Supplemental information

A mixture-of-experts deep generative

model for integrated analysis

of single-cell multiomics data

Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura
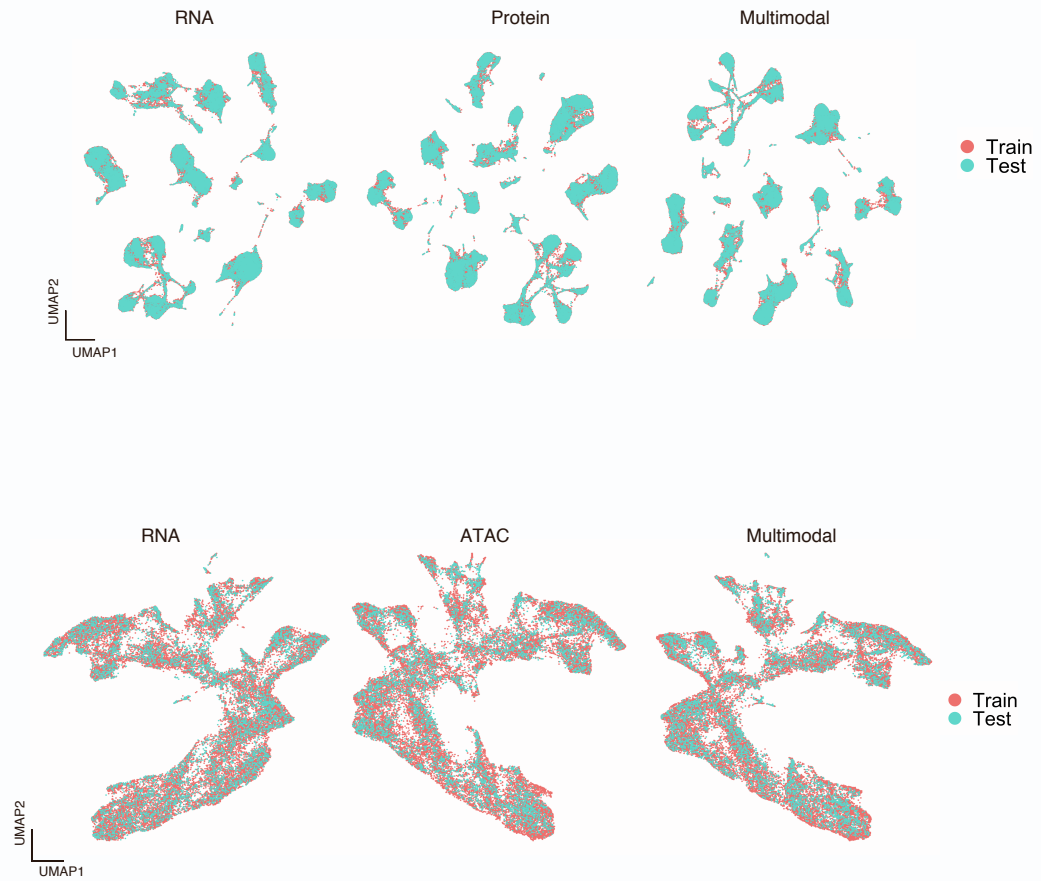
Fig S1. scMM embeds train and test datasets into the shared latent space, Related to Figure 2 and 4. UMAP visualization of unimodal and multimodal latent variables color-coded by assignments to either train or test dataset. Results for (**a**) human PBMC CITE-seq and (**b**) mouse skin SHARE-seq datasets are shown.

Fig S2. scMM revealed previously undiscovered heterogeneity in the PBMC dataset, Related to Figure 2. a, UMAP visualization of multimodal latent variables colored according to the expression levels of surface proteins differentially expressed in subsets of CD8 and CD4 T populations. **b**, CD14 Mono populations annotated in scMM and Seurat clustering analysis are indicated.
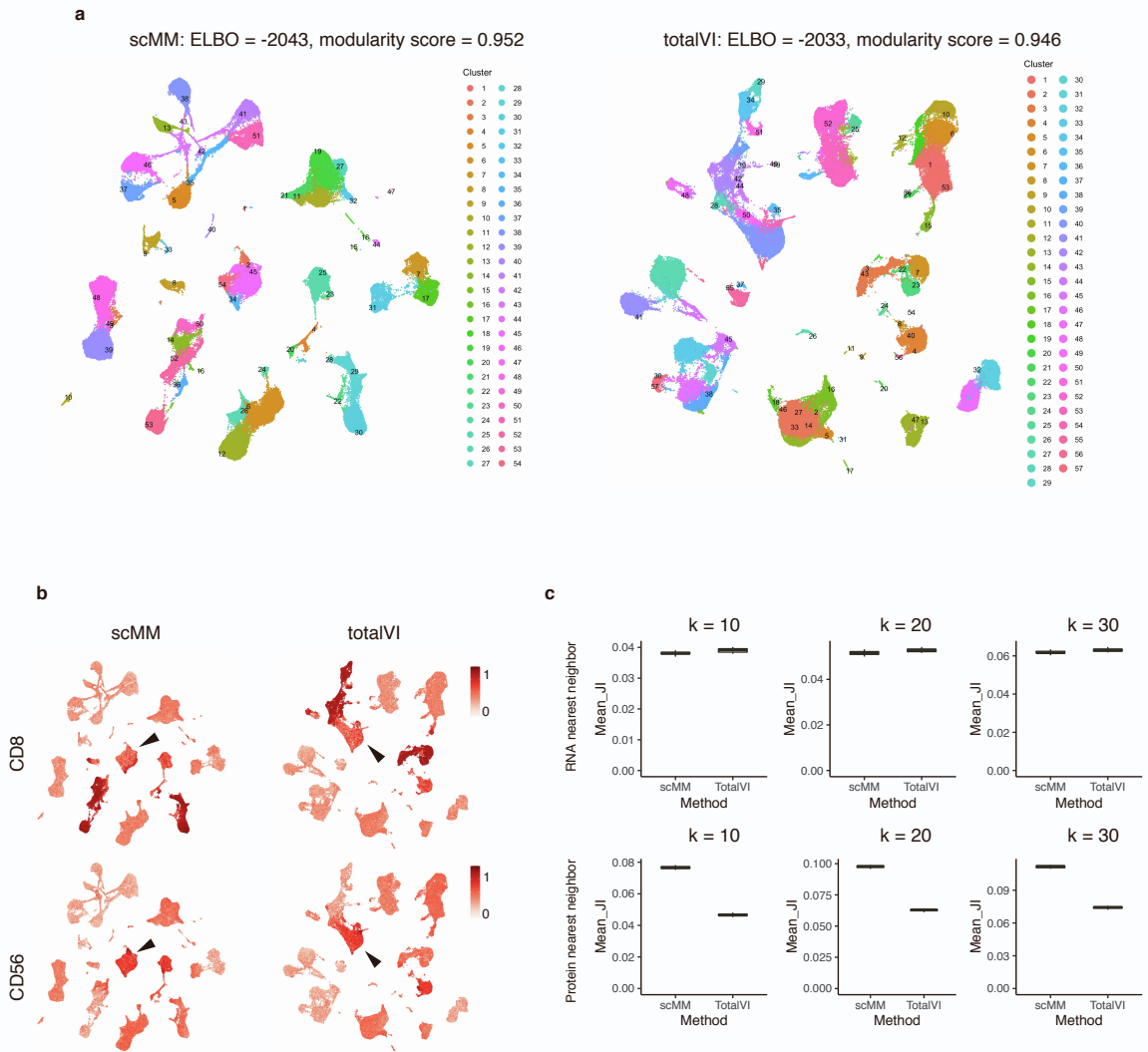
Fig S3. Comparison of scMM and totalVI performance, Related to Figure 2. (a) UMAP projection of latent variables inferred by scMM and totalVI. ELBO of each model and modularity scores calculated by PhenoGraph are shown above the projection. Each cell is colored according to the PhenoGraph cluster. **(b)** Cells are colored according to the expression levels of surface protein marker CD8 (CD8 T cell) and CD56 (NK cell). Black arrowheads indicate a NK population that is separated in UMAP projection of scMM but not in totalVI. **(c)** Mean JI calculated using transcriptome and surface protein nearest neighbors. Results for different number of neighbors *k* are shown.
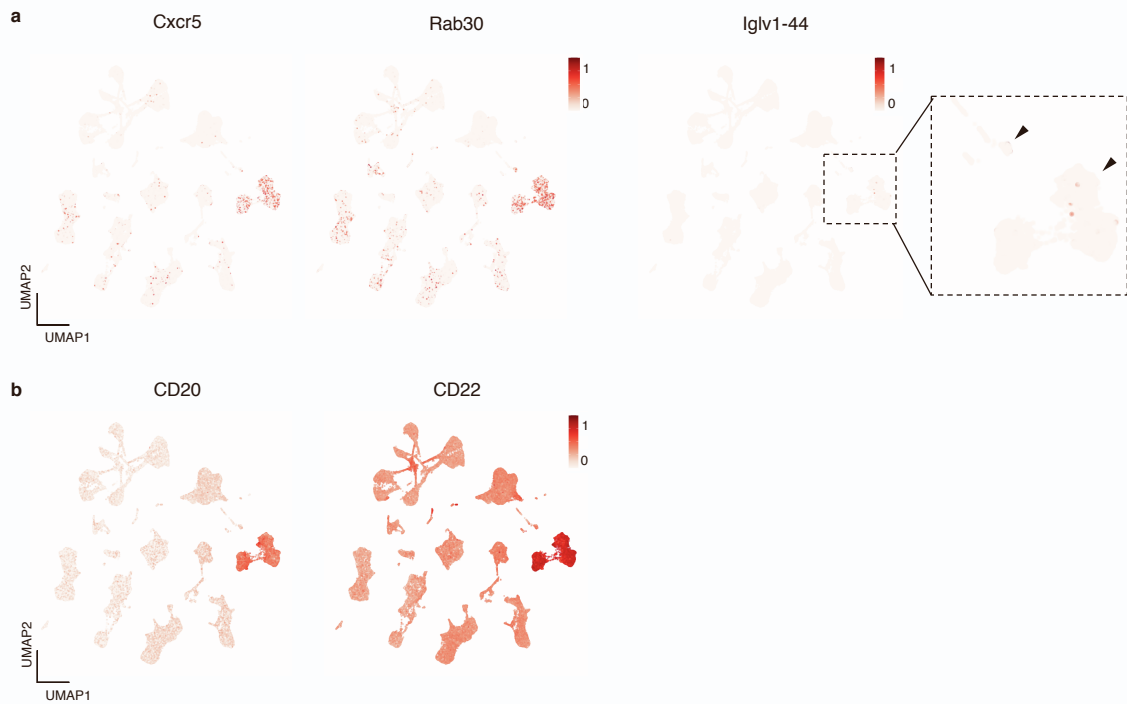
Fig S4. Genes and surface proteins detected to be associated with the latent dimension 1 enriches in B cell and plasmablast populations, Related to Figure 2. a, UMAP visualization of multimodal latent variables colored by expression levels of Cxcr5, Rab30, and Iglv1-44. Black arrowheads in enlarged panel indicate expression of Iglv-44. **b**, UMAP visualization of multimodal latent variables colored according to the expression levels of CD20 and CD22.
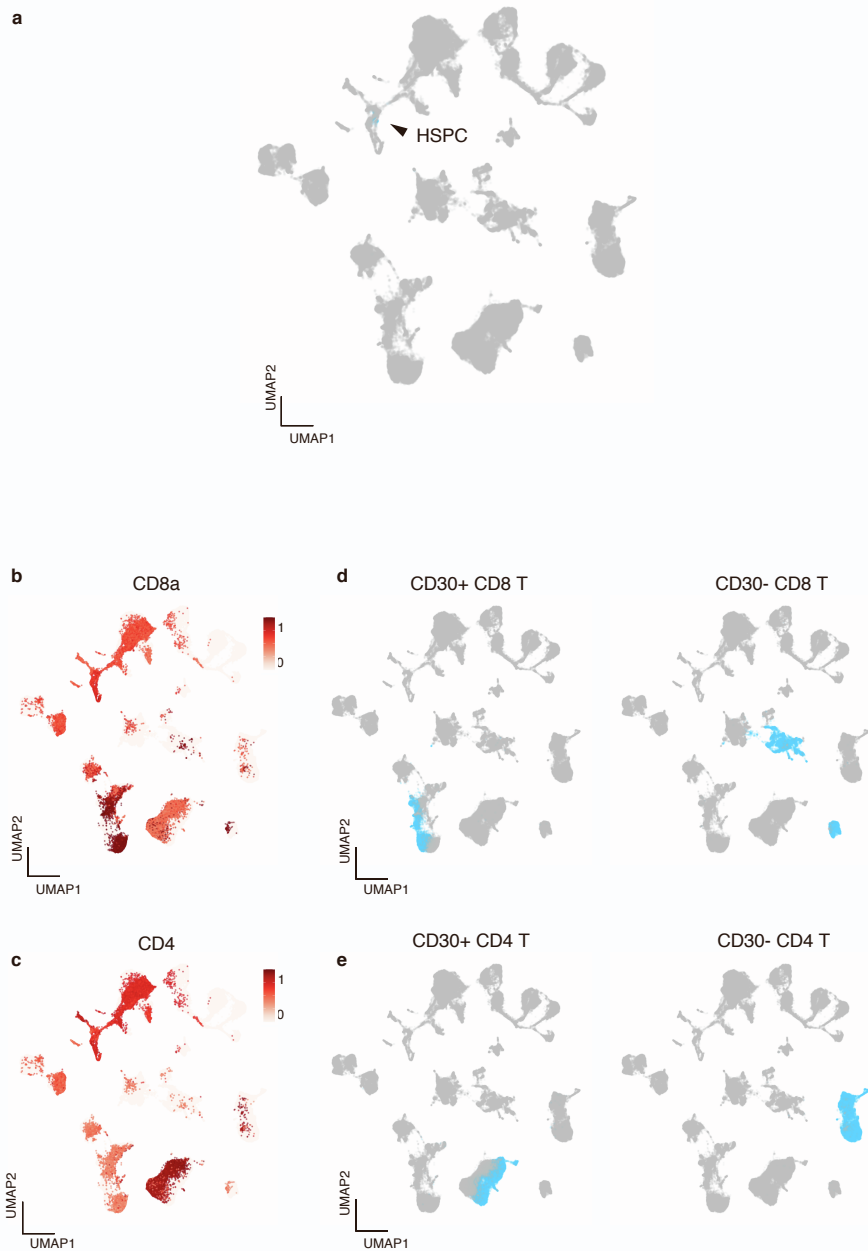
Fig S5. Embeddings of BMNC cells into the latent space by scMM captures the biological properties of bone marrow populations, Related to Figure 3. a, Joint UMAP visualization of BMNC and PBMC transcriptome unimodal latent variables. HSPC population annotated in PBMC dataset is color-coded and indicated by black arrowhead. **b**, **c**, UMAP embeddings for BMNC dataset is colored according to the expression levels of CD8 and CD4, respectively. **d**, **e**, CD8 and CD4 subsets in the PBMC dataset with and without CD30 expression are color-coded.
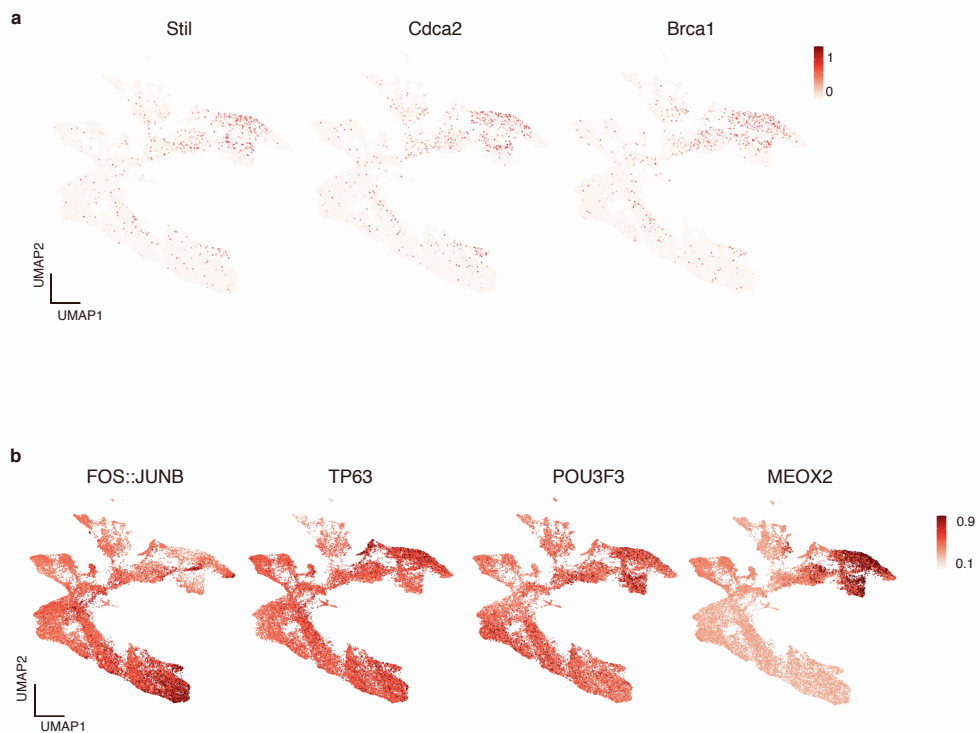
Fig S6. Gene and motifs detected to be associated with latent dimension 9 enriches in proliferative keratinocyte subsets, Related to Figure 4. a, UMAP visualization of multimodal latent variables colored according to the expression levels of Stil, Cdca2, Brca1. **b**, UMAP visualization of multimodal latent variables colored according to the motif scores for FOS::JUNB, TP63, POU3F3, and MEOX2.