x libraries were sequenced by an Ultima sequencer, and as most libraries were
501 processed through a standard Illumina pipeline, we wished to minimize batch effects related to

502 small variation in sequencing platform. To that end we used libraries that were re-sequenced in
503 both platforms. We calculated an Illumina-Ultima correction factor per gene as the mean log2-
504 fold change in expression of the gene across re-sequenced libraries. We then normalized each
505 Ultima-sequenced library by downsampling genes with at least 0.28 log2-fold Ultima
506 overexpression, and resampling genes with at least 0.2 Illumina overexpression. The
507 downsampling and resampling were performed for each gene independently, across all cells in a
508 library. The thresholds for downsampling and resampling were chosen such that the overall
509 number of UMIs per cells remained similar. 87 genes with at least 4-fold change between Ultima
510 and Illumina were excluded from further processing.

511 **Computing the reference metacell model**

512 The model was built with metacell2, with a target metacell size of 200 cells. We marked
513 forbidden genes such as histone genes, cell cycle related genes, ribosomal genes, stress
514 response genes (including e.g., FOS, JUN), sex-linked genes, genes with high or inconsistent
515 differences between Illumina- and Ultima-sequenced technical replicates, and other genes that
516 we found to have high technical variation. These genes were not used by metacell2 when
517 calculating gene-gene similarities, but were included in downstream analysis. We annotated the
518 metacells using known markers as illustrated in **EDF 1C**. We excluded from most downstream
519 analyses metacells from cell types with low CD34 expression (monocytes, B cells, T cells, NK
520 cells, DCs, endothelial cells), and metacells which expressed marker genes from distinct cell
521 types that we marked as doublets.

522 For visualization of the metacell manifold we used UMAP projection of the metacell expressioin
523 vector over genes with specific enrichment over cell types . The cell types considered were
524 either all cell types (for **EDF 1B**), or CD34+ cell types (for **Fig 1B**).

525 **BM comparison and projections**

526 We used three BM datasets: a dataset including two CD34-enriched bone marrow cells that we
527 collected for this study, the Human Cell Atlas (HCA) dataset[19], and a CD34+ bead-enriched BM
528 dataset from[30].

529 We built a metacell model for the two CD34-enriched BM samples we collected using metacell2.
530 Since these samples only collected one sample per library, we did not run Vireo and Souporcell.

531 We marked metacells expressing marker genes of distinct populations as doublets, and excluded
532 them from further analysis.

533 We have previously processed and annotated the HCA dataset in a metacell model. We
534 downloaded the Setty et al. sequencing data and processed it by running cellranger and created
535 a metacell model. To project our PB data, our BM data, and the Setty dataset on the HCA
536 dataset, we correlated between the HCA metacells and the projected metacells over genes
537 showing high variance in the HCA metacell model. We annotated each Setty metacell using the
538 mode of the 5 most correlated HCA metacells. We annotated our BM data using the mode of
539 the 5 most correlated HCA metacells, and using expression of gene markers. We projected
540 metacells (from each respective model) on the HCA UMAP using the mean x and y values of the
541 5 most correlated HCA metacells. To compare S-phase genes between the PB and BM (**EDF 3B**),
542 we calculated for each PB and HCA metacell its S-phase signature (mean expression of six cell
543 cycle genes: TYMS, H2AFZ, PCNA, MCM4, HELLS, MKI67), and plot the distribution of this score
544 across metacells for each cell type.

545 **HSC differentiation gene programs**

546 To visualize transcriptional dynamics in HSC cells, we sorted MEBEMP and CLP metacells based
547 on their AVP expression. To calculate differential expression between HSC and neighboring cell
548 types (**EDF 5B**), we calculated the geometric mean of each gene across HSCs, CLP-M and
549 MEBEMP-E metacells, and took the difference between HSC and MEBEMP-E, and between HSC
550 and CLP-M.

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

552 To create **EDF4 A,B** we compared each individual's pooled expression profile to a matched
553 expression profile based on the individual's distribution across metacells. We performed the
554 analysis separately for MPP / MEBEMPs (BEMP, EP, MEBEMP-E/L, GMP-E and MPP) and CLPs
555 (CLP-E/M/L, NKTDP). In each group of cell types, we downsampled each cell to have 500 UMIs
556 and summed the UMIs across all cells of each individual, normalized the sum to 1 and calculated
557 log2, to obtain the observed expression. To compute matched expression, we downsampled
558 each metacell to have 90K UMIs and summed all UMIs of the metacell each cell belongs to for
559 each individual. We normalized this matched expression to sum to 1, and took log2. For **EDF**
560 **4A,B** we plotted all genes that were expressed in either the observed or matched expression in

561 at least one individual (log2 expression > 2^-14.5 for MPP / MEBEMPs, > 2^-13.5 for the CLPs
562 which had less UMIs), and that had at least 2-fold change between observed and matched in at
563 least one individual. We excluded genes that exhibited strong batch effects.

564 **HSPC compositional analysis**

565 To explore variance in cell type composition between individuals, we first calculated the
566 distribution of each individual's cells across the CD34+ cell types. To perform compositional
567 analysis at higher resolution than cell types, we partitioned cells from CD34+ cell types into finer
568 grained bins. We used one HSC bin, four CLP bins, and ten MEBEMP / MPP bins, for a total of 15
569 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
570 based on decreasing AVP expression of their metacells, such that bins 2-4 had the same number
571 of cells. We similarly assigned MPP and MEBEMP-E/L cells into 10 bins based on AVP such that
572 these bins had an equal number of cells.

573 For **Fig 2D** bottom panel, we calculated the enrichment of each individual's cells in each bin
574 (log2 of the ratio compared to the median across individuals). We partitioned individuals into
575 three group with different CLP numbers based on each individual's mean enrichment across CLP
576 bins 2-4 – those with mean enrichment > 0.5 are high CLP, those with < -0.5 are low, and the
577 rest are intermediate. We next defined the stemness score as the ratio between the number of
578 cells in MPP / MEBEMP bins 1-5 and the total MPP / MEBEMP number (cells in bins 1-10).
579 Individuals with stemness score > 0.5 had enriched stemness. The combinations of CLP
580 enrichment and stemness define the six classes shown in the figure. For visualization we further
581 sorted individuals within each cluster based on their stemness score.

582 In some cases, cells from one blood sample were divided to two different 10x libraries. In this
583 case cells from both libraries were considered together to represent the individual. Biological
584 replicates were only considered when patients were sampled on two different days. For
585 technical replicates, we only considered cells that appeared in a specific 10x library, even if an
586 individual had other cells in another library from the same day. This policy holds for all analyses
587 related to biological and technical replicates in this study.

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

589 We used permutation tests to test the relation between cell type distribution and a label (age,
590 CBC, sync-score or *LMNA* signature). We sorted 11 CD34+ cell types from late MEBEMP

591 differentiation through HSC and to late CLP differentiation (cell types are displayed by this order
592 in **Fig 2B**). We looked at triplets of adjacent cell types in this ordering, and calculated for each
593 triplet the total frequency each individual has from these cell types, obtaining a vector of length
594 9 per individual. We then correlated each of these 9 sums to the label, and took the maximal
595 absolute value from all these correlation values as a test statistic. We repeated this process after
596 permuting the label 1000 or 10000 times, and used the test statistics from the permutations to
597 derive a p-value.

598 **Variably expressed gene modules**

599 We detected genes modules with high variance across individuals while controlling as much as
600 possible for compositional variation. This involved the following procedure, applied separately
601 for myeloid and lymphoid states (described first for myeloids):

602 A. we calculated for each individual the 5th percentile of the number of UMIs across their cells,
603 and downsampled the cells to this number. We then pooled all downsampled cells for each
604 individual from the MPP metacells, normalized to sum to 1 and took log2. This gave us the
605 observed expression profile of each individual.

606 B. We created an expected expression profile for each individual as follows. We partitioned the
607 MPP metacells into 30 bins based on their AVP expression, and downsampled the metacells to
608 90K UMIs. For each of the 30 bins, we averaged the expression of all genes across downsampled
609 metacells in the bin. This defined an expression profile for each of the 30 bins. To obtain an
610 individual's expected expression, we calculated a weighted mean of the 30 bins' expression
611 profiles, where the weight of each bin is proportional to the fraction of the individual's cells
612 from that bin, normalizing to sum to 1 and taking log2. We then calculated the difference
613 between the observed and expected expression profiles.

614 C. Our data show some batch effect distinguishing samples collected in two calendric periods. As
615 this effect could introduce co-variation between genes across individuals, we applied a
616 correction controlling for it. This was done using a linear model fitting each gene to the patient
617 sample collection period, its age and sex. We then subtracted the inferred period factor from
618 the samples that were collected in the second period. We found that this approach significantly
619 reduced emergence of gene clusters linked with calendric sample collection bias.

620 D. We screened for genes with high variance that were unlikely to be affected residually by the
621 main manifold differentiation process. We removed genes with high batch effects, genes with
622 high AVP correlation (absolute value Pearson correlation > 0.65), and genes highly correlated
623 (absolute value Pearson correlation > 0.5) with a module of genes that were differential
624 between the first and second library collection periods. We then calculated each gene's variance
625 in the difference between the observed and expected expression across individuals. As some of
626 the variance can be explained due to sampling noise, we plotted each gene's variance across
627 individuals compared to its mean expression across individuals. We sorted genes by this
628 expression value and subtracted from the variance of each gene a rolling mean of the variances
629 of 100 neighboring genes in that ordering. We chose genes with variance at least 0.08 higher
630 than the rolling mean variance.

631 E. We calculated a gene-gene Spearman correlation matrix for high variance genes, and
632 clustered the correlation profiles using hierarchical clustering. We removed genes with low
633 mean correlation (< 0.2) to their cluster's genes, and then removed gene clusters with low mean
634 correlation between their genes (≤ 0.25 mean correlation of all gene pairs). We also computed
635 the gene-gene correlations while using only samples from our first library collecting period, and
636 required gene clusters to have high mean correlation (> 0.25) between their genes using only
637 these early samples. We removed additional gene modules that we observed were due to batch
638 effect, or the represented traces of MEBEMP differentiation that were not normalized by
639 approach. This resulted in **Fig 3F**.

640 We performed a similar (steps A-E) analysis for CLPs (**EDF 8C**), with a few differences. The
641 analysis included cells from CLP-M metacells. The cells were partitioned into 6 bins, and the
642 partitioning was based on the average of their DNTT and VPREB1 expression. Genes with high
643 absolute correlation to the average of DNTT and VPREB1 were excluded. After clustering the
644 gene-gene correlation profiles and removing genes as described for the MEBEMPs, gene clusters
645 with remaining correlation to CLP differentiation were removed.

646 **Age regression**

647 We developed age regression models for MEBEMP and CLP expression separately. To predict
648 age, we used the difference between an individual's observed and expected gene expression as
649 described above. We only used genes whose minimal expression across individuals is at least 2^-

650 14.5 for MEBEMPs and 2^-15.5 for CLPs. We trained a LASSO model using nested leave-out-
651 sample-out cross validation: For each sample we left out, we performed a leave-one-sample-out
652 cross validation on the remaining samples to select LASSO's λ parameter, trained a model using
653 the selected λ and made a prediction on the left-out sample.

654 **LMNA signature**

655 We used the difference between an individual's observed and expected gene expression, and
656 correlated all genes to LMNA separately in the MEBEMP expression and CLP expression
657 matrices. We then summed the MEBEMP and CLP correlation values, and kept genes whose sum
658 was > 0.7. We further removed genes that exhibited high technical variance, and were left with
659 17 genes in the signature. To calculate each individual's LMNA signature, we calculated for each
660 individual the average value of the 17 genes in the observed – expected matrix, for MEBEMPs
661 and CLPs separately.

662 To plot **EDF 8E**, we calculated the geometric mean of the LMNA signature genes for each
663 individual separately in each one of the 10 MPP / MEBEMP bins described earlier for **Fig 2D**.

664 **GoT Analysis**

665 GoT[24] performed on sample #122 allowed us to mark this individual's cells as wild-type or
666 mutated. Due to the low VAF of #122's DNMT3A mutation, and in order to increase power, we
667 marked cells whose DNMT3A mutation status could not be determined by GoT as wild-type
668 cells. For **Fig 2G**, we examined #122's cells' distribution across cell types.
669

670 **Sync-score**

671 We defined the AVP signature to include genes with high correlation (> 0.6) to AVP across HSC,
672 MPP and MEBEMP metacells, and the GATA1 signature to include those with correlation > 0.7 to
673 GATA1. We filtered genes with mean relative expression > 2^-10 in these metacells, to preclude
674 a small number of genes from dominating the signatures. We then selected HSC, MPP,
675 MEBEMP-E and MEBEMP-L cells, scored them by the fraction of UMIs from the AVP and GATA1
676 signatures, and partitioned them into 20 bins of AVP signature expression and 20 bins of GATA1
677 signature expression, such that all AVP bins and all GATA1 bins had the same number of cells.
678 The sync-score is then defined as the fraction of cells in GATA1 bin 13 and above (upper two

679 quintiles of GATA1) that are in AVP bin 9 and above (upper three quintiles of AVP expression).
680 To visualize the sync scores (**Fig 3I**), we normalized the 20x20 bin matrix to sum to 1, smoothed
681 the obtained matrix by averaging cells using a running window of length 3, and taking log2.

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

683 Differential expression was performed separately for MPP / MEBEMP cells, and for CLPs. The
684 MPP and CLP-M matrices that were used to detect variant gene modules were used here for
685 differential expression. Differential expression was performed separately for males and females.
686 Each gene's expression value was correlated with max VAF, age and 20 CBC values using
687 Spearman correlation, and the correlation was tested for significance. p-values were FDR-
688 corrected (Benjamini-Hochberg) for each label separately. For max VAF we also performed a
689 Mann-Whitney test comparing individuals with and without a detected mutation. Differential
690 expression between sexes was done using Mann-Whitney test on the same expression matrices.

691 **Myeloid malignancies scRNA-seq initial processing**

692 We processed additional 10x libraries using cellranger as described previously. As the myeloid
693 malignancies data was multiplexed with other samples, including some that are not part of this
694 study, we detected doublets using Vireo and Souporcell[27], and assigned cells into individuals as
695 we described above. All the malignancies data was sequenced with the Ultima platform and we
696 corrected the UMI matrix by downsampling and resampling UMIs as described above. We then
697 created a metacell model for each of 12 samples separately: 2 healthy individuals, 3 CMML
698 patients, 2 MDS patients (one patient has two samples), 1 MDS / MPN patient, 1 myelofibrosis
699 patient, and 2 AML patients. As before, we excluded cells with less than 500 UMIs, with more
700 than 20% expression of mitochondrial genes, or with high expression of megakaryocyte genes.
701 We used the same set of ignored genes as we use for the healthy PB metacell model. We set the
702 target number of cell per metacell such that a metacell will have about 300K UMIs.

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

704 To project each individual's metacells on the healthy reference, we correlated the query
705 metacells with the reference's metacells. The correlation was performed in log2 scale and after
706 downsampling the query metacells to have 150K UMIs, using genes that were variable in the
707 reference. Query metacells were then annotated using the most common cell type among the 5
708 atlas metacells they were most correlated with. Metacells that mapped to CD34- reference

709 metacells were discarded for the following analyses. **Fig 4B** reports the mean correlation
710 between each query metacell and its 5 most correlated reference metacells, and places each
711 query metacell on the mean x and y coordinates of its 5 most highly correlated atlas metacells.
712 To plot **Fig 4E**, we calculated the observed expression of each gene as the geometric mean
713 expression across all metacells of an individuals that were annotated as MPP/MEBEMP-
714 E/MEBEMP-L. We looked at the 5 most correlated atlas metacells for each one of these query
715 metacells, and calculated for each gene its geometric mean to obtain the expected expression.
716 Due to systematic differences between the query and atlas in some samples, we normalized the
717 expected values by sorting genes based on their mean expression in the observed and expected,
718 and subtracting from the expected value a rolling mean across 500 genes of the difference
719 between the expected and observed. We defined differentially expressed (DE) genes as having
720 an at least 2-fold difference between observed and expected.

721 **Figs 4C,D** are based on projection of single cells. Cells were correlated to metacells
722 downsampled to 150K UMIs, and the raw UMI vectors were used for correlation. **Fig 4C** then
723 shows the distribution of the annotation of the metacells that cells were projected onto. For **Fig**
724 **4D**, each query cell was classified to the bin (as in **Fig 2D**) that was most common among cells in
725 the metacell to which it mapped.
726

727 **Karyotype analysis**

728 To perform karyotype analysis, we subtracted from each metacell's expression (log2 after
729 normalizing to sum to 1) the geometric mean expression of the 5 atlas metacell it was most
730 correlated with. For each metacell we then substracted the median difference value over all
731 genes. We then partitioned each chromosome into bins with an equal number of genes such
732 that all bins will have at least 40 genes, and computed the median difference across all genes in
733 the bin. For this analysis, we only considered genes with an average expression of at least 2^-
734 15.5 in either the pool of query metacells or the pool of the matched atlas metacells. This
735 analysis provides metacell resolution karyotypes, as shown in **EDF 9**. To plot **Fig 4F** we calculated
736 the median difference value for each gene across metacells that were projected to CD34+ cell
737 types.

738 **Profiling signatures in disease cases**

739 We separated AML-1 metacells into AML-1-1 and AML-1-2 by their karyotype as shown in **EDF 9**.
740 We created lists of genes overexpressed in healthy NKTDPT, CLP, HSC and MEBEMP metacells.
741 We scored each AML metacell by the geometric mean of all genes in each of the above gene
742 lists, as well as in the LMNA signature genes. We set thresholds for a metacell to express a
743 particular gene program by observing the distribution of the gene program's expression across
744 reference metacells in the relevant cell population (e.g., NKTDP metacells for the NKTDP gene
745 program, see dashed lines in **EDF 10A**).

746 For the heatmap in **EDF 11A**, we selected genes from the above gene programs, as well as genes
747 higher in AML-1 or in AML-2 compared to the reference across all cell types, and genes higher in
748 AML-1-2 compared to AML-1-1 in all the following cell types: CLP-E, MPP and MEBEMP-E.

749 To discover de-novo gene programs in the AML samples (**EDF 10C**), we selected genes that are
750 highly variable in each AML metacell model, calculated their correlation across metacells,
751 clustered their correlation profiles, and filtered clusters with low mean correlation (less than
752 0.4), or genes with low mean correlation (less than 0.4) to their cluster. We manually added
753 several genes of interest to the displayed matrices (**EDF 10C**).

754

755 **Figure Legends**

756 **Figure 1A** – experimental design, **1B** – annotated 2D UMAP projection of our metacell manifold
757 following filtration of metacells with low CD34 expression. Symmetric (**1C**) and asymmetric (**1D**)
758 regulation of specific HSC transcription factors upon bifurcation to the CLP (right) and MEBEMP
759 (left) lineages. Each panel shows the expression of one gene (y axis). Metacells in all panels are
760 ordered (left to right) by increasing AVP expression in the MEBEMP lineage and decreasing AVP
761 expression in the CLP lineage. Units for gene expression in all figure panels are log2 of each
762 gene's fractional expression. **1E** – the metacell population of interest (dotted line) linking BEMPs
763 to their MEBEMP-L precursors. **1F** – positively and negatively regulated TFs involved in early
764 BEMP differentiation. **1G** – gene-gene plot of *IRF8* against *TCF7* expression as hallmark markers
765 of DC and T cell differentiation respectively. The high *ACY3* NKTDP metacell population of
766 interest is depicted (dotted line). This population exhibits high expression of both T and
767 dendritic cell regulators, forming a gradient consisting of NK/T cell-like progenitors exhibiting a
768 high *TCF7/IRF8* expression ratio along with high expression of other T cell hallmarks such as *CD7*,

769 *MAF*, *IL7R*, *TRBC2*, and DC-like progenitors exhibiting a low *TCF7/IRF8* expression ratio, along
770 with high expression of other DC hallmarks, such as the myeloid TF *PU.1* and the MHC class II
771 gene *CD74* (**1H**).

772 **Figure 2A** - characterization of inter-individual HSPC compositional state variation (scheme). **2B**
773 – boxplots of cell state frequency distributions across individuals (logarithmic scale). Percents
774 calculated out of *CD34*+ population. Boxplot centers, hinges and whiskers represent median,
775 first and third quartiles and 1.5× interquartile range, respectively. Numbers represent mean +/-
776 SD for each distribution. **2C -** comparison of cell state frequencies between 20 biological
777 replicates and their original samples, for CLP (CLP-E, CLP-M, CLP-L, NKTDP - top) and MEBEMP
778 (MEBEMP-E, MEBEMP-L, ERYP, BEMP - bottom) populations. All biological replicates were
779 sampled 1 year following original blood draw. **2D** (**top**) - individual cell state frequency profiles
780 over the HSC-MEBEMP and HSC-CLP differentiation gradients of 6 subjects (colored lines), each
781 representing one of six archetypes (classes) of HSPC composition in healthy individuals. Dashed
782 lines represent the median (black) and 5th and 95th percentiles (grey) of the studied population.
783 **2D** (**bottom**) cell state enrichment map over 15 differentiation bins (rows), for all studied
784 individuals (columns) clustered into 6 classes. Classes I & II represent individuals relatively
785 enriched in lymphoid progenitors, whereas classes V & VI represent individuals with relative
786 depletion of lymphoid progenitors. Individuals are sorted by stemness within each class. Age
787 and sex are denoted for each individual. **2E** – CBC correlations to cell state frequencies:
788 %Lymphocytes (from WBC, calculated for entire cohort, left), HCT (males only, center), RDW
789 (males only, right). Missing individuals lacked sufficient cells for analysis. Permutation test p
790 values are displayed for each correlation. See Methods for details on the permutation-based
791 test. **2F** – boxplots showing CLP frequency distributions in individuals with (right) and without
792 (left) clonal hematopoiesis. **2G** – Relative cell state frequencies in mutant (right) and non-
793 mutant (left) cells following GoT of sample #122 (DNMT3A R882 mutated, VAF = 0.07). **2H** – CH
794 frequency (by gene) in age- and sex-matched high (red) and normal (black) RDW individuals in a
795 cohort of 18,147 individuals.

796 **Figure 3A** – Analysis of age-linked compositional differences in MEBEMP (MEBEMP-E, MEBEMP-
797 L, ERYP, BEMP, left) and CLP (CLP-E, CLP-M, CLP-L, right) populations, comparing specific cell
798 state frequencies (out of total *CD34*+ population) in young (<50 years) vs. old (>60-years)
799 individuals without clonal hematopoiesis, performed for males (blue) and females (red)

800 separately. Kruskal-Wallis p values for group differences are denoted on top. **3B** - Analysis of
801 age-linked compositional differences within the MEBEMP differentiation trajectory, comparing
802 the abundance of more (MEBEMP-L) to less (MPP) differentiated states in young (<50 years) vs.
803 old (>60-years) individuals, performed for males (blue) and females (red) separately. The
804 Kruskal-Wallis p value for difference among these groups is denoted on top. **3C** – as in 3A, for
805 the HSC population. **3D** – Analysis of age-linked differences in (CD34+) cHSPC frequencies from
806 total PBMCs in a recent cohort of 1000 healthy individuals undergoing PBMC scRNAseq[28]. **3E** –
807 True age (x) vs age predicted based on composition-controlled MEBEMP expression (y). **3F** –
808 gene-gene correlation heatmap, calculated over individual-level HSC-MEBEMP gene expression
809 controlled for HSC-MEBEMP composition. **3G** - intra-individual correlation of *LMNA* signatures in
810 CLPs and MEBEMPs, for both males (blue) and females (red). **3H** – Analysis of age-linked
811 differences in LMNA signature expression for CLP (right) and MEBEMP (left) populations in
812 young (<50 years) vs. old (>60-years) individuals. Y axis denotes log2(observed/expected
813 expression) normalized for composition. Boxplot centers, hinges and whiskers represent
814 median, first and third quartiles and 1.5× interquartile range, respectively. **3I** – individual
815 heatmaps of single cell counts over 20 bins of stemness (*AVP* signature, y axis) and MEBEMP
816 differentiation (*GATA1* signature, x axis). Individual identifier, as well as his/her RBC, and MCV
817 are denoted on top. **3J** – comparison between individual sync scores and clinical parameters
818 (RBC/MCV) across males. High and low sync scores (denoted by red and black dots respectively)
819 define clinically distinct populations. **3K** – correlation between individual sync scores and cell
820 state compositions. Permutation test p value denoted on top.

821 **Figure 4A** – diagnostic approach to leukemia analysis using our cHSPC reference atlas (scheme):
822 1. scRNAseq on CD34-enriched PB and construction of a patient-specific metacell model, 2.
823 Projection of patient derived metacells on the healthy reference atlas. 3. Compositional (relative
824 cell state frequency) analysis and 4. Composition-controlled differential gene expression
825 analysis. 5. Mutational and CNV analysis using targeted DNA sequencing and RNA-based
826 karyotyping. **4B** –projection of metacells derived from 2 new healthy individuals (not included in
827 the reference model), 3 CMML patients, 2 MDS patients (1 of them prior to and following
828 treatment), 1 MDS/MPN overlap patient, 1 myelofibrosis patient and 2 AML patients on our
829 healthy HSPC reference metacell model. Gene expression correlations between patient
830 (projected) and reference metacells are color coded according to the legend on the right. **4C,D** –
831 individual cell state frequency profiles over the HSC-MEBEMP and HSC-CLP differentiation

832 gradients for 2 healthy and 7 patient samples (1 of them prior to and following treatment
833 initiation). Dashed lines in Figure 4D represent the median (black) and 5th and 95th percentiles
834 (grey) of the healthy population. **4E** – Compositional-controlled differential gene expression of 2
835 healthy and 7 patient samples (1 of them prior to and following treatment initiation) to the
836 normal reference, quantifying the number of differentially expressed genes and identifying
837 specific genes recurrently induced or repressed in disease. **4F** - scRNAseq karyotyping for 2 AML
838 patients and 1 MDS patient prior to and following treatment initiation. Metacell models were
839 created for each MDS/AML patient and projected over our healthy reference map. Coupled
840 reference and projected (patient) metacells were then used for calculating expression ratios
841 over all expressed genes in all chromosomes. Log2 fold-change expression (patient/healthy) for
842 all expressed genes across all chromosomes is shown. Red lines represent the median of each
843 chromosomal fold-change distribution.

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

846 **1A** – age distribution (decimals) of studied population by sex. **1B** – 2D UMAP projection of our
847 metacell model prior to *CD34* metacell filtering. **1C** - relative expression heatmap of cell states
848 (columns) and gene markers used for cell state annotation (rows). **1D** – Metacell filtration on
849 CD34 levels. **1E** - expression plot of *MPO* and *GATA1/VPREB1* expression showing all 3
850 differentiation routes (GMPs, CLPs, MEBEMPs) from HSCs. **1F -** gene-gene expression plot of
851 *DNTT* and *RUNX3*, showing early CLP differentiation and their bifurcation into late CLPs and
852 NKTDPs. All gene expression values are obtained by normalizing gene expression to sum to 1
853 and taking log2.

854 **EDF 2 – BM comparisons I**

855 **2A** – 2D UMAP projection of a non-*CD34*-enriched BM metacell model from the Human Cell
856 Atlas[16], colored by a BM-specific cell state annotation. **2B** - projection of our PB *CD34*+ derived
857 metacells on the non-*CD34* enriched BM metacell model. **2C** – projection of our BM *CD34*+
858 derived metacells on the non-*CD34* enriched BM metacell model. **2D -** projection of BM *CD34*+
859 derived metacells from Setty et al. on the non-*CD34* enriched BM metacell model. **2E** - gene-
860 gene expression plots comparing PB *CD34*+ derived metacells with their BM *CD34*+ counterparts
861 from our study and from Setty et al. for all differentiation trajectories. Panels (top to bottom)

862 represent CLP differentiation, MEBEMP differentiation, GMP differentiation, BEMP
863 differentiation, DC differentiation, and MPO/CLP/MEBEMP trifurcation from HSCs. PB / BM
864 metacells are colored by PB / BM annotations, respectively.

865 **EDF 3** – **BM comparisons II – CLP differentiation and cell-cycling**

866 **3A** – gene-gene expression plots comparing PB *CD34*+ derived metacells with their BM *CD34*+
867 counterparts from our study and from Setty et al. for markers and regulators of CLP
868 differentiation and bifurcation, highlighting differences between PB and BM. **3B** – cell state
869 specific comparison of S-phase signatures in circulating (left) vs. BM (HCA[16], right) HSPCs.

870 **EDF 4 – Individual-specific composition-controlled differential gene expression,**
871 **megakaryocytic contamination**

872 **4A,B –** Individual-specific differential gene expression after controlling for each individual's
873 distribution across the *CD34*+ PB manifold in MEBEMPs (top) and CLPs (bottom). **4C** - relative
874 expression heatmap of the megakaryocytic markers *PF4* and *PPBP* and cell state specific
875 markers, across metacells with high megakaryocytic signature. The figure shows an abnormally
876 high doublet rate involving megakaryocytes. Cells contained in such metacells were accordingly
877 excluded from the final metacell model.

878 **EDF 5 – Circulating HSCs**

879 **5A** - gene-gene expression plots comparing the PB high *AVP*/*HLF* HSC population (left) with that
880 found in two BM metacell models [16,17] and in our CD34+ BM data. PB and BM metacells are
881 colored by PB and BM annotations, respectively. **5B** – map of transcriptionally activated genes
882 upon exit from the HSC state and differentiation toward lymphoid (CLP) and non-lymphoid
883 (MEBEMP) fates. Dots represent genes. HSC/CLP and HSC/MEBEMP gene expression ratios are
884 depicted on the y and x axis respectively. Class I genes are representative of the HSC state; Class
885 II genes exhibit symmetric transcriptional activation upon exit from the HSC state towards CLP
886 and MEBEMP fates, whereas Class III, IV, V, VI exhibit asymmetrical transcriptional activation
887 upon exit from the HSC state towards CLP (class III, V) and MEBEMP (Class IV, VI) fates. n is the
888 number of genes in each class.

889 **EDF 6 - Factors involved in BEMP and NKTDP differentiation**

890 **6A** - factors positively and negatively regulated in the early stages of BEMP specification. **6B** –
891 gene-gene expression plots of *DNTT* and *ACY3* comparing *CD34*-enriched (Setty et al.[17]) and non-
892 enriched (HCA[16]) BM to our *CD34*+ BM model (top), as well as non-enriched (Yazar et al. [28]) and
893 partially enriched PB (Zheng et al.[36]) to our *CD34*+ PB model (bottom). Metacells are color-coded
894 by *SYT2* expression. The *SYT2* high, *ACY3* high, *DNTT* intermediate population clearly seen in our
895 data is completely lacking from the BM datasets. **6C** – anti-correlation of the DC *IRF8-MHC-II*
896 coupled dynamics and the T cell regulator *TCF7*, involved in the bifurcation of the NKTDP cell
897 state to its sub-populations.

898 **EDF 7 – Stability of cell state compositions across technical and biological replicates**

899 **7A** – comparison of Illumina and Ultima Genomics sequenced data. Each panel represents one
900 library that was sequenced by both technologies. Points represent genes, and each gene's
901 expression level across all cells in the library as determined by Illumina (x) and by Ultima
902 Genomics (y) is shown. **6B** – Cell state frequency comparisons between 39 technical & 20
903 biological replicates and their original samples. Each pair of panels represents one cell state,
904 denoted on top. Panels on the left of each pair compare the cell state frequency in the original
905 sample, sequenced by Illumina (x), to its technical replicate frequency, sequenced by Ultima
906 Genomics (y). Panels on the right of each pair compare the cell state frequency in the original
907 sample (x) to its biological replicate frequency (y). All biological replicates were sampled 1 year
908 following original blood draw.

909 **EDF 8 – Composition-controlled transcriptional variation: the *LMNA* signature and sync score**

910 **8A –** CLP frequency against maximal VAF (%) for individuals with CH. **8B** – True age (x) vs age
911 predicted based on composition-controlled CLP expression (y). **8C** - gene-gene correlation
912 heatmap, calculated over individual-level CLP gene expression controlled for CLP composition.
913 **8D** – the *LMNA* signature – co-variation of *LMNA* expression with *ANXA1/2*, *AHNAK*, *MYADM,*
914 *TSPAN2* and *VIM*. **8E** – heatmap of individual *LMNA* signature expression across the MEBEMP
915 trajectory. Individual age and sex are color-coded on top. **8F** – *LMNA* signature expression in
916 HSCs (denoted by high *AVP*) and throughout MPP / MEBEMP (left) and lymphoid (right)
917 differentiation. **8G** - *LMNA* signature expression correlations between 39 technical & 20
918 biological replicates and their original samples. **8H** - sync score correlations between 39

919 technical & 20 biological replicates and their original samples. All biological replicates were
920 sampled 1 year following original blood draw.

921 **EDF 9 - scRNAseq karyotyping**

922 **9A** - scRNAseq karyotyping for 2 AML patients (N186, N205) and 1 MDS patient (N211) with
923 isolated del5q (bottom), prior to and following treatment with lenalidomide. Metacell models
924 were created for each MDS/AML patient and projected over our healthy reference map.
925 Coupled reference and projected (patient) metacells were then used for calculating expression
926 ratios over all expressed genes in all chromosomes. Heatmap of log2 expression fold-change
927 (patient/reference) per metacell over all genes expressed in each chromosome is shown. **9B** –
928 scRNAseq karyotyping for 2 healthy (N294, N207), 3 CMML (N192, N235, N48), 1 MDS (N249), 1
929 MDS/MPN (N251) and 1 Myelofibrosis (N198) patient. Karyotyping is performed as in 9A. Log2
930 fold-change expression (patient/healthy) for all expressed genes across all chromosomes is
931 shown. Red lines represent the median of each chromosomal fold-change distribution.

932 **EDF 10 – AML subpopulation analysis and characterization**

933 **10A -** each of 4 panels refers to a different cell state gene signature as noted on the x-axis. Panel
934 top - boxplots of gene module expression distributions for different cell states in our reference
935 atlas. Panel bottom - Gene signature expression density plots for each of the AML subclones.
936 Reference gene signature distributions (panel top) were used to identify subpopulations of AML
937 cells with CLP, MEBEMP, HSC and NKTDP characteristics (panel bottom). Dashed lines represent
938 the threshold for expressing a gene signature, and the fraction of cells expressing a signature
939 per AML clone is listed. **10B** - same as 10A for *LMNA* expression across AML clones, showing
940 high variability in *LMNA* expression across AML metacells. **10C** - left – correlation heatmap of
941 differentially expressed gene signatures for AML-1 (N186, top) and AML-2 (N205, bottom). The
942 malignant state is characterized by multiple novel gene signatures in addition to aberrant
943 expression of "healthy" differentiation-related modules, right – UMAP projection of the metacell
944 models of AML-1 (top) and AML-2 (bottom), colored by relative expression of differentially
945 expressed genes. Overexpression of BCL2 in AML-1-2 compared to AML-1-1 can be seen.

946 **EDF 11 - The transcriptional landscape of AML: aberrant expression of reference genes and**
947 **multiple novel gene signatures**

948 **11A** – Expression heatmap of several genes across reference cell states and AML subclones. The
949 malignant state differs greatly from the healthy state both in the expression of reference genes
950 and by multiple additional gene expression signatures. **11B** - panels compare gene expression of
951 major differentiation regulators associated with lymphoid (top) and MPP / MEBEMP (bottom)
952 differentiation, in reference (left), AML-1 (N186, middle, color coded by subclones) and AML-2
953 (N205, right) metacells.

954

955 **Data Availability**

956 All data can be explored in: https://apps.tanaylab.com/MCV/blood_aging/

957

958 **Code Availability**

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

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

961

962 **Author Information**

963

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

965

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

967

968 Authors and Affiliations

969

970 **Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel**
971 Nili Furer, Nimrod Rappoport, Adi Danin, Nathali Kaushansky, Noa Chapal-Ilani, Amos Tanay
972 Liran Shlush

973

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

976

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

Declaration of 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.

We thank all members of the Shlush and Tanay laboratories for their constructive comments.

1068 **Supplementary Information**

1069 **Supplementary Table S1 – Basic participant demographics**. This table contains data on sex and
1070 age of all studied participants at sampling time, along with their allocated study IDs.

1071 **Supplementary Table S2 – Longitudinal CBC's**. This table contains median values for each CBC
1072 parameter for each sampled individual across 5 years prior to sampling date. WBC – white blood
1073 cells, RBC – red blood cells, MCV – mean corpuscular volume, MCH - mean corpuscular
1074 hemoglobin, MCHC - mean corpuscular hemoglobin concentration, RDW - RBC width
1075 distribution, MPV – mean platelet volume.

1076 **Supplementary Table S3 – ARCH sequencing**. This table contains variant calling data on all ARCH
1077 positive individuals, including their Individual ID, specific gene involved, chromosome number,
1078 position, reference and alternate alleles, and the reported average variant allele frequency
1079 (VAF) from single (amplicon) or duplicate (MIP) sequencing instances.

1080 **Supplementary Table S4 – HSC expression differences**. This table contains each gene's mean
1081 expression across metacells in HSCs, MEBEMP-Es, and CLP-Ms, and the differences in mean
1082 expression between HSCs and the latter two cell types. These expression differences further
1083 classify genes as belonging to each of VIII groups as described in EDF 5B.

1084 **Supplementary Table S5 – high RDW case-control longitudinal CBC's**. This table contains all
1085 blood count instances of high RDW individuals and controls. WBC – white blood cells, RBC – red
1086 blood cells, MCV – mean corpuscular volume, MCH - mean corpuscular hemoglobin, MCHC -
1087 mean corpuscular hemoglobin concentration, RDW - RBC width distribution, MPV – mean
1088 platelet volume.

1089 **Supplementary Table S6 – high RDW case-control ARCH sequencing**. This table contains variant
1090 calling data on all ARCH positive high-RDW cases and controls, including their Individual ID,
1091 specific gene involved, chromosome number, position, reference and alternate alleles, whether
1092 this alternation is considered a leukemic hotspot, and the reported average variant allele
1093 frequency (VAF) from single (amplicon) or duplicate (MIP) sequencing instances.

1094 **Supplementary Table S7 – MEBEMP differential expression screen**. This table contains
1095 correlations between gene expression and different clinical labels across studied individuals.
1096 Correlations are between expression values normalized for distribution across the MPP-

1097 MEBEMP population, and the following labels:
1098 - maximal mutant VAF. The Spearman correlation between the maximal VAF and the gene
1099 expression is reported, as well as the p-value and BH corrected q-value for the correlation's
1100 equality to 0 (two-sided test). These values are reported in columns starting with "vaf_cor". The
1101 p-value and q-value are also reported for Mann-Whitney two-sided test, comparing the
1102 expression of the gene in individuals with and without mutations. These values are reported in
1103 columns starting with "vaf_mann_whitney".
1104 - age. The Spearman correlation, p-value for correlation's equality to 0 (two-sided) and BH-
1105 corrected q-value are reported for males and females separately. These are reported in the
1106 columns starting with "age".
1107 - sex. The p-value and BH corrected q-value for a Mann-Whitney two-sided test comparing
1108 expression in males and females are reported. These are reported in the columns starting with
1109 "sex".
1110 - CBCs. For each of 20 CBC values, and for males and females separately, three values are
1111 reported: the Spearman correlation between gene expression and CBC value, and p-value for
1112 that correlation being equal to 0, and the BH-corrected q-values. These values are reported in
1113 the remaining columns.
1114 Only genes whose expression levels exceeded a minimal threshold, and who displayed small
1115 technical variation, are listed.

1116 **Supplementary Table S8 – CLP differential expression screen**. Similar to Supplementary Table 7,
1117 using CLP cell types.

1118 **Supplementary Table S9 – LMNA gene correlations**. This table contains the Spearman
1119 correlation of each gene to *LMNA* across MEBEMP and CLP metacells (mc_cor columns), and
1120 across individual composition-controlled MEBEMP and CLP expression profiles (indiv_cor
1121 columns). The last column states whether the gene is part of the LMNA signature or not.

1122 **Supplementary Table S10 – sync genes**. This table contains data for how the genes used to
1123 calculate the sync-scores were selected. It lists each gene's correlation to AVP and GATA1 across
1124 the HSC-MPP-MEBEMP trajectory (avp_cor and gata1_cor), whether the gene is considered
1125 lateral and therefore excluded from the signature, mean expression across MEBEMP metacells,
1126 and whether the gene is part of the AVP and GATA1 gene signatures (is_avp_sig and
1127 is_gata1_sig).

1128 **Supplementary Table S11 – Individual scores and signatures**. Key scores and signatures are
1129 listed for each individual: *LMNA* signature in MEBEMPs (lmna_mebemp_score), *LMNA* signature
1130 in CLPs (lmna_clp_score), sync-score, and the number of cells belonging to each cHSPC cell state
1131 (under columns whose names start with "num_").

1132

1133 **References**

1134

1135 1. Osgood, E. E. Normal hematologic standards. *Arch Intern Med* **56**, 849 (1935).

1136 2. Cohen, N. M. *et al.* Personalized lab test models to quantify disease potentials in healthy
1137 individuals. *Nat Med* (2021) doi:10.1038/S41591-021-01468-6.

1138 3. Giorno, R., Clifford, J. H., Beverly, S. & Rossing, R. G. Hematology Reference Values
1139 Analysis by Different Statistical Technics and Variations with Age and Sex. *Am J Clin*
1140 *Pathol* **74**, 765–770 (1980).

1141 4. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute
1142 leukaemia. *Nature* **506**, 328–333 (2014).

1143 5. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R.
1144 Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators
1145 and persist in remission. *Proc NatI Acad Sci U S A* **111**, 2548–2553 (2014).

1146 6. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes.
1147 *New England Journal of Medicine* **371**, 2488–2498 (2014).

1148 7. Busque, L., Buscarlet, M., Mollica, L. & Levine, R. L. Concise Review: Age-Related Clonal
1149 Hematopoiesis: Stem Cells Tempting the Devil. *Stem Cells* vol. 36 1287–1294 Preprint at
1150 https://doi.org/10.1002/stem.2845 (2018).

1151 8. Cohen, K. S. *et al.* Circulating CD34+ progenitor cell frequency is associated with clinical
1152 and genetic factors. *Blood* **121**, e50–e56 (2013).

1153 9. Mende, N. & Laurenti, E. Hematopoietic stem and progenitor cells outside the bone
1154 marrow: where, when, and why. *Exp Hematol* **104**, 9–16 (2021).

1155 10. Ainciburu, M. *et al.* Uncovering perturbations in human hematopoiesis associated with
1156 healthy aging and myeloid malignancies at single-cell resolution. *Elife* **12**, (2023).

1157 11. Bender, J. *et al.* Identification and comparison of CD34-positive cells and their
1158 subpopulations from normal peripheral blood and bone marrow using multicolor flow
1159 cytometry. *Blood* **77**, 2591–2596 (1991).

1160 12. Mende, N. *et al.* Unique molecular and functional features of extramedullary
1161 hematopoietic stem and progenitor cell reservoirs in humans. *Blood* (2022)
1162 doi:10.1101/2020.01.26.919753.

1163 13. Goodman, J. W. & Hodgson, G. S. Evidence for stem cells in the peripheral blood of mice.
1164 *Blood* **19**, 702–14 (1962).

1165 14. Biezuner, T. *et al.* An improved molecular inversion probe based targeted sequencing
1166 approach for low variant allele frequency. *NAR Genom Bioinform* **4**, (2022).

1167 15. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer
1168 metacell algorithm for scalable scRNA-seq analysis. *Genome Biol* **23**, 100 (2022).

1169 16. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas
1170 bone marrow single-cell interactive web portal. *Exp Hematol* **68**, 51–61 (2018).

1171 17. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir.
1172 *Nat Biotechnol* **37**, 451–460 (2019).

1173 18. Lehnertz, B. *et al. HLF* expression defines the human hematopoietic stem cell state. *Blood*
1174 **138**, 2642–2654 (2021).

1175 19. Frelin, C. *et al.* GATA-3 regulates the self-renewal of long-term hematopoietic stem cells.
1176 *Nat Immunol* **14**, 1037–1044 (2013).

1177 20. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on
1178 transcriptional landscapes links state to fate during differentiation. *Science (1979)* **367**,
1179 (2020).

1180 21. Lavaert, M. *et al.* Resource Integrated scRNA-Seq Identifies Human Postnatal Thymus
1181 Seeding Progenitors and Regulatory Dynamics of Differentiating Immature Thymocytes II
1182 Resource Integrated scRNA-Seq Identifies Human Postnatal Thymus Seeding Progenitors
1183 and Regulatory Dynami. 1088–1104 (2020) doi:10.1016/j.immuni.2020.03.019.

1184 22. Scoville, S. D. *et al.* A Progenitor Cell Expressing Transcription Factor RORγt Generates All
1185 Human Innate Lymphoid Cell Subsets. *Immunity* **44**, 1140–1150 (2016).

1186 24. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*
1187 **559**, (2018).

1188 25. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the
1189 causes and consequences of clonal hematopoiesis. *Nat Genet* **54**, 1155–1166 (2022).

1190 26. Nam, A. S. *et al.* Somatic mutations and cell identity linked by Genotyping of
1191 Transcriptomes. *Nature* **571**, 355–360 (2019).

1192 27. Sudo, K., Ema, H., Morita, Y. & Nakauchi, H. Age-Associated Characteristics of Murine
1193 Hematopoietic Stem Cells. *J Exp Med* **192**, 1273–1280 (2000).

1194 28. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of
1195 autoimmune disease. *Science (1979)* **376**, (2022).

1196 29. Gao, T. *et al.* Haplotype-aware analysis of somatic copy number variations from single-
1197 cell transcriptomes. *Nat Biotechnol* **41**, 417–426 (2023).

1198 30. Boyle, E. A., O'Roak, B. J., Martin, B. K., Kumar, A. & Shendure, J. MIPgen: Optimized
1199 modeling and design of molecular inversion probes for targeted resequencing.
1200 *Bioinformatics* **30**, 2670–2672 (2014).

1201 31. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-
1202 cell RNA-seq data without genotype reference. *Genome Biol* **20**, 273 (2019).

1203 32. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype
1204 without reference genotypes. *Nat Methods* **17**, 615–620 (2020).

1205 33. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell
1206 Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291.e9 (2019).

1207 34. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in
1208 Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-
1209 337.e4 (2019).

1210 35. Zhang, P. *et al.* Single-cell RNA sequencing to track novel perspectives in HSC
1211 heterogeneity. *Stem Cell Res Ther* **13**, 39 (2022).

1212

Figure 1

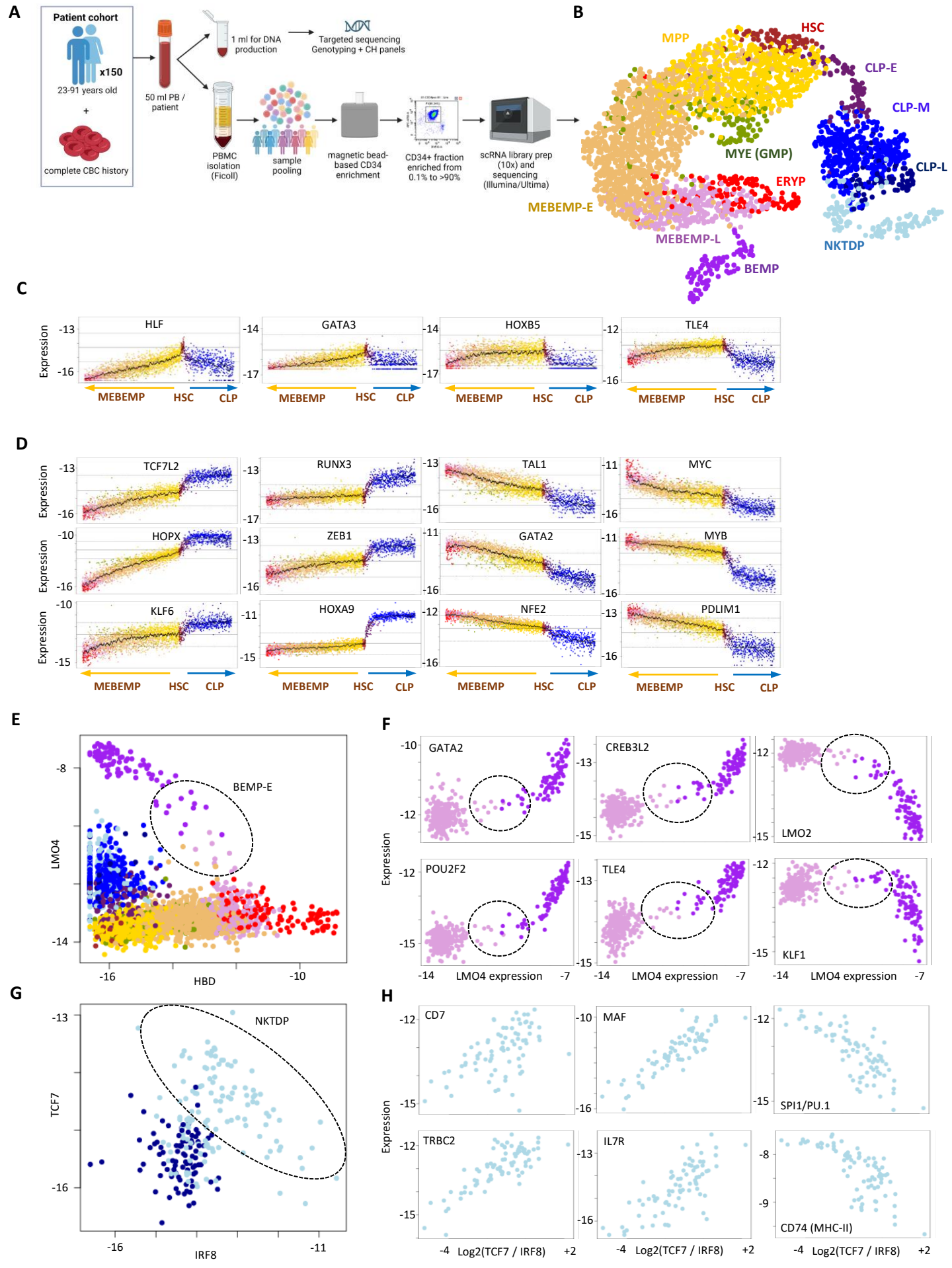

A
Patient cohort
x150
23-91 years old
+
complete CBC history
50 ml PB / patient
1 ml for DNA production
Targeted sequencing
Genotyping + CH panels
PBMC isolation (Ficoll)
sample pooling
magnetic bead-based CD34 enrichment
CD34+ fraction enriched from 0.1% to >90%
scRNA library prep (10x) and sequencing (Illumina/Ultima)
B
MPP
HSC
CLP-E
CLP-M
CLP-L
MYE (GMP)
ERYP
MEBEMP-E
MEBEMP-L
BEMP
NKTDP
C
HLF
GATA3
HOXB5
TLE4
Expression
MEBEMP
HSC
CLP
D
TCF7L2
RUNX3
TAL1
MYC
HOPX
ZEB1
GATA2
MYB
KLF6
HOXA9
NFE2
PDLIM1
E
BEMP-E
LMO4
HBD
F
GATA2
CREB3L2
LMO2
POU2F2
TLE4
KLF1
LMO4 expression
G
NKTDP
TCF7
IRF8
H
CD7
MAF
SPI1/PU.1
TRBC2
IL7R
CD74 (MHC-II)
Log2(TCF7 / IRF8)

Figure 2

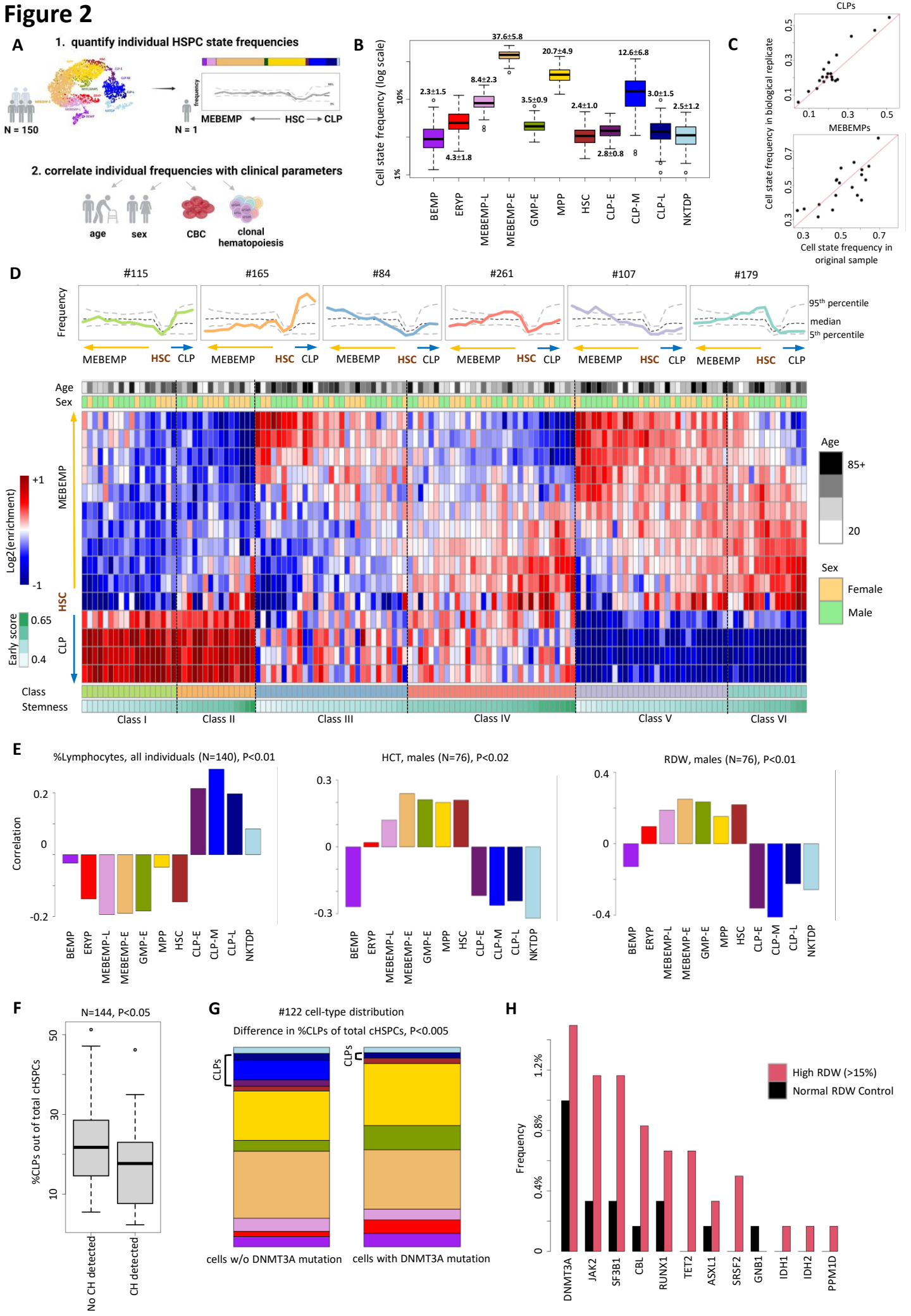

A
1. quantify individual HSPC state frequencies
N = 150
N = 1
MEBEMP
HSC
CLP
2. correlate individual frequencies with clinical parameters
age
sex
CBC
clonal hematopoiesis
B
Cell state frequency (log scale)
10%
1%
2.3±1.5
4.3±1.8
8.4±2.3
37.6±5.8
3.5±0.9
20.7±4.9
2.4±1.0
2.8±0.8
12.6±6.8
3.0±1.5
2.5±1.2
BEMP
ERYP
MEBEMP-L
MEBEMP-E
GMP-E
MPP
HSC
CLP-E
CLP-M
CLP-L
NKTDP
C
CLPs
MEBEMPs
Cell state frequency in biological replicate
Cell state frequency in original sample
D
#115
#165
#84
#261
#107
#179
Frequency
95th percentile
median
5th percentile
MEBEMP
HSC
CLP
Age
Sex
Log2(enrichment)
+1
-1
Early score
0.65
0.4
Class
Stemness
Class I
Class II
Class III
Class IV
Class V
Class VI
85+
20
Female
Male
E
%Lymphocytes, all individuals (N=140), P<0.01
HCT, males (N=76), P<0.02
RDW, males (N=76), P<0.01
Correlation
F
N=144, P<0.05
%CLPs out of total cHSPCs
No CH detected
CH detected
G
#122 cell-type distribution
Difference in %CLPs of total cHSPCs, P<0.005
CLPs
cells w/o DNMT3A mutation
cells with DNMT3A mutation
H
High RDW (>15%)
Normal RDW Control
Frequency
DNMT3A
JAK2
SF3B1
CBL
RUNX1
TET2
ASXL1
SRSF2
GNB1
IDH1
IDH2
PPM1D

Figure 3

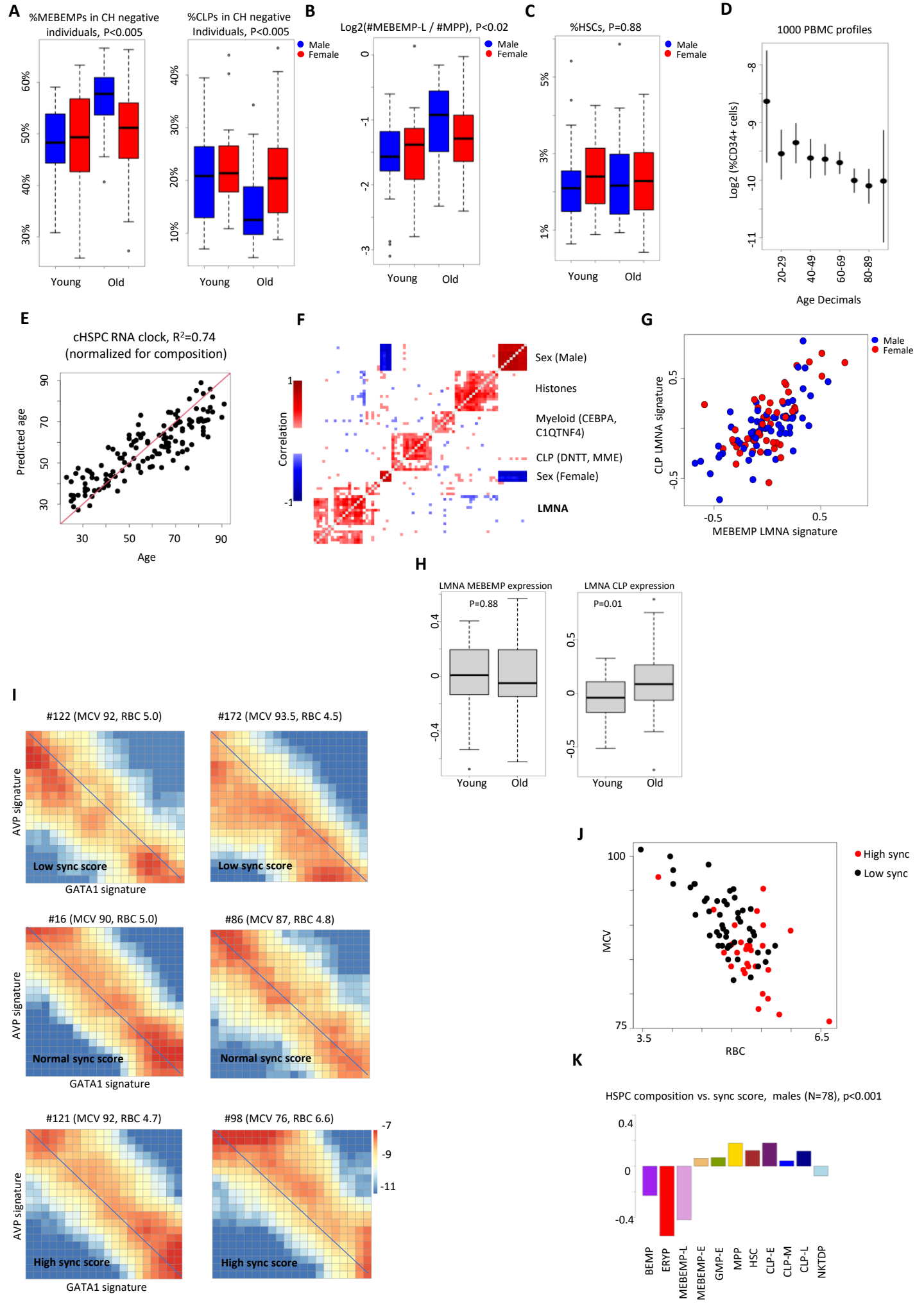
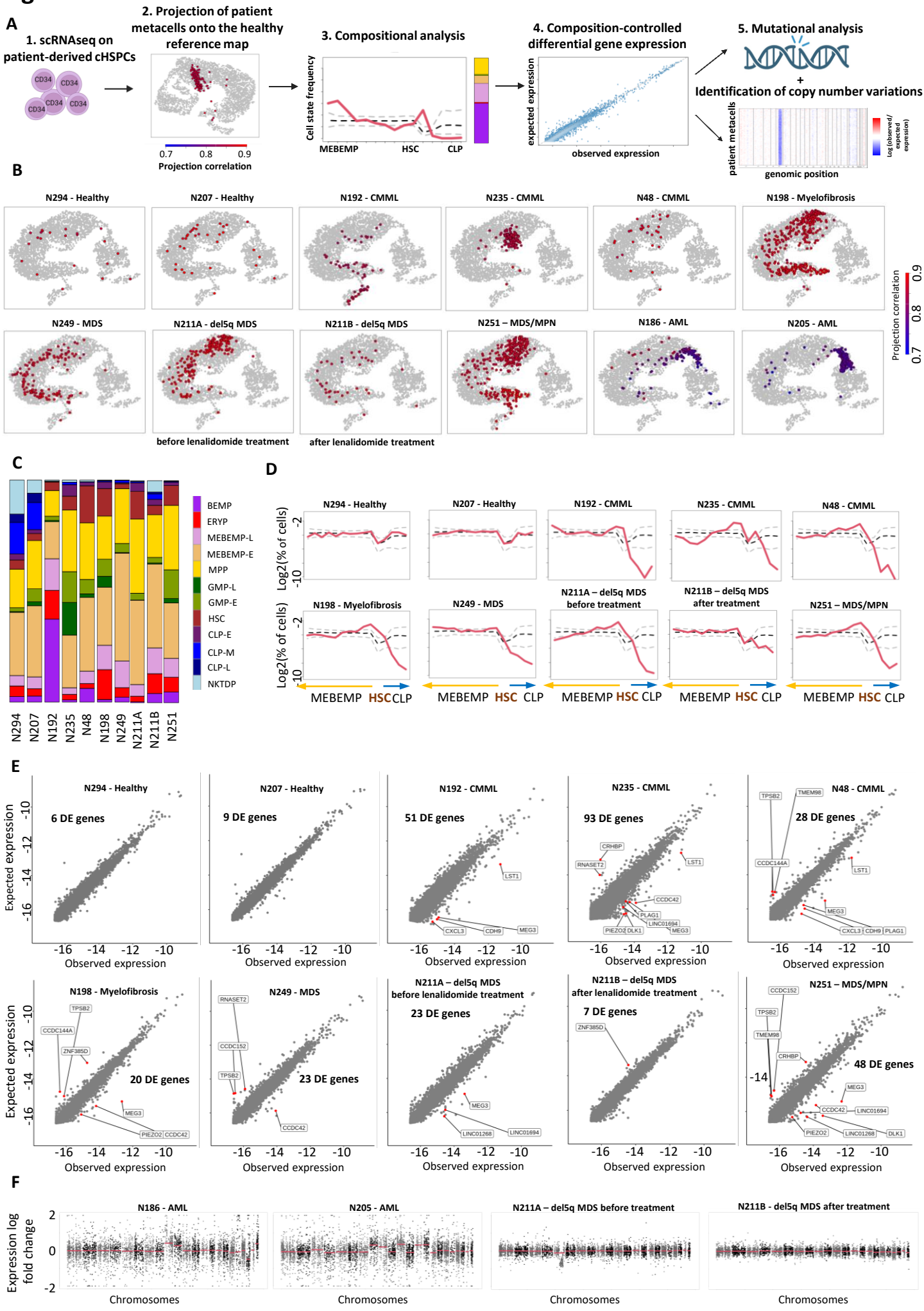
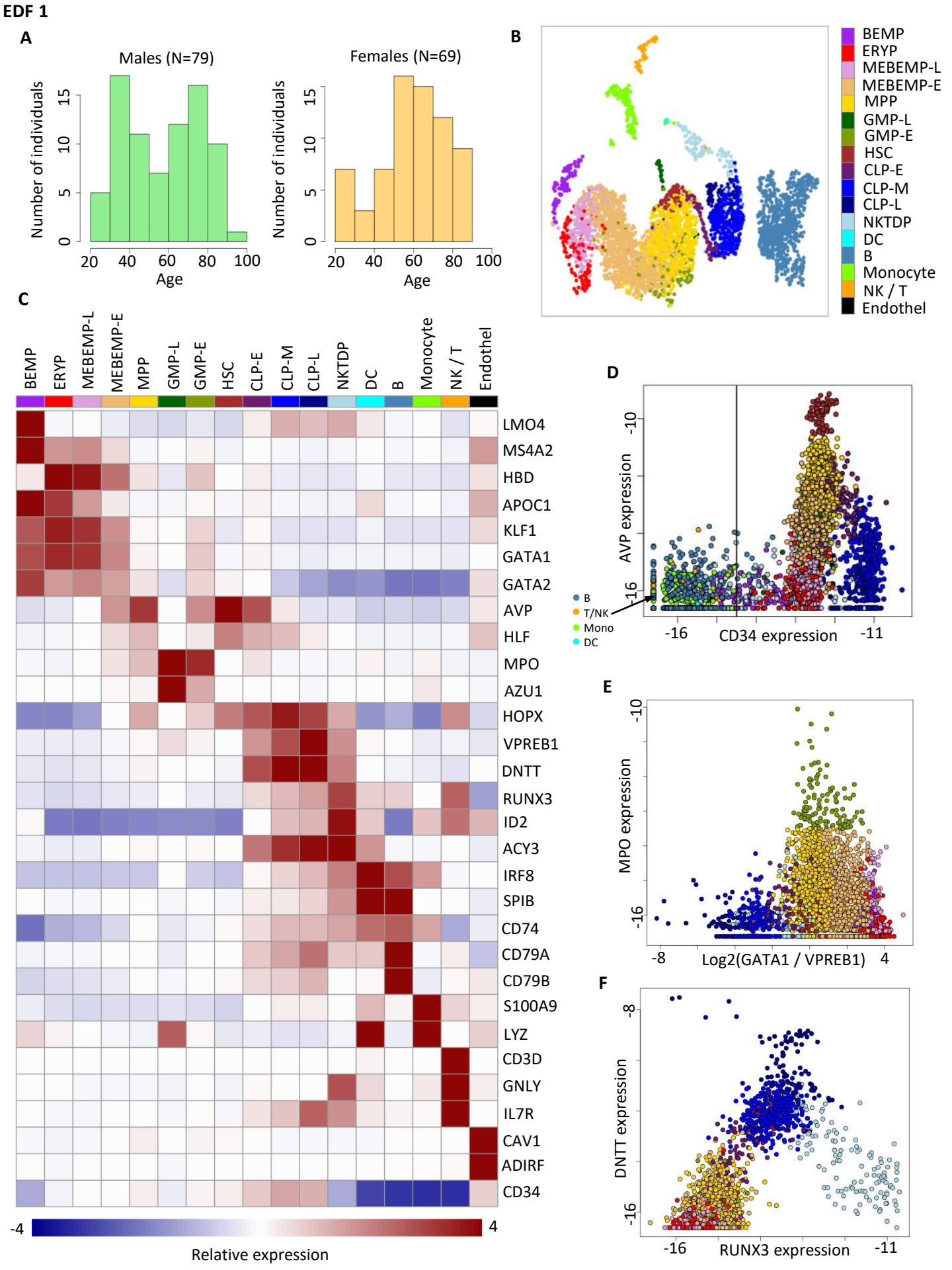

A
%MEBEMPs in CH negative individuals, P<0.005
%CLPs in CH negative Individuals, P<0.005
Male
Female
Young
Old
B
Log2(#MEBEMP-L / #MPP), P<0.02
C
%HSCs, P=0.88
D
1000 PBMC profiles
Log2 (%CD34+ cells)
Age Decimals
E
cHSPC RNA clock, R²=0.74
(normalized for composition)
Predicted age
Age
F
Correlation
Sex (Male)
Histones
Myeloid (CEBPA, C1QTNF4)
CLP (DNTT, MME)
Sex (Female)
LMNA
G
CLP LMNA signature
MEBEMP LMNA signature
H
LMNA MEBEMP expression
P=0.88
LMNA CLP expression
P=0.01
I
#122 (MCV 92, RBC 5.0)
#172 (MCV 93.5, RBC 4.5)
#16 (MCV 90, RBC 5.0)
#86 (MCV 87, RBC 4.8)
#121 (MCV 92, RBC 4.7)
#98 (MCV 76, RBC 6.6)
AVP signature
GATA1 signature
Low sync score
Normal sync score
High sync score
J
High sync
Low sync
MCV
RBC
K
HSPC composition vs. sync score, males (N=78), p<0.001
BEMP
ERYP
MEBEMP-L
MEBEMP-E
GMP-E
MPP
HSC
CLP-E
CLP-M
CLP-L
NKTDP

Figure 4


A
1. scRNAseq on patient-derived cHSPCs
CD34
2. Projection of patient metacells onto the healthy reference map
0.7 0.8 0.9
Projection correlation
3. Compositional analysis
Cell state frequency
MEBEMP HSC CLP
4. Composition-controlled differential gene expression
expected expression
observed expression
5. Mutational analysis
+
Identification of copy number variations
patient metacells
genomic position
Log (observed/ expected expression)
B
N294 - Healthy
N207 - Healthy
N192 - CMML
N235 - CMML
N48 - CMML
N198 - Myelofibrosis
N249 - MDS
N211A - del5q MDS
N211B - del5q MDS
N251 – MDS/MPN
N186 - AML
N205 - AML
before lenalidomide treatment
after lenalidomide treatment
Projection correlation
0.7 0.8 0.9
C
BEMP
ERYP
MEBEMP-L
MEBEMP-E
MPP
GMP-L
GMP-E
HSC
CLP-E
CLP-M
CLP-L
NKTDP
N294 N207 N192 N235 N48 N198 N249 N211A N211B N251
D
Log2(% of cells)
-2
-10
N294 - Healthy
N207 - Healthy
N192 - CMML
N235 - CMML
N48 - CMML
N198 - Myelofibrosis
N249 - MDS
N211A – del5q MDS before treatment
N211B – del5q MDS after treatment
N251 – MDS/MPN
MEBEMP HSC CLP
E
Expected expression
Observed expression
-10 -12 -14 -16
N294 - Healthy
6 DE genes
N207 - Healthy
9 DE genes
N192 - CMML
51 DE genes
LST1
CXCL3
CDH9
MEG3
N235 - CMML
93 DE genes
CRHBP
RNASET2
LST1
CCDC42
PLAG1
LINC01694
PIEZO2 DLK1
MEG3
N48 - CMML
28 DE genes
TPSB2
TMEM98
CCDC144A
LST1
MEG3
CXCL3 CDH9 PLAG1
N198 - Myelofibrosis
20 DE genes
TPSB2
CCDC144A
ZNF385D
MEG3
PIEZO2 CCDC42
N249 - MDS
23 DE genes
RNASET2
CCDC152
TPSB2
CCDC42
N211A – del5q MDS before lenalidomide treatment
23 DE genes
MEG3
LINC01268
LINC01694
N211B – del5q MDS after lenalidomide treatment
7 DE genes
ZNF385D
N251 – MDS/MPN
48 DE genes
CCDC152
TPSB2
TMEM98
CRHBP
MEG3
CCDC42
LINC01694
PIEZO2
LINC01268
DLK1
-16 -14 -12 -10
F
Expression log fold change
2
0
-2
N186 - AML
N205 - AML
N211A – del5q MDS before treatment
N211B - del5q MDS after treatment
Chromosomes

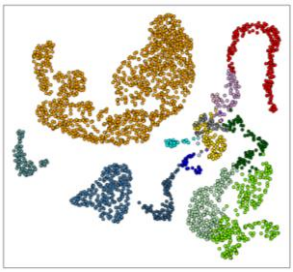
EDF 1


A
Males (N=79)
Females (N=69)
Number of individuals
Age
B
BEMP
ERYP
MEBEMP-L
MEBEMP-E
MPP
GMP-L
GMP-E
HSC
CLP-E
CLP-M
CLP-L
NKTDP
DC
B
Monocyte
NK / T
Endothel
C
LMO4
MS4A2
HBD
APOC1
KLF1
GATA1
GATA2
AVP
HLF
MPO
AZU1
HOPX
VPREB1
DNTT
RUNX3
ID2
ACY3
IRF8
SPIB
CD74
CD79A
CD79B
S100A9
LYZ
CD3D
GNLY
IL7R
CAV1
ADIRF
CD34
-4
4
Relative expression
D
AVP expression
CD34 expression
B
T/NK
Mono
DC
E
MPO expression
Log2(GATA1 / VPREB1)
F
DNTT expression
RUNX3 expression

EDF 2

A

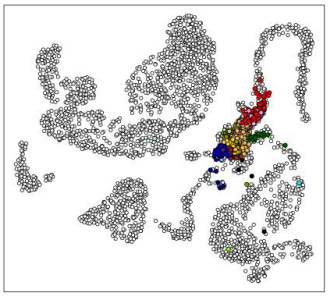
non-CD34-enriched BM metacell model
(Human cell Atlas)



- HSC / MEBEMP-E
- MPP
- MEP
- MK-P
- ERYP
- BEMP
- NK / T
- CMP
- GMP
- Monocyte
- NKTDP
- MDP
- LMPP
- Pro-B
- B
- Plasma

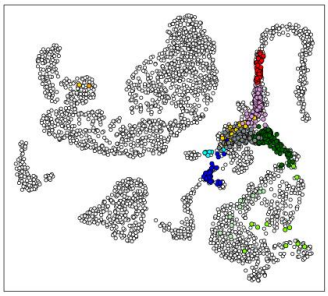
B

UMAP Projection CD34+ PB (this study)



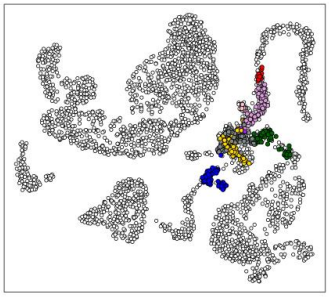
C

UMAP Projection CD34+ BM (this study)

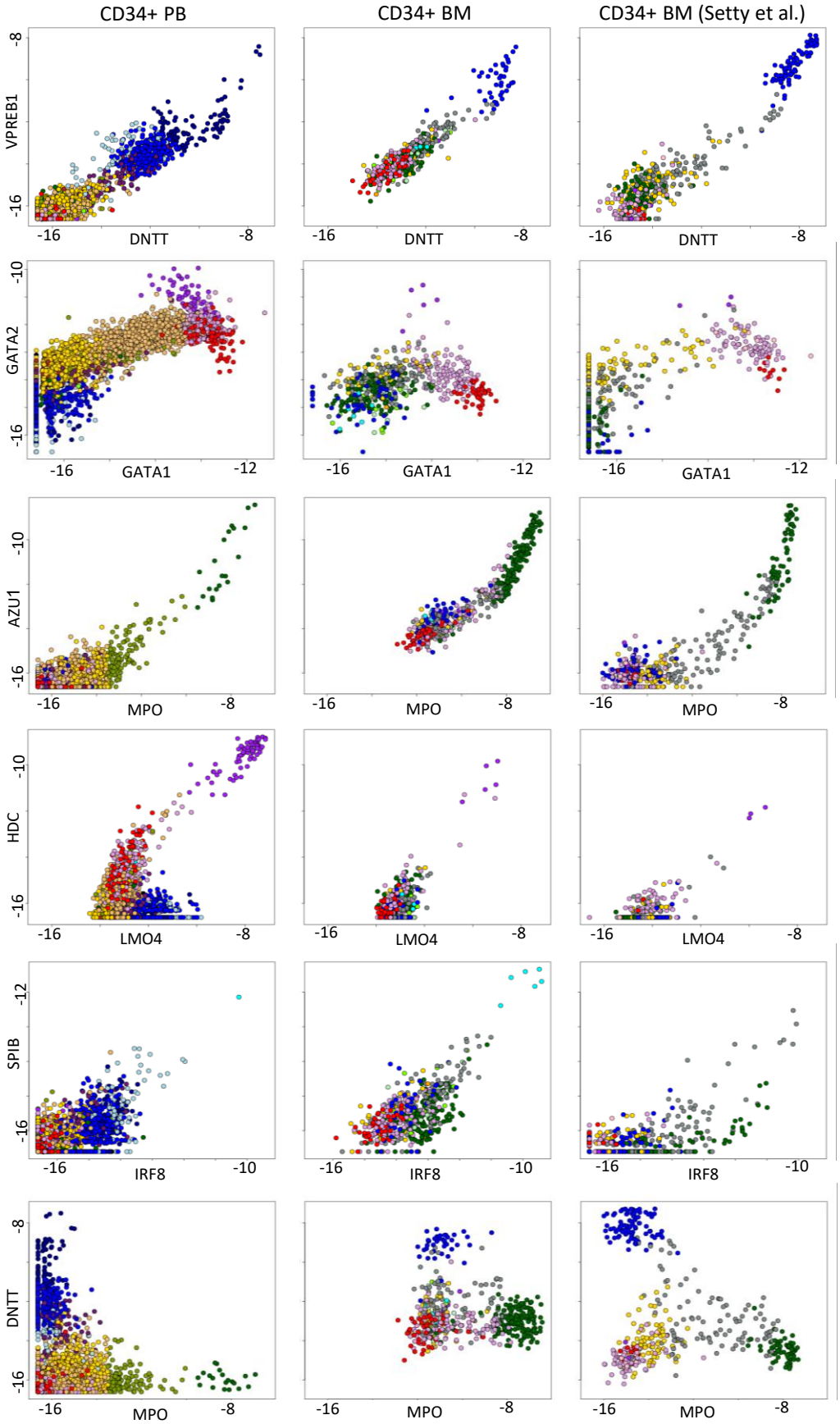


D

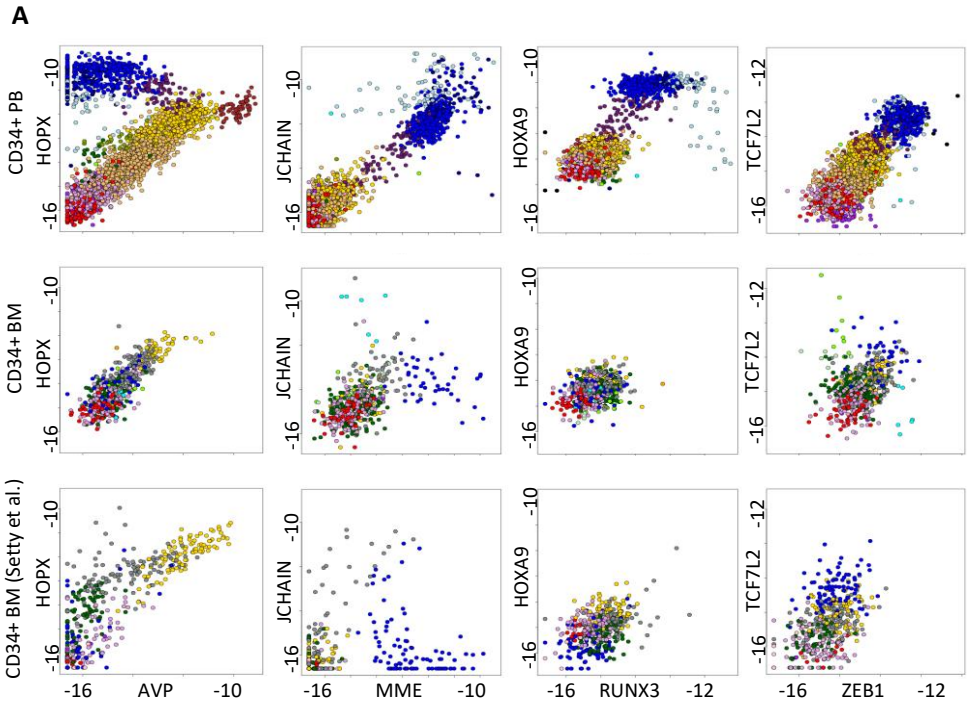
UMAP Projection CD34+ BM (Setty et al.)

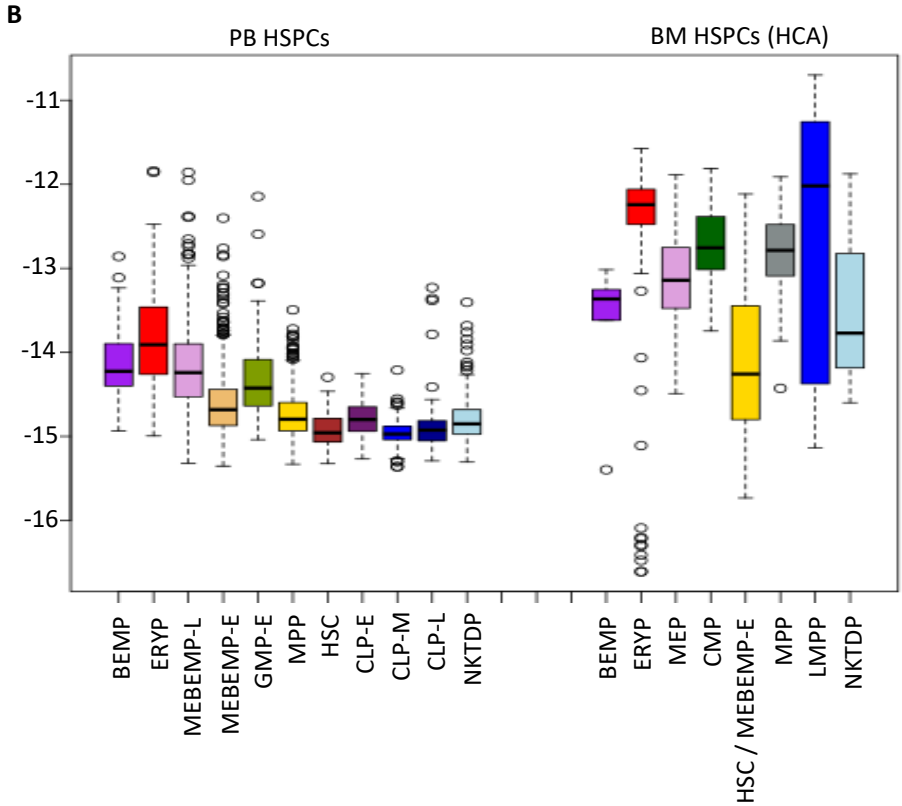


E

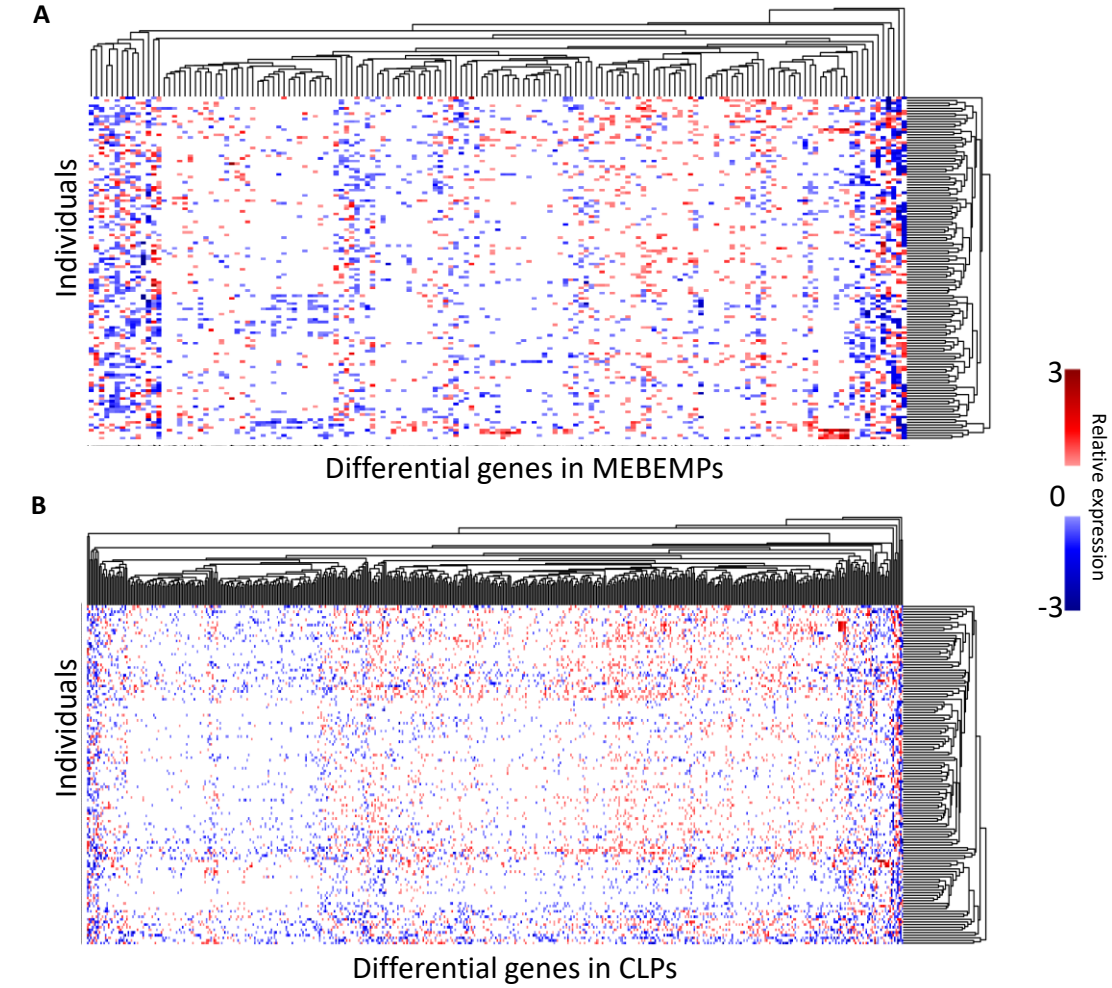


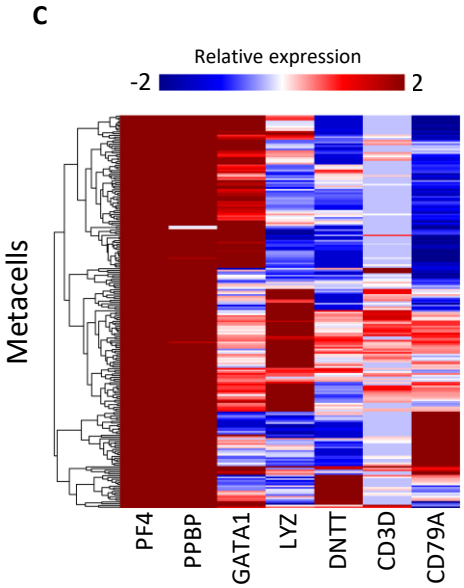
EDF 3


A
CD34+ PB
CD34+ BM
CD34+ BM (Setty et al.)
HOPX
JCHAIN
HOXA9
TCF7L2
-10
-16
-12
AVP
MME
RUNX3
ZEB1


B
PB HSPCs
BM HSPCs (HCA)
-11
-12
-13
-14
-15
-16
BEMP
ERYP
MEBEMP-L
MEBEMP-E
GMP-E
MPP
HSC
CLP-E
CLP-M
CLP-L
NKTDP
MEP
CMP
HSC / MEBEMP-E
LMPP

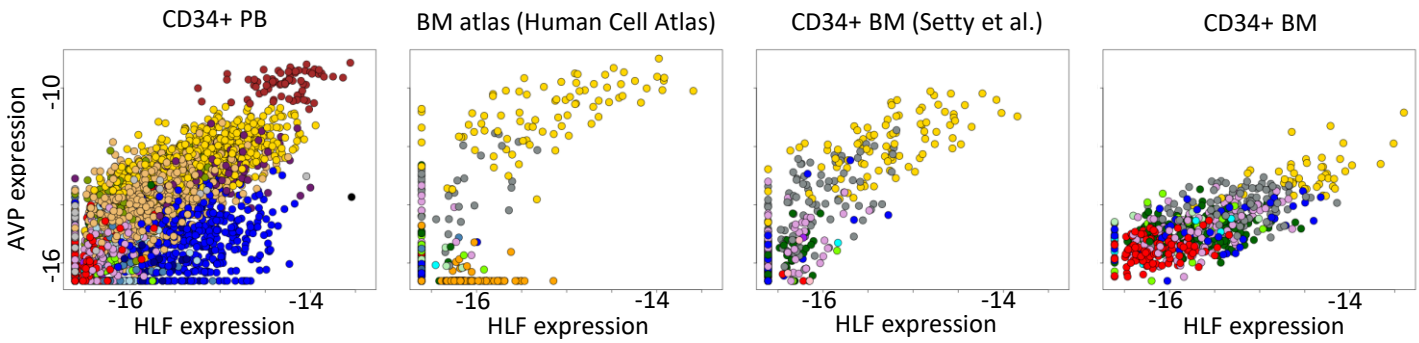
EDF 4


A
Individuals
Differential genes in MEBEMPs
B
Individuals
Differential genes in CLPs
3
0
-3
Relative expression

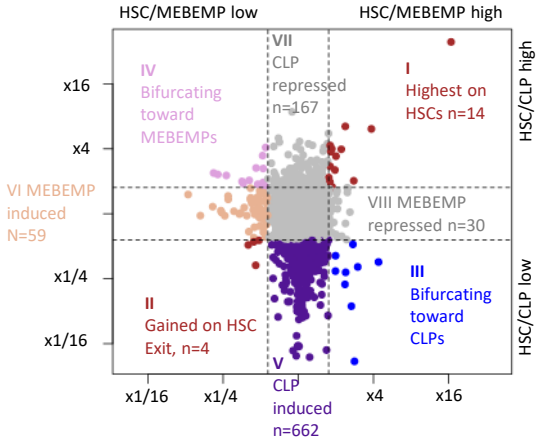

C
Relative expression
-2
2
Metacells
PF4
PPBP
GATA1
LYZ
DNTT
CD3D
CD79A

EDF 5

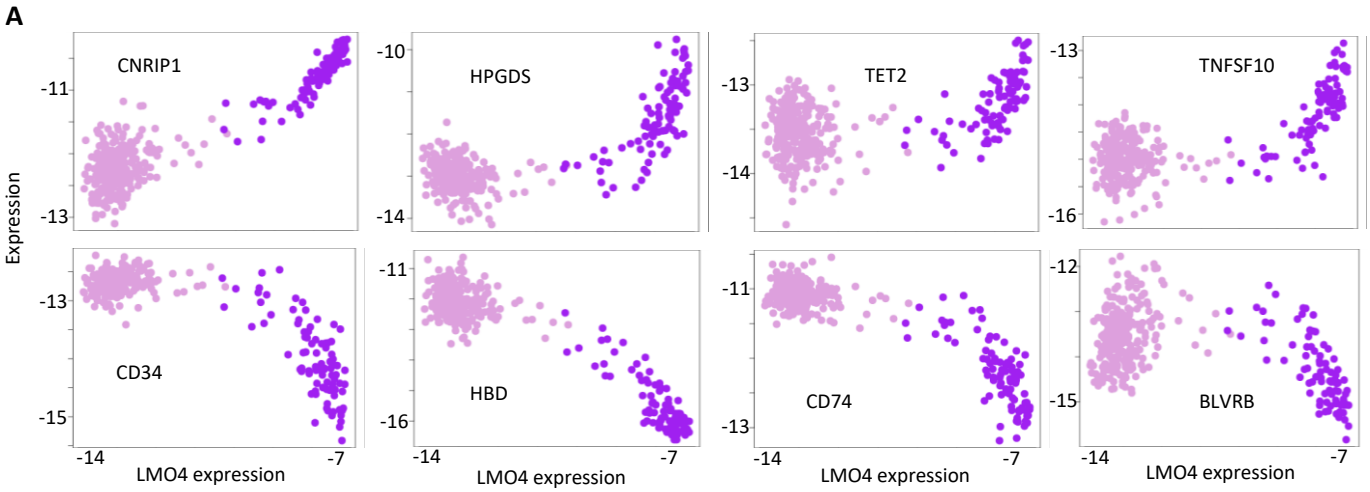
A

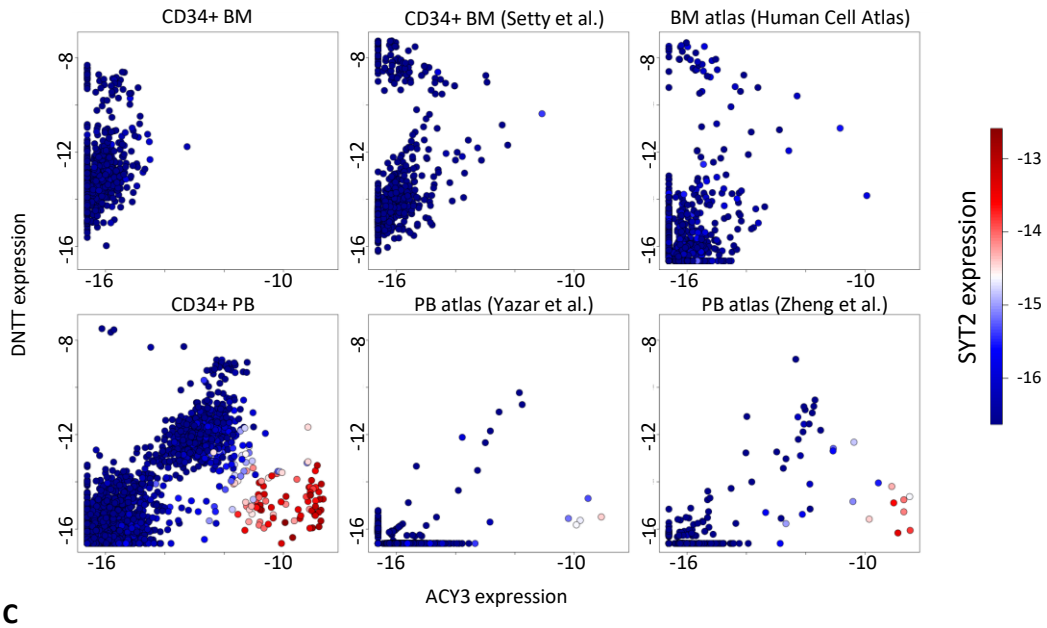

CD34+ PB
BM atlas (Human Cell Atlas)
CD34+ BM (Setty et al.)
CD34+ BM
AVP expression
-10
-16
-16
-14
HLF expression

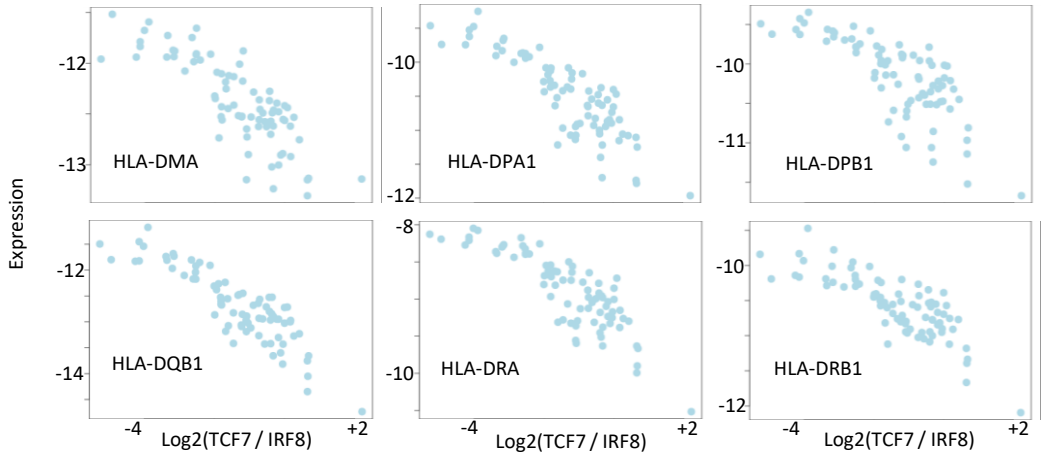
B


HSC/MEBEMP low
HSC/MEBEMP high
HSC/CLP high
HSC/CLP low
x16
x4
x1/4
x1/16
IV
Bifurcating
toward
MEBEMPs
VII
CLP
repressed
n=167
I
Highest on
HSCs n=14
VI MEBEMP
induced
N=59
VIII MEBEMP
repressed n=30
III
Bifurcating
toward
CLPs
II
Gained on HSC
Exit, n=4
V
CLP
induced
n=662

EDF 6

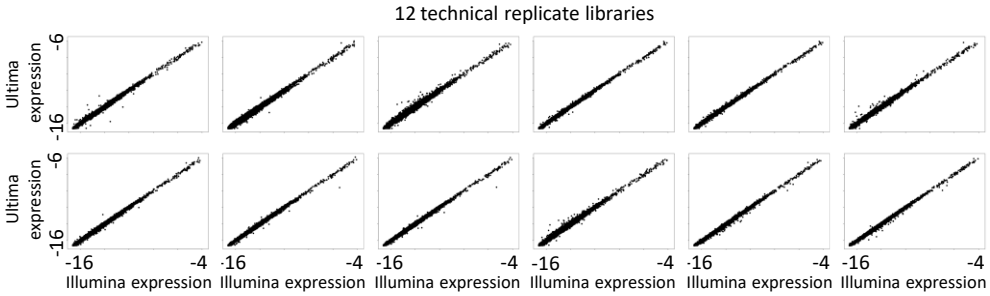

A
CNRIP1
HPGDS
TET2
TNFSF10
CD34
HBD
CD74
BLVRB
Expression
LMO4 expression
-14
-7


B
CD34+ BM
CD34+ BM (Setty et al.)
BM atlas (Human Cell Atlas)
CD34+ PB
PB atlas (Yazar et al.)
PB atlas (Zheng et al.)
DNTT expression
ACY3 expression
SYT2 expression
-16
-10

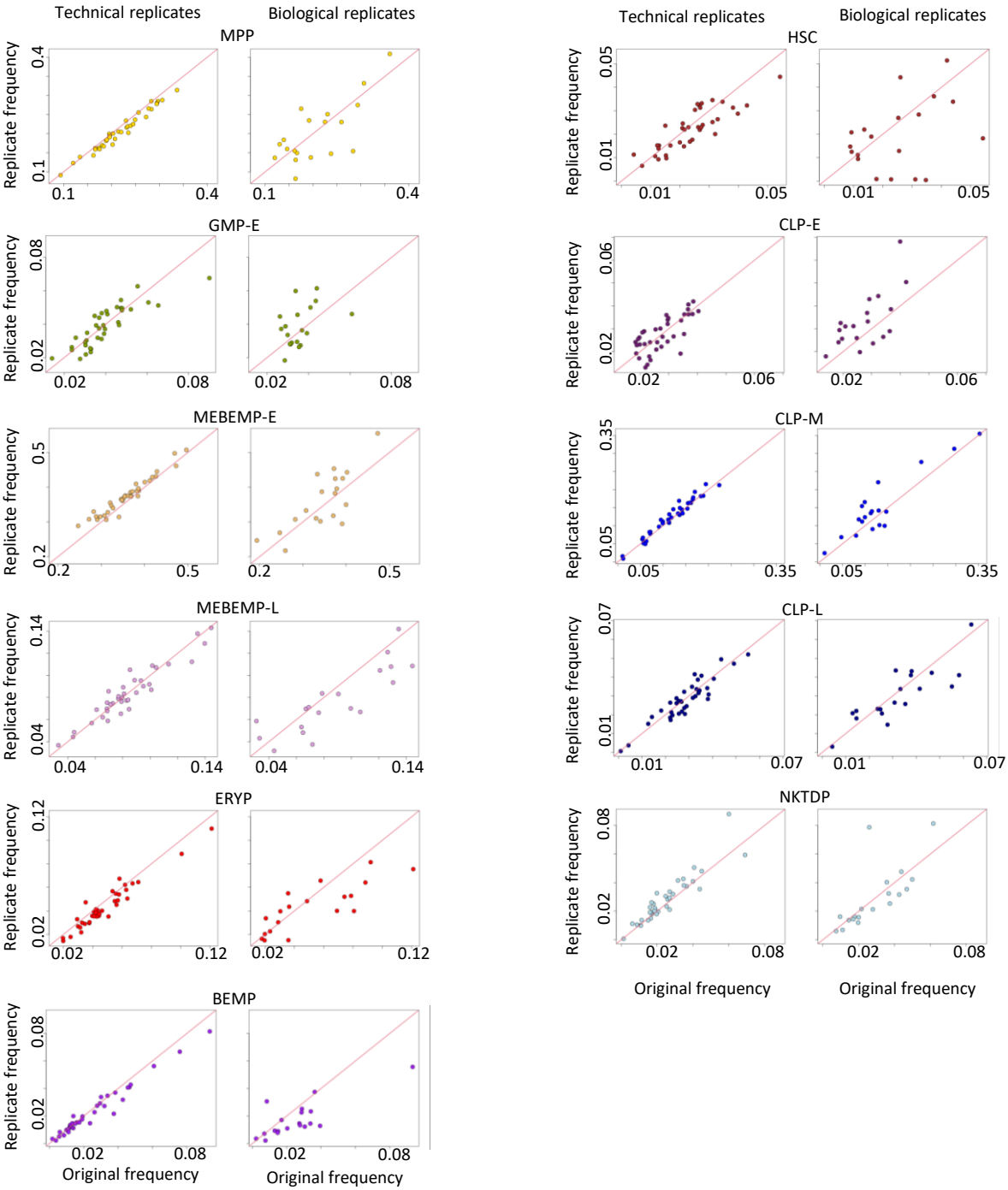

C
HLA-DMA
HLA-DPA1
HLA-DPB1
HLA-DQB1
HLA-DRA
HLA-DRB1
Expression
Log2(TCF7 / IRF8)
-4
+2

EDF 7

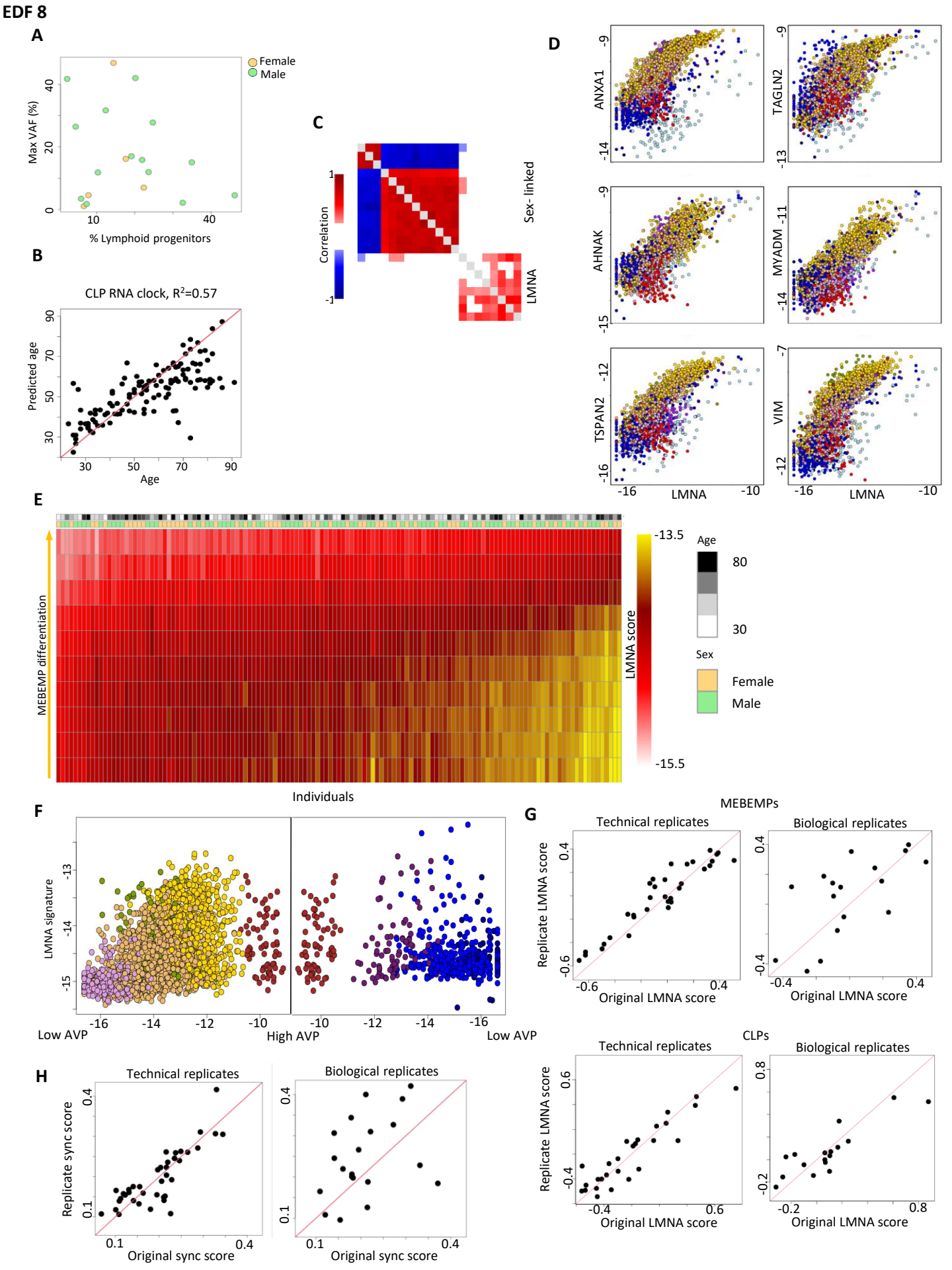
A


12 technical replicate libraries
Ultima expression
-6
-16
-16
-4
Illumina expression

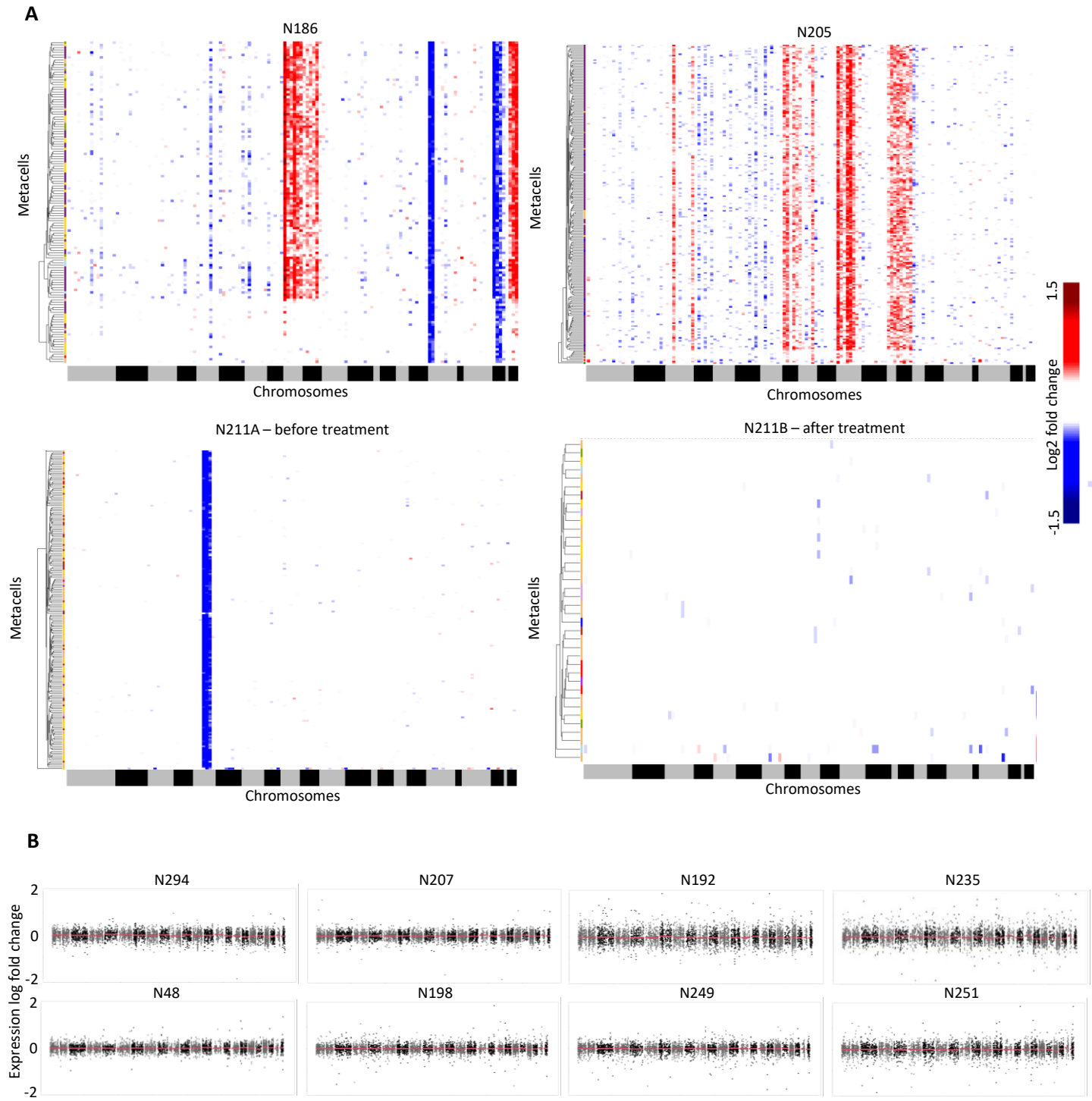
B


Technical replicates
Biological replicates
MPP
HSC
GMP-E
CLP-E
MEBEMP-E
CLP-M
MEBEMP-L
CLP-L
ERYP
NKTDP
BEMP
Replicate frequency
Original frequency

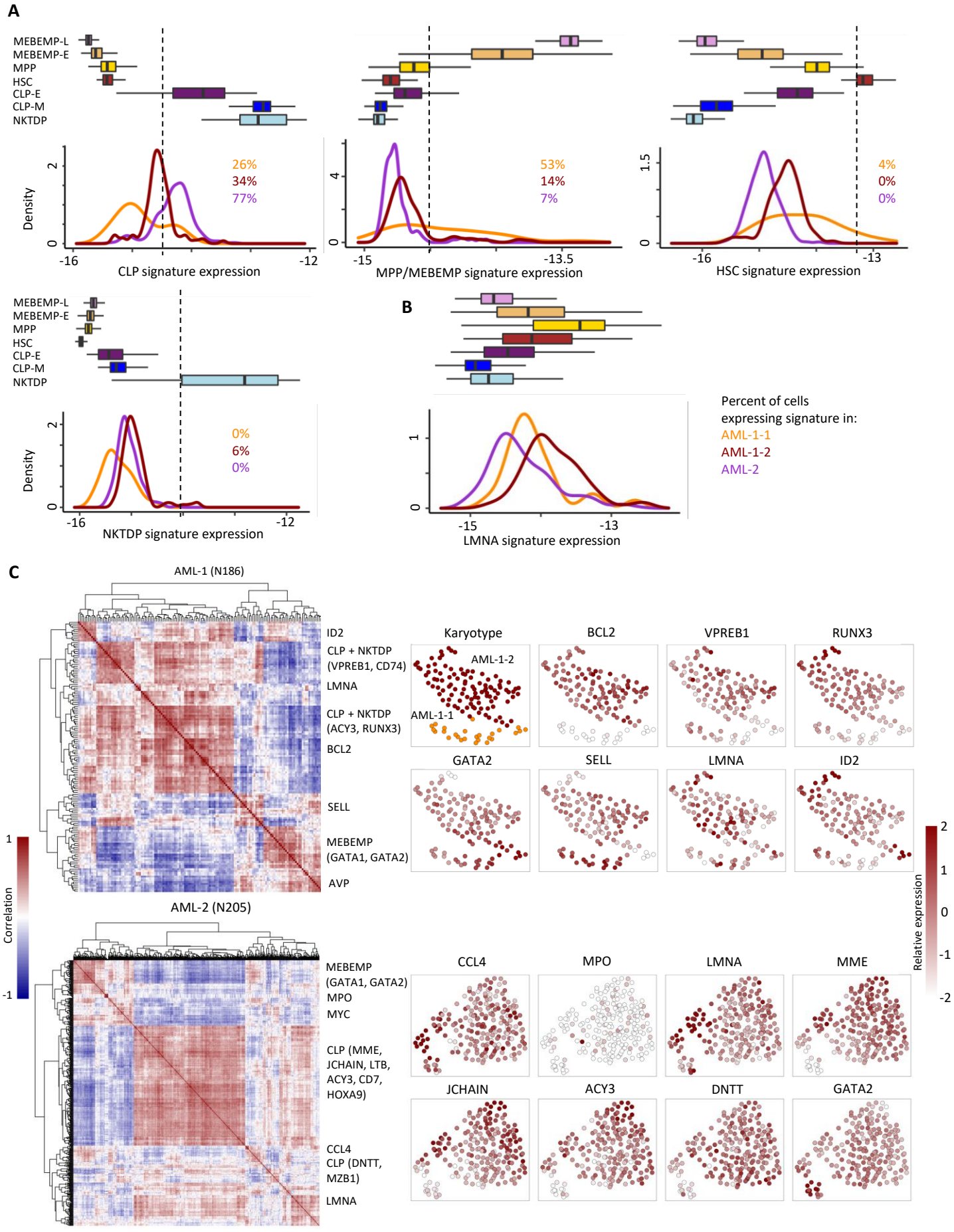
EDF 8


A
Female
Male
Max VAF (%)
% Lymphoid progenitors
B
CLP RNA clock, R²=0.57
Predicted age
Age
C
Correlation
Sex- linked
LMNA
D
ANXA1
TAGLN2
AHNAK
MYADM
TSPAN2
VIM
LMNA
E
MEBEMP differentiation
Individuals
LMNA score
Age
Sex
Female
Male
F
LMNA signature
Low AVP
High AVP
Low AVP
G
MEBEMPs
Technical replicates
Biological replicates
Replicate LMNA score
Original LMNA score
CLPs
H
Technical replicates
Biological replicates
Replicate sync score
Original sync score

EDF 9

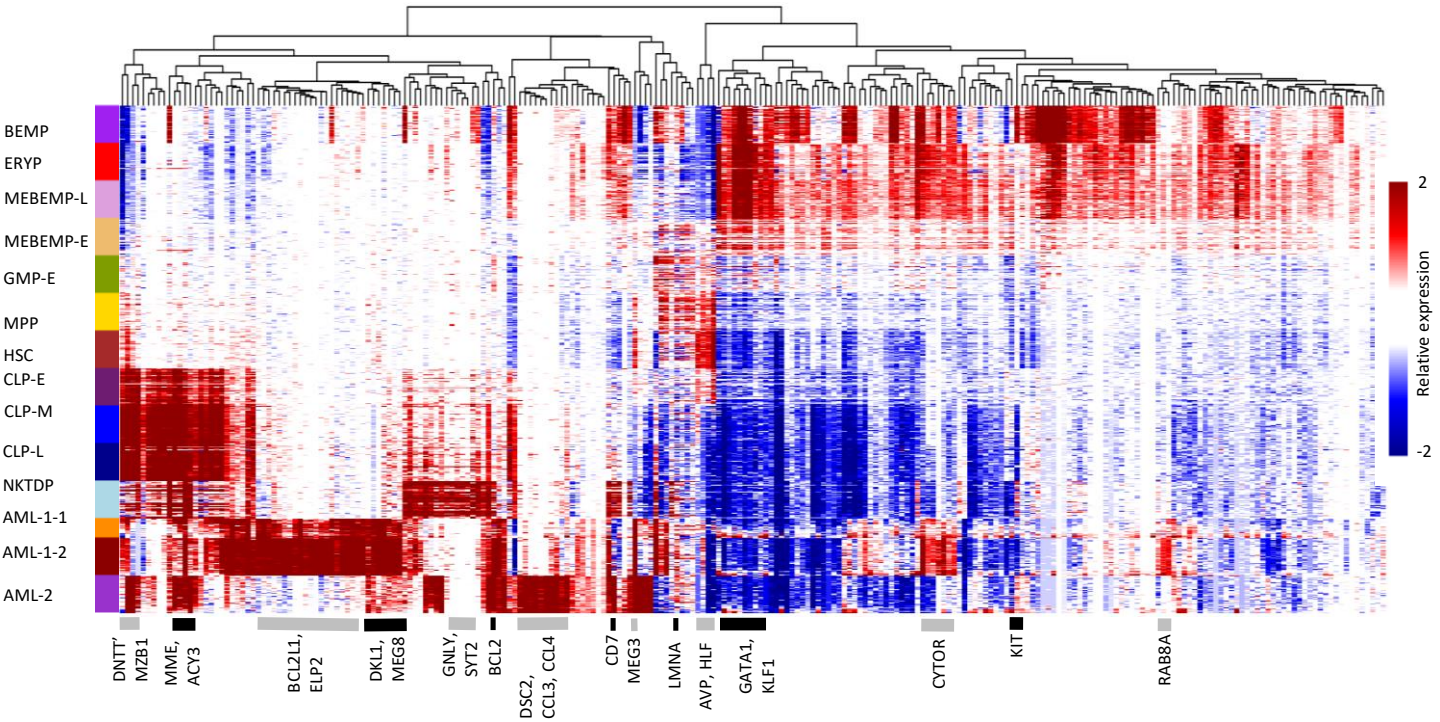

A
N186
N205
N211A – before treatment
N211B – after treatment
Metacells
Chromosomes
Log2 fold change
1.5
-1.5
B
N294
N207
N192
N235
N48
N198
N249
N251
Expression log fold change
2
0
-2

EDF 10

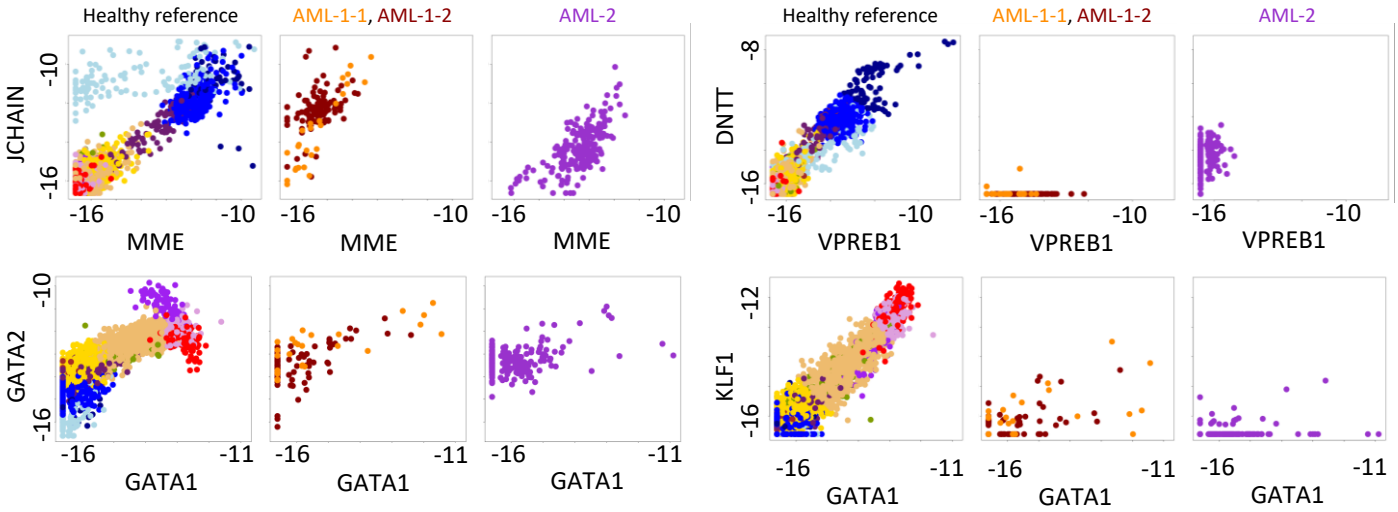

A
MEBEMP-L
MEBEMP-E
MPP
HSC
CLP-E
CLP-M
NKTDP
Density
26%
34%
77%
CLP signature expression
53%
14%
7%
MPP/MEBEMP signature expression
4%
0%
0%
HSC signature expression
0%
6%
0%
NKTDP signature expression
B
LMNA signature expression
Percent of cells expressing signature in:
AML-1-1
AML-1-2
AML-2
C
AML-1 (N186)
ID2
CLP + NKTDP (VPREB1, CD74)
LMNA
CLP + NKTDP (ACY3, RUNX3)
BCL2
SELL
MEBEMP (GATA1, GATA2)
AVP
Karyotype
AML-1-2
AML-1-1
BCL2
VPREB1
RUNX3
GATA2
SELL
LMNA
ID2
Correlation
Relative expression
AML-2 (N205)
MEBEMP (GATA1, GATA2)
MPO
MYC
CLP (MME, JCHAIN, LTB, ACY3, CD7, HOXA9)
CCL4
CLP (DNTT, MZB1)
LMNA
CCL4
MPO
LMNA
MME
JCHAIN
ACY3
DNTT
GATA2

EDF 11

A


BEMP
ERYP
MEBEMP-L
MEBEMP-E
GMP-E
MPP
HSC
CLP-E
CLP-M
CLP-L
NKTDP
AML-1-1
AML-1-2
AML-2
2
Relative expression
-2
DNTT' MZB1
MME, ACY3
BCL2L1, ELP2
DKL1, MEG8
GNLY, SYT2
BCL2
DSC2, CCL3, CCL4
CD7
MEG3
LMNA
AVP, HLF
GATA1, KLF1
CYTOR
KIT
RAB8A

B


Healthy reference
AML-1-1, AML-1-2
AML-2
JCHAIN
MME
DNTT
VPREB1
GATA2
GATA1
KLF1
-10
-16
-8
-11
-12