Supplementary Figures for Alignment of single-cell RNA-seq samples without overcorrection using kernel density matching

Mengjie Chen[1,2*], Qi Zhan[3], Zepeng Mu[3], Lili Wang[3], Zhaohui Zheng[4,5], Jinlin Miao[4,5], Ping Zhu[4,5*], Yang I Li[1,2*]

[1] Section of Genetic Medicine, Department of Medicine, University of Chicago, IL

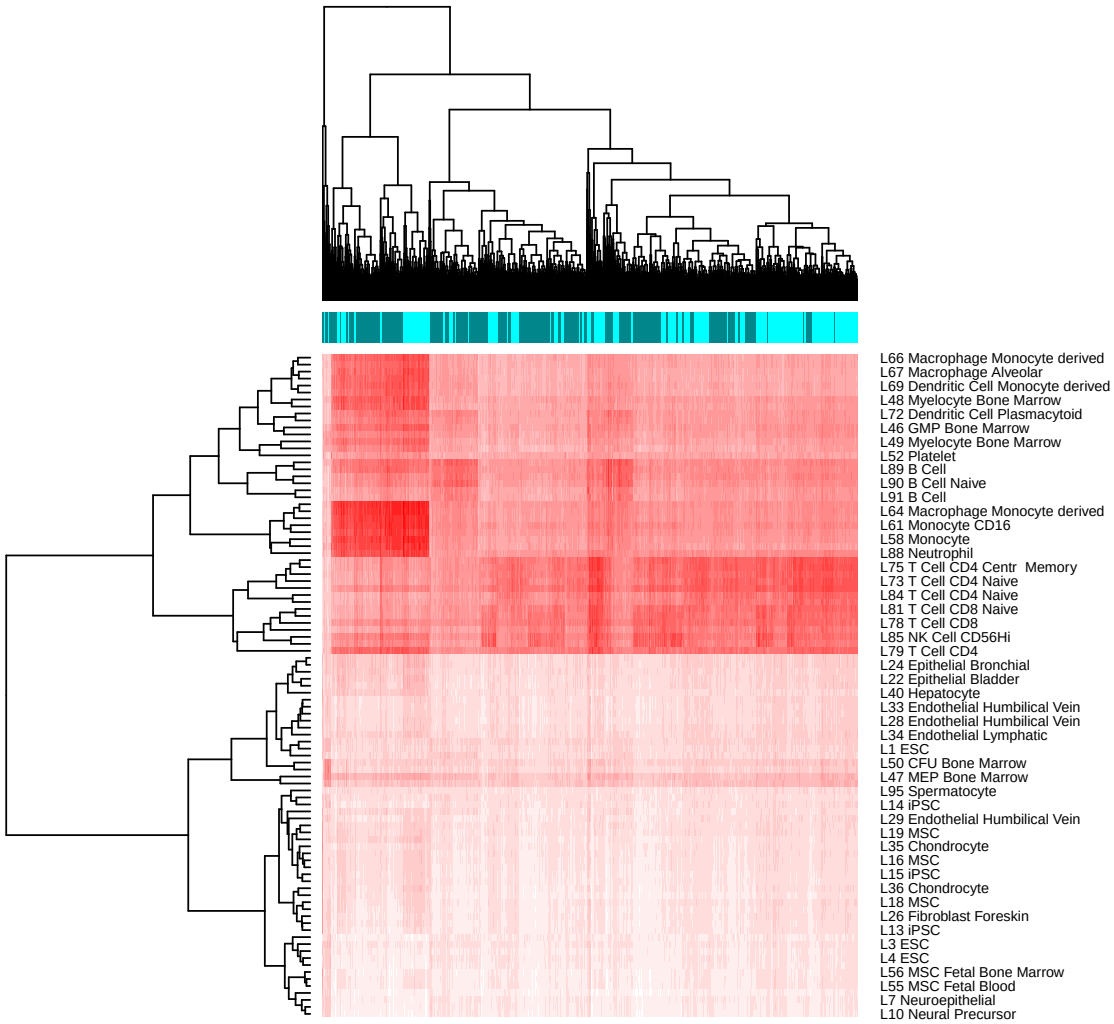
[2] Department of Human Genetics, University of Chicago, IL

[3] Committee on Genetics, Genomics & Systems Biology, University of Chicago, IL

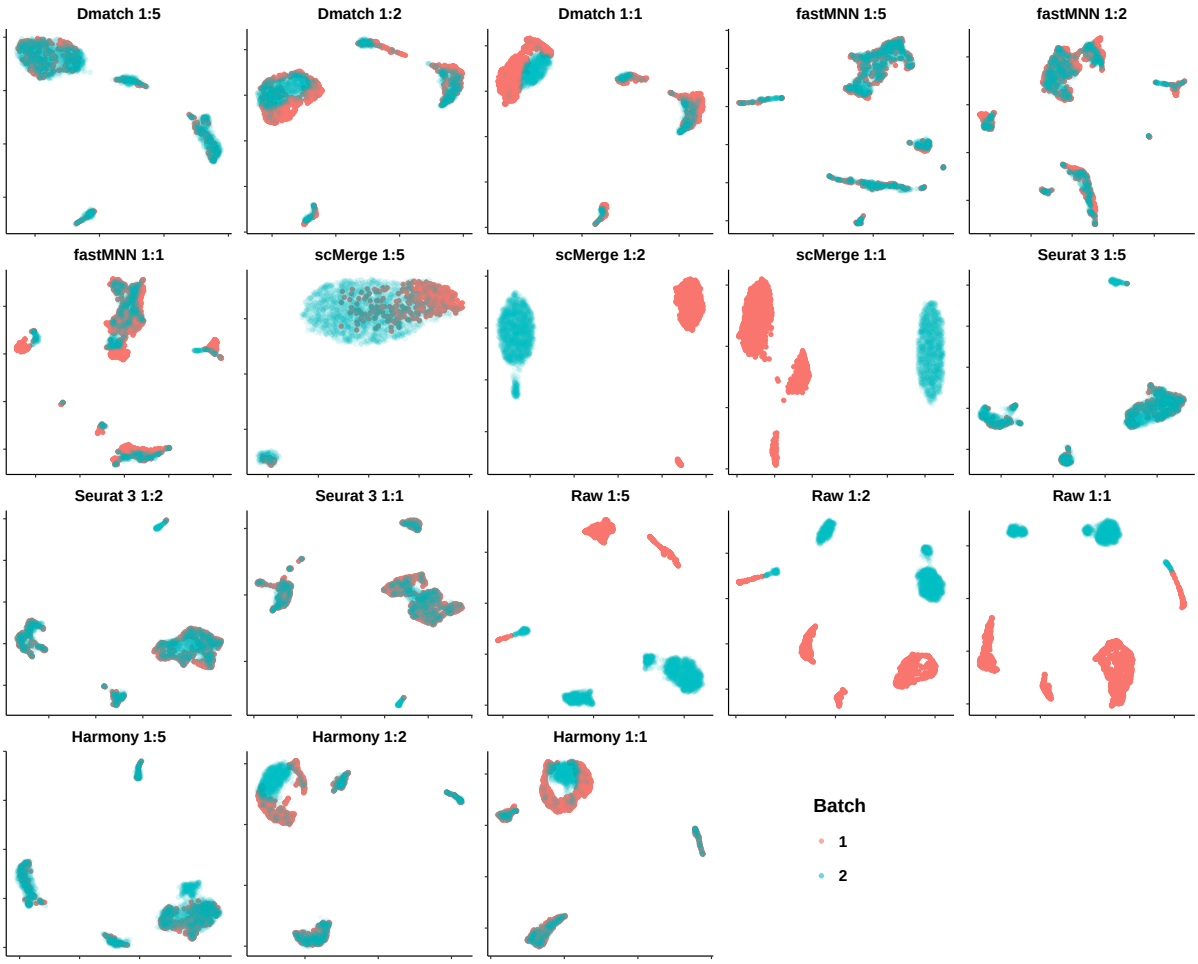
[4] Department of Clinical Immunology, Xijing Hospital, Xi'an, China

[5] National Translational Science Center for Molecular Medicine, Xi'an, China

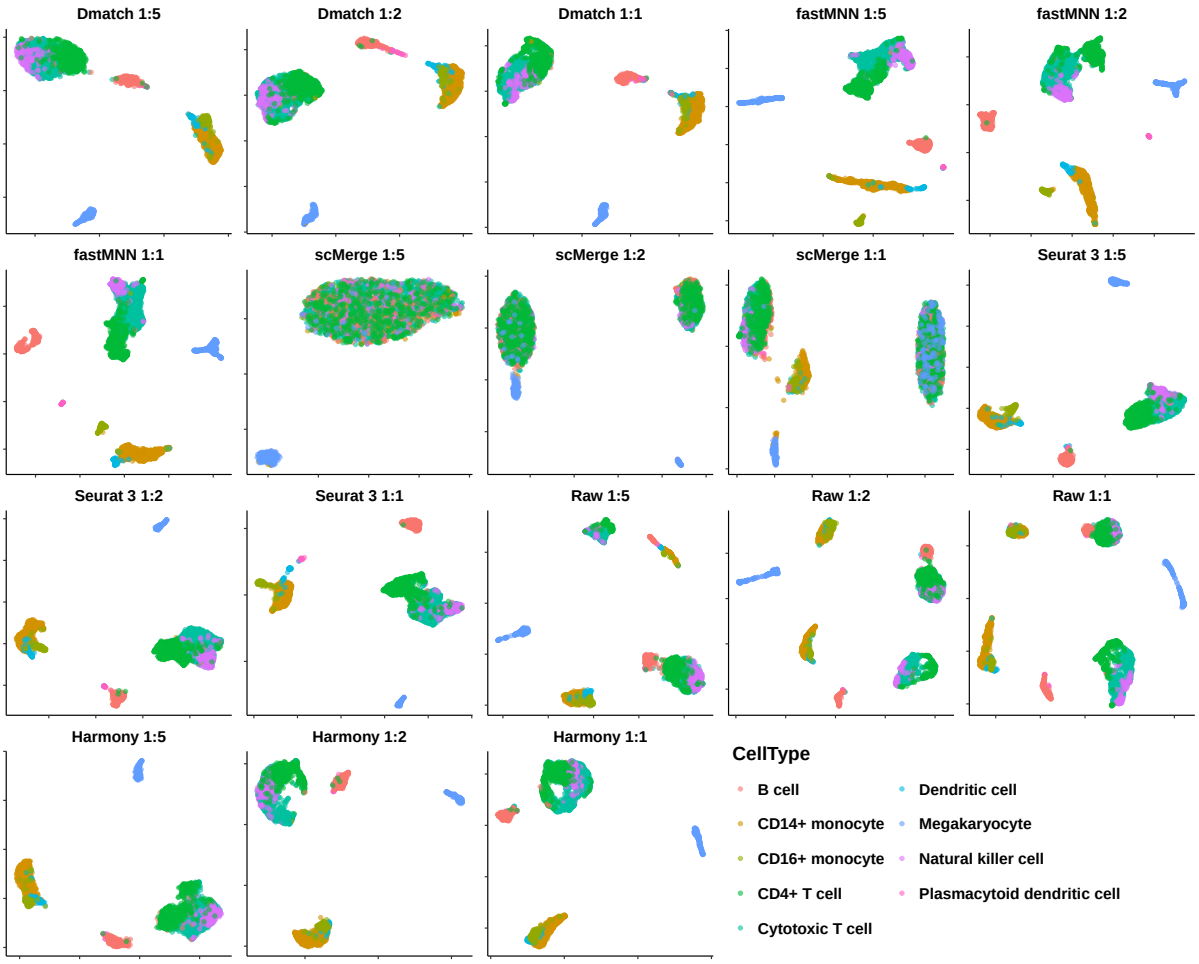
Correspondence should be addressed to M.C. (mengjiechen@uchicago.edu), Z.P. (zhuping@fmmu.edu.cn) or Y.I.L. (yangili1@uchicago.edu)



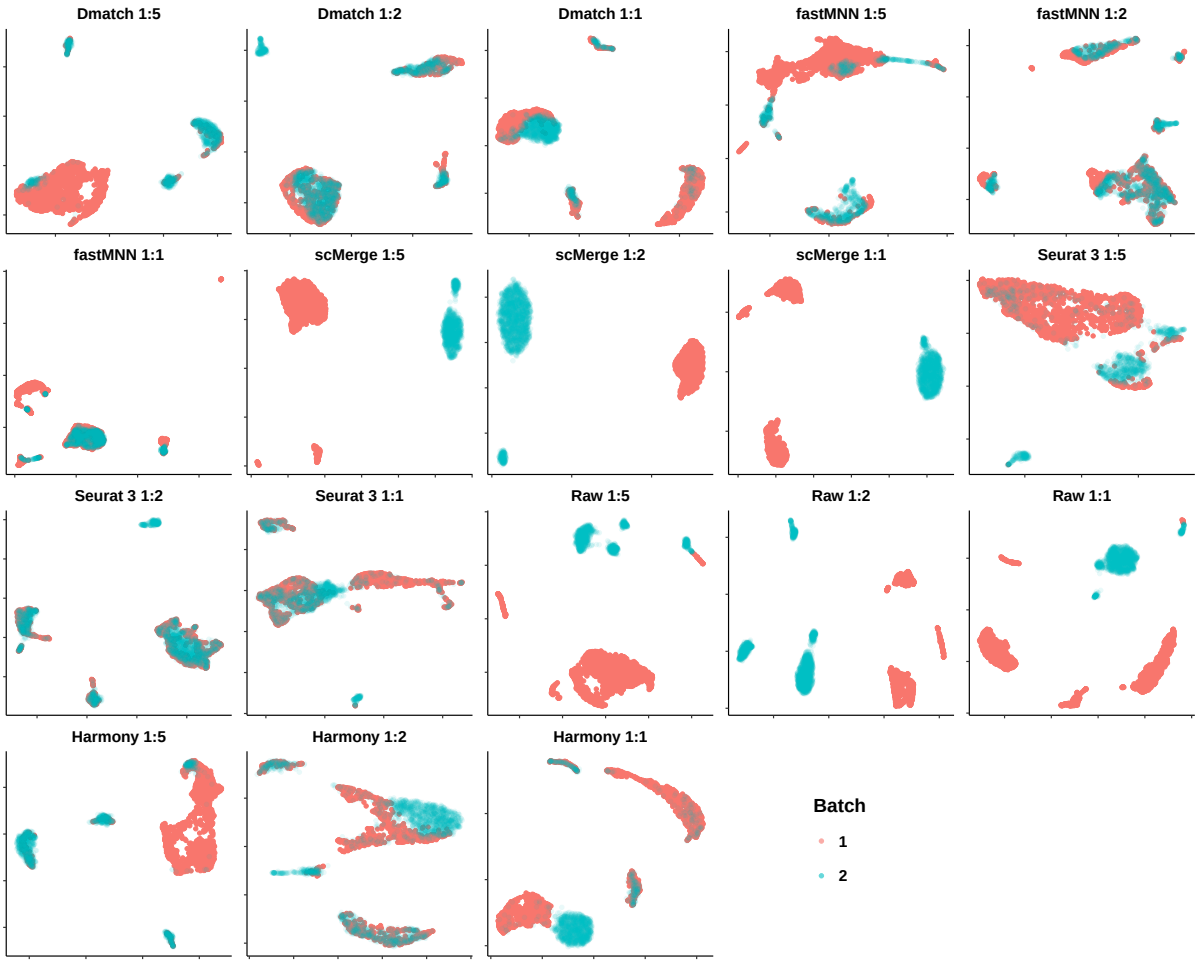
Supplemental Fig. S1: `Dmatch` uses a large reference transcriptomes from the Primary Cell Atlas to identify subpopulations from the observed cells based on the projection. We found that the consistency of cell type assignment was reduced if Pearson's correlations of the 95-dimensional vectors were not set to zero except for those between the cell and the top five reference atlas cell types with the highest Pearson's correlation coefficient. When this sparsity was not induced, the biclustering becomes noisier and the anchors less apparent (compared to Fig. 1B).



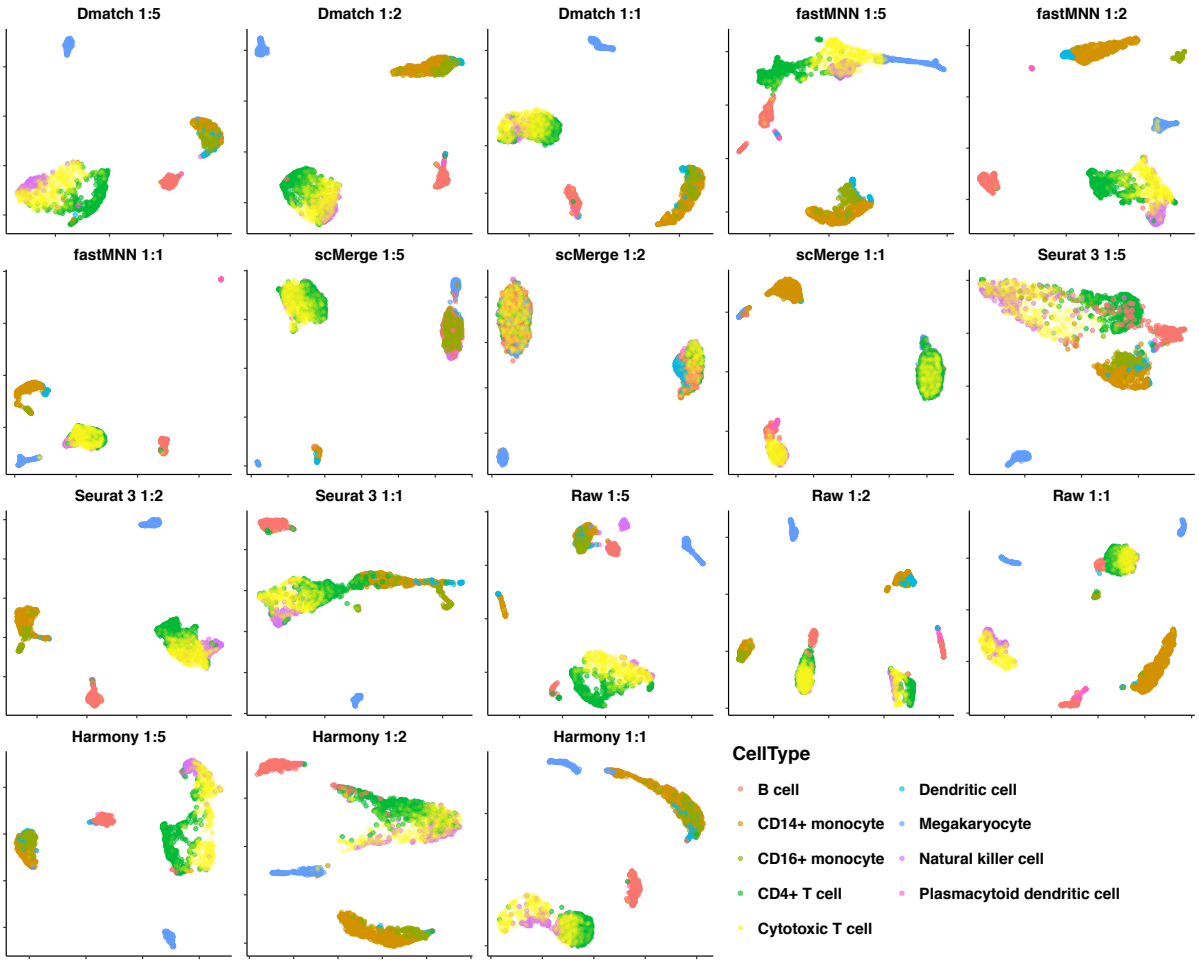
Supplemental Fig. S2: UMAP of aligned simulated data or unaligned (raw) colored by simulated batch (1 or 2) for one simulation scenario (All cell types shared and medium batch effect). The 1:1, 1:2, and 1:5 ratios denote the ratios of numbers of cells in the two batches to represent scenarios with equal number of cells, moderately unequal number of cells, and very unbalanced number of cells, respectively.



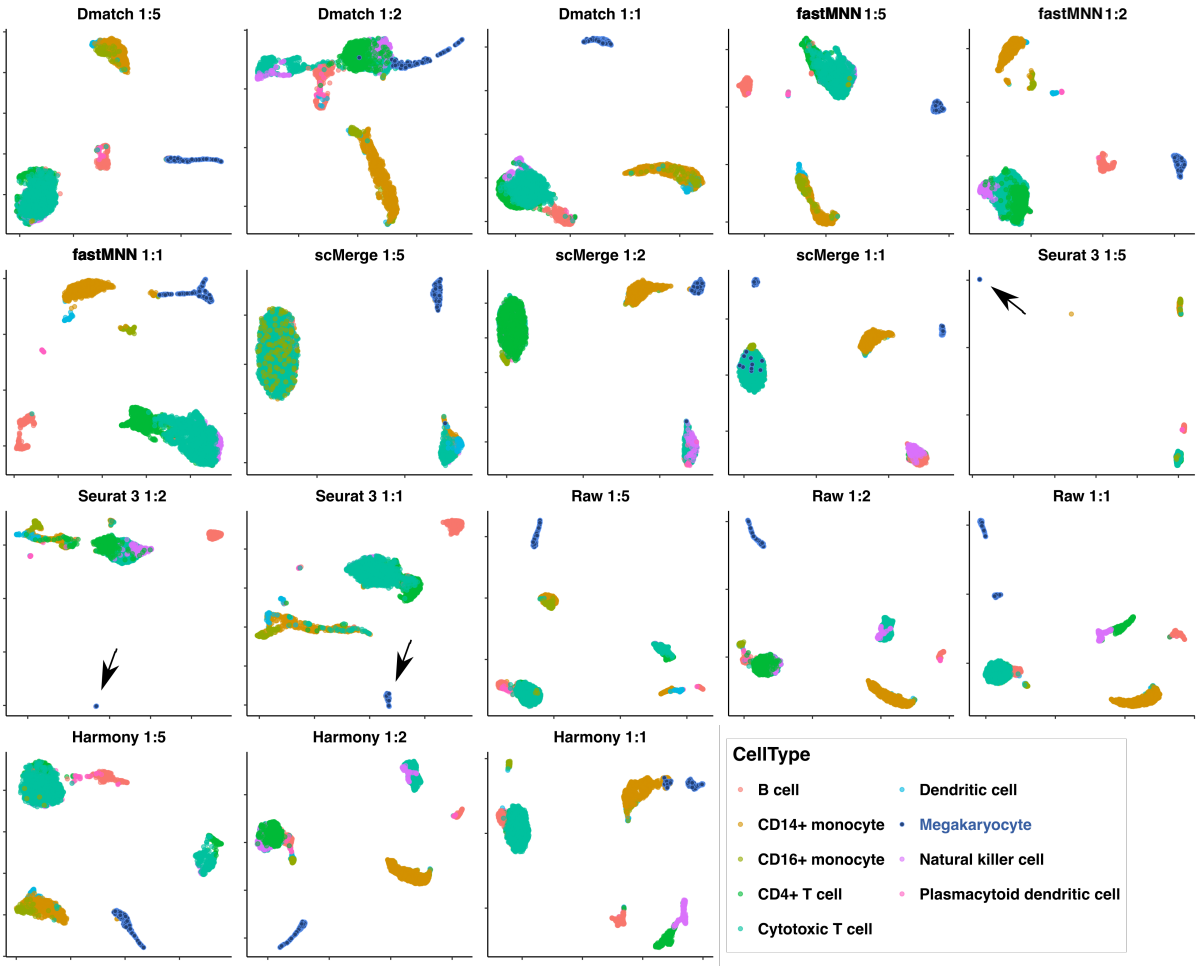
Supplemental Fig. S3: UMAP of aligned simulated data or unaligned (raw) colored by assigned cell type from the original data for one simulation scenario (All cell types shared and medium batch effect). The 1:1, 1:2, and 1:5 ratios denote the ratios of numbers of cells in the two batches to represent scenarios with equal number of cells, moderately unequal number of cells, and very unbalanced number of cells, respectively.



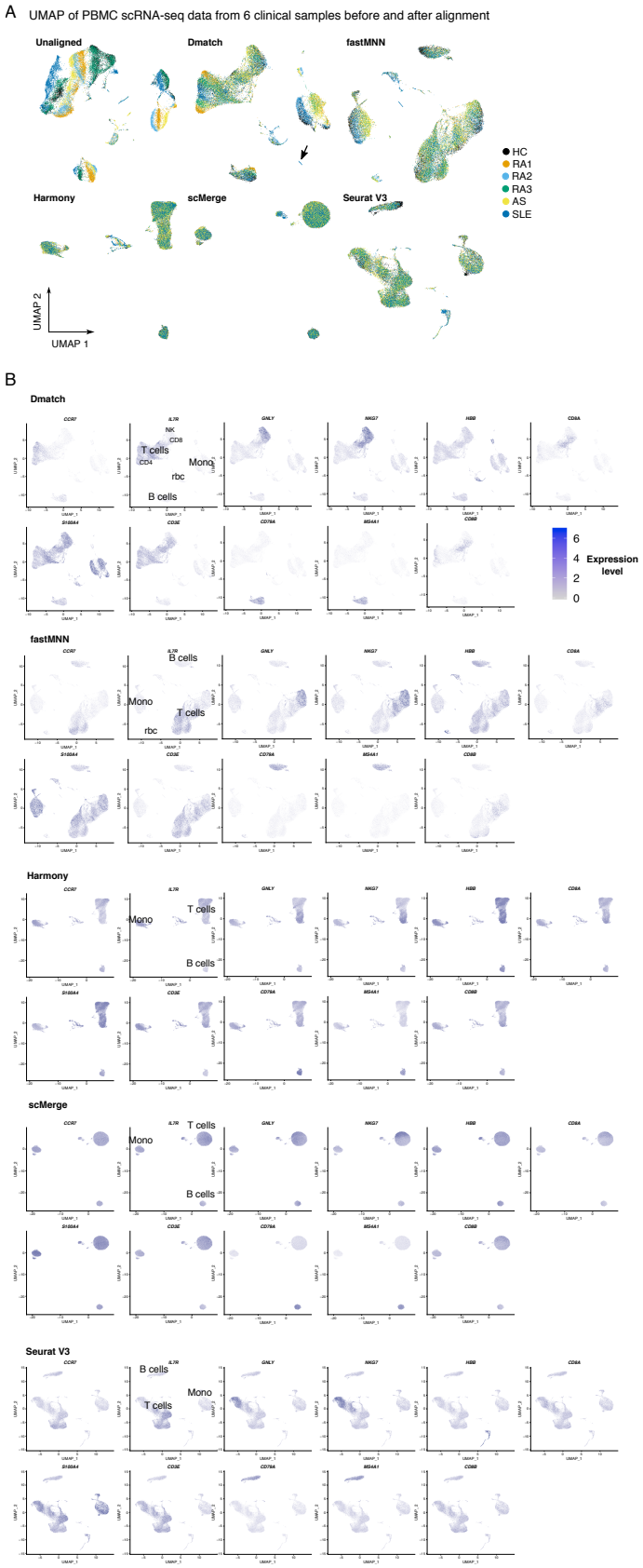
Supplemental Fig. S4: UMAP of aligned simulated data or unaligned (raw) colored by simulated batch (1 or 2) for one simulation scenario (Six cell types out of 9 shared and medium batch effect). The 1:1, 1:2, and 1:5 ratios denote the ratios of numbers of cells in the two batches to represent scenarios with equal number of cells, moderately unequal number of cells, and very unbalanced number of cells, respectively.



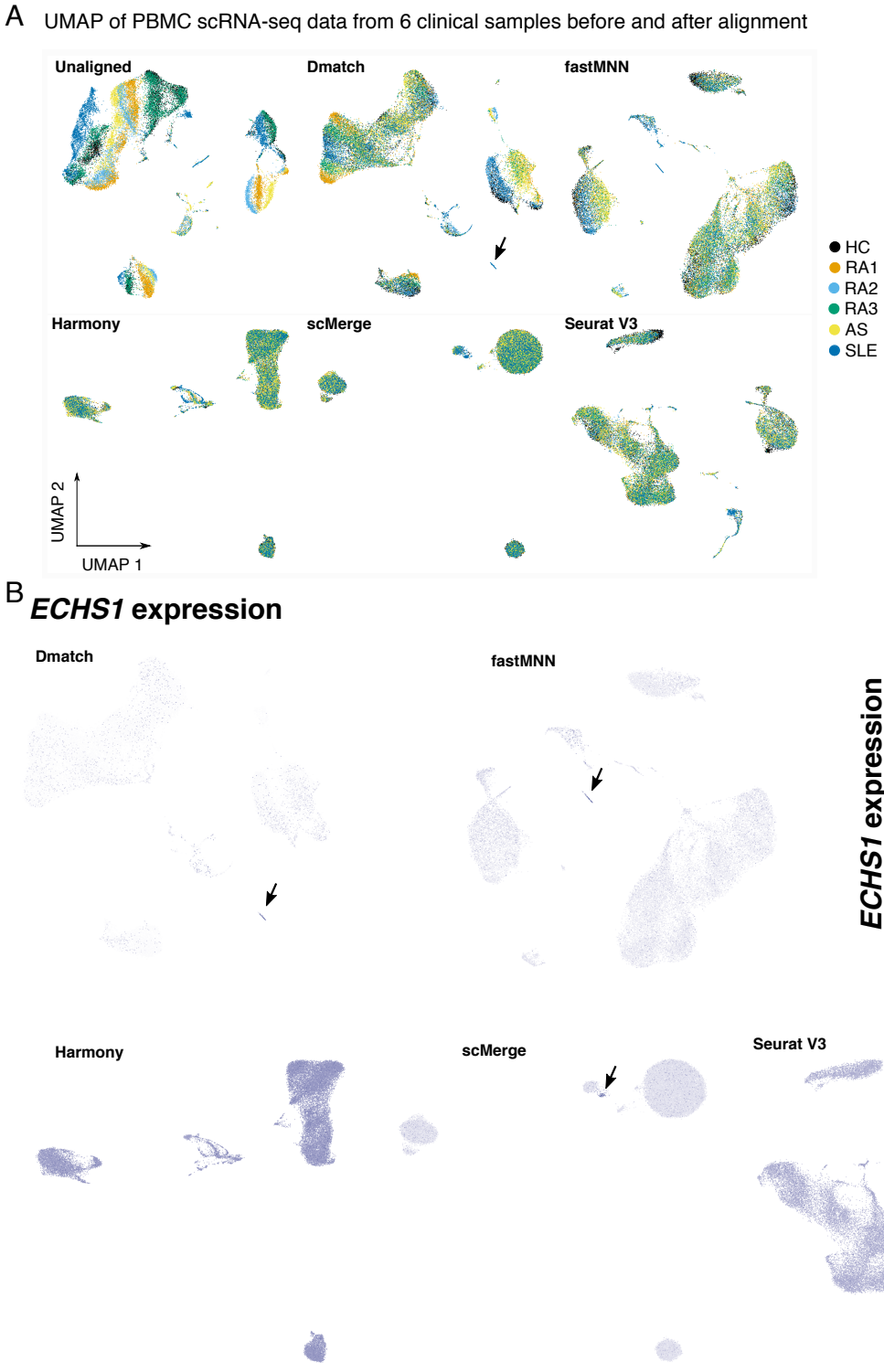
Supplemental Fig. S5: UMAP of aligned simulated data or unaligned (raw) colored by assigned cell type from the original data for one simulation scenario (Six cell types out of 9 shared and medium batch effect). The 1:1, 1:2, and 1:5 ratios denote the ratios of numbers of cells in the two batches to represent scenarios with equal number of cells, moderately unequal number of cells, and very unbalanced number of cells, respectively. $CD4^+$ T cells and Cytotoxic T cells show separation in `Dmatch`-, fastMNN-, and Harmony-aligned data. However, Cytotoxic T cells are largely overlapping in scMerge, and Seurat V3-aligned data indicating over-correction.



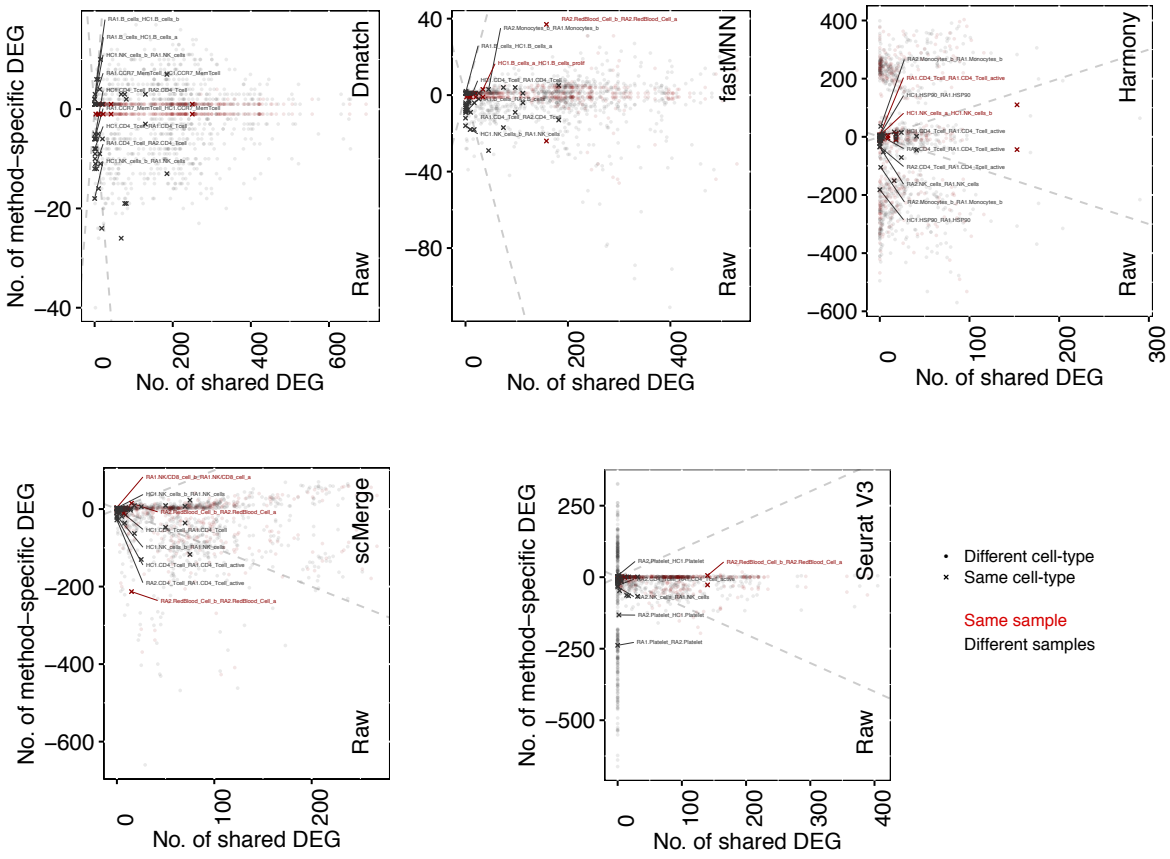
Supplemental Fig. S6: UMAP of aligned simulated data or unaligned (raw) colored by assigned cell type from the original data for one simulation scenario (Six cell types out of 9 shared and medium batch effect). The 1:1, 1:2, and 1:5 ratios denote the ratios of numbers of cells in the two batches to represent scenarios with equal number of cells, moderately unequal number of cells, and very unbalanced number of cells, respectively. Large reduction of megakaryocytes heterogeneity after correction using Seurat V3, indicating over-correction.



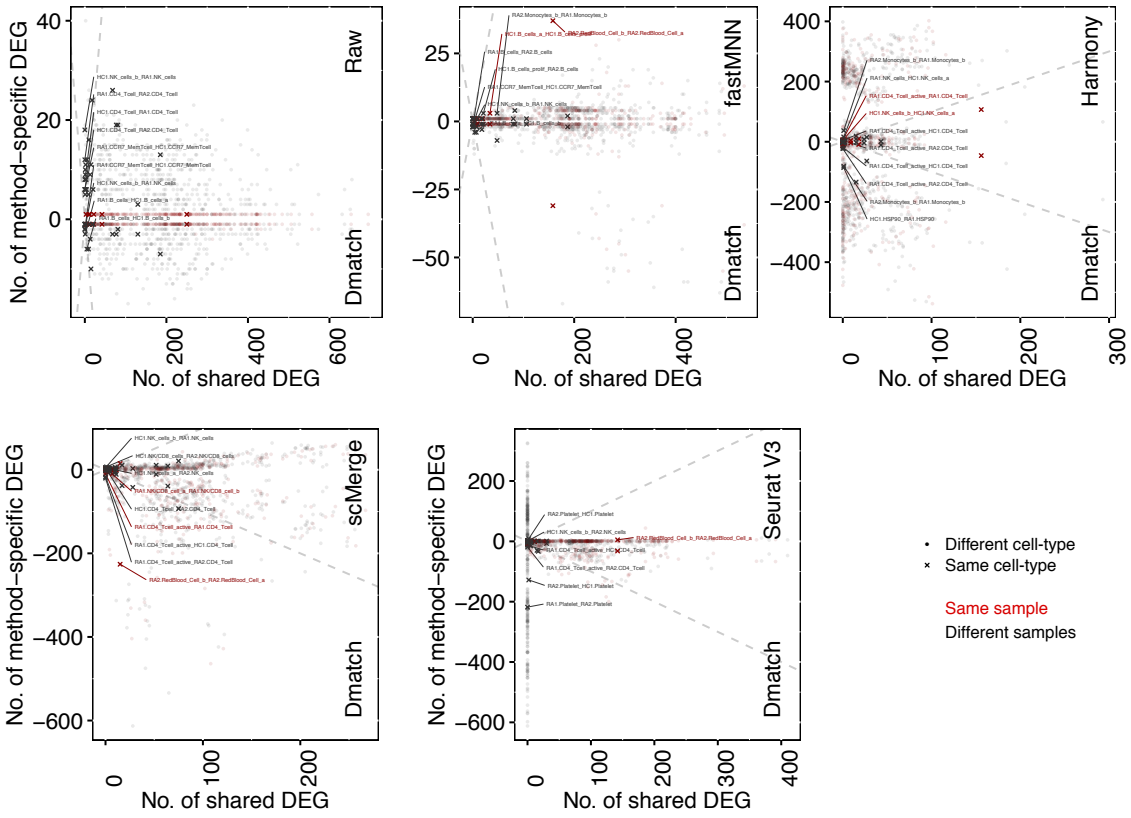
Supplemental Fig. S7: **(A)** UMAP dimensionality reduction of PBMC scRNA-seq data from 6 clinical samples before and after alignment comparing `Dmatch`, fastMNN, Harmony, scMerge, and Seurat V3. **(B)** Monocytes, T cells and B cells are clearly separated in the UMAP clustering for all methods. Shades of purple represent marker expression in each cell relative to other cells in the same UMAP (arbitrary unit).



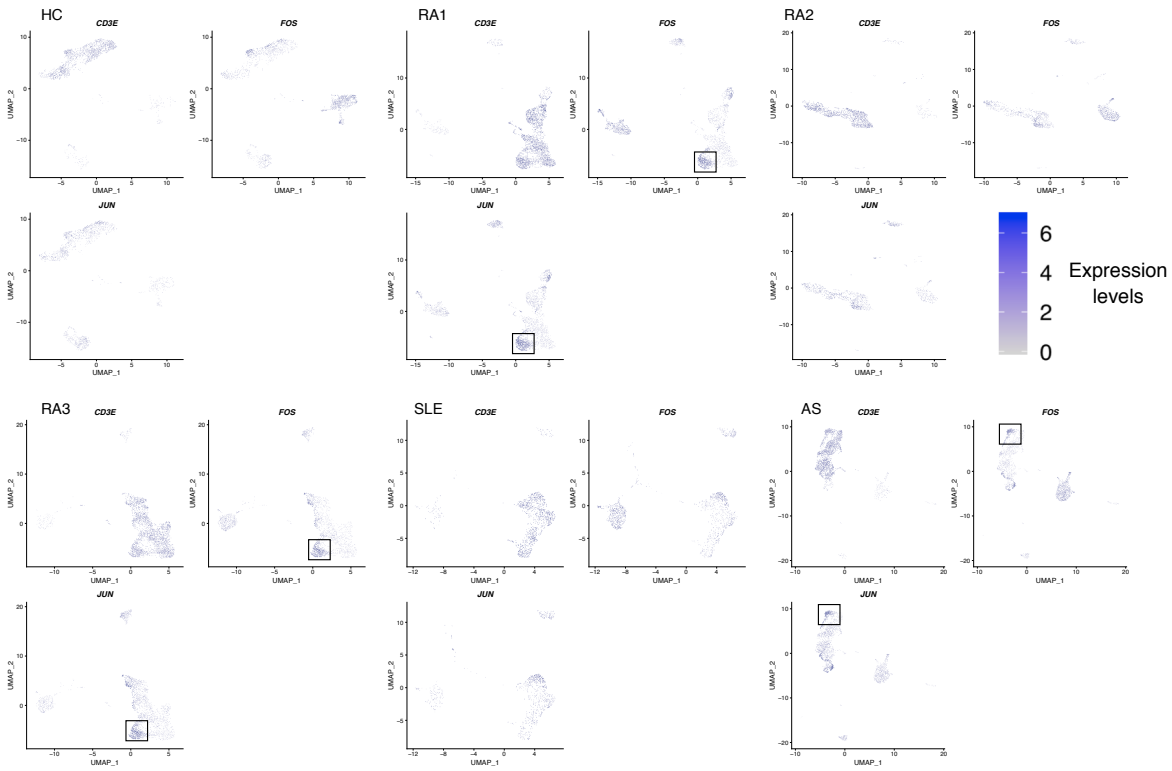
Supplemental Fig. S8: **(A)** UMAP dimensionality reduction of PBMC scRNA-seq data from 6 clinical samples before and after alignment comparing `Dmatch`, fastMNN, Harmony, scMerge, and Seurat V3. **(B)** The marker *ECHS1* for the "kidney" cells cluster tags the SLE-specific cluster only in data aligned using `Dmatch`, fastMNN, and to some extent scMerge. Shades of purple represent marker expression in each cell relative to other cells in the same UMAP.



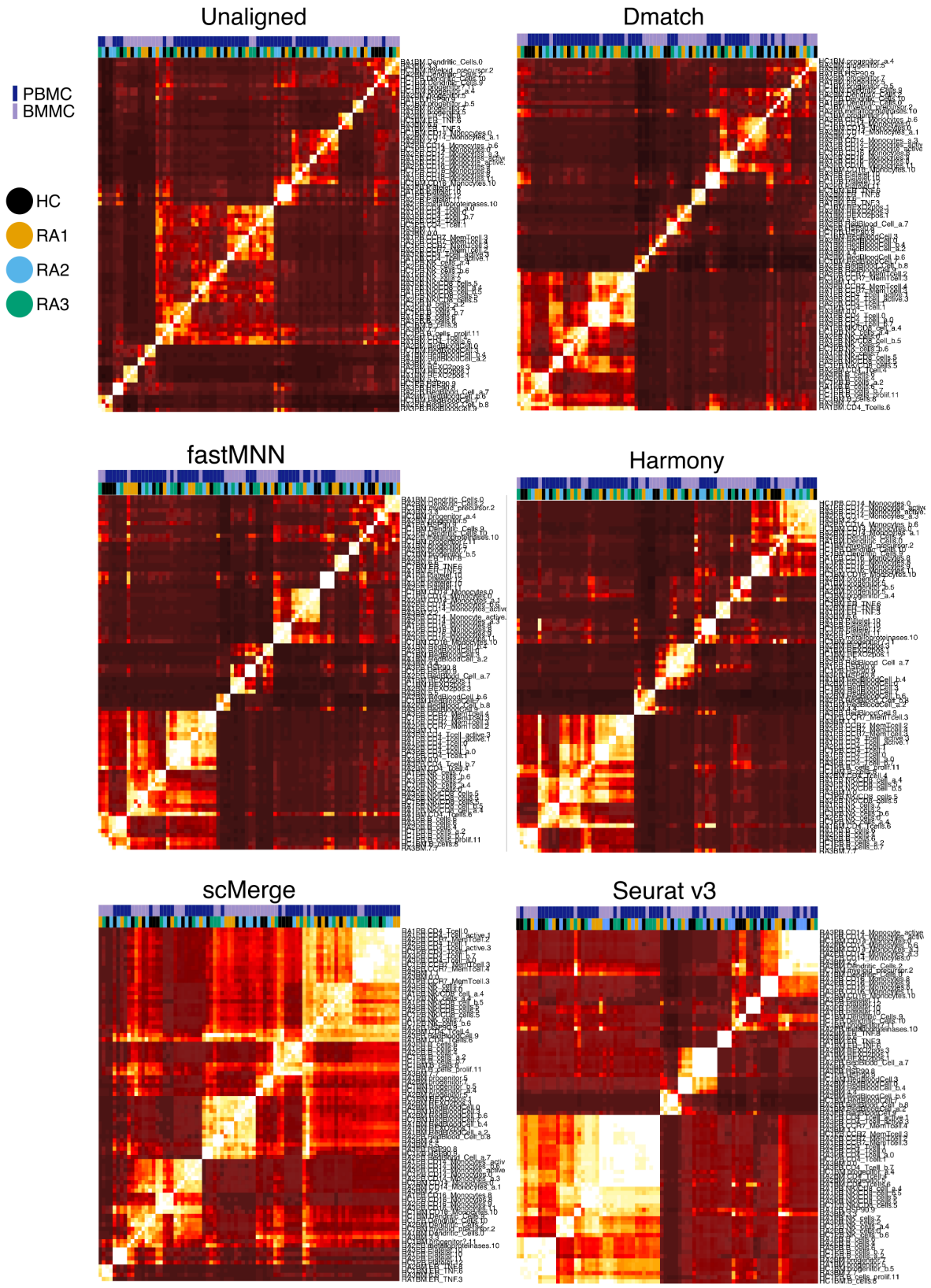
Supplemental Fig. S9: Scatter plot showing the number of differentially expressed genes (DEG) that are identified using unaligned data (positive y-axis) or using data aligned using different methods (negative y-axis), versus the number of DEG that are identified in both datasets (x-axis). Each point represents a comparison between cell type clusters from the same or different cell type, or from the same or different sample. Successful removal of batch effect is supported by smaller numbers of DEGs resulting from aligned data in same cell type comparisons, compared to unaligned data (smaller positive y-axis than negative y-axis for "x"'s). However, for most methods other than `Dmatch`, there were considerable variation in DEG between unaligned and aligned data for comparisons across clusters from the same sample (red "x"'s or points).



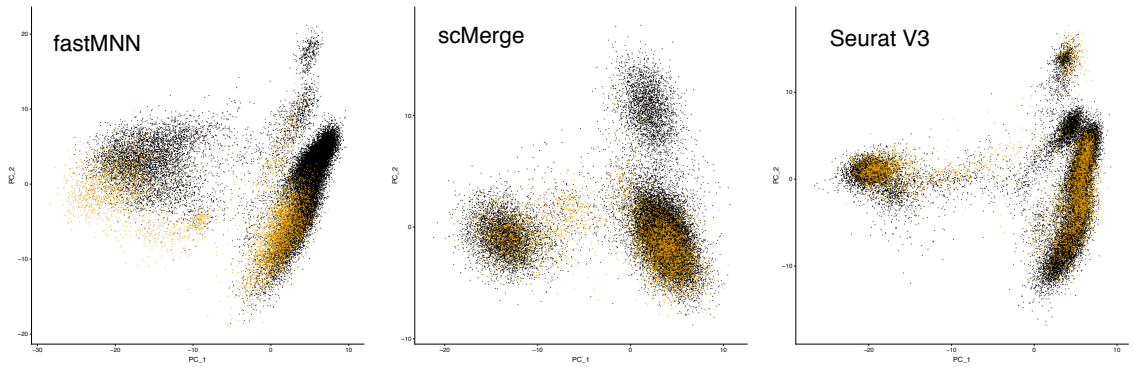
Supplemental Fig. S10: Scatter plot showing the number of differentially expressed genes (DEG) that are identified using `Dmatch`-aligned data (negative y-axis) or using data aligned using different methods (positive y-axis), versus the number of DEG that are identified in both datasets (x-axis). Each point represents a comparison between cell type clusters from the same or different cell type, or from the same or different sample. We observed substantial differences between DEGs identified after alignment from `Dmatch` and other alignment methods.



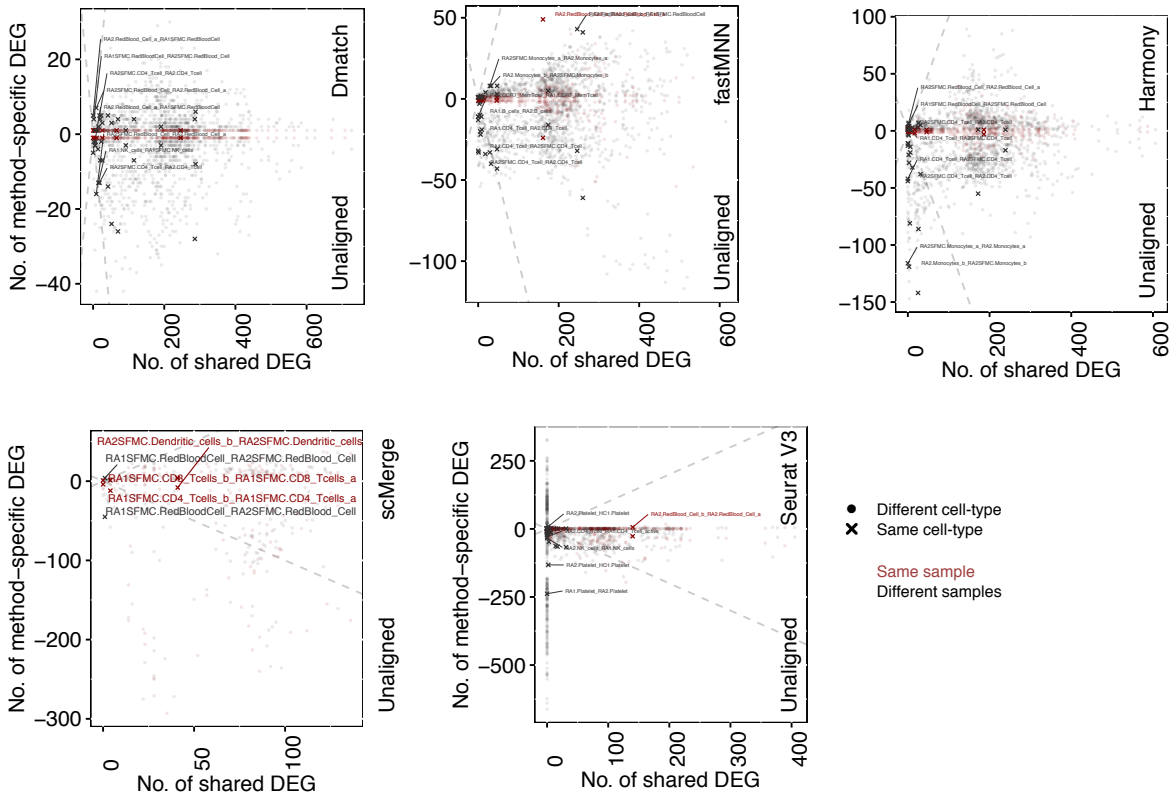
Supplemental Fig. S11: UMAP of dimensionality reduction of PBMC scRNA-seq data from 6 PBMC samples, separately. We observed *JUN* and *FOS* in a subcluster of T-cells in PBMC of patients with autoimmune disease (RA1, RA3, AS) whose monocytes also expressed *IL1B*. Shades of purple represent marker expression in each cell relative to other cells in the same UMAP.



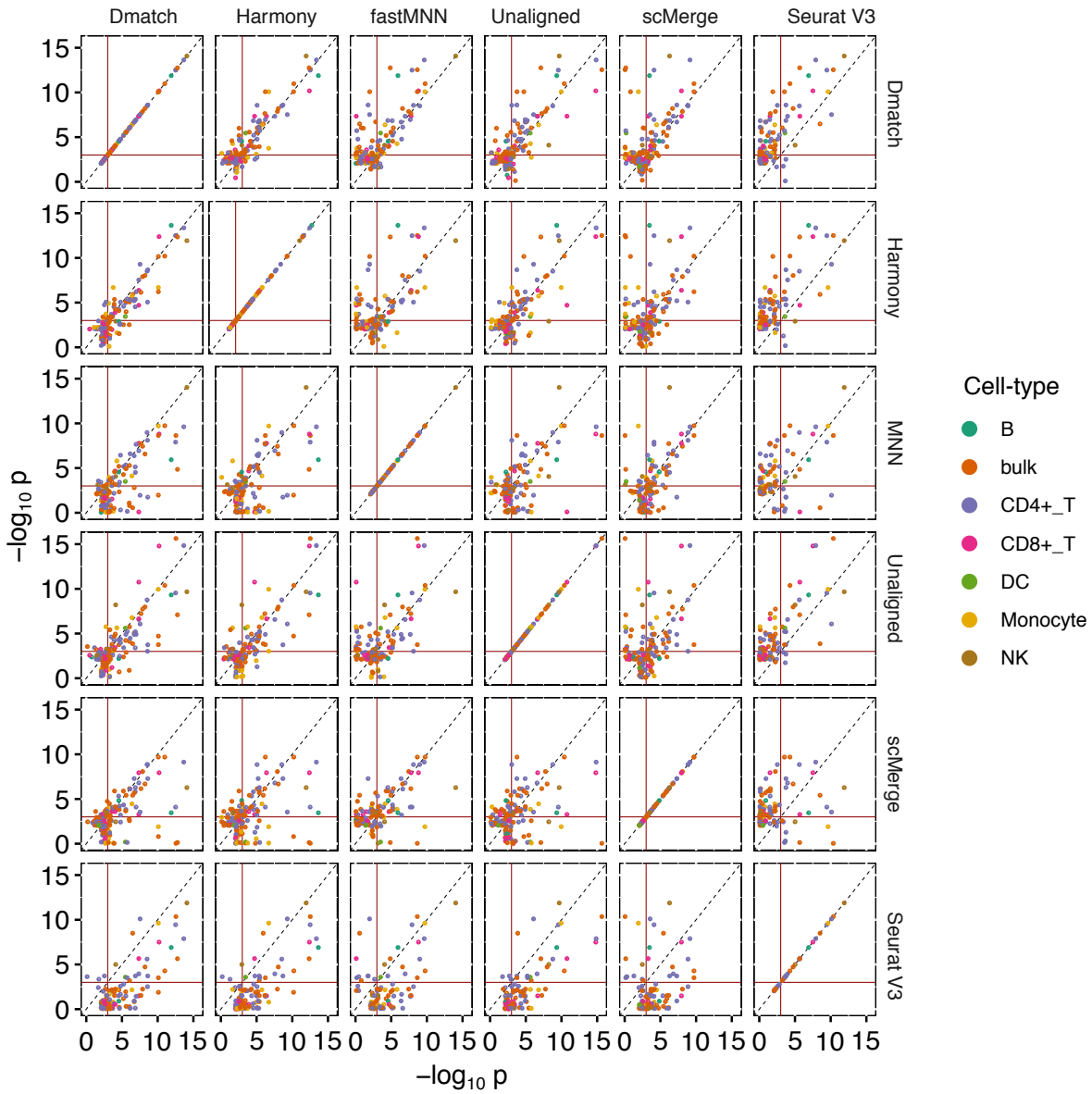
Supplemental Fig. S12: Heatmaps showing the number of DEGs identified across Seurat clusters in PBMC and BMMC samples using aligned data from different methods.



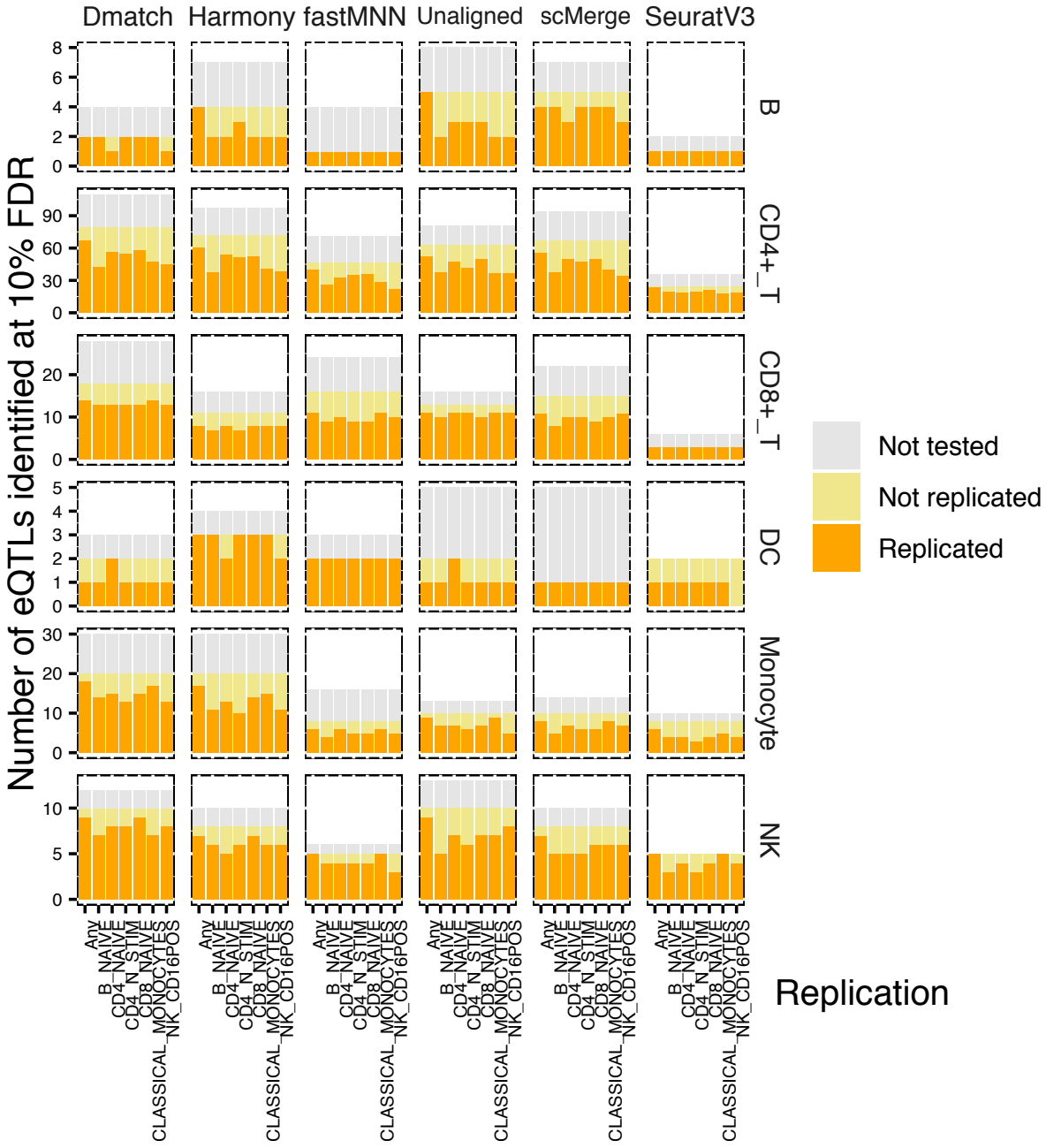
Supplemental Fig. S13: PCA of scRNA-seq samples from PBMC and SF after alignment using different methods. Batch effect is reduced after alignment using all methods.



Supplemental Fig. S14: Scatter plot showing the number of differentially expressed genes (DEG) that are identified using unaligned data (negative y-axis) or using data aligned using different methods (positive y-axis), versus the number of DEG that are identified in both datasets (x-axis). Each point represents a comparison between cell type clusters from the same or different cell type, or from the same or different sample. Successful removal of batch effect is supported by smaller numbers of DEGs resulting from aligned data in same cell type comparisons, compared to unaligned data (smaller positive y-axis than negative y-axis for "x"'s).



Supplemental Fig. S15: Scatter plot showing the linear regression p-values comparing eQTLs (best gene-SNP pair) identified in different cell types (colors) using unaligned or aligned data from different methods. While the p-values were generally highly correlated, in many cases they were more significant for `Dmatch`-aligned data compared to data aligned using other methods.



Supplemental Fig. S16: Replication of eQTLs identified in a bulk RNA-seq study of immune cell type eQTLs from the DICE consortium. The eQTLs that were not tested correspond to SNPs whose genotype could not be determined or to genes filtered out due to low expression.