Cell Reports Methods, Volume *2*

Supplemental information

PeakVI: A deep generative model for single-cell chromatin accessibility analysis

Tal Ashuach, Daniel A. Reidenbach, Adam Gayoso, and Nir Yosef

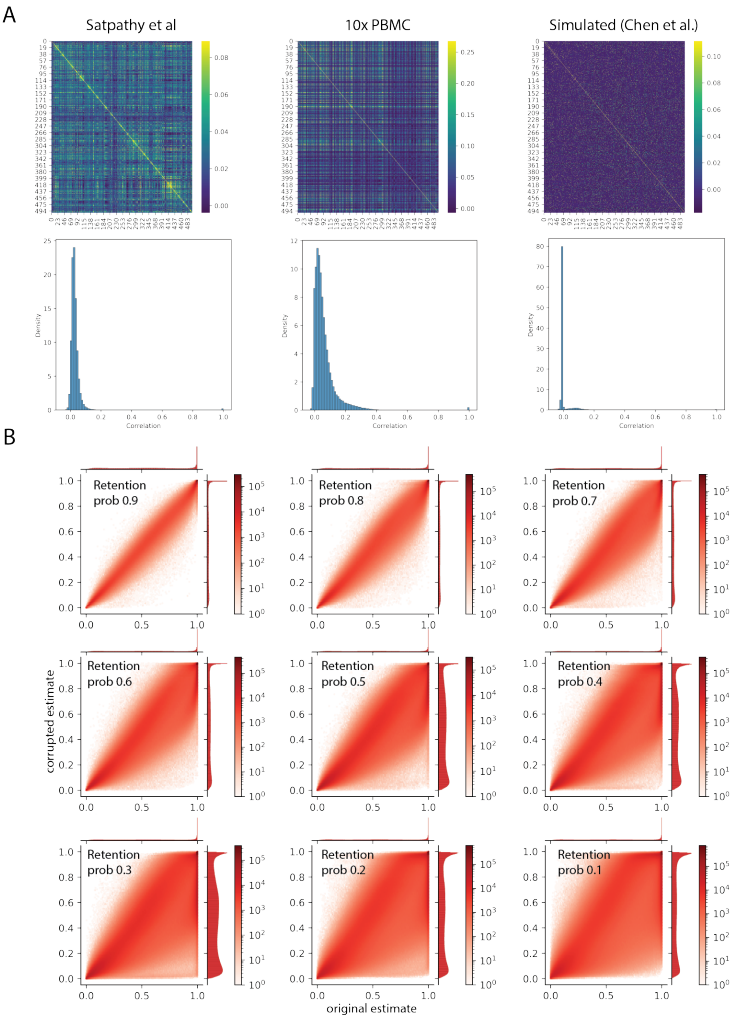

Figure S1: scATAC properties and PeakVI robustness, related to Figure 1. **(A)** Pearson correlation matrix (top) and distribution of correlation coefficients (bottom) of regions in three datasets: the immune cell dataset from Satpathy et al (Satpathy et al. 2019) (left); the sample multi-omics 10K cells PBMC dataset from 10x Genomics (center); and a simulated Bone Marrow dataset generated by Chen et al (Chen et al. 2019). For visual purposes, figures were generated using only the first 500 regions in each dataset, and across all available cells. Simulated data does not adequately represent the covariance structure of real scATAC-seq data. **(B)** Robustness analysis. Corruption analysis, in which observations were randomly replaced by zeros. Visualization is limited only to corrupted indices, showing that while increased corruption destabilizes the model, PeakVI is overall highly robust to the sparsity of low quality data.

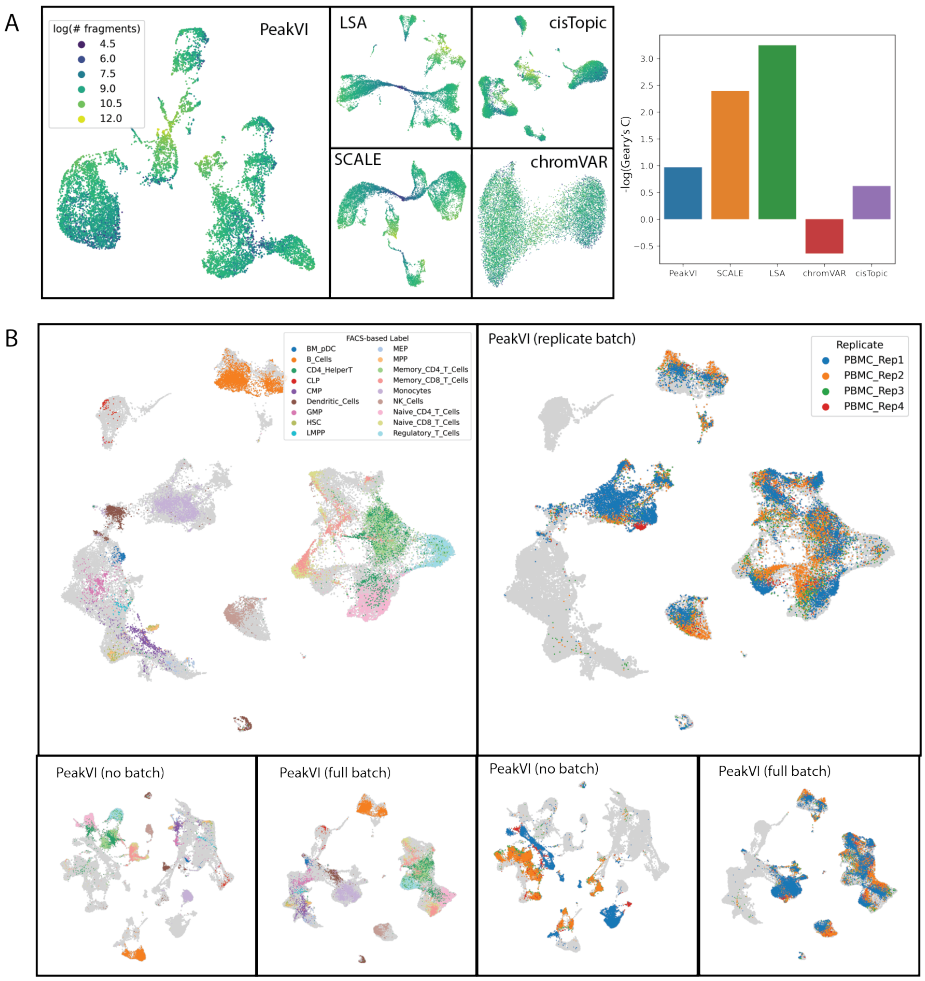

Figure S2: Additional Latent Visualizations, related to Figure 2. **(A)** Library size technical confounding. UMAPs of the sample paired scRNA and scATAC-seq PBMC data from 10X genomics, colored by the number of fragments mapped for each cell (left) and the spatial autocorrelation measured using Geary's C(Geary 1954) (right). LSA and SCALE are most impacted by library size effects, PeakVI and cisTopic are robust, and chromVAR is negatively correlated. **(B)** Visualizations of the Hematopoiesis data using three configurations of PeakVI, related to Figure 2. Treating replicates of multi-replicate samples as separate batches (replcate batch); without batch correction (no batch); treating each sample as a separate batch (full batch). Colored by FACS-based labels (top) and replicates of the unsorted PBMC samples (bottom). Unlabelled cells are colored in light gray.

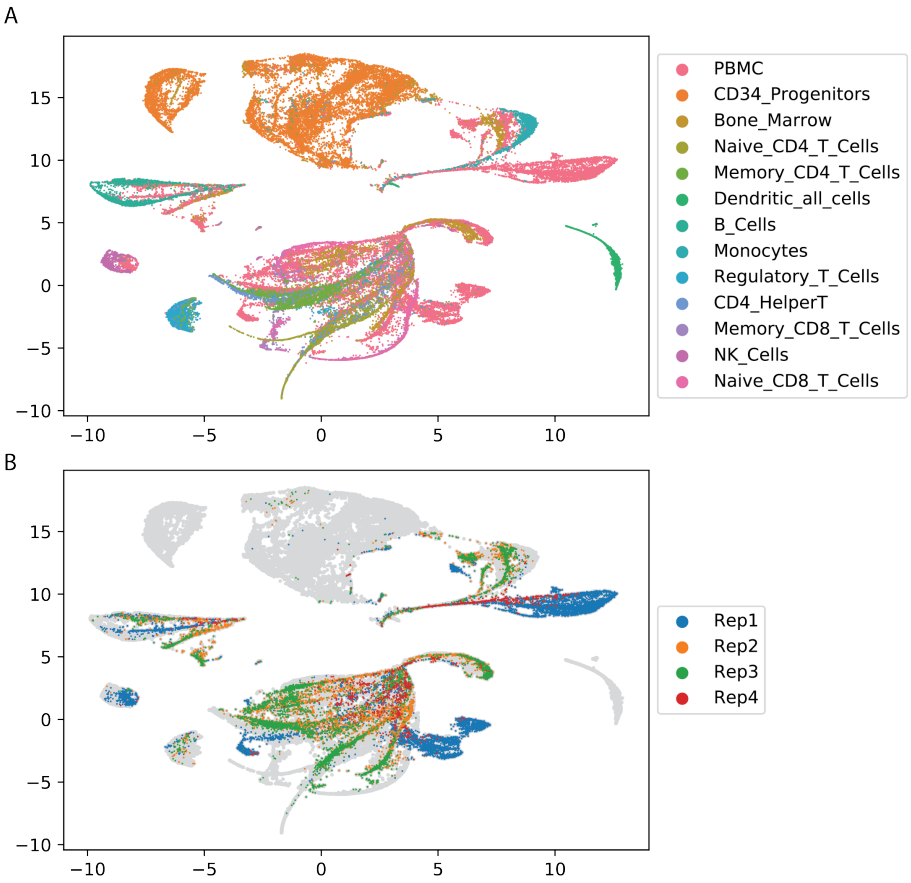

Figure S3: Visualization of the Hematopoiesis data using the ArchR dimensioanlity reduction (Iterative LSA), colored by cell type (A) and batch (B). Related to figure 2. Related to FIgure 2.

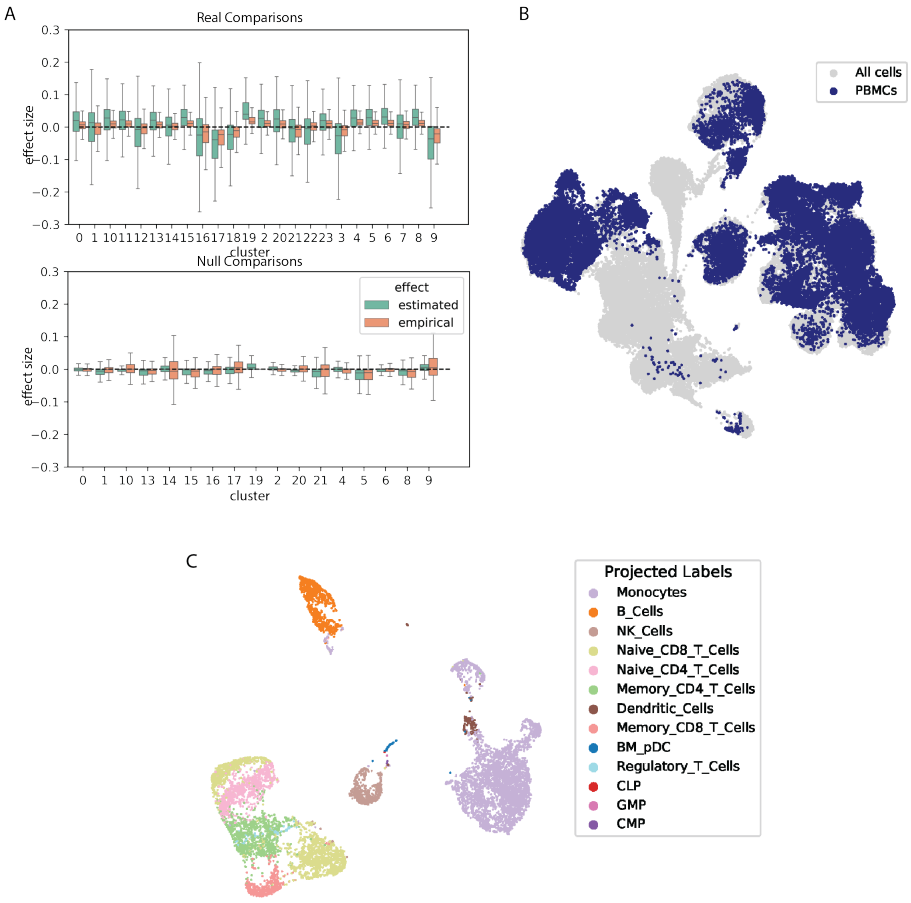

Figure S4: Differential Accessibility and Annotation, related to Figure 3 and Figure 4. **(A)** The effect size distribution for each real (top) and null (bottom) comparison. PeakVI estimated effects are amplified compared with the empirical effect in real comparisons, but the opposite is true for null comparisons. Overall PeakVI consistently has a better signal-to-noise ratio. Data represented as the mean $\pm$ 1SD, whiskers are $1.5 \cdot IQR$. **(B)** The low-dimensional representation of the Hematopoiesis data, trained in a scArches-compatible manner, with cells from PBMC samples in dark blue, showing how PBMCs are distributed in the space. **(C)** The low-dimensional representation of the Sample 10X PBMC data, with labels transferred from the hematopoiesis data.