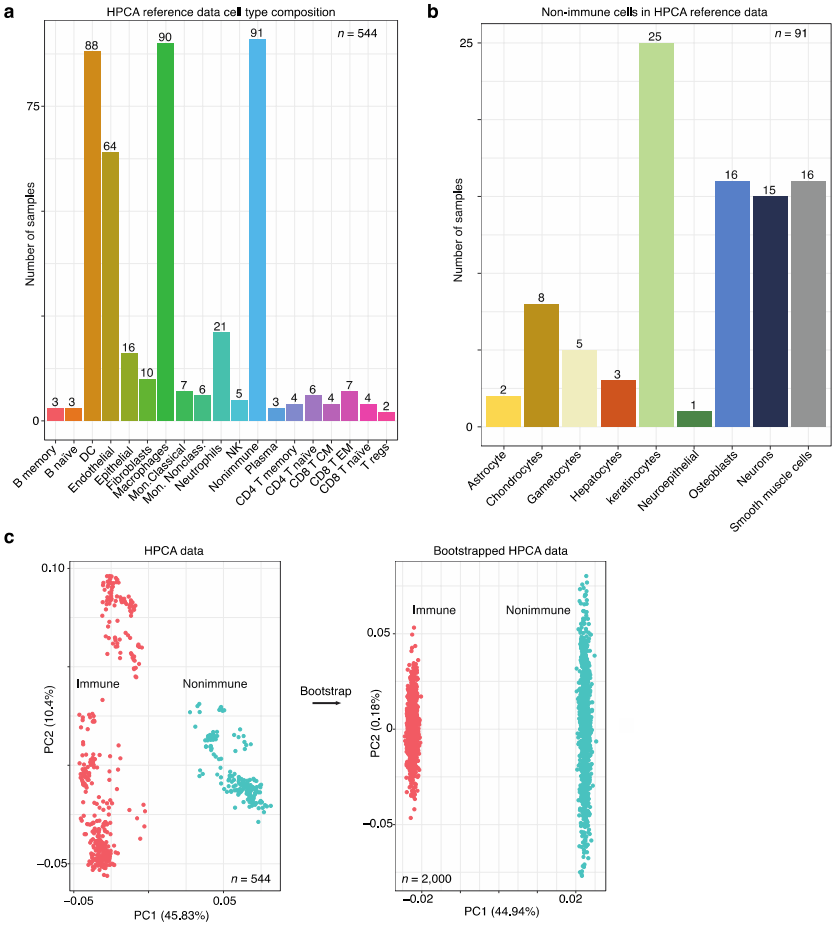
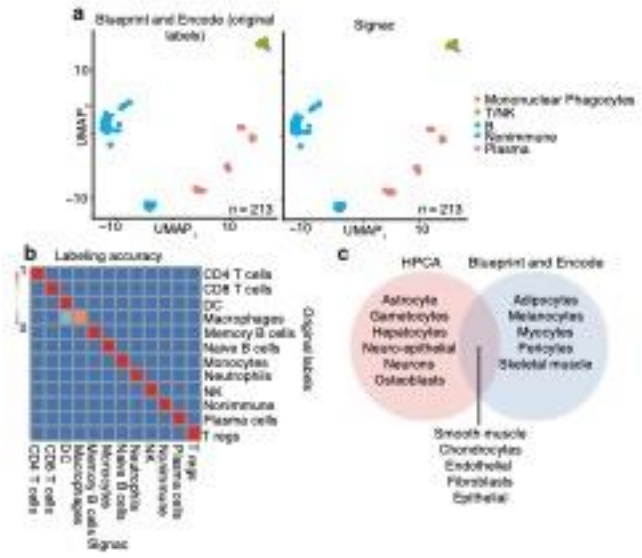
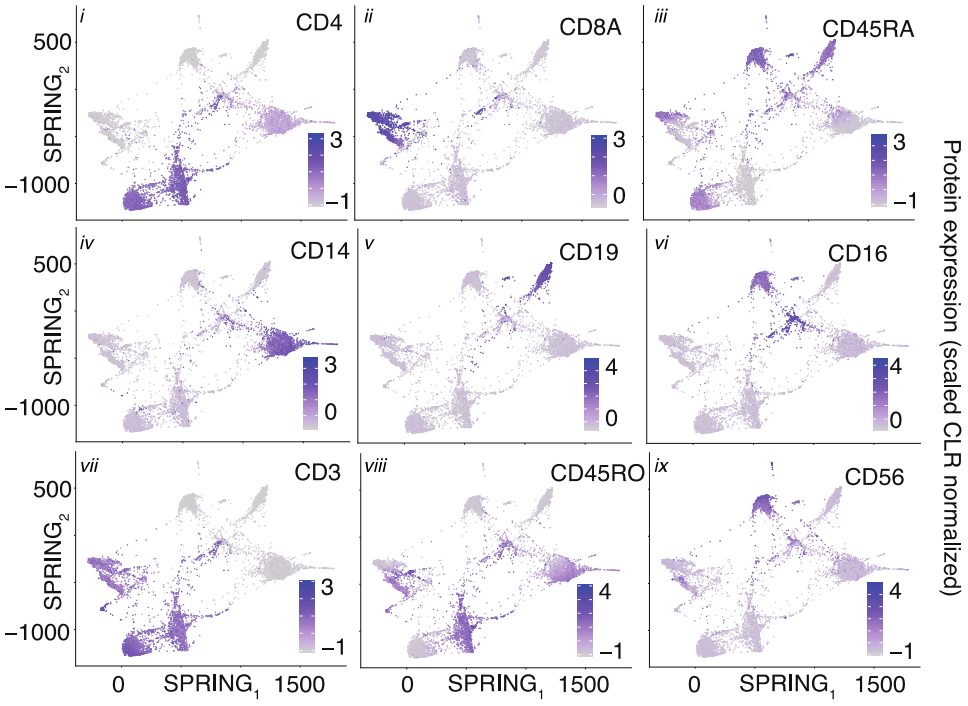


Supplemental Figure 1: Overview of the HPCA reference data. A, Number of samples bar plot for each annotated cell type. Bar plot indicates the number of samples (y-axis; numbers) for each cellular phenotype (x-axis) that was in the HPCA reference data set. **B, Number of samples bar plot for the "nonimmune" cellular phenotypes.** Bar plot as depicted in Supplemental Figure 1A for the $n = 91$ samples in the "nonimmune" cell type category. **C, Principal component analysis (PCA) plots revealed that the HPCA reference data and the bootstrapped training data were separable by immune and nonimmune gene markers.** PCA

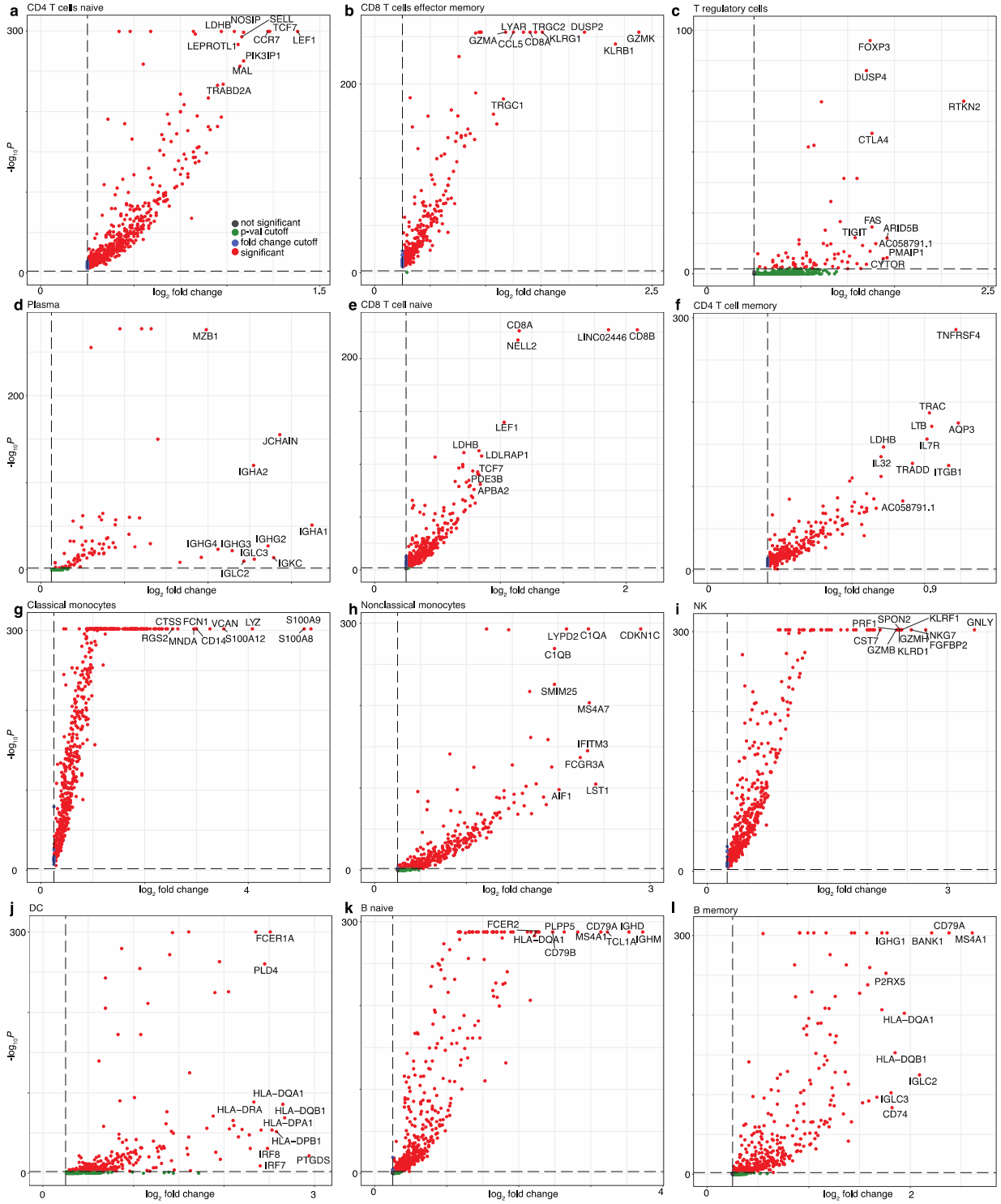
plot (left) shows a gene expression sample (each dot) from the HPCA reference data that is either labeled as immune (red) or nonimmune (teal) based on the classification hierarchy established here and the annotated cell type established experimentally (**Fig. 1A**). PCA was performed on the marker gene that were determined by differential gene expression analysis comparing the samples annotated as immune and nonimmune. After bootstrapping (arrow); PCA was performed with the same marker genes, except with data that was bootstrapped from the immune and nonimmune samples. We note that the structure in the PCA plot prior to bootstrapping was largely removed, consistent with the view that bootstrapping generated what can be thought of as composite or as average immune and nonimmune gene expression samples.



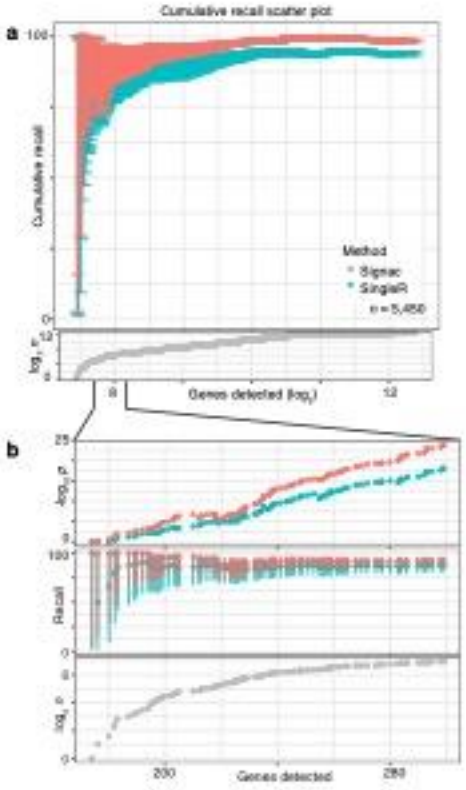
Supplemental Figure 2: Neural networks robustly classified validation data from the Blueprint and Encode consortia. A, UMAP plots show that cellular phenotypes were accurately classified by Signac. In the scatter plot (left), each dot is a sample of pure cell type gene expression data that was amended labels for cellular phenotypes (colors; see legend). The data represented here are in a two-dimensional embedding (axes), in which distances correspond to transcriptional similarities between samples (closer samples are more similar); we determined this embedding with UMAP. Cellular phenotype labels (colors) were established either by the Blueprint and Encode consortium (left) with empirical measurements or by our computational approach (right). **B, Heatmap shows that distinct immune cell phenotypes were accurately classified by Signac**. Heatmap displays the fraction of the samples within each cellular phenotype category (axes) that were accurately classified by our approach (scale bar; red is more accurate; blue is less accurate). **C, Venn diagram shows that Signac accurately identified the transcriptomes of nonimmune cellular phenotypes despite never being trained to recognize them**. Venn diagram depicts the nonimmune cellular phenotypes that were specific to the HPCA reference data (left; red), that were shared between the HPCA reference and the Blueprint and Encode data (middle; purple), and that were distinct to the Blueprint and Encode data (right; blue).



Supplemental Figure 3: Protein expression SPRING plots validated the cellular phenotypes classified by Signac for the CITE-seq PBMCs data. Each SPRING plot (*i-ix*) displays the z-score transformed CLR normalized protein expression (colors) generated for each individual cell.

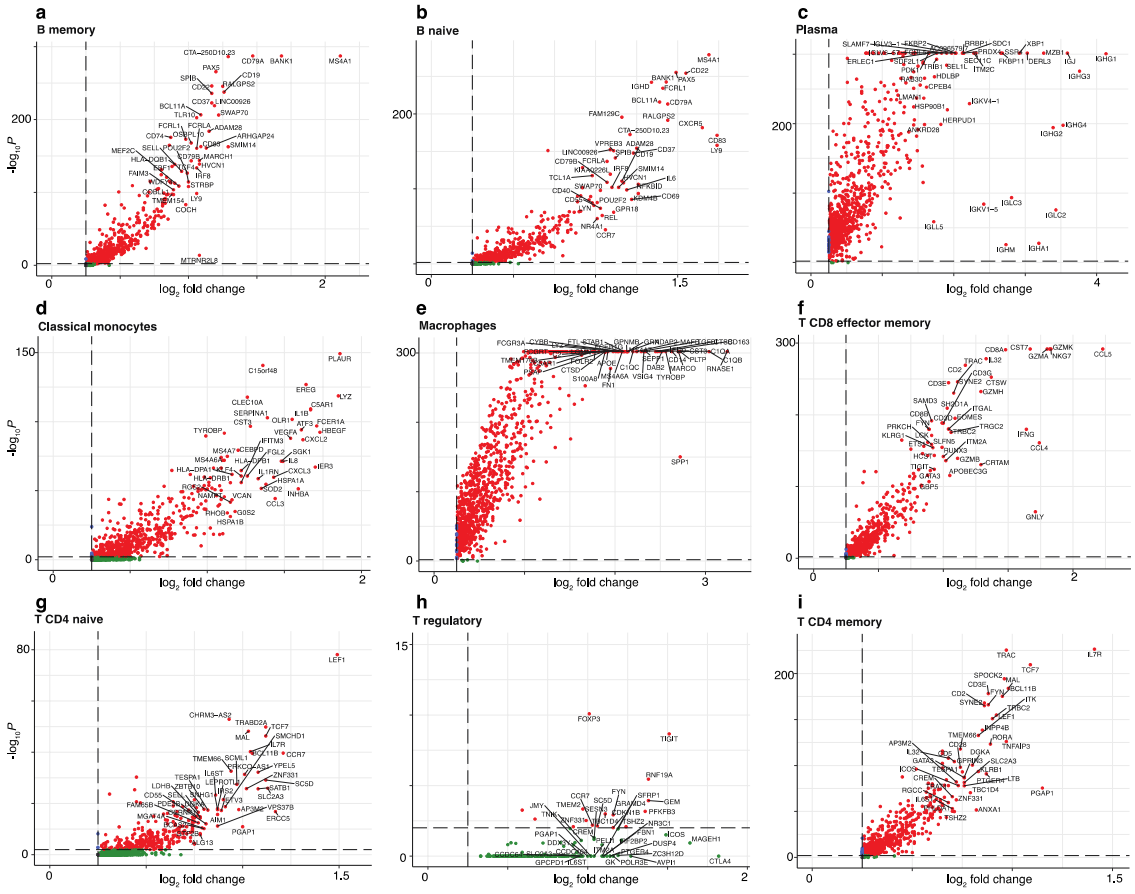

a
CD4 T cells naive
LDHB
NOSIP
SELL
TCF7
CCR7
LEF1
LEPROTL1
PIK3IP1
MAL
TRABD2A
not significant
p-val cutoff
fold change cutoff
significant
-log10 P
log2 fold change
b
CD8 T cells effector memory
LYAR
TRGC2
DUSP2
GZMA
CCL5
CD8A
KLRG1
GZMK
KLRB1
TRGC1
log2 fold change
c
T regulatory cells
FOXP3
DUSP4
RTKN2
CTLA4
FAS
TIGIT
ARID5B
AC058791.1
PMAIP1
CYTOR
log2 fold change
d
Plasma
MZB1
JCHAIN
IGHA2
IGHA1
IGHG4
IGHG3
IGHG2
IGLC3
IGKC
IGLC2
-log10 P
log2 fold change
e
CD8 T cell naive
CD8A
NELL2
LINC02446
CD8B
LEF1
LDHB
LDLRAP1
TCF7
PDE3B
APBA2
log2 fold change
f
CD4 T cell memory
TNFRSF4
TRAC
LTB
AQP3
LDHB
IL7R
IL32
TRADD
ITGB1
AC058791.1
log2 fold change
g
Classical monocytes
CTSS
FCN1
VCAN
LYZ
S100A9
RGS2
MNDA
CD14
S100A12
S100A8
-log10 P
log2 fold change
h
Nonclassical monocytes
LYPD2
C1QA
CDKN1C
C1QB
SMIM25
MS4A7
IFITM3
FCGR3A
LST1
AIF1
log2 fold change
i
NK
PRF1
SPON2
KLRF1
GNLY
CST7
GZMH
NKG7
FGFBP2
GZMB
KLRD1
log2 fold change
j
DC
FCER1A
PLD4
HLA-DQA1
HLA-DRA
HLA-DQB1
HLA-DPA1
HLA-DPB1
IRF8
IRF7
PTGDS
-log10 P
log2 fold change
k
B naive
FCER2
PLPP5
CD79A
IGHD
HLA-DQA1
MS4A1
TCL1A
IGHM
CD79B
log2 fold change
l
B memory
CD79A
IGHG1
BANK1
MS4A1
P2RX5
HLA-DQA1
HLA-DQB1
IGLC2
IGLC3
CD74
log2 fold change

Supplemental Figure 4: IMAGES identified from immune phenotypes in CITE-seq PBMCs single cell transcriptomes. A-I, Volcano plots demonstrate the IMAGES identified in each cell population. Each scatter plot depicts the statistical association (y-axis) and the average fold-change of IMAGES (each dot is a unique gene) for immune cell phenotypes. Colors (red) indicate IMAGES that passed the thresholds applied to the fold-change and adjusted p-values.

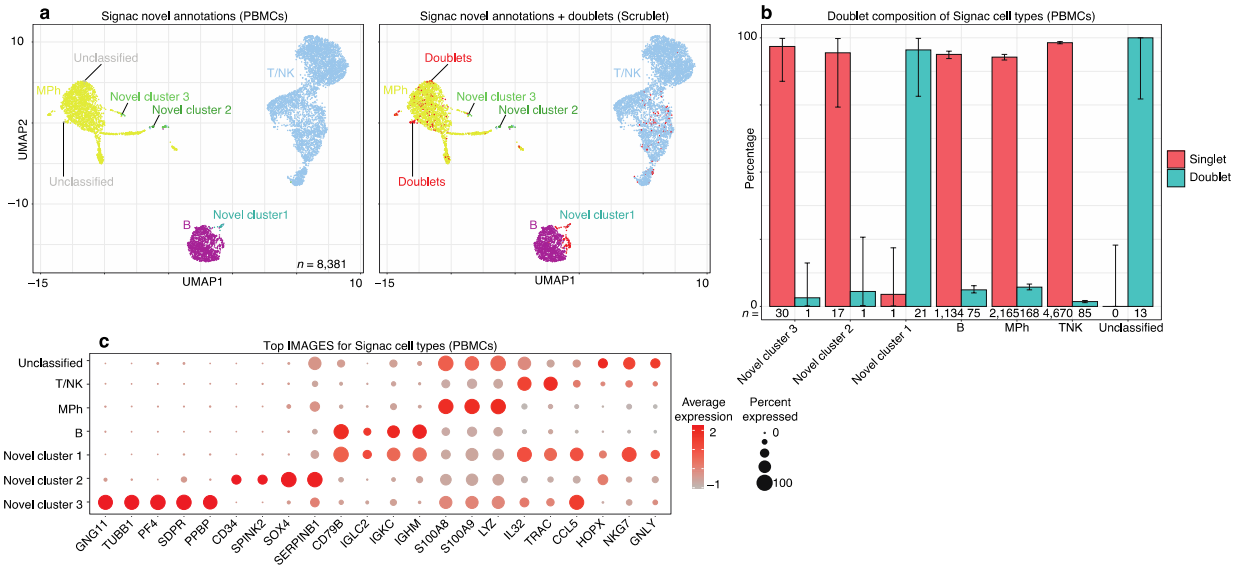


Supplemental Figure 5: Signac accurately recalled flow cytometry labels with few genes detected. A, Cumulative recall scatter plot showed that Signac outperformed SingleR as a function of genes detected in immune cell transcriptomes. Plot depicts the cumulative recall

(y-axis) of immune cell type labels. The labels were originally determined by flow cytometry (T cell, B cell or monocyte). Recall of these labels was calculated cumulatively as a function of the number of genes detected (x-axis) by either Signac (red) or SingleR (teal). Fibroblasts were omitted from this analysis due to broad misclassification by SingleR. Scatter plot, bottom depicts the number of single cell transcriptomes (*n*) as a function of genes detected. Error bars (top) are 95% C.I.s determined by two-sided binomial testing. **B, Inset shows stronger Signac performance at low genes detected**. Scatter plot (top) depicts the *p*-value for the two-sided binomial test and showed that Signac (red) outperformed SingleR (teal) at low sequencing depths. Scatter plots (middle; bottom) are close-ups of **Supplemental Figure 5A** for data with less than 300 genes detected.

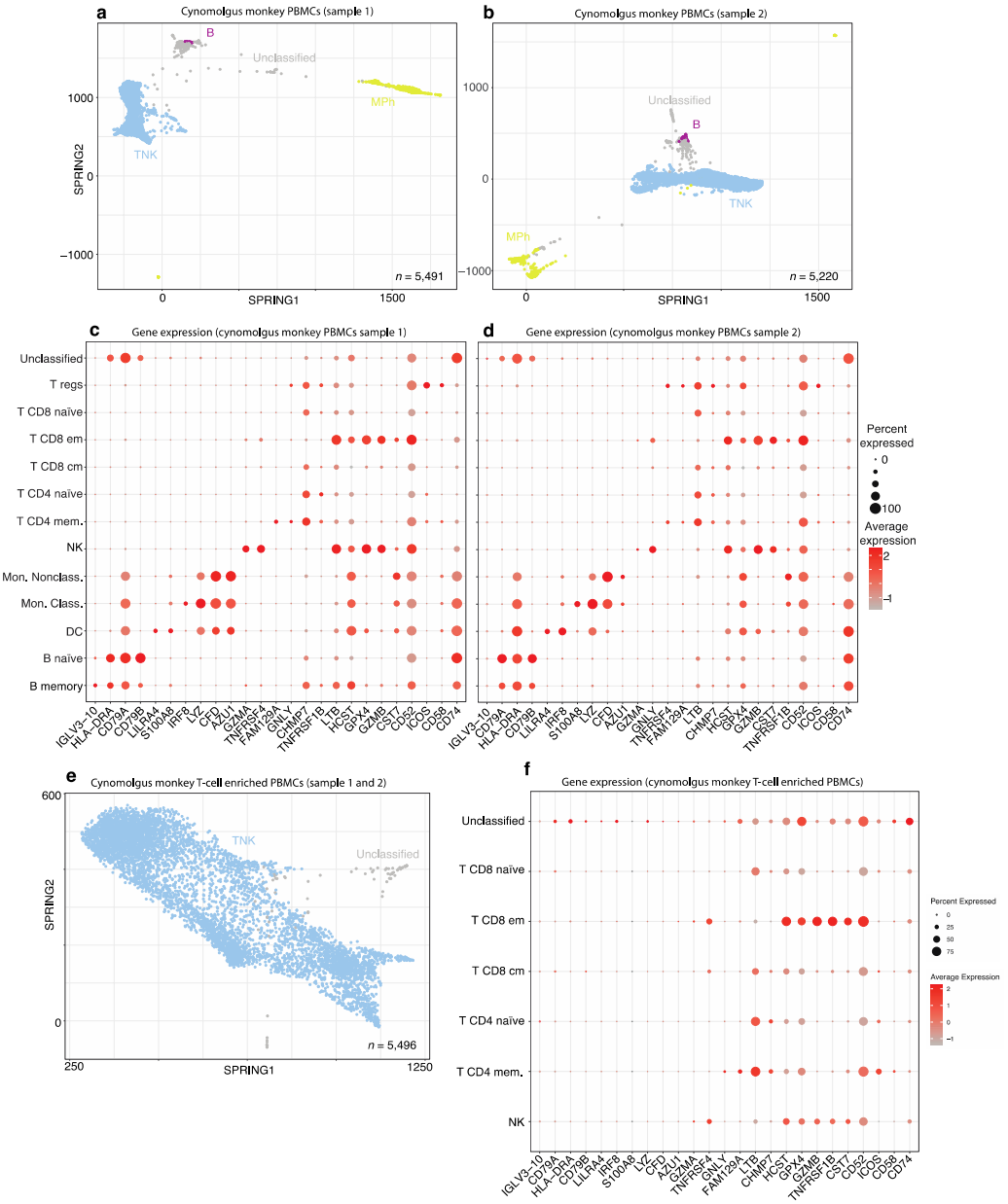

a
B memory
b
B naive
c
Plasma
d
Classical monocytes
e
Macrophages
f
T CD8 effector memory
g
T CD4 naive
h
T regulatory
i
T CD4 memory
-log10 P
log2 fold change

Supplemental Figure 6: IMAGES identified from immune phenotypes in synovium single cell transcriptomes. A-I, Volcano plots demonstrate the IMAGES identified in each cell population. Each scatter plot depicts the statistical association (y-axis) and the average fold-change of IMAGES (each dot is a unique gene) for immune cell phenotypes. Colors (red) indicate IMAGES that passed the thresholds applied to the fold-change and adjusted p-values.



Supplemental Figure 7: Signac-generated "unclassified" PBMCs were all identified as doublets (Scrublet), novel and classified cell type populations were mostly singlets. A-I, Volcano plots demonstrate the IMAGES identified in each cell population. **A, UMAP plot displays the cell type annotations of Signac for PBMCs (left), and the same data but with doublets classified by Scrublet (right)**. Each cell barcode (*n* = 8,381 from one human donor) was classified by Signac (left), and then doublet labels were amended to each cell with Scrublet (right; red); see Methods: Signac classification. **B, Bar plot reveals that unclassified cells were entirely composed of doublets, whereas novel cell populations and classified cells were mostly composed of singlets**. Each bar shows the percentage (y-axis) of each cell type (x-axis) that is a doublet (teal) or a singlet (red). Error bars correspond to 95% confidence intervals, two-sided binomial test. **C, IMAGE expression dot plot shows that unclassified cells and novel cluster 2 were doublet-like, whereas novel cluster 2 and 3 were singlet-like and enriched for**

known platelet and hematopoietic stem-cell gene markers. Dot plot shows the percentage (size) of single-cell transcriptomes within a cell type (y-axis) for which non-zero expression of marker genes was observed (x-axis). Color displays the average gene expression (red indicates more expression) in each cell type category. Novel cluster 1 was enriched for IMAGES that were typically enriched in either B cells or T cells, but not both (i.e., these cells expressed both *CD79B* and *TRAC*), consistent with the view that these cells were doublets. Novel cluster 3 images ($GNG11^{+}$ $TUBB1^{+}$) suggested platelet-like cells, and novel cluster 2 images ($CD34^{+}$ $SPINK2^{+}$) suggested hematopoietic stem cell-like cells.


a
Cynomolgus monkey PBMCs (sample 1)
B
Unclassified
MPh
TNK
SPRING2
SPRING1
n = 5,491
b
Cynomolgus monkey PBMCs (sample 2)
Unclassified
B
TNK
MPh
n = 5,220
c
Gene expression (cynomolgus monkey PBMCs sample 1)
Unclassified
T regs
T CD8 naïve
T CD8 em.
T CD8 cm.
T CD4 naïve
T CD4 mem.
NK
Mon. Nonclass.
Mon. Class.
DC
B naïve
B memory
IGLV3–10 HLA–DRA CD79A CD79B LILRA4 S100A8 IRF8 LYZ CFD AZU1 GZMA TNFRSF4 FAM129A GNLY CHMP7 TNFRSF1B LTB HCST GPX4 GZMB CST7 CD52 ICOS CD58 CD74
d
Gene expression (cynomolgus monkey PBMCs sample 2)
IGLV3–10 CD79A HLA–DRA CD79B LILRA4 IRF8 S100A8 LYZ CFD AZU1 GZMA GNLY TNFRSF4 FAM129A LTB CHMP7 HCST GPX4 GZMB CST7 TNFRSF1B CD52 ICOS CD58 CD74
Percent expressed
0
100
Average expression
2
−1
e
Cynomolgus monkey T-cell enriched PBMCs (sample 1 and 2)
TNK
Unclassified
SPRING2
SPRING1
n = 5,496
f
Gene expression (cynomolgus monkey T-cell enriched PBMCs)
Unclassified
T CD8 naïve
T CD8 em
T CD8 cm
T CD4 naïve
T CD4 mem.
NK
IGLV3–10 CD79A HLA–DRA CD79B LILRA4 IRF8 S100A8 LYZ CFD AZU1 GZMA TNFRSF4 GNLY FAM129A LTB CHMP7 HCST GPX4 GZMB TNFRSF1B CST7 CD52 ICOS CD58 CD74
Percent Expressed
0
25
50
75
Average Expression
2
1
0
−1

Supplemental Figure 8: Signac accurately classified primate PBMCs without any species-specific training. **A, SPRING plot for PBMCs from cynomolgus monkey donor 3003. B, SPRING plot for PBMCs from cynomolgus monkey donor 3004. C-D, IMAGE expression dot plot shows that Signac classifications for immune phenotypes were consistent with known gene markers**. **E, SPRING plot for PBMCs from cynomolgus monkey samples that were enriched for T cells during sequencing.** See Methods: Cross-species classification of single cell data from cynomolgus monkey PBMCs with human reference data. **F, IMAGE expression dot plot shows that Signac classifications for immune phenotypes were consistent with T cell type enrichment assay.**