

SUPPLEMENTARY MATERIALS

TITLE: Progenitor hierarchy of chronic myelomonocytic leukemia identifies inflammatory monocytic-biased trajectory linked to worse outcomes

AUTHORS: Meghan C. Ferrall-Fairbanks^{*1,2,3}, Abhishek Dhawan^{*4}, Brian Johnson³, Hannah Newman⁴, Virginia Volpe⁴, Christopher Letson⁴, Markus Ball⁴, Anthony M. Hunter⁵, Maria E. Balasis⁴, Traci Kruer⁴, Nana Adjoa Ben-Crentsil⁴, Jodi L. Kroeger⁶, Robert Balderas⁷, Rami S. Komrokji⁴, David A Sallman⁴, Jing Zhang⁸, Rafael Bejar⁹, Philipp M. Altrock^{**3,10}, and Eric Padron^{**4}

AFFILIATION: ¹University of Florida Health Cancer Center, University of Florida, Gainesville, FL; ²J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL; ³Department of Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, FL; ⁴Department of Malignant Hematology, Moffitt Cancer Center, Tampa, FL; ⁵Department of Hematology and Medical Oncology, Winship Cancer Institute of Emory University, Atlanta, GA; ⁶Flow Cytometry Core Facility, Moffitt Cancer Center, Tampa, FL; ⁷BD Biosciences, San Jose, CA; ⁸McArdle Laboratory for Cancer Research, University of Wisconsin-Madison, Madison, WI ⁹Moore's Cancer Center, University of California San Diego Health, La Jolla, CA; ¹⁰Department of Evolutionary Theory, Max Planck Institute for Evolutionary Biology, Ploen, Germany

*- authors contributed equally to this work

** - senior authors contributed equally to this work

Corresponding author: Eric Padron

Mailing Address: Moffitt Cancer Center, 12902 USF Magnolia Drive, Tampa, Florida-33617, USA

Email: Eric.Padron@moffitt.org

Phone Number: +1-813-745-8264

SUPPLEMENTARY METHODS

31
32

33 **Software Version**

34 Analysis in R used version 3.6.0 or greater. Seurat package version used was 3.1.2 or
35 greater. Other package versions are specified when discussed.

36

37 **Projection onto Palantir t-SNE, Setty Rep 1**

38 Differentiation trajectories for each sample were calculated using previously computed
39 normal hematopoietic trajectories from Palantir (6), a tool which uses marker genes to
40 assign probabilities of differentiation. While Palantir succeeds with normal samples, the
41 use of very few marker genes made the tool prone to inaccuracies or uninterpretable
42 results stemming from the aberrations introduced by malignant samples. Therefore,
43 instead of relying on single genes in our malignant samples, each malignant cell was
44 assigned the branch probabilities and t-SNE coordinates of nearest-neighbor reference
45 cells using the first 50 dimensions of Harmony-adjusted PCA space.

46 To leverage all three replicates from the Setty paper, the three samples were
47 merged into one Seurat object. The samples were then normalized and scaled using
48 LogNormalize() and ScaleData() Seurat functions. ScaleData() by default uses the top
49 2000 variable features. PCA was performed on the scaled data using the RunPCA()
50 Seurat function. Harmony was used (parameters theta = 1, max.iter.harmony = 20,
51 group.by.vars = sample) to reduce sample-to-sample variation. Then, replicate 2 and 3
52 were assigned the t-SNE coordinates of the nearest replicate 1 cell in the first 50
53 dimensions of Harmony-adjusted PCA space. The result produced t-SNE coordinates in

54 the same replicate 1 embedding for all 3 replicates. Additionally, branch probabilities of
55 each cell from all three replicates were taken directly from the Setty data.

56 For each malignant sample, Quality Control (QC) was performed as in the Setty
57 paper, to eliminate possible non-biological sources of difference without removing any
58 additional reference cells. Cells were removed if they had <1000 UMI count, < 315
59 genes per cell, or mitochondrial percentage > 20%, as was done in the Setty paper. For
60 each malignant sample, a combined object was created by merging the three Setty
61 replicates and the malignant sample. This combined object (four samples, 3 Setty, 1
62 malignant) was log-normalized using the LogNormalize() function in Seurat and then
63 scaled using the ScaleData() Seurat function. PCA and Harmony were performed on
64 the scaled object, using the RunPCA() and RunHarmony() functions in Seurat (Harmony
65 parameters: theta = 2, Max.iter.harmony = 20). In this run of Harmony, differences due
66 to lab protocols were removed and therefore the malignant sample was integrated
67 against the other three replicates as a group, instead of integrating all of them
68 individually. Then, each malignant cell was assigned the t-SNE coordinates of the
69 nearest replicate 1, 2 or 3 cell in the first 50 dimensions of Harmony-adjusted PCA
70 space. Additionally, each malignant cell was assigned a weighted average of the branch
71 probabilities of the 30 nearest neighbors from replicate 1, 2, and 3, with the weighting
72 calculated as the inverse of the distance in 50-dimensional Harmony-adjusted PCA
73 space. After this was run for each malignant cell in each malignant sample, the result
74 gave t-SNE coordinates and branch probabilities for each malignant sample.

75 All Seurat functions were run with default parameters, except where otherwise
76 noted.

77 **Density Visualization for lineage trajectories**

78 Given the t-SNE coordinates for each sample from the Palantir projection, the density of
79 cells along each branch could be visualized in the t-SNE space. The reference density
80 was again set as the grouping of all three Setty replicates. First, the `kde2d()` function
81 (from MASS R package version 'MASS_7.3-53.1', $n=200$, $h=3$) was run on the t-SNE
82 embeddings of both the reference and each malignant sample separately. Then, the
83 results were log-transformed separately with a scale factor of 1000 and a pseudo-count
84 of 1. The reference was subtracted from each malignant sample, and the `melt()` function
85 (from R package 'reshape2_1.4.4') was run to format the data. The resulting data was
86 plotted using `ggplot2` (`geom_tile()` and `geom_point()`) to show over- (blue) and under-
87 (red) densities relative to the reference.

88

89 **Single-cell RNA sequencing Quality Control (QC)**

90 Data from publicly available normal samples was imported and a Seurat object was
91 created with the eight normal samples. A lower cutoff for the number of features was set
92 to 450. The percent of mitochondrial RNA was set to 0.05 for normal samples to remove
93 dead cells. For each normal dataset (6-8), cells with number of features greater than 2
94 standard deviations above the dataset mean were removed to account for possible
95 doublets. From the Setty dataset, 25,041 of 41,331 (60.5%) cells were kept. From the
96 Zheng dataset, 8,799 of 9,262 (95%) cells were kept. From the Hua dataset, 29,832 of
97 32,289 (92.4%) cells were kept. Zheng and Hua datasets were previously filtered for
98 mitochondrial content, explaining the higher percentage of quality cells in those
99 datasets.

100 For the CMML samples, a lower feature cutoff of 450 was also used. For
101 mitochondrial content, 25% was used as the lower cutoff due to the higher percentage
102 of mitochondrial RNA in cancer cells. Any cells with a feature count greater than 2
103 standard deviations above the mean were also excluded to account for doublets. Of
104 182,189 initial cells, 137,578 (75.5%) high-quality cells were kept.

105

106 **Pseudo-Bulk Aggregation, pseudo-bulk UMAP, Ward clustering, and Signature** 107 **Heatmap**

108 The input for pseudo-bulk aggregation was all quality-controlled and log-normalized
109 scRNAseq data (39 Moffitt + 8 Normal). `ScaleData()` was run on this data, again
110 keeping the first 2000 variable features. The dataset was then divided by sample. For
111 each sample, the arithmetic mean of the scaled data was calculate for each of the 2000
112 features. The result is a matrix with 47 rows (one for each sample) and 2000 columns
113 (one for each variable feature/transcript). UMAP was performed on this matrix using the
114 `umap()` function (parameters: `n_neighbors = 39`, `metric = 'euclidean'`, `min_dist = 0.05`) of
115 the `uwot` package (version: `uwot_0.1.10`).

116 Additionally, hierarchical clustering was performed on this pseudo-bulk matrix.
117 Distances were computed using the `dist()` function with “euclidean” method from the
118 `stats` package (version: 4.0.2). The clustering was performed on the distances using the
119 `hclust()` function from the `stats` package with method “ward.D2” (version: 4.0.2). The
120 resulting dendrogram was divided into four groups, corroborating the groupings based
121 on bulk UMAP and lineage bias.

122 Using the groupings defined by the Ward hierarchical clustering, the heatmap
123 shown in **Fig. 1E** was constructed from a similar pseudo-bulk matrix to the one detailed
124 above. Instead of using the top 2000 variable features, this matrix was constructed
125 using only the 180 features (top 60 from each of HSC, GMP, MEP) from the Wu gene
126 signatures (13). The arithmetic mean of the scaled expression (z-score) is plotted for
127 each sample, and each gene in the signature.

128

129 **PCA, Harmony, UMAP, and Louvain Clustering**

130 All high-quality cells (n = 201,250) were used in both UMAP projection and clustering,
131 as shown in **Fig. 3A**. Steps were performed in R (version >= 3.6.0) using R package
132 Seurat (version >= 3.1.2). First, quality-controlled count data was log-normalized using
133 Seurat function `NormalizeData()` with default parameters stated here for redundancy
134 (“`normalization.method = 'LogNormalize'`” and “`scale.factor = 10000`”). Next, Seurat
135 function `FindVariableFeatures()` identified the top 2000 features using default
136 parameters (“`selection.method = 'vst'`”). Normalized data was then scaled using Seurat
137 function `ScaleData()`. `ScaleData()` scales the top 2000 features identified using
138 `FindVariableFeatures()`. `ScaleData()` allows the option of controlling for variables using
139 the “`vars.to.regress`” parameter. To eliminate differences due to dead cells (high
140 mitochondrial count) and differing read targets across datasets, we specified
141 “`vars.to.regress = c('nCount_RNA', 'percent.mito')`”. PCA was then performed on the
142 scaled data using the top 2000 features and the default parameters of Seurat function
143 `RunPCA()`. Harmony (version 0.1.0) (9) was then used to correct for batch effects due to
144 different datasets. Seurat function `RunHarmony()` was used as a wrapper to Harmony.

145 RunHarmony() parameter “group.by.vars” was set to variable “tech” in the metadata of
146 the Seurat object which specifies the dataset of the each individual cell. Other
147 RunHarmony() parameters were set to default (sigma = 0.1, reduction = “pca”). The
148 output from RunHarmony() is a corrected PCA embedding which is then used for further
149 analysis.

150 Using the first 50 dimensions of the “harmony” reduction, neighbor graph
151 construction was performed using the default parameters of the FindNeighbors() Seurat
152 function (k.param = 20, reduction = “harmony”, dims = 1:50, n.trees = 50). Using this
153 neighbor graph, clusters were constructed at various resolutions using Louvain
154 clustering as implemented in the FindClusters() Seurat function with other parameters
155 set to default (algorithm = 1, n.start = 10, n.iter = 10). Resolutions used were: (0.025,
156 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2) which identified between 9 (0.025) and 21 (0.2)
157 communities. Clustree (version 0.4.3) was used to visualize cells moving between
158 clusters at various resolutions. We used clustree visualization to identify the resolution
159 of 0.05, with 13 communities, as having the optimal tradeoff between resolution and
160 noise. Next, the RunUMAP() function in Seurat was used for visualization. The first 50
161 dimensions of the “harmony” reduction were used for the RunUMAP() function. Other
162 parameters were set to default values (reduction = “harmony”, dims = 1:50,
163 umap.method = “uwot”, n.neighbors = 30, metric = “cosine”, min.dist = 0.3).

164

165 **SingleR**

166 SingleR (version 1.6.1) (20) is a tool used to assign cell type status to cells profiled
167 using single-cell RNA sequencing. It leverages bulk RNA sequencing from flow-

168 cytometry sorted references to map each individual cell in a query dataset to a cell type
169 in the reference dataset. We use three built-in references from the “SingleR” package;
170 “NovershternHematopoieticData()” (55), “HumanPrimaryCellAtlasData()” (56), and
171 “BlueprintEncodeData()” (57, 58), all of which have several hematopoietic progenitor
172 cell types. We also use an additional reference (GSE42519) which we call the Rapin
173 dataset, originating from published work in Rapin et al. *Blood*. 2014 (21). We restrict
174 the reference cell types to those possibly observed within our CD34+ bone marrow
175 scRNAseq dataset, and then group them into six broader categories: Hematopoietic
176 Stem Cell (HSC), Granulocyte-Macrophage Progenitor (GMP), Megakaryocyte-
177 Erythroid Progenitor (MEP), Common Lymphoid Progenitor (CLP), Multi-Potent
178 Progenitor (MPP), and Common Myeloid Progenitor (CMP). We use the main cell types
179 from the Novershtern and Human Primary Cell Atlas datasets and the fine cell types
180 from the Blueprint Encode dataset. The “main” cell types are broader categories
181 whereas the “fine” cell types are more specific. “Main” and “fine” labels for each
182 reference were chosen as such to ensure that the six broader categories were
183 represented. For example, in the Blueprint Encode dataset, the “main” label grouped
184 MEP and HSC, prompting us to elect the “fine” label, which distinguished between the
185 two cell types. B-cells, T-cells, NK cells and their respective progenitors were grouped
186 with CLP. Erythroblasts were grouped with MEPs. Pro-myelocytes were grouped with
187 GMPs.

188 Cell type assignment was computed using SingleR independently for each
189 reference. Following single reference cell type assignment, assignments were
190 compared across references. If three or four references agreed on a cell type for a

191 given cell from our dataset, that cell type would be assigned to the individual cell. If less
192 than three of the references agreed, the cell type would be classified as “No
193 Consensus”. As seen on the UMAP in **Fig. 3C**, many of the cells with no consensus cell
194 type appear between HSCs and MEPs, indicating that these cells may just have been
195 “caught in between” while undergoing the process of differentiation. Still, only 10.2 % of
196 cells had no consensus cell type. SingleR results also used to show CLP and HSC
197 depletion in CMML, as detailed in **Fig. 1I-J, Supplementary Fig. S5**.

198

199 **Single-cell Pathway Scores**

200 As an orthogonal approach to evaluating cell type, specifically for Clus2 cells, at the
201 single cell level, we used the Wu GMP signature (13) to generate a score for each
202 individual cell. Using all 100 genes from the GMP signature, Seurat function
203 AddModuleScore() was used to assign each cell a score. The scores for cells in cluster
204 2 are shown in (**Supplementary Fig. S10**). This approach was also applied to evaluate
205 upregulation of WNT/ β -catenin signaling by creating a score based on the Gene Set
206 Enrichment Analysis Geneset for unstimulated and WNT pathway stimulated
207 hematopoietic progenitor geneset (GSE26351; **Supplementary Fig. S11C-D**) (54).
208 UMAP of WNT signature score (**Supplementary Fig. S11C**) is made using Seurat
209 FeaturePlot() with “min.cutoff” set to 0 for visualization.

210

211 **mitoClone**

212 mitoClone (version 1.0) (25) uses mitochondrial reads, which typically have better
213 coverage, to infer clonal composition of cells within a sample. From the single cell bam

214 files, the mitoClone function baseCountsFromBamList() with specification “sites = MT:1-
215 16569”, creates count tables. Then, the count tables are used as input into the
216 mutationCallsFromBlacklist() function with parameter: min.af = 0.1, min.num.samples =
217 $0.01 * (\# \text{ cells})$, universal.var.cells = $0.9 * (\# \text{ cells})$, max.var.na = 0.5, max.cell.na = 0.75.
218 The parameters are a balance between the resolution and noise. This choice of
219 parameters is slightly lower resolution but gives greater confidence in the clonal
220 breakdown observed.

221 With multiple samples from the same patient, which are the cases we show in
222 **Fig. 4L-S** and **5A-H**, we run the sequential samples together in the
223 mutationCallsFromBlacklist() function. Then, the phylogenetic reconstruction is done in
224 the muta_cluster() function with default parameters. This step requires a gurobi license,
225 which is free for academic users. The output of muta_cluster gives clonal information
226 and a confidence estimate for each single cell, which can then be used for visualization
227 and clonal distribution calculation. Clonal distribution across samples from the same
228 patient remains remarkably similar, lending confidence that these are observed
229 phylogenies and not simply noise.

230 There are several cases where the clonal reconstruction finds only one clone.
231 This is to be expected with CMML, as it is a clonal disease. There are a few cases in
232 which there are no selected sites for clonal reconstruction, and in those cases, we
233 assume clonality.

234

235

236

237 **COMET**

238 COMETSC (version 0.1.13) (24) is a python package used to identify markers from
239 scRNAseq data to be used in flow cytometry. Currently, there is a limit to the number of
240 cells used for COMET, which is 65,000. For our purposes (to identify markers for Clus2
241 cells), we include all cluster 2 cells and the remaining cells of the 65,000-cell allotment
242 are randomly sampled from non-Clus2 cells. In order to find markers for cluster 2 only, a
243 cell that is in cluster 2 is assigned a 1 and non-cluster 2 cells are assigned 0 for the
244 cluster input “.txt” file. We run COMET with “-K 3” to look for “panels” that are up to 3
245 combinations of individual markers, though we only use a single marker. The output of
246 COMET gives a true positive and true negative, indicating the accuracy of using the
247 given markers to identify the population. Due to expected dropout in scRNAseq data,
248 we prioritize a high true negative value, that is, we want markers which are only present
249 in cluster 2, even if they are not present in every cluster 2 cell (high specificity, low
250 sensitivity).

251

252 **Pathway Analysis**

253 Enrichr (53) was used with differentially expressed features ($p < 0.05$) between cluster 2
254 and other cells. “Panther 2016” (29) pathway from Enrichr is shown in **Fig. 6A**.

255

256 **Dimensionality reduction and unsupervised clustering of high parameter flow** 257 **cytometry data**

258 The unsupervised clustering analysis was performed using FlowJo version 10. The HSC
259 and myeloid progenitors were identified using the gating strategy explained in the

260 results. The fluorescent data from stem and myeloid progenitors (identified by manual
261 gating) of patients and healthy subjects were concatenated. UMAP (UMAP, version
262 2.1) plugin in FlowJo was used for dimensionality reduction. The following parameters
263 were used for dimensionality reduction: nearest neighbors-30, minimum distance-0.5,
264 distance function-Euclidean, 22 fluorescent parameters representing the compensated
265 cytokine receptors in HSCs and 21 fluorescent parameters representing the CRs in
266 myeloid progenitors. Phenograph (version 0.2) (36) plugin in FlowJo was used for
267 clustering as previously described²³. The following parameters were used for clustering:
268 k-nearest neighbors=30. 22 and 21 fluorescent parameters representing the
269 compensated CRs. Manual gating of each of the CRs was also performed to calculate
270 MFI and percentage positive data.

271

272

SUPPLEMENTARY TABLES

273

274 **Supplementary Table S1.** Comparison of clinical baseline characteristics between
275 patients in scRNA-Seq and FCM cohorts. The comparisons were made using non-
276 parametric Mann-Whitney test, Fischer's exact test and Chi-square analysis. The
277 comparisons revealed comparable clinical baseline characteristics between the 2
278 patient cohorts.

279

280 **Supplementary Table S2. Panther Pathways from Genes Upregulated in Cluster 2**
281 **Compared to Cluster 0.**

282

283 **Supplementary Table S3. Panther Pathways from Genes Downregulated in**
284 **Cluster 2 Compared to Cluster 0.**

285

286 **Supplementary Table S4. FPKM values of 51 receptors in Healthy and CMML**
287 **CD34+ cells extracted from bulk RNA-Seq datasets.**

288

289 **Supplementary Table S5. Baseline characteristics of patient samples used in**
290 **scRNA-Seq cohort.** The table shows baseline characteristics of each of the samples
291 used in scRNA-Seq study.

292

293 **Supplementary Table S6. Baseline characteristics of patient samples used in flow**
294 **cytometry cohort.** The table shows baseline characteristics of each of the samples
295 used in FCM study.

296

297 **Supplementary Table S7. Summary statistics of samples used for scRNA-Seq.**

298

299 **Supplementary Table S8. TotalSeq™-D Human Heme Oncology Cocktail, V1.0.**
300 The table details the specificity, clone, barcode sequence of each of the 45 antibodies
301 used in the TotalSeq study.

302

303 **Supplementary Table S9. Myeloid Panel (45 Genes, 312 Amplicons).** The table lists
304 the genes profiled in the myeloid panel.

305

306 **Supplementary Table S10. Reagent information for PE-conjugated flow cytometry**
307 **screen.**

308

309 **Supplementary Table S11. Reagent information for CRD flow cytometry panel.**

310

311 **Supplementary Table S12. Reagent information for murine stem and progenitor**
312 **flow cytometry panel.**

313

314 **Supplementary Table S13. Reagent information for PDX flow cytometry panel.**

315

316 **Supplementary Table S14. Baseline characteristics of publicly available normal**
317 **samples.**

318 **SUPPLEMENTARY FIGURES**

319
320 **Suppl. Fig. S1.** Consort Diagram of CMML patient samples evaluated with single-cell
321 RNA sequencing and high parameter flow cytometry (FCM) in this study.

322
323 **Suppl. Fig. S2.** Pseudo-bulk aggregation analysis of scRNAseq showed distinct three
324 differentiation trajectories.

325
326 **Suppl. Fig. S3.** Three distinct trajectories were confirmed from projection of CMML
327 samples onto a single-cell proteo-genomic reference map of hematopoiesis.

328
329 **Suppl. Fig. S4.** Clinical parameter associations with monocytic-bias, MEP-biased, and
330 normal-like patient groupings showed no significant differences in blast percentage,
331 platelets, WBC, ALC, ANC, and absolute monocytosis.

332
333 **Suppl. Fig. S5.** CMML patients show HSC depletion as compared to normals.

334
335 **Suppl. Fig. S6.** Single-cell gene expression of HSC signatures show depletion in HSCs
336 in CMML.

337
338 **Suppl. Fig. S7.** Gating strategy used for identification of stem and myeloid progenitor
339 populations in CMML patients and controls.

340
341 **Suppl. Fig. S8.** Clinical characteristics of patients with HSC depletion.

342
343 **Suppl. Fig. S9.** Cluster 2 drives Mono-bias assignment.

344
345 **Suppl. Fig. S10.** Gene expression analysis of Clus2 cells showed GMP like signature.

346
347 **Suppl. Fig. S11.** Expression of CTNNB1, IRF8, and WNT pathway signature score in
348 CMML GMPs in scRNAseq cohort.

349
350 **Suppl. Fig. S12.** Expression of Fc gamma receptors in scRNAseq cohort.

351
352 **Suppl. Fig. S13.** CD120b expression across stem and progenitor populations.

353
354 **Suppl. Fig. S14.** Merged survival analysis of the single-cell RNA sequencing and flow
355 cytometry cohorts.

356
357 **Suppl. Fig. S15.** Clus2 characterized by CD284 expression.

358
359 **Suppl. Fig. S16.** Palantir mappings with mitoClone clonal information indicated by color
360 for all samples run individually.

361

362 **Suppl. Fig. S17.** Development and optimization of CRD flow panel.

363

364 **Suppl. Fig. S18.** Distribution of HSPCs in competitive BMT studies in *NRAS* model.

365

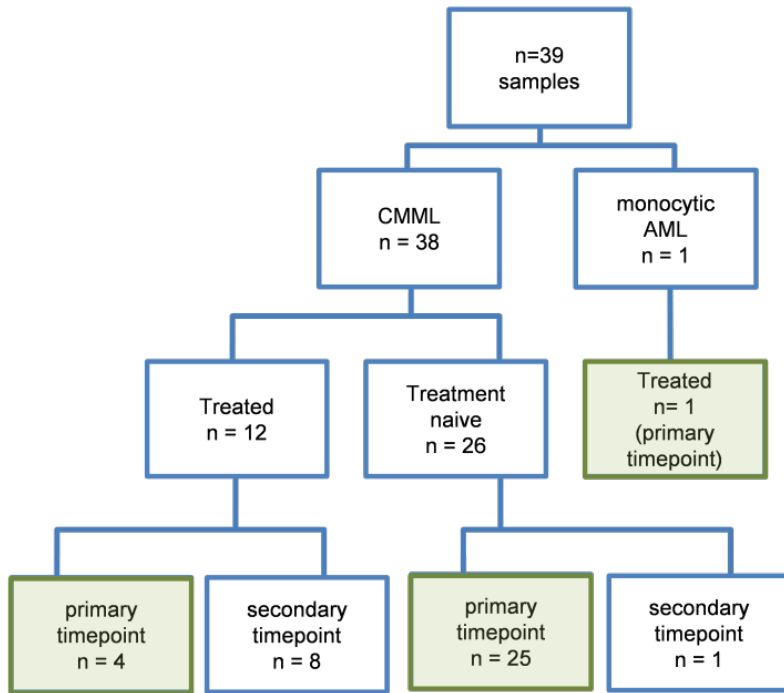
366 **Suppl. Fig. S19.** Plasma cytokine levels 6 hours post injection of LPS or vehicle.

367

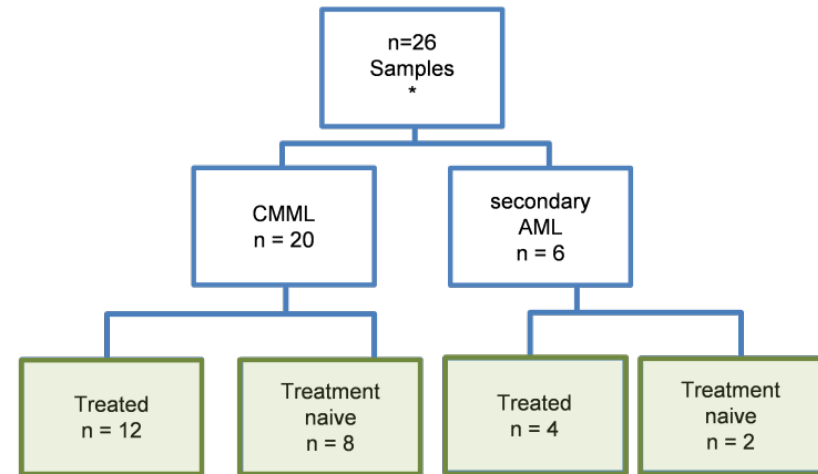
368 **Suppl. Fig. S20.** Cellular density in Palantir pseudotime across differentiation

369 trajectories in normal samples.

single-cell RNA sequencing study



flow cytometry study



unique patients
 scRNAseq: $4+25+1 = 30$
 FCM: $12+8+4+2 = 26$

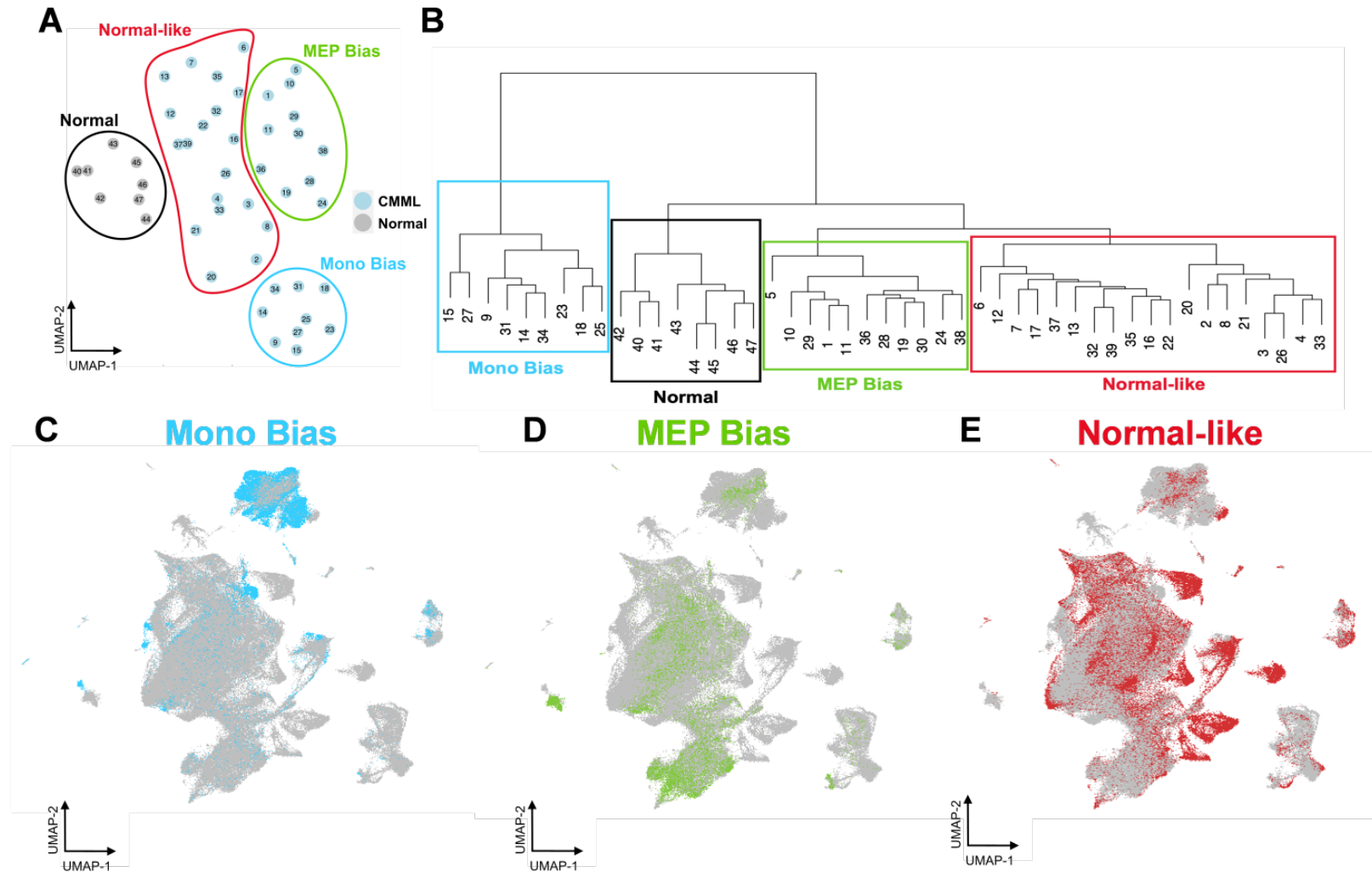
scRNAseq+FCM-overlap:
 $30+26-1 = 55$ unique cases

*one patient sample
 included in both single-cell
 RNA sequencing & flow
 cytometry studies*

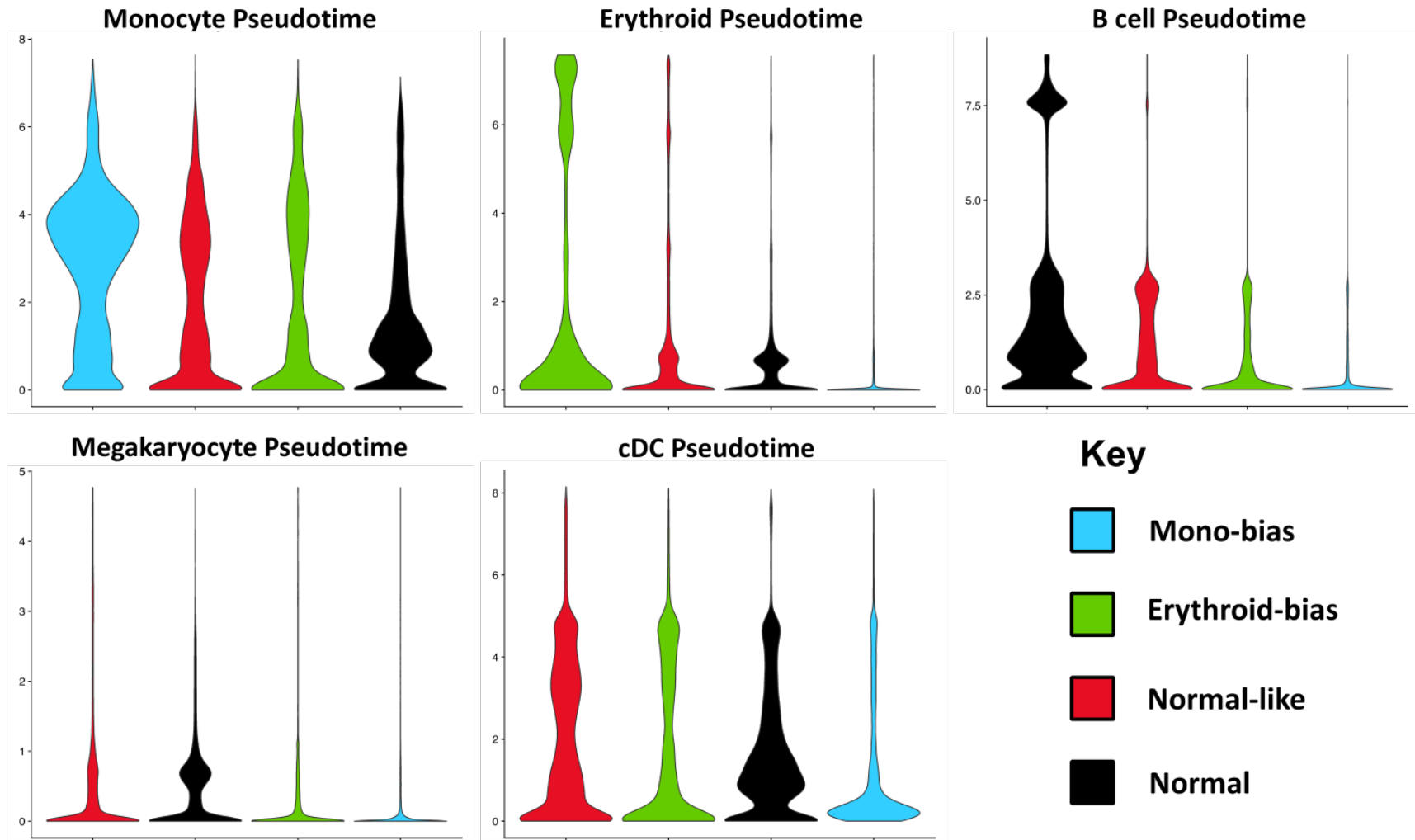
*- samples used for Total Seq

370
 371
 372

Supplementary Figure S1. Consort Diagram of CMML patient samples evaluated with single-cell RNA sequencing and high parameter flow cytometry (FCM) in this study.

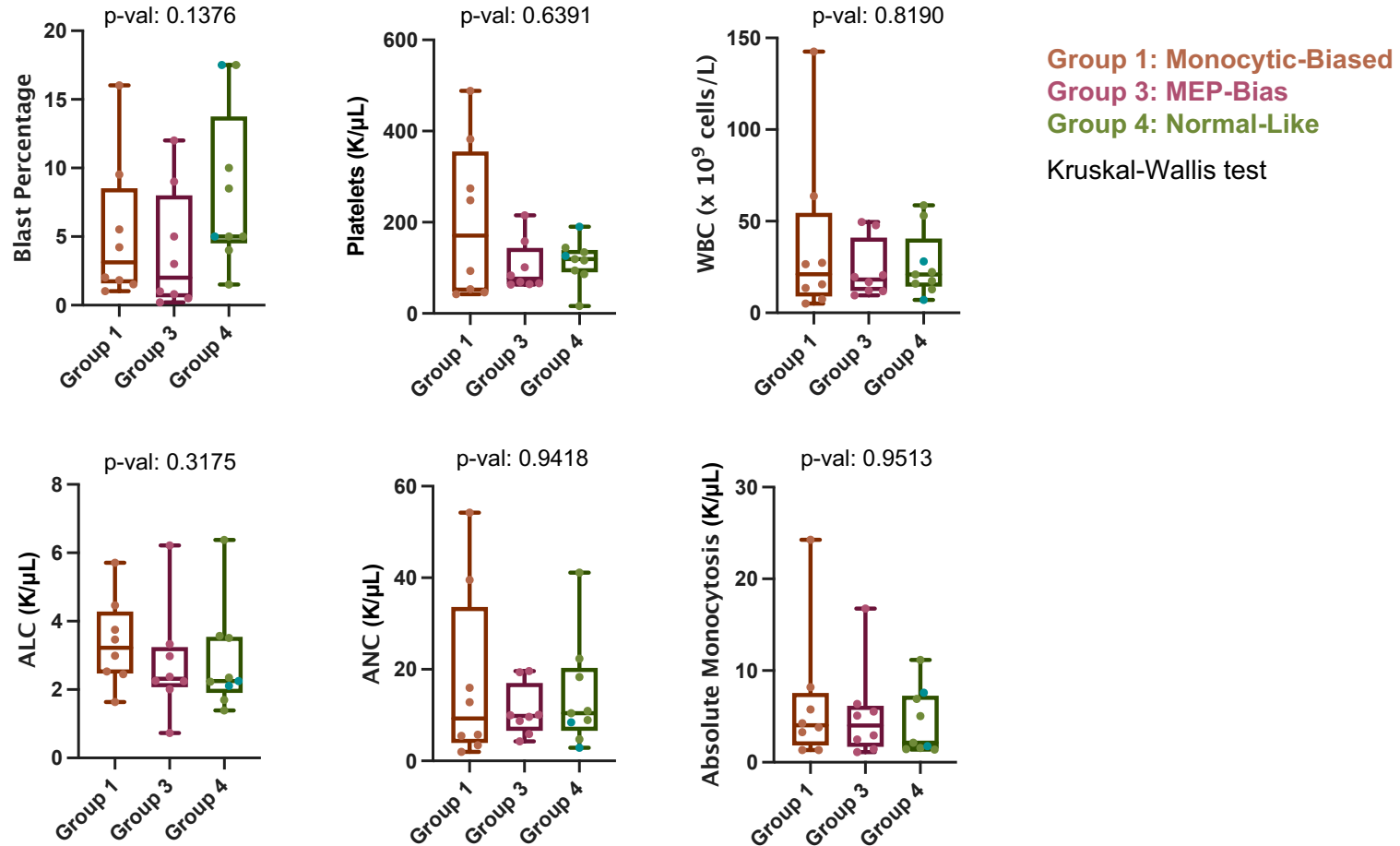


373
 374 **Supplementary Figure S2. Pseudo-bulk aggregation analysis of scRNAseq showed distinct three differentiation**
 375 **trajectories. (A)** Pseudo-bulk aggregation of CMML scRNAseq cohort visualized with UMAP projections. **(B)** Ward
 376 hierarchical clustering of CMML and normal samples identifies the three distinct trajectories. **(C)** Single-cell UMAP
 377 projections highlighting cells from Mono-Bias samples, **(D)** MEP-Bias samples, and **(E)** Normal-like samples.



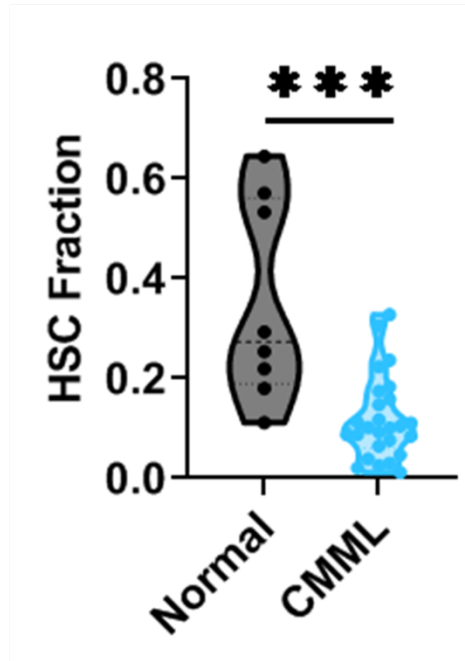
378
 379
 380
 381
 382
 383

Supplementary Fig S3. Three distinct trajectories were confirmed from projection of CMML samples onto a single-cell proteo-genomic reference map of hematopoiesis. Patients categorized as mono-bias had elevated monocyte pseudotime when mapped to the single-cell proteo-genomic reference published by Triana et al. in *Nature Immunology* (2021)



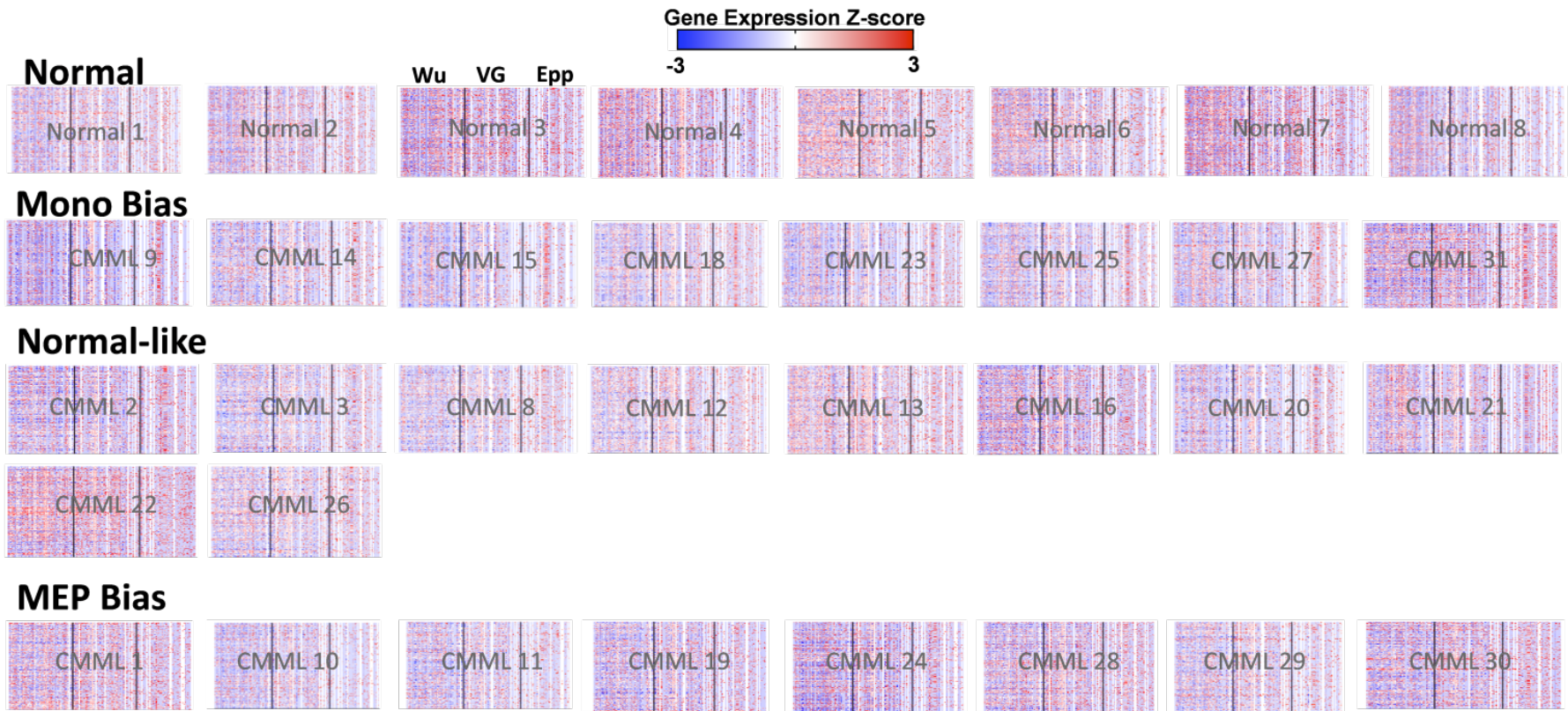
384
 385
 386
 387
 388
 389
 390
 391

Supplementary Fig S4. Clinical parameter associations with monocytic-bias (Mono-Bias), MEP-biased, and normal-like patient groupings showed no significant differences in blast percentage, platelets, WBC, ALC, ANC, and absolute monocytosis. Non-parametric Kruskal-Wallis test was used to compared continuous variables across treatment naïve patients aggregated based on the three distinct trajectories identified. p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001.



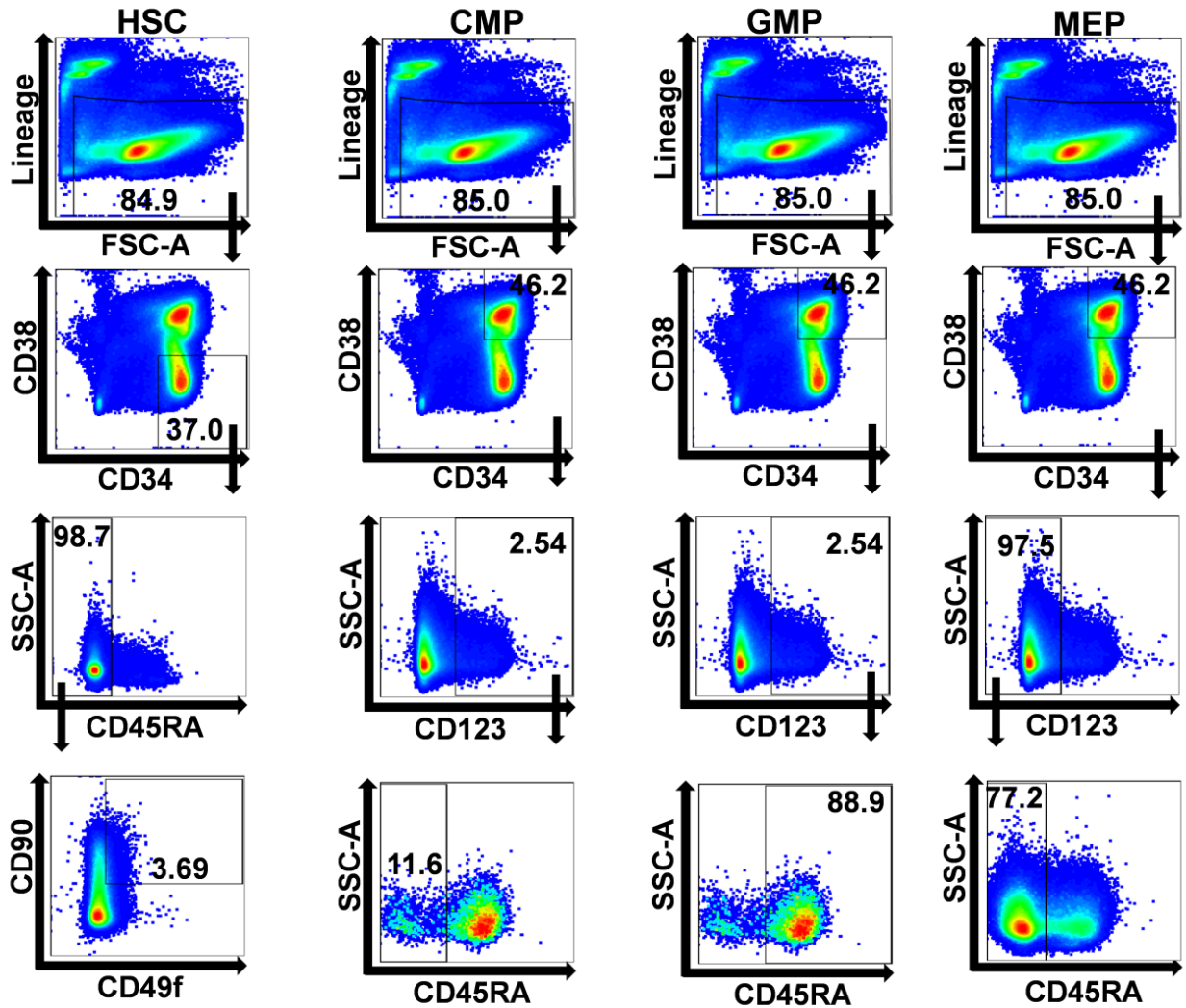
392
393
394
395
396
397
398
399
400

Supplementary Figure S5. CMML patients show HSC depletion as compared to normals. There was also a depletion in the SingleR assignment of HSC cell type in treatment naïve CMML samples (p-value: 0.0004; Mann-Whitney test). p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001.



401
402
403
404
405
406
407
408
409
410

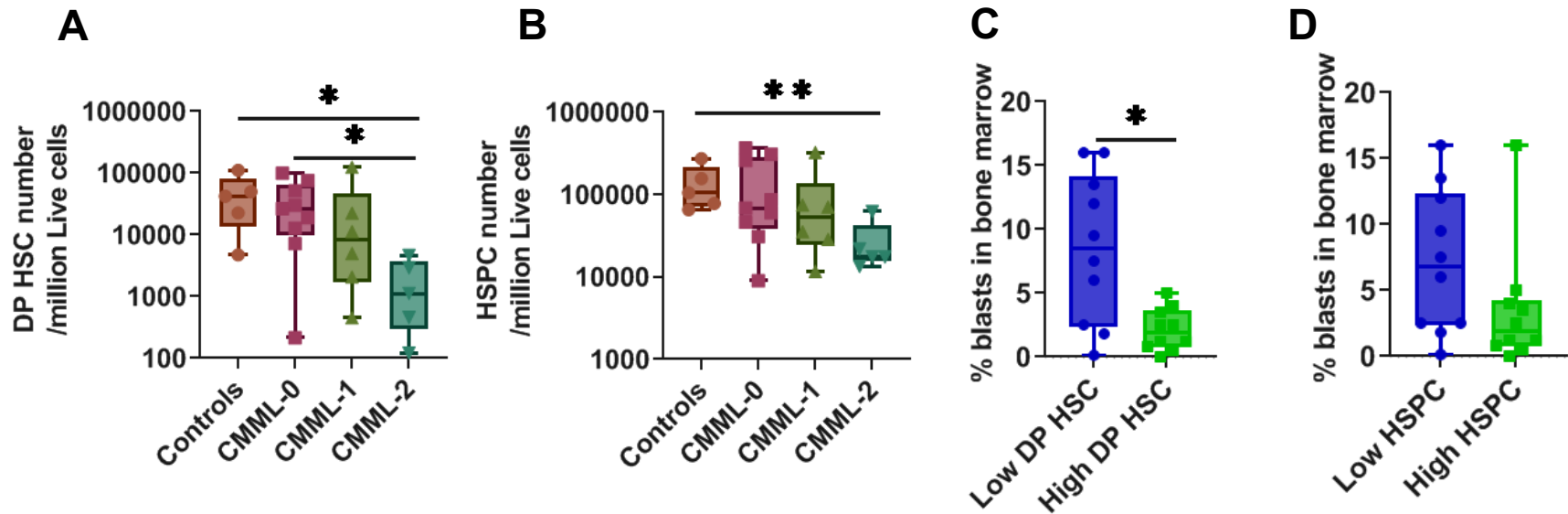
Supplementary Figure S6. Single-cell gene expression of HSC signatures show depletion in HSCs in CMML. HSC depletion was robust and replicated across three single-cell derived HSC signatures (Wu = Wu et al. *Blood Advances* 2020; VG = Van Galen et al. *Cell* 2019; and Epp = Eppert et al. *Nature Medicine* 2011). Treatment-naïve samples separated by trajectory bias show.



411
 412
 413
 414
 415
 416
 417
 418
 419
 420

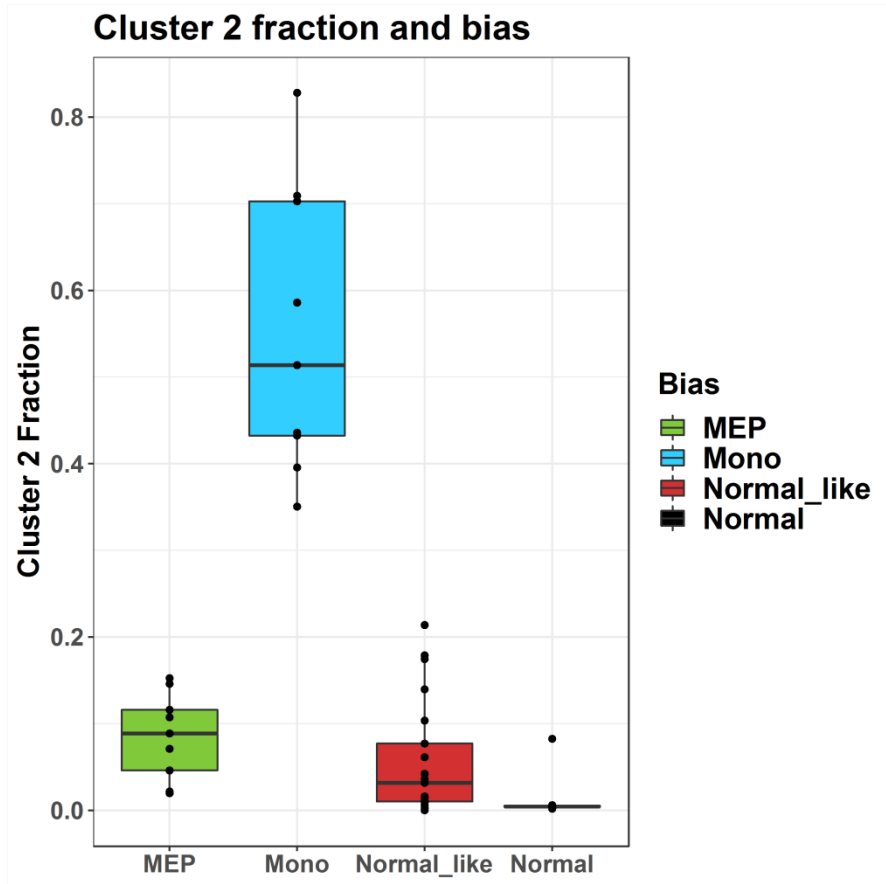
Supplementary Figure S7. Gating strategy used for identification of stem and myeloid progenitor populations in CMML patients and controls.

Triple positive HSCs: Lin⁻CD34⁺CD38⁻CD45RA⁻CD90⁺CD49F⁺,
 CMP: Lin⁻CD34⁺CD38⁺CD123⁺CD45RA⁻,
 GMP: Lin⁻CD34⁺CD38⁺CD123⁺CD45RA⁺,
 MEP: Lin⁻CD34⁺CD38⁺CD123⁻CD45RA⁻.



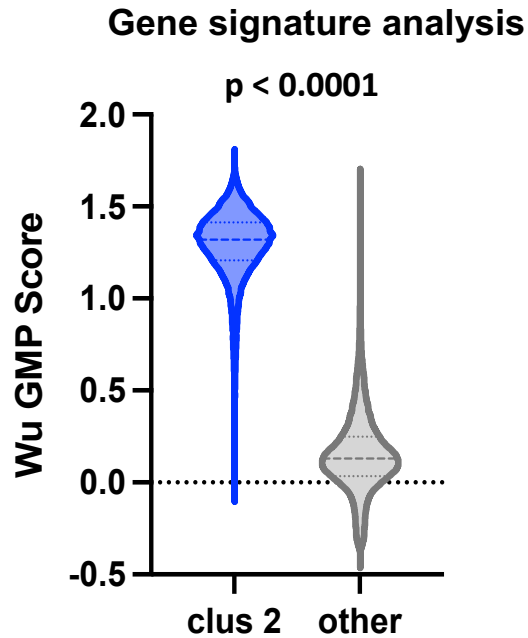
421
 422
 423
 424
 425
 426
 427
 428
 429
 430
 431
 432

Supplementary Fig S8. Clinical characteristics of patients with HSC depletion. (A) Comparison of HSC frequency between controls and WHO-classified CMML stages using the flow-cytometry identified HSC immunophenotypes showed HSC depletion with disease progression in double positive HSCs, (B) single positive HSCs also known as HSPCs, n=20 patient cases and 5 control cases. (C) Evaluation of bone marrow blast content between low HSC and high HSC group of patients showed that blast content was inversely correlated with HSC numbers in double positive HSCs, (D) single positive HSCs, n=20 patient cases. Data was analyzed using Mann-Whitney test; p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001.



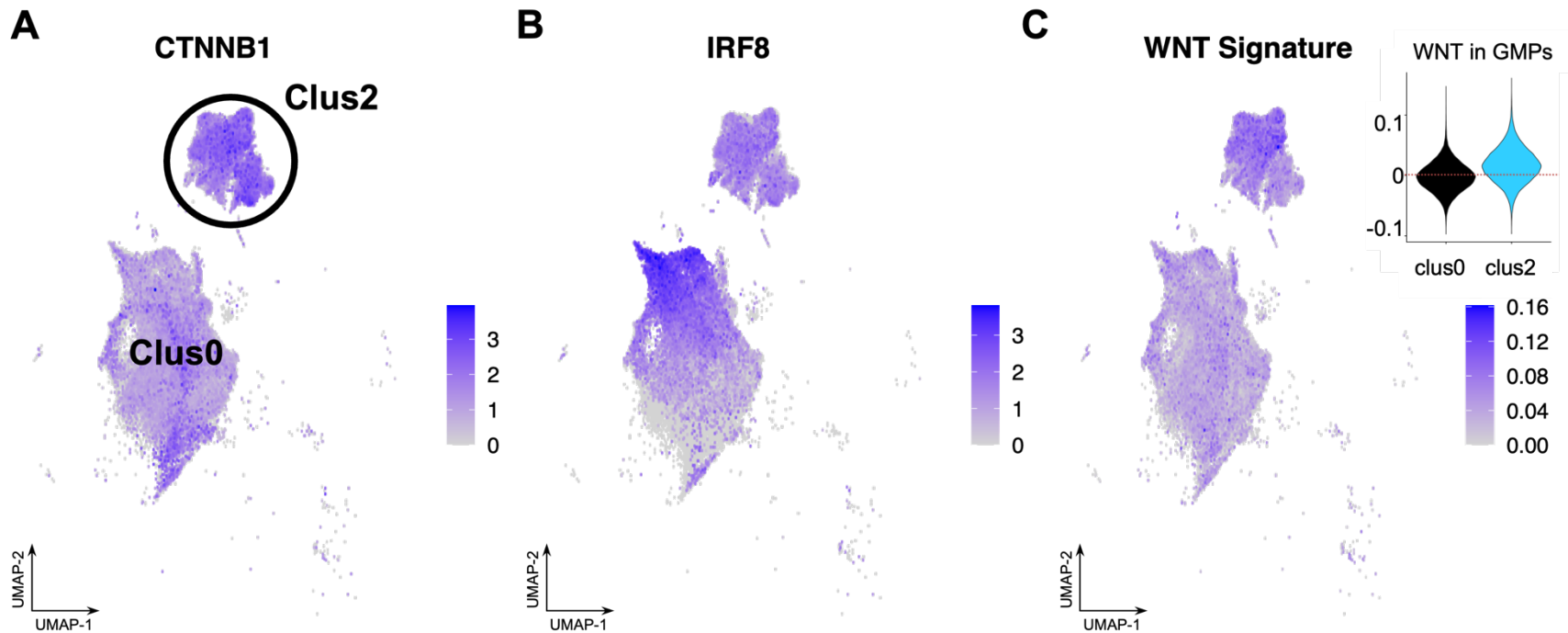
433
 434
 435
 436
 437
 438
 439

Supplementary Figure S9. Cluster 2 drives Mono-bias assignment. Fraction of cells assigned to cluster 2 in each sample, with samples grouped by differentiation bias.



440
441
442
443
444
445
446
447
448
449
450

Supplementary Figure S10. Gene expression analysis of Clus2 cells showed GMP like signature. The SingleR results were validated by scoring each cell with a previously published GMP gene signature score (from Wu *Blood Advances* 2020) and cells in Clus2 had significantly higher GMP scores than cells not in Clus2 (mean score of 0.5504 in Clus 2 and 0.0649 not in Clus 2; p-value: <0.0001). Nonparametric Mann-Whitney tests were used to compare two group data. p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001.



451

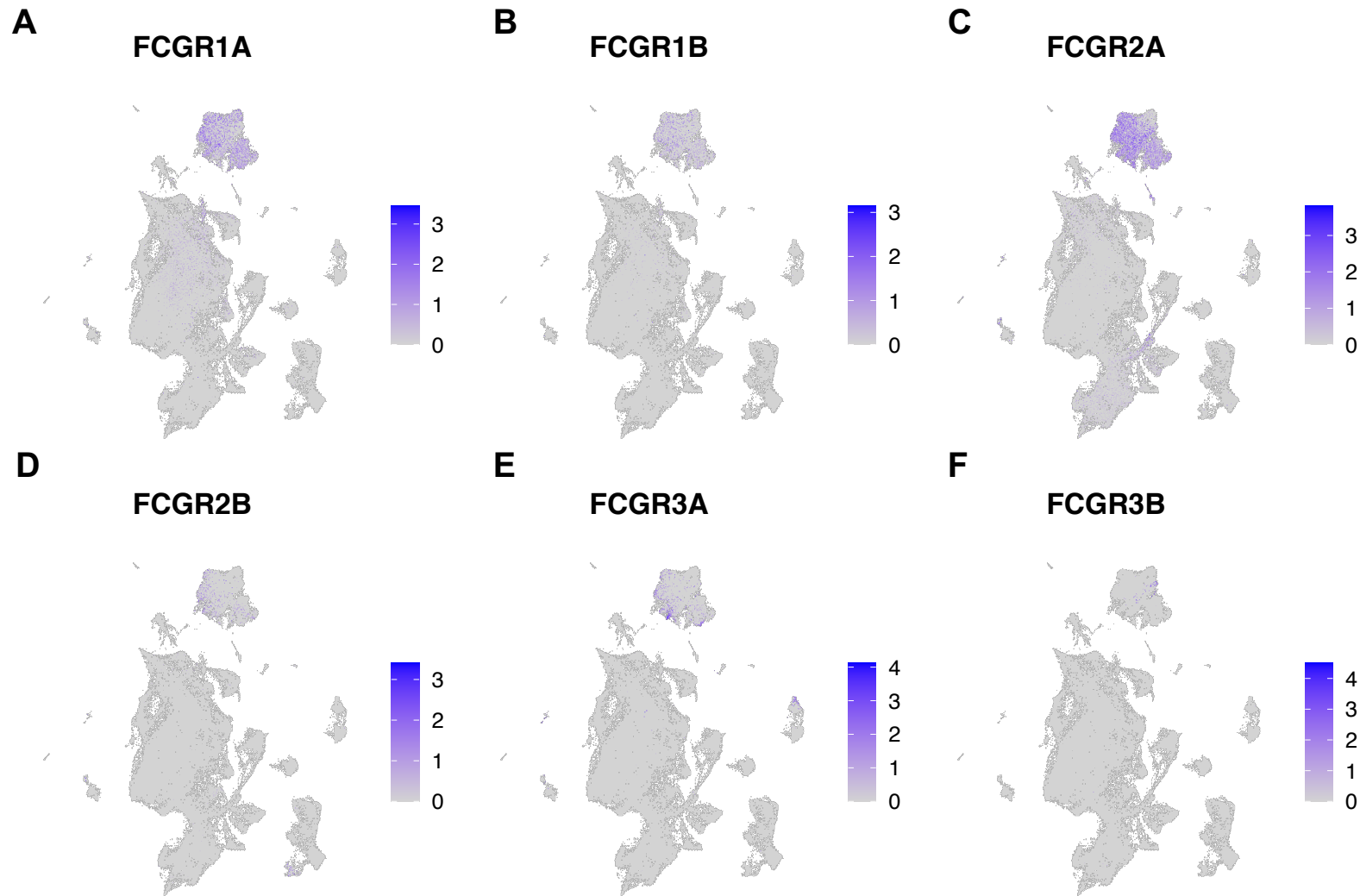
452

453 **Supplementary Figure S11. Expression of CTNNB1, IRF8, and WNT pathway signature score in CMML GMPs in**
 454 **scRNAseq cohort.** Gene expression per cell was visualized on UMAP projections of all single cells in cohort with Seurat
 455 featurePlot() function of (A) *CTNNB1* and (B) *IRF8*. (C) WNT pathway up-regulation was scored using Seurat
 456 AddModuleScore() and Gene Set Enrichment Analysis GeneSet GSE26351 describing WNT pathway stimulation in
 457 human CD34+ hematopoietic progenitor populations (established by Trompouki et al *Cell*. 2011 (54)).

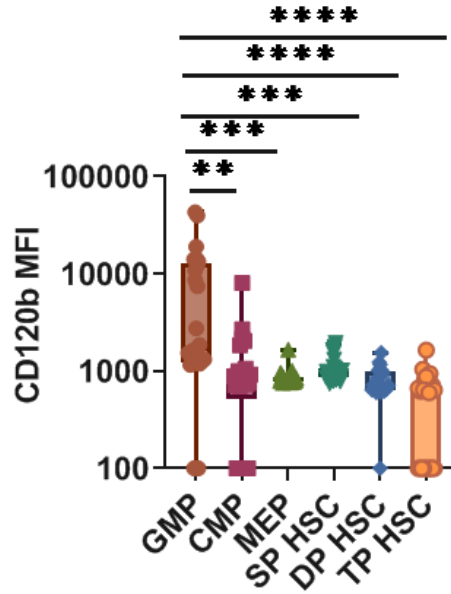
458

459

460

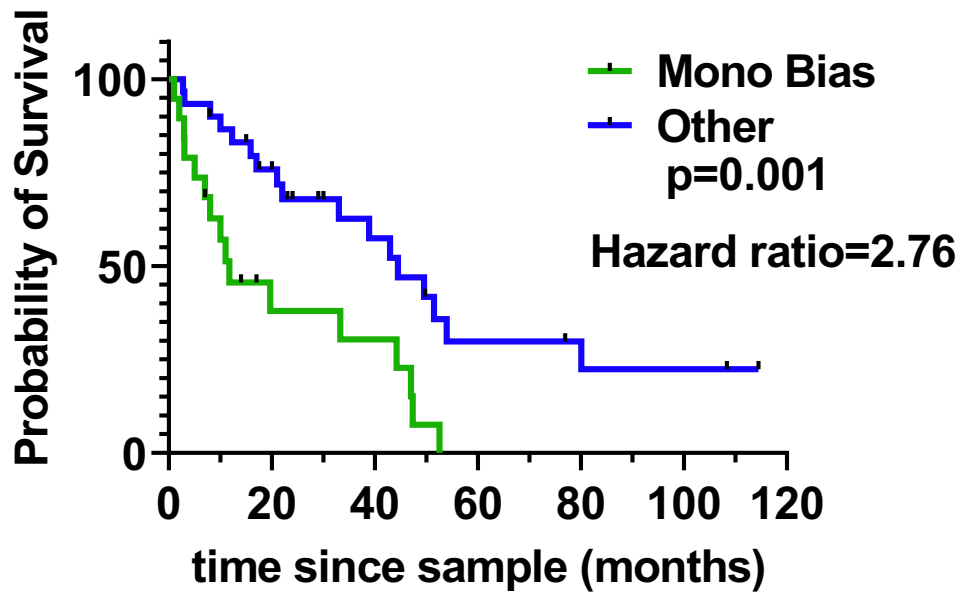


461
 462 **Supplementary Figure S12. Expression of Fc gamma receptors in scRNAseq cohort.** Gene expression per cell was
 463 visualized on UMAP projections of all single cells in cohort with Seurat featurePlot() function of (A) *FCGR1A*, (B)
 464 *FCGR1B*, (C) *FCGR2A*, (D) *FCGR2B*, (E) *FCGR3A*, and (F) *FCGR3B*. Elevated expression in Clus2 cells indicates a
 465 possible state of myelopoiesis induced by stress (23).



466
 467 **Supplementary Figure S13. CD120b expression across stem and progenitor**
 468 **populations.** CD120b expression across stem and progenitor cells as determined by
 469 flow cytometry. Data was analyzed using Mann-Whitney test. p-value significance
 470 represented by * < 0.05, ** < 0.01, *** < 0.001.

471



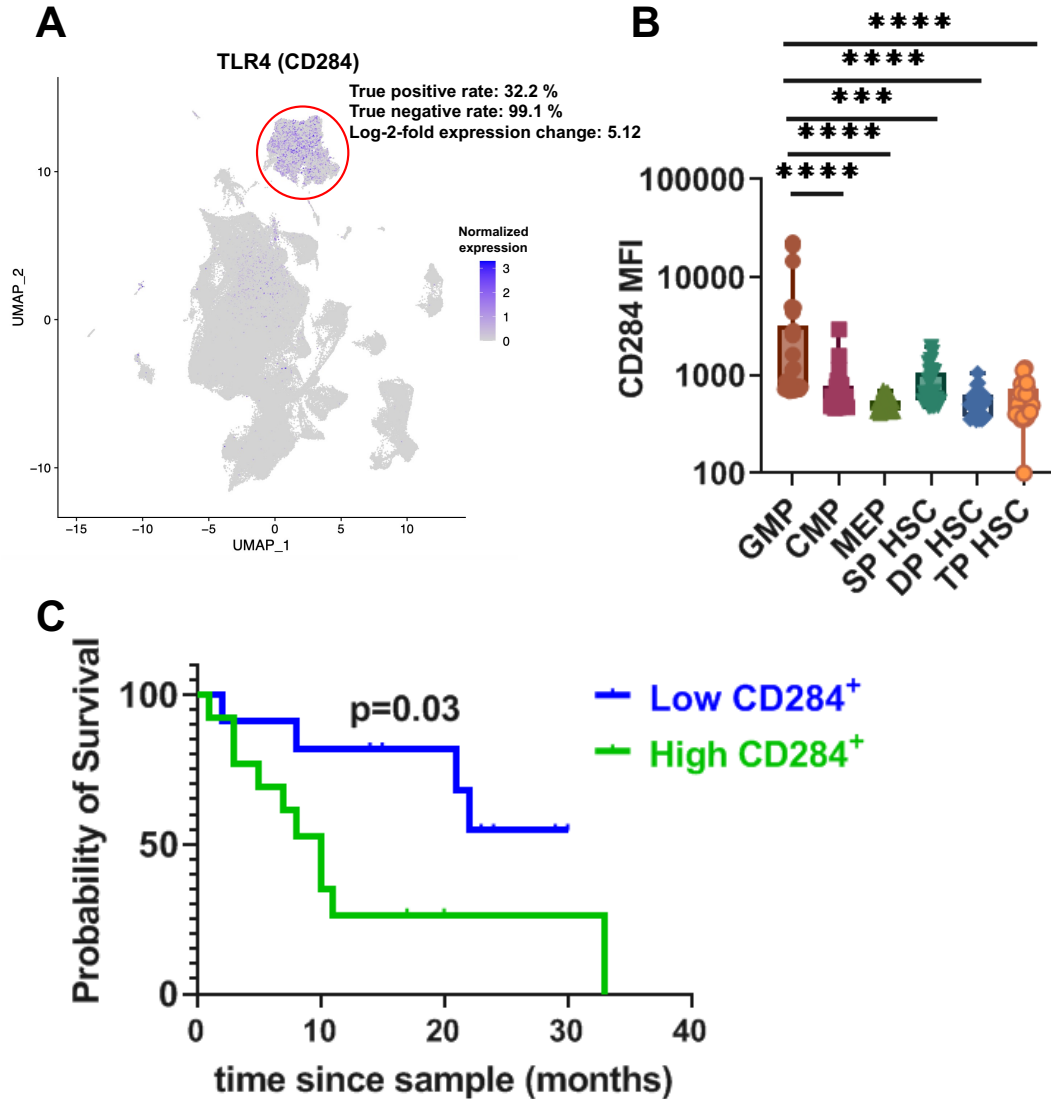
472

473

474 **Supplementary Figure S14. Merged survival analysis of the single-cell RNA**
475 **sequencing and flow cytometry cohorts.** KM survival analysis showed patients with
476 monocytic-bias had inferior survival (n=55; log-rank p-value: 0.001). p-value significance
477 represented by * < 0.05, ** < 0.01, *** < 0.001.

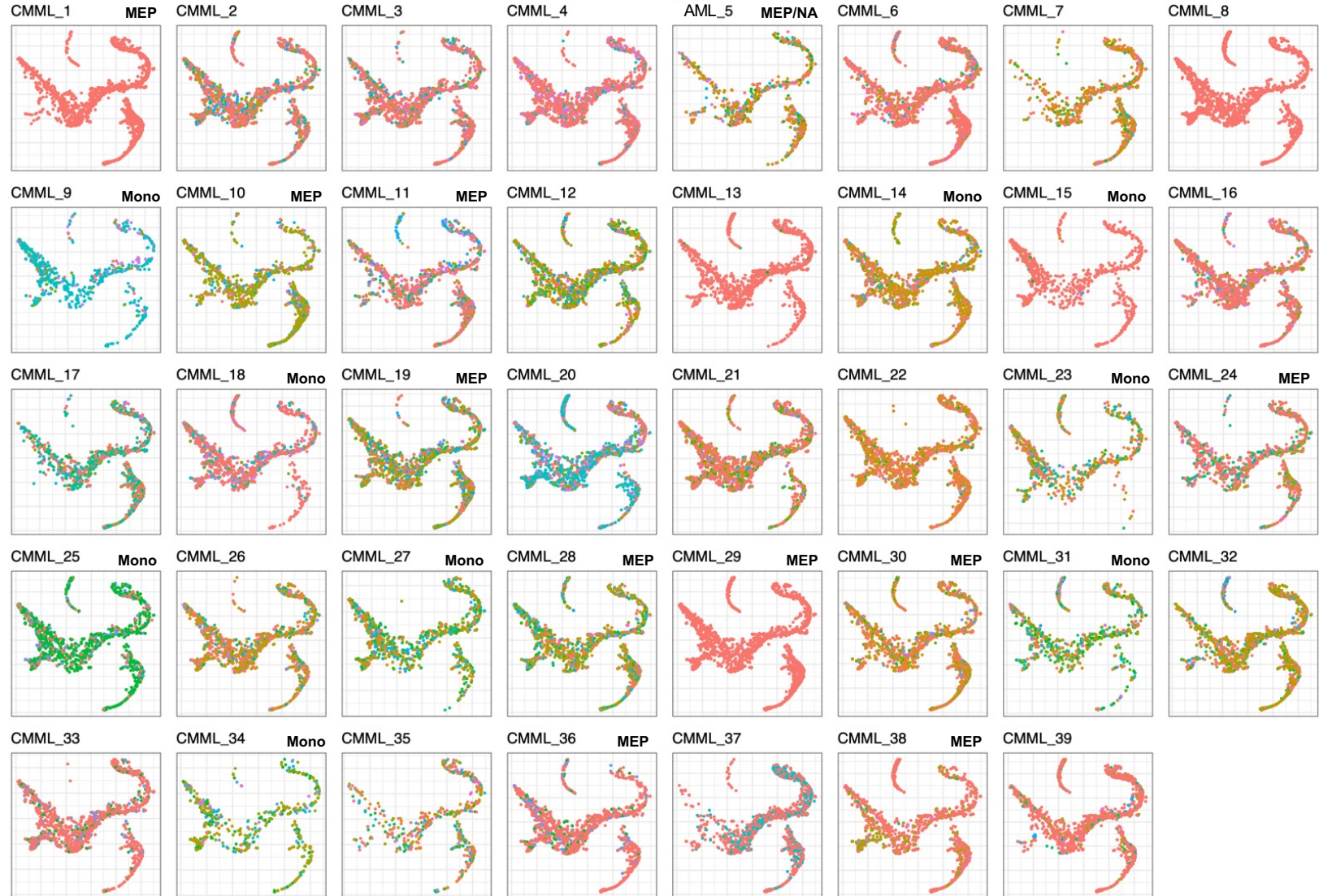
478

479



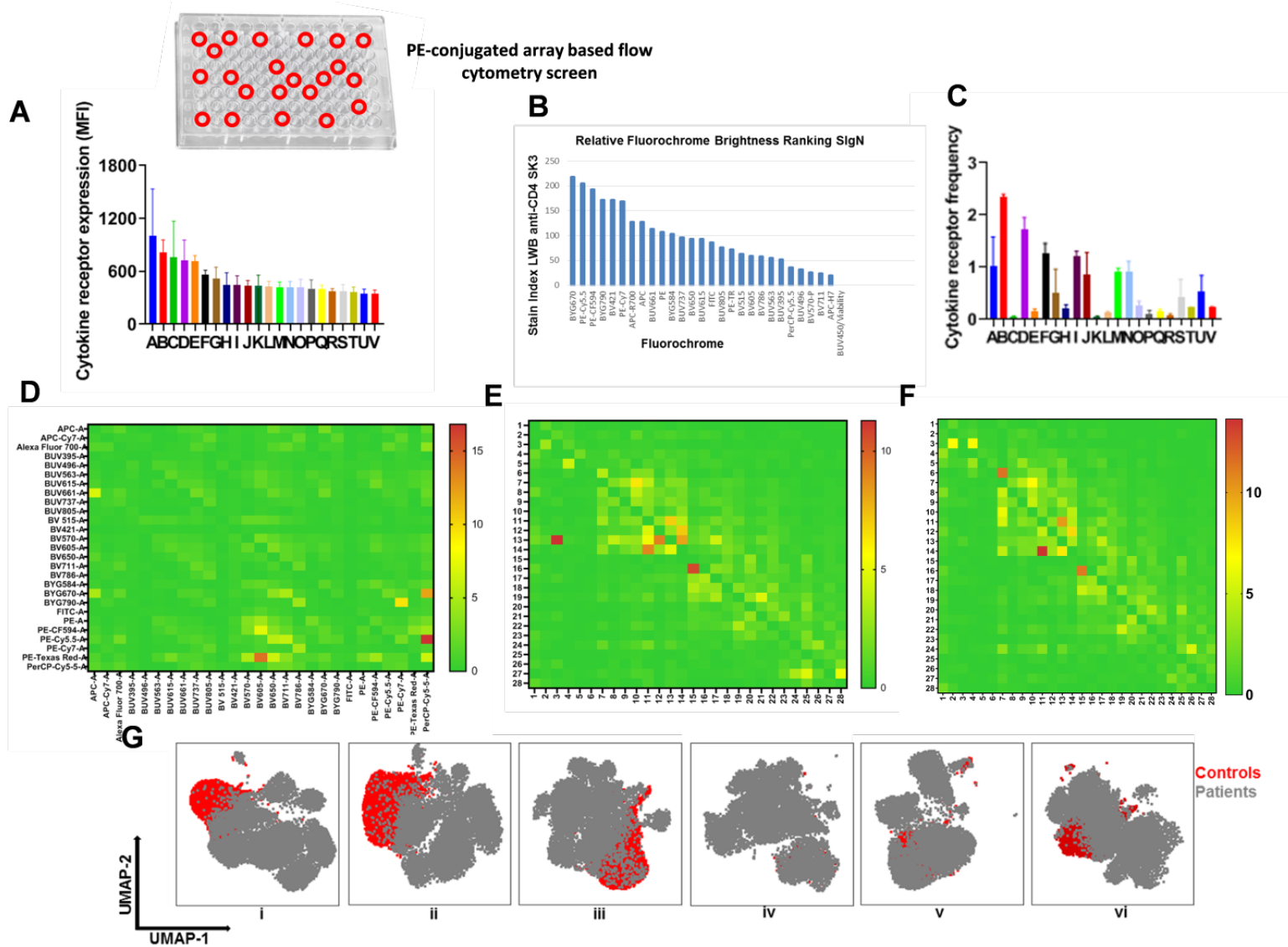
480
 481
 482
 483
 484
 485
 486
 487
 488
 489
 490
 491

Supplementary Figure S15. Clus2 characterized by CD284 expression. (A) COMET was used to identify differential gene expression markers well-suited for validation with flow-cytometry. COMET identified TLR4 (encoded cell surface marker CD284) as a marker for identifying Clus2 cells with a true positive performance of 32.2% and true negative performance of 99.1%. (B) CD284 expression across stem and progenitor cells as determined by flow cytometry. Data was analyzed using Mann-Whitney test. (C) KM survival analysis showed patients with high CD284⁺ expression had inferior survival (n=26; log-rank p-value: 0.03). p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001.



492
493
494
495

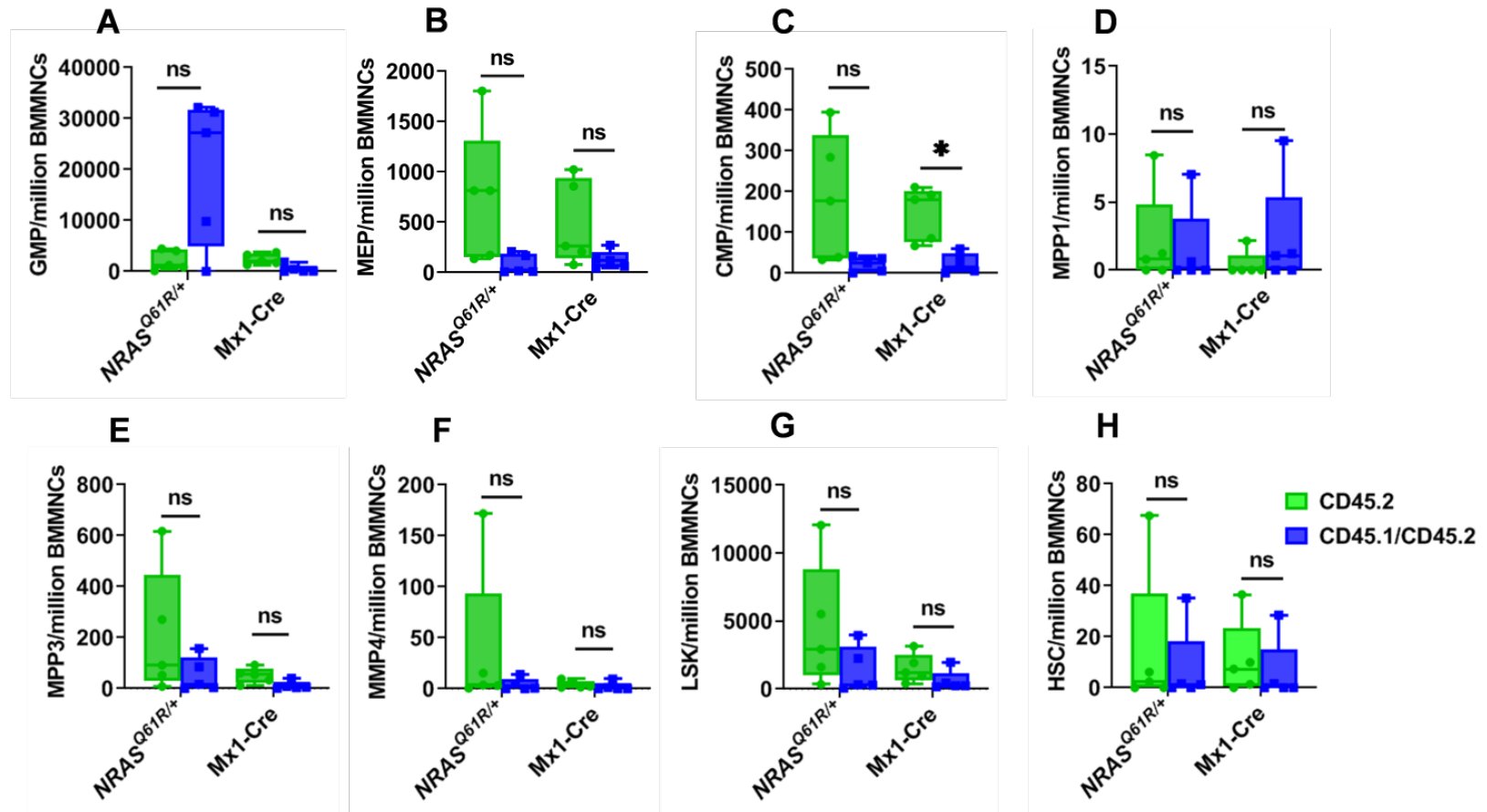
Supplementary Figure S16. Palantir mappings with mitoClone clonal information indicated by color for all samples run individually. Mono and MEP bias samples labeled; all others are Normal-like. No trend associating clonality with differentiation trajectories.



496
497
498
499

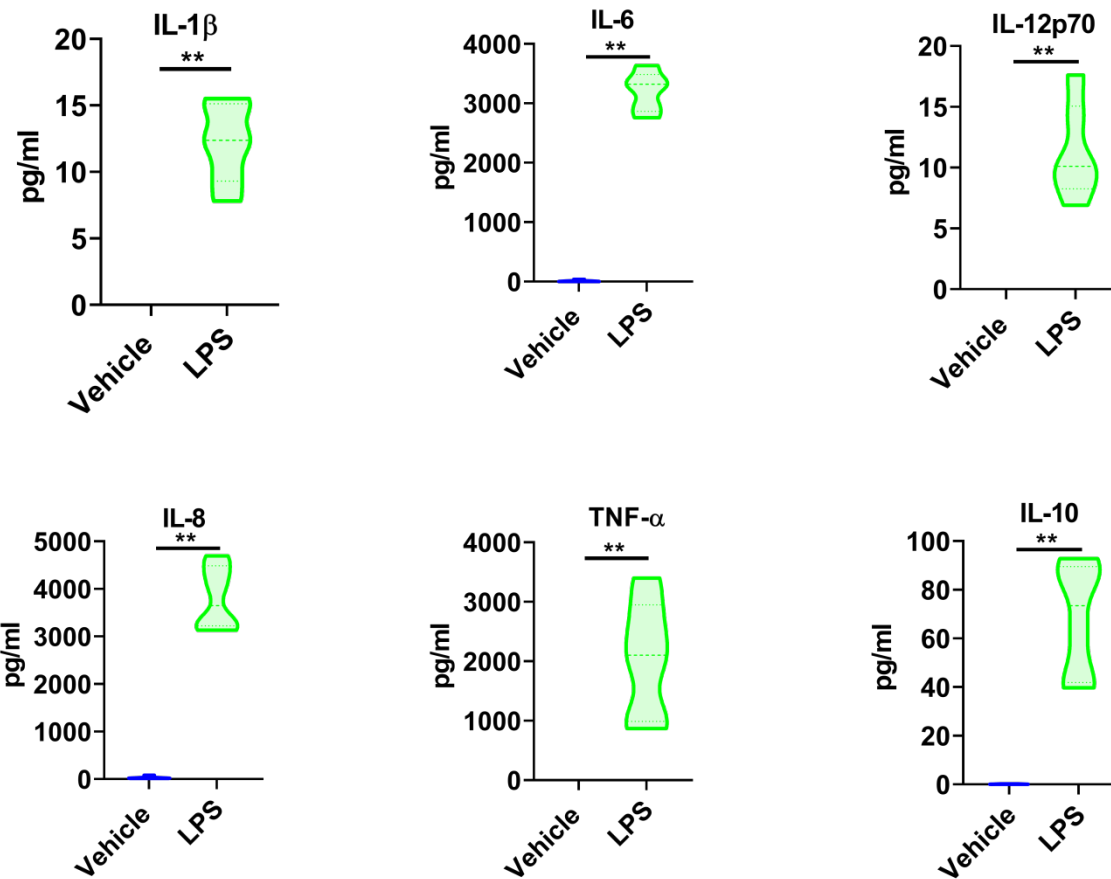
Supplementary Figure S17. Development and optimization of CRD flow panel. (A) The MFI of cytokine receptors based on the expression data generated from the PE-conjugated flow cytometry screen. (B) The stain index for the

500 commercially available fluorophores. **(C)** The frequency (percentage of total) data for receptors generated from the PE-
501 conjugated flow cytometry screen. **(D)** SSM generated by using LWB stained with CD4 antibody conjugates on Symphony
502 A5. The cytokine receptors were conjugated with appropriate fluorophores based on expression, frequency, and SSE
503 data. **(E)** The first iteration of the SSM specific to our panel generated by staining compensation particles and cells with
504 titred volume of respective 28 antibodies/dyes (single cell stain controls). **(F)** The revised SSM generated post
505 optimization of spillover sources identified in Fig **E**. **(G)** UMAP visualization of Patients and Controls in i) triple positive
506 HSCs ii) double positive HSCs iii) single positive HSCs (HSPCs) iv) CMPs v) GMPs vi) MEPs. The following codes have
507 been used for the cytokine receptors in figures **A** and **C**) A:TIM3, B:CD123, C:CD284, D:CD117, E:CD215, F:CD132,
508 G:CD126, H:CDw125, I:CD114, J:CD282, K:CD181, L:CD182, M:CD135, N:CD110, O:CD115, P:CD120b, Q:CD218a,
509 R:CD192, S:CD184, T:CD120a, U:CD119, V:CD116. The following codes have been used for the cytokine receptors in
510 figures **D**, **E** and **F**) 1:TIM3, 2:CD123, 3:CD284, 4:CD117, 5:CD215, 6:CD132, 7:CD126, 8:CDw125, 9:CD114, 10:CD282,
511 11:CD181, 12:CD182, 13:CD135, 14:CD110, 15:CD115, 16:CD120b, 17:CD218a, 18:CD192, 19:CD184, 20:CD120a,
512 21:CD119, 22:CD116.
513
514



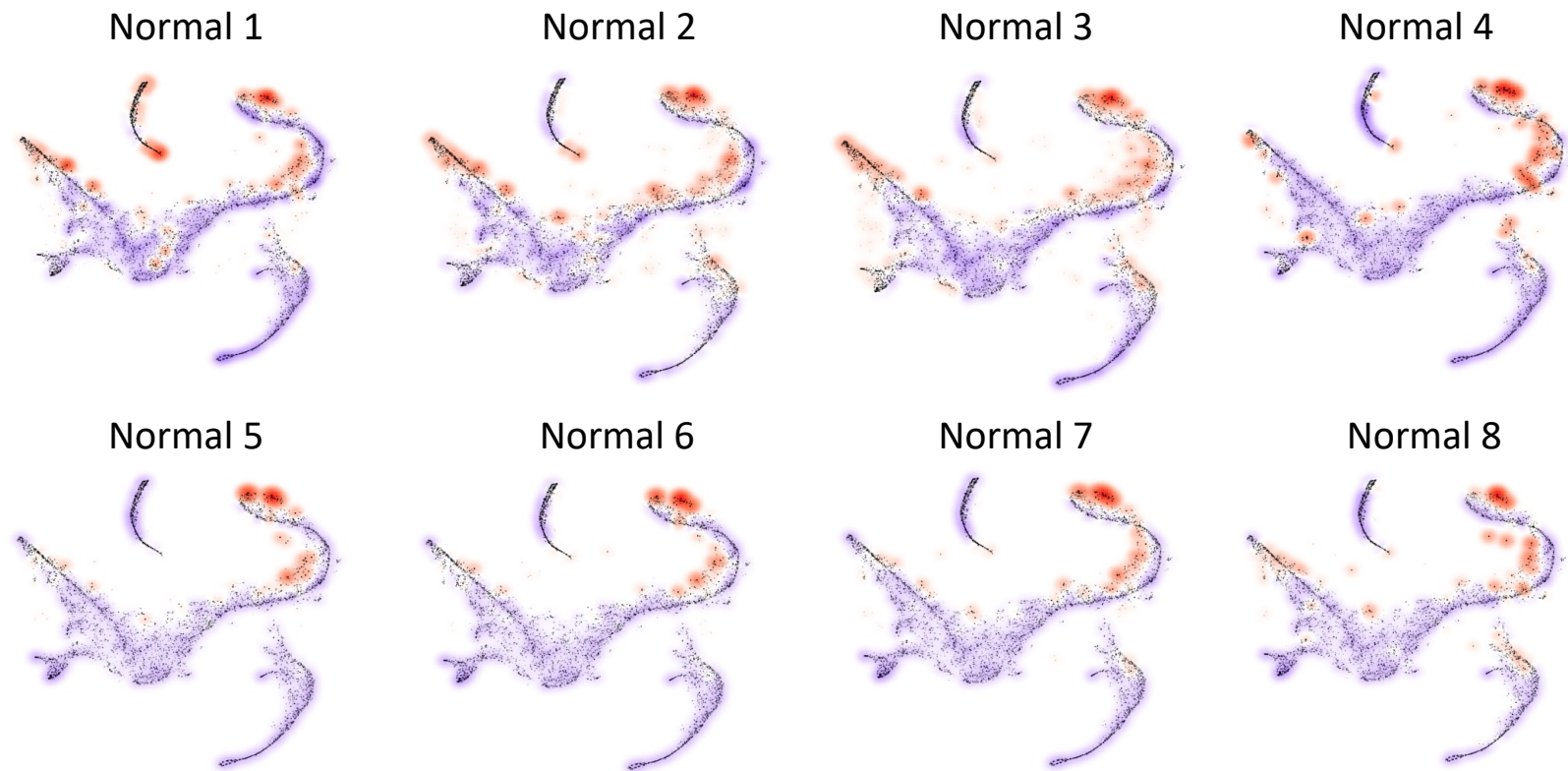
515
516
517
518
519
520

Supplementary Figure S18. Distribution of HSPCs in competitive BMT studies in NRAS model: (A) GMPs (B) MEPs (C) CMPs (D) MPP1 (E) MPP3 (F) MPP4 (G) LSKs (H) HSCs. Support marrow-(CD45.1/CD45.2) vs CD45.2 Nras^{Q61R/+}; Mx1-Cre, n=5 Nras^{Q61R/+}; mice and 5 Mx-1Cre mice. Data was analyzed using multiple paired t-test.



521 **Supplementary Figure S19. Plasma cytokine levels 6 hours post injection of LPS (n=6) or vehicle (n=6).** Plasma
 522 was obtained from submandibular bleeds. Data was analyzed using non-parametric Mann-Whitney test. p-value
 523 significance represented by * < 0.05, ** < 0.01, *** < 0.001.
 524

525
 526



527
528
529
530
531
532

Supplementary Figure S20. Cellular density in Palantir pseudotime across differentiation trajectories in normal samples. Samples display HSC enrichment but no clear trajectory bias.