Supplementary Material

SUPPLEMENTARY TEXT

Application of kNN-smoothing to scRNA-Seq data of mouse myeloid progenitor cells

To further compare our method to a previously proposed approach (Dijk et al. 2017), we applied our smoothing algorithm to a scRNA-Seq dataset of mouse myeloid progenitor cells (Paul et al. 2015). We generated a heatmap of characteristic genes for 19 clusters identified by the authors of the original study, as well as for important cell surface markers, in a way that allows a direct comparison to the results obtained by Dijk et al. (2017) (see Figure S6a,b). We found that even though k-nearest neighbor smoothing is much simpler than their approach, our method performed similarly well in generating smooth expression profiles for cells belonging to the same cluster, while respecting cluster boundaries.

We similarly examined the pairwise correlations of cell surface markers, and obtained qualitatively similar results to Dijk et al. (2017) (see Figure S6c-e). As in their study, recovering cell type-specific co-expression patterns depended on the amount of smoothing applied. Some differences were observed in the precise shapes of the associations, but it was not clear how much of this was due to differences in normalization and/or scaling used for visualization. In summary, for this particular dataset, the diffusion-based approach by Dijk et al. (2017) and our algorithm gave qualitatively similar results, although there were some quantitative differences.

SUPPLEMENTARY FIGURES

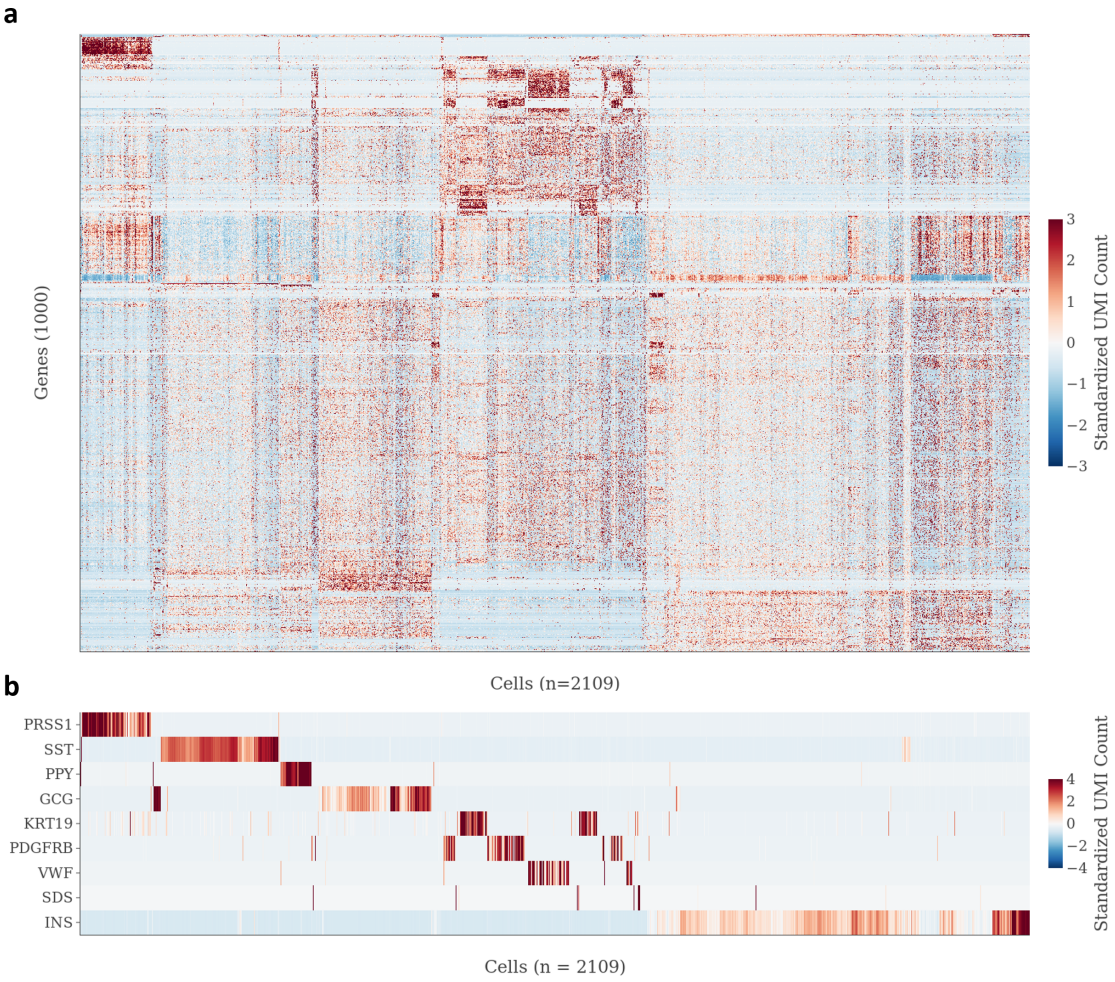

Figure S1. Hierarchical clustering of unsmoothed scRNA-Seq data from human pancreatic islet tissue. Shown is the PANCREAS dataset, from a study by Baron et al. (2016). **a** Heatmap showing the results of hierarchical clustering of genes and cells performed on the unsmoothed data, after filtering for the 1,000 most variable genes, as in Figure 4b). **b** Expression of cell type-specific marker genes, as in Figure 4d, but with genes reordered to accommodate the new clustering results.

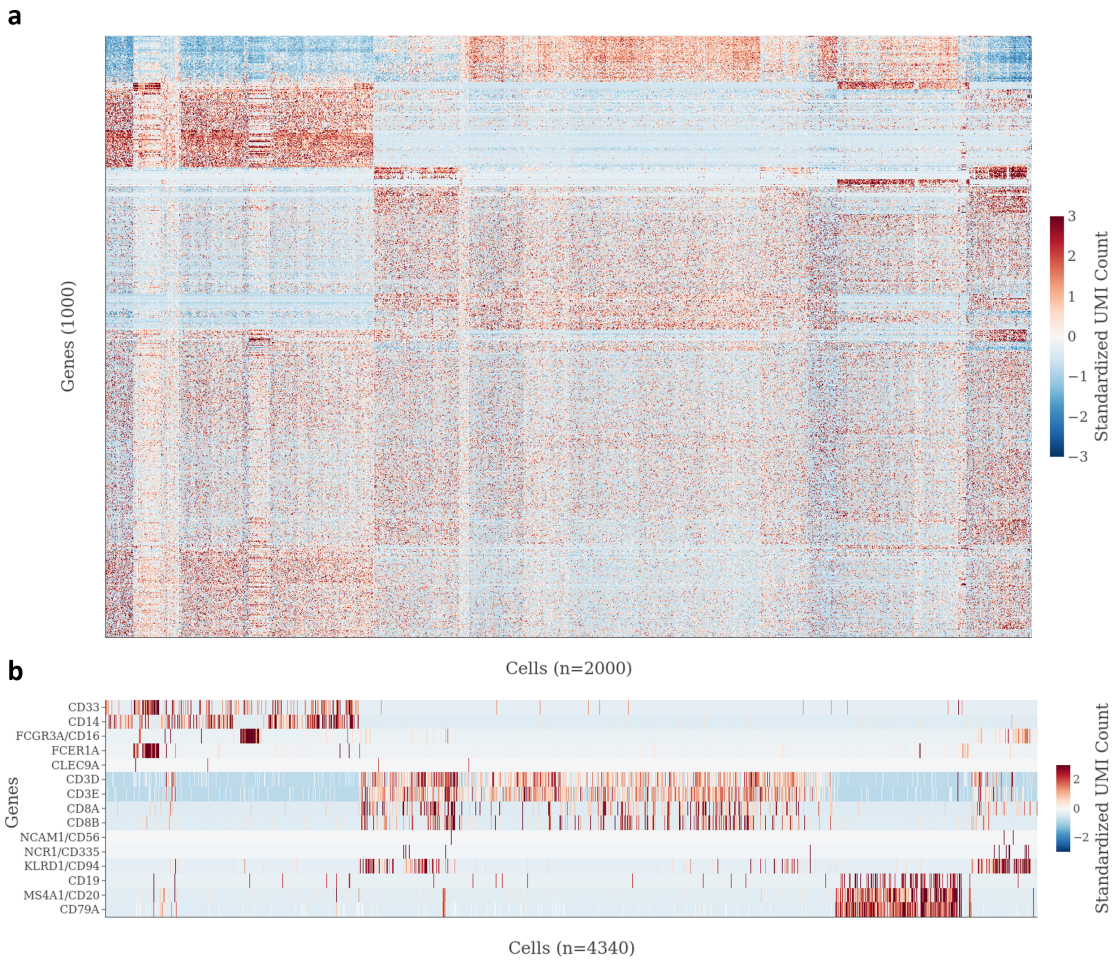

Figure S2. Hierarchical clustering of unsmoothed scRNA-Seq data from human peripheral blood mononuclear cells. Shown is the `PMBC` dataset. **a** Heatmap showing the results of hierarchical clustering of genes and cells performed on the unsmoothed data, after filtering for the 1,000 most variable genes, as in Figure 5b). **b** Expression of cell type-specific marker genes, as in Figure 5d.

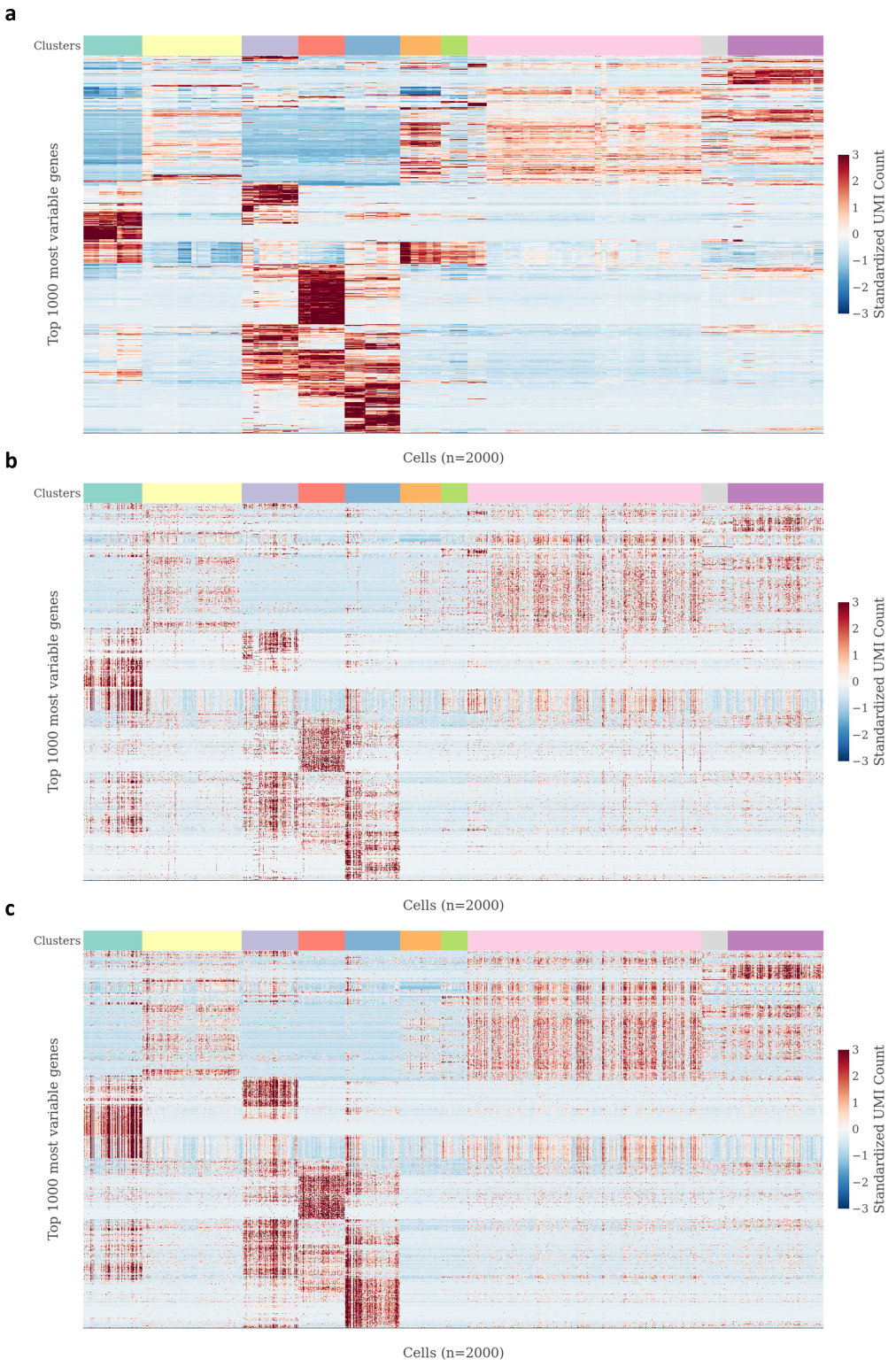

Figure S3. Comparison of smoothed, unsmoothed, and simulated scRNA-Seq data for human pancreatic islet tissue. All panels show heatmaps of the 1,000 most variable genes, where genes and cells are ordered according to hierarchical clustering results, obtained using the 2,000 most variable genes in the smoothed `PANCREAS` data (with k=15). Assignments of cells to one of 10 clusters (based on the same hierarchical clustering results) are shown on top of each heatmap. **a** Smoothed data. **b** Unsmoothed data. **c** Simulated data. Only a random subset of 2,000 cells (out of 2,109 cells) is shown in each heatmap.

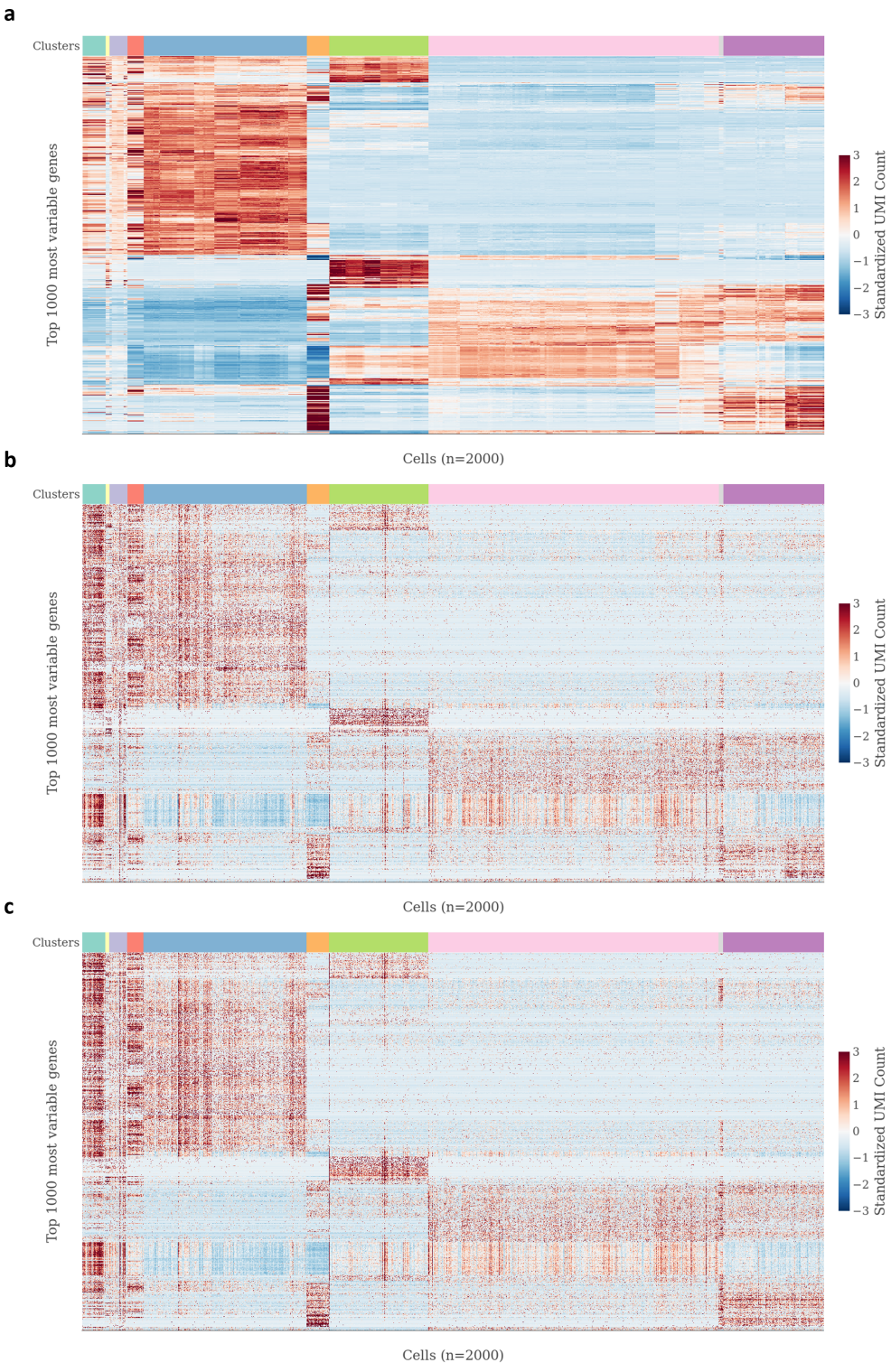

Figure S4. Comparison of smoothed, unsmoothed, and simulated scRNA-Seq data for human peripheral blood mononuclear cells. See Figure S3 for descriptions of panels (**a**)-(**c**).

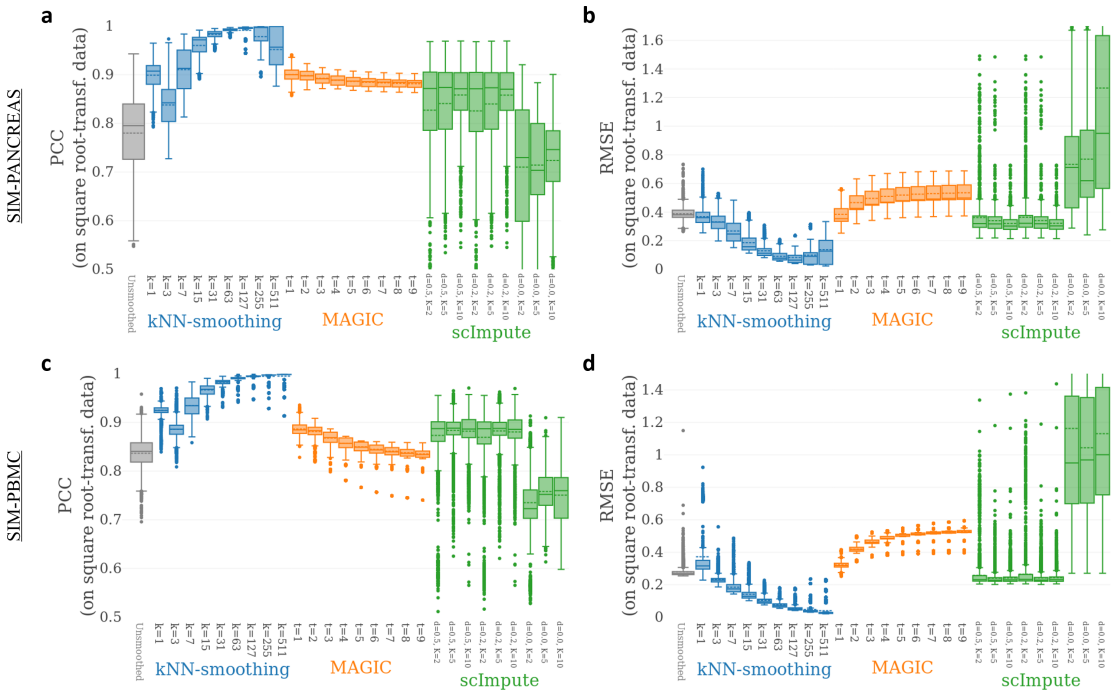

Figure S5. Accuracy of kNN-smoothing in comparison to other smoothing methods for simulated scRNA-Seq datasets. a, b Accuracy on `SIM-PANCREAS`dataset. **c, d**. Accuracy on `SIM-PBMC`dataset. Figure mirrors Figure 6, but shown are accuracy measures calculated on square root-transformed data, not on $\log_2$-transformed data (see Methods for details).

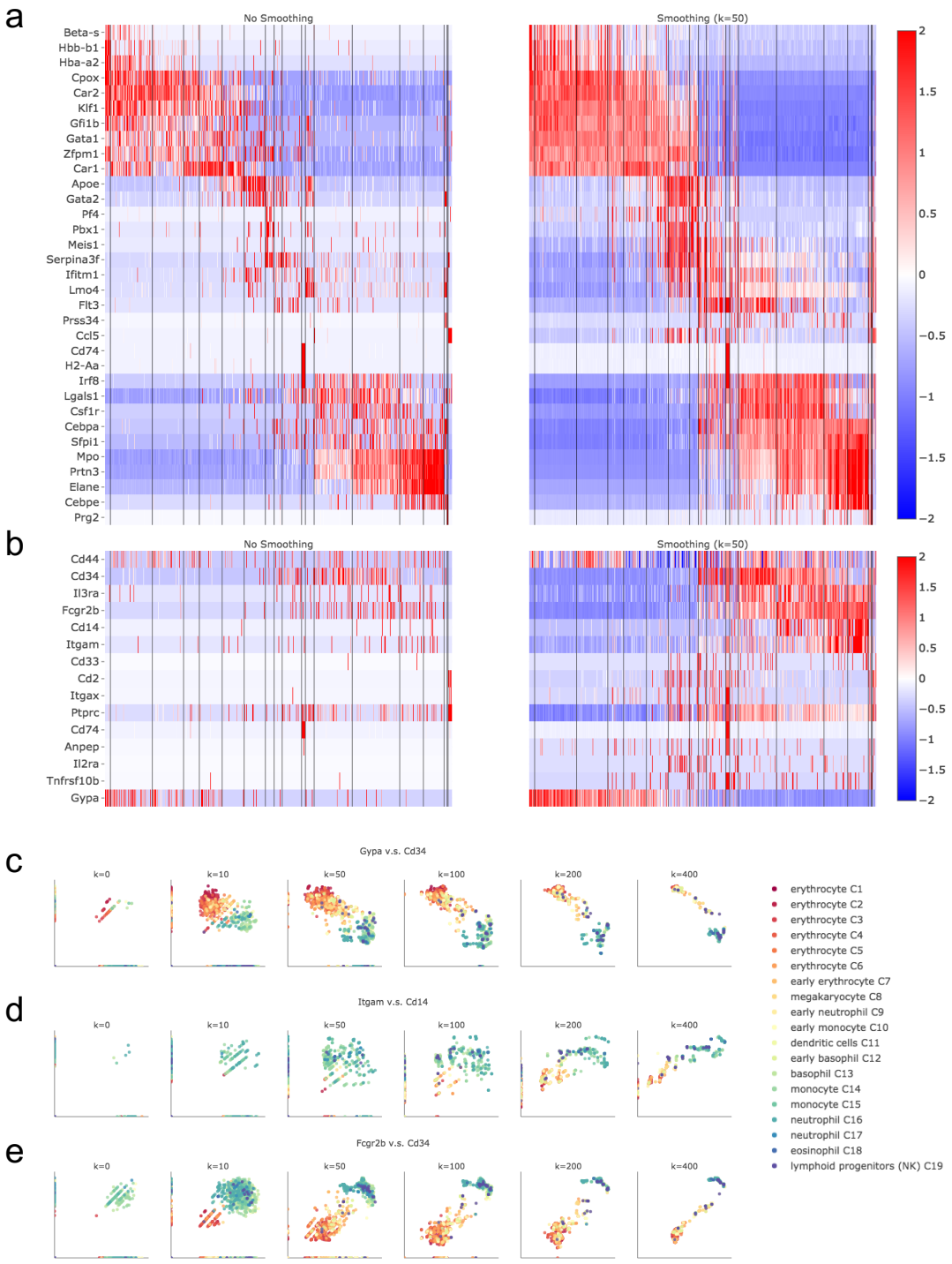

Figure S6. Application of k-nearest neighbor smoothing to scRNA-Seq data of mouse myeloid progenitors. This figure is directly comparable to Figure 3 from Dijk et al. (2017). (**a, b**) Heatmaps of the expression matrices for (**a**) 33 key hematopoietic genes, and **b**) 15 surface marker genes of immune cells, as defined in Paul et al. (2015), before smoothing (left) and after smoothing (right). Gene are ordered as same as shown in Dijk et al. (2017), Figure 3. Cells from left to right are ordered in clusters (C1-C19) as defined in Paul et al. (2015). **c**-**e**) Scatter plots of expressions showing the recovery of relationships of three pairs of immune marker genes after smoothing with different k (k=0, 10, 50, 100, 200, 400). Each dot is an individual cell colored by the 19 clusters used in **a**. See Methods for details.