**Supplementary Data for**

# CyTOFmerge: Integrating mass cytometry data across multiple panels

Tamim Abdelaal[1,2], Thomas Höllt[2,3], Vincent van Unen[4], Boudewijn P.F. Lelieveldt[1,2,5], Frits Koning[4], Marcel J.T. Reinders[1,2], Ahmed Mahfouz[1,2,*]

[1]Delft Bioinformatics Lab, Delft University of Technology, 2628 XE Delft, The Netherlands.

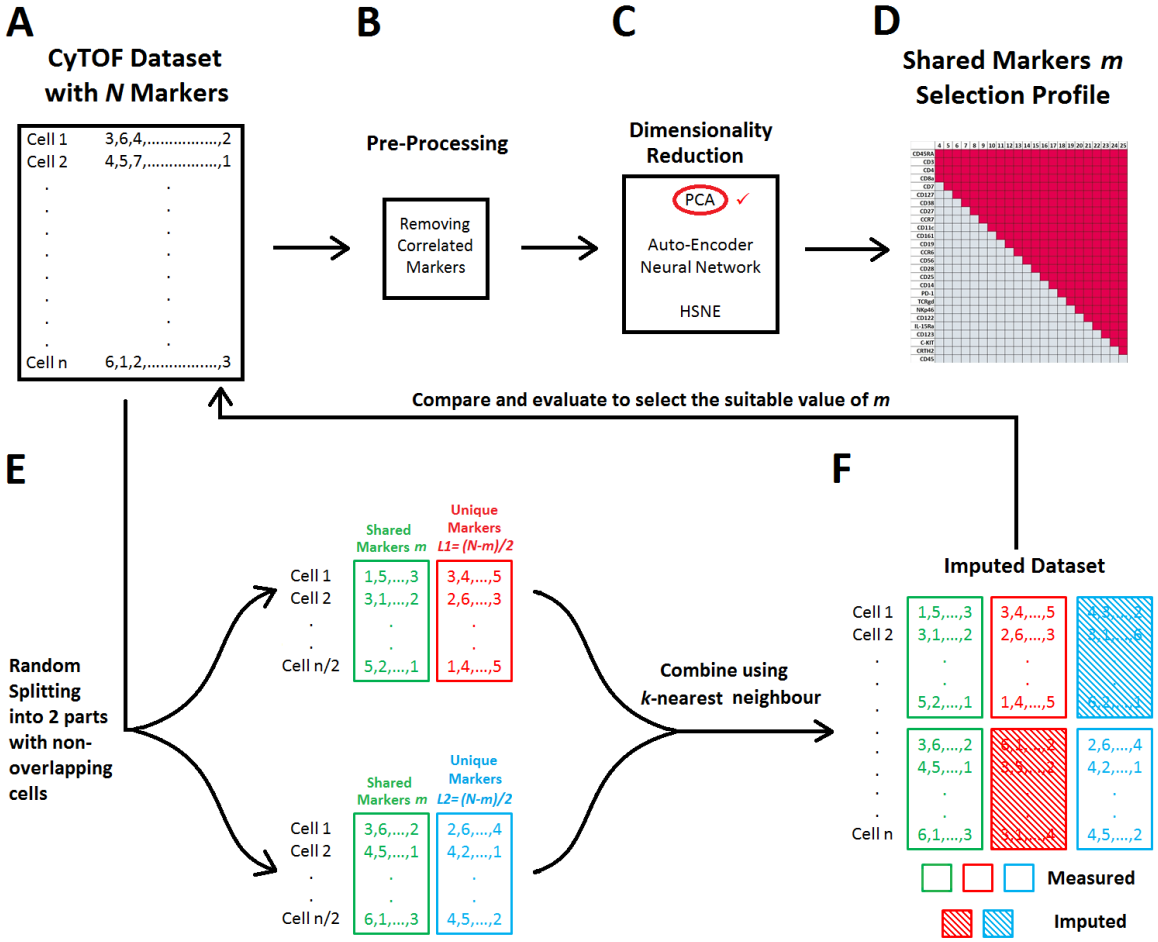
[2]Leiden Computational Biology Center, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands.

[3]Computer Graphics and Visualization Group, Delft University of Technology, 2628 XE Delft, The Netherlands.

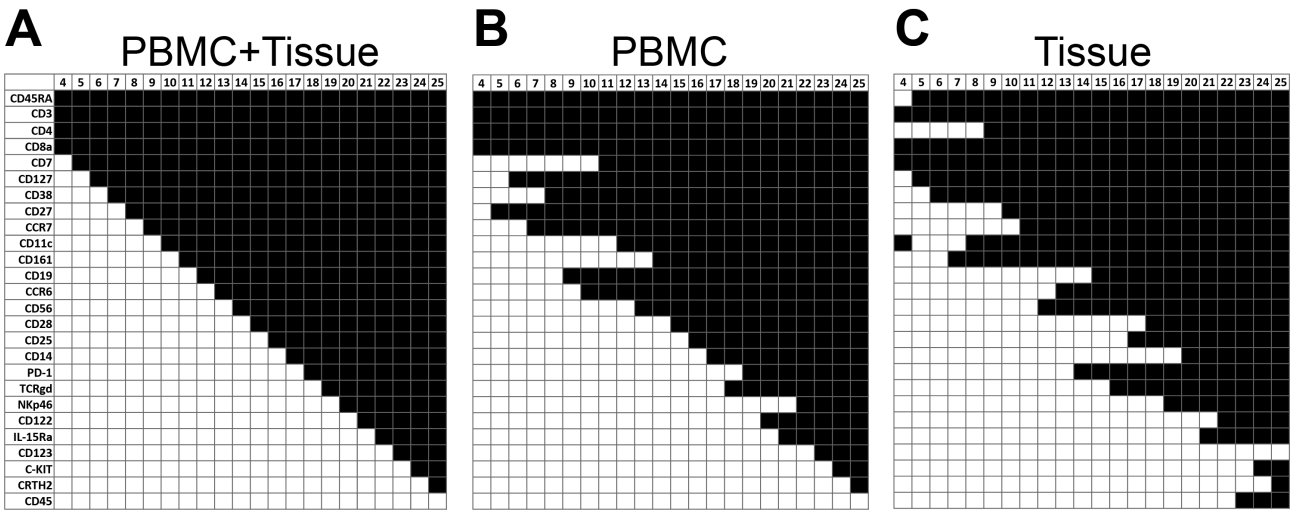
[4]Department of Immunohematology and Blood Transfusion, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands.

[5]Department of Radiology, Leiden University Medical Center, 2333 ZA Leiden, The Netherlands.

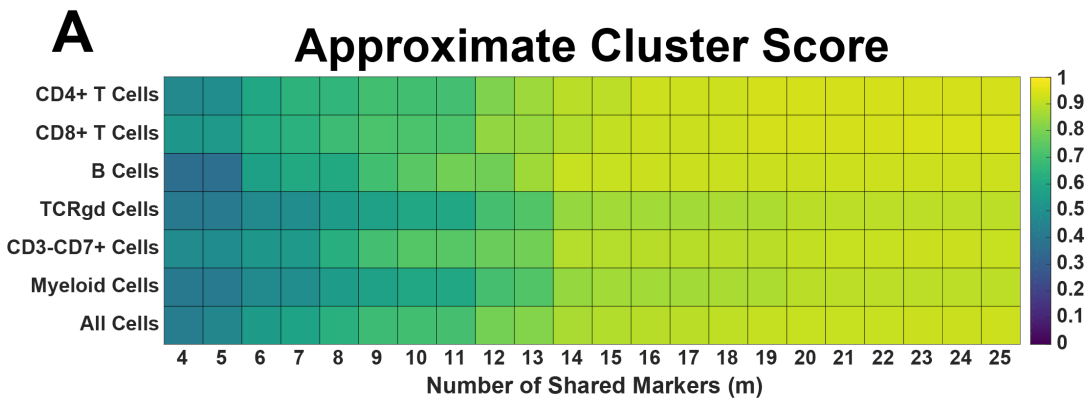
Correspondence to: Dr. Ahmed Mahfouz (a.mahfouz@lumc.nl)

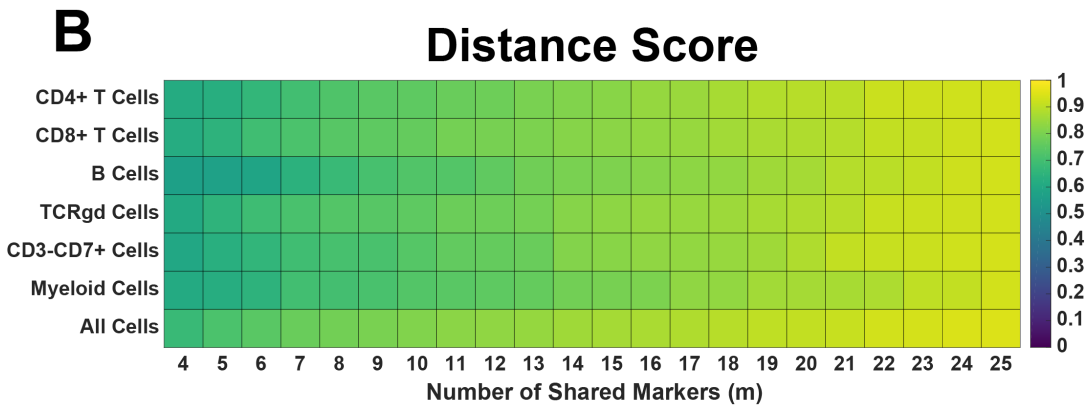


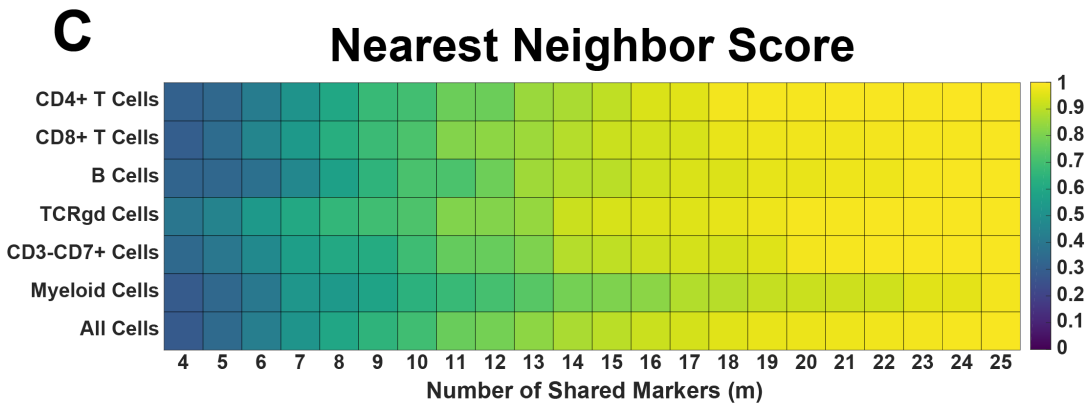

**Supplementary Fig. S1 Selection pipeline for shared markers (A)** CyTOF dataset measured by a single panel, **(B)** calculate markers pairwise correlation and remove highly correlated markers, **(C)** apply dimensionality reduction to obtain the selection profile **(D)** for a wide range of *m* (red is selected, grey is not selected). **(E)** Evaluation of all values of *m* by randomly splitting the data, recombine it, and then comparing the imputed data **(F)** with the original data to select the minimal *m* with accepted performance.

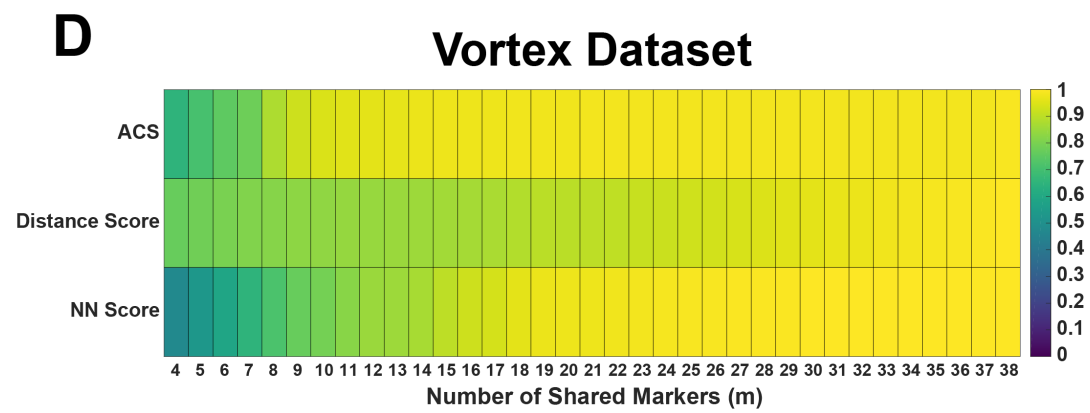



**Supplementary Fig. S2 Selected shared markers for PBMC and tissue:** The PCA-based selected markers using **(A)** all samples (PBMC+Tissue), **(B)** using only PBMC samples and **(C)** using only tissue samples. (Marker ordering is based on the PCA selection profile, black is selected, white is not selected)

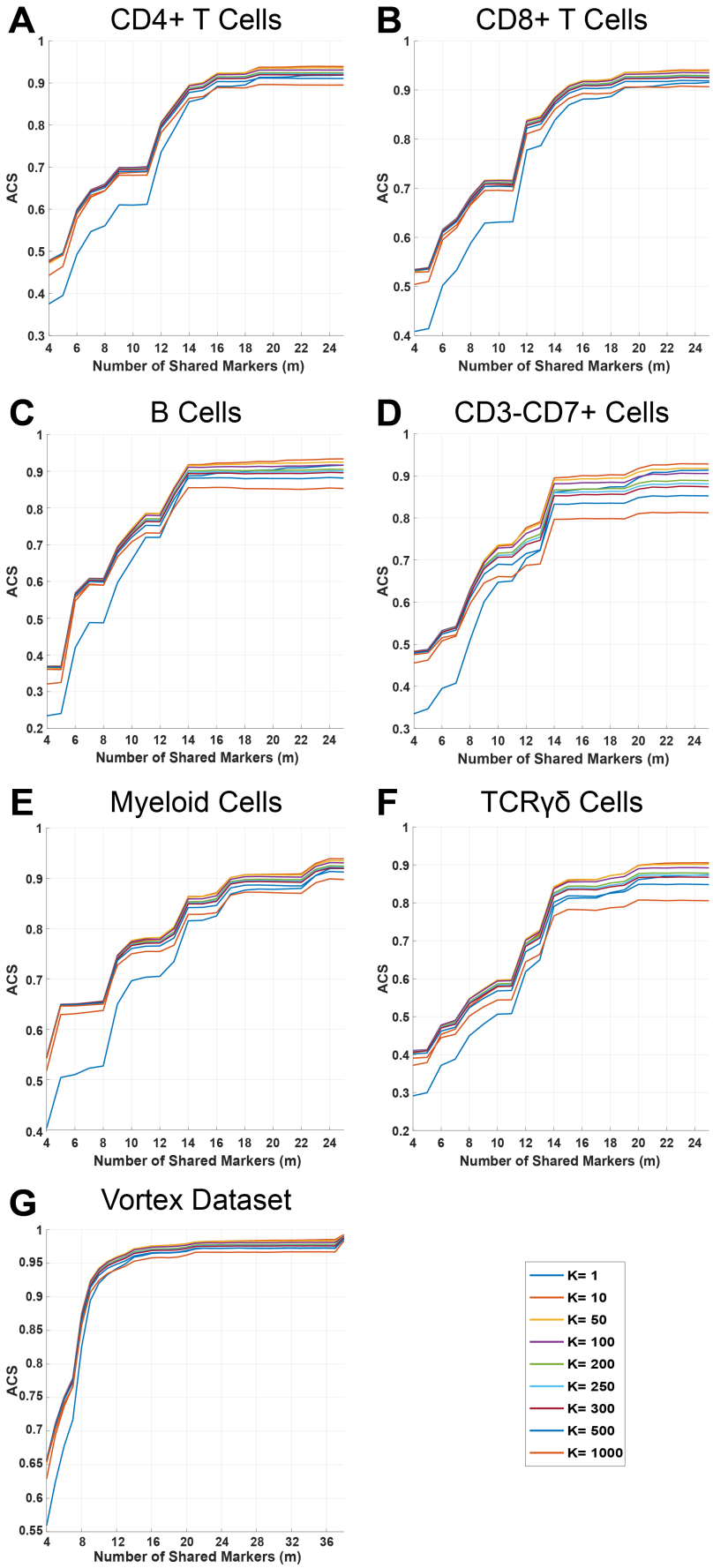


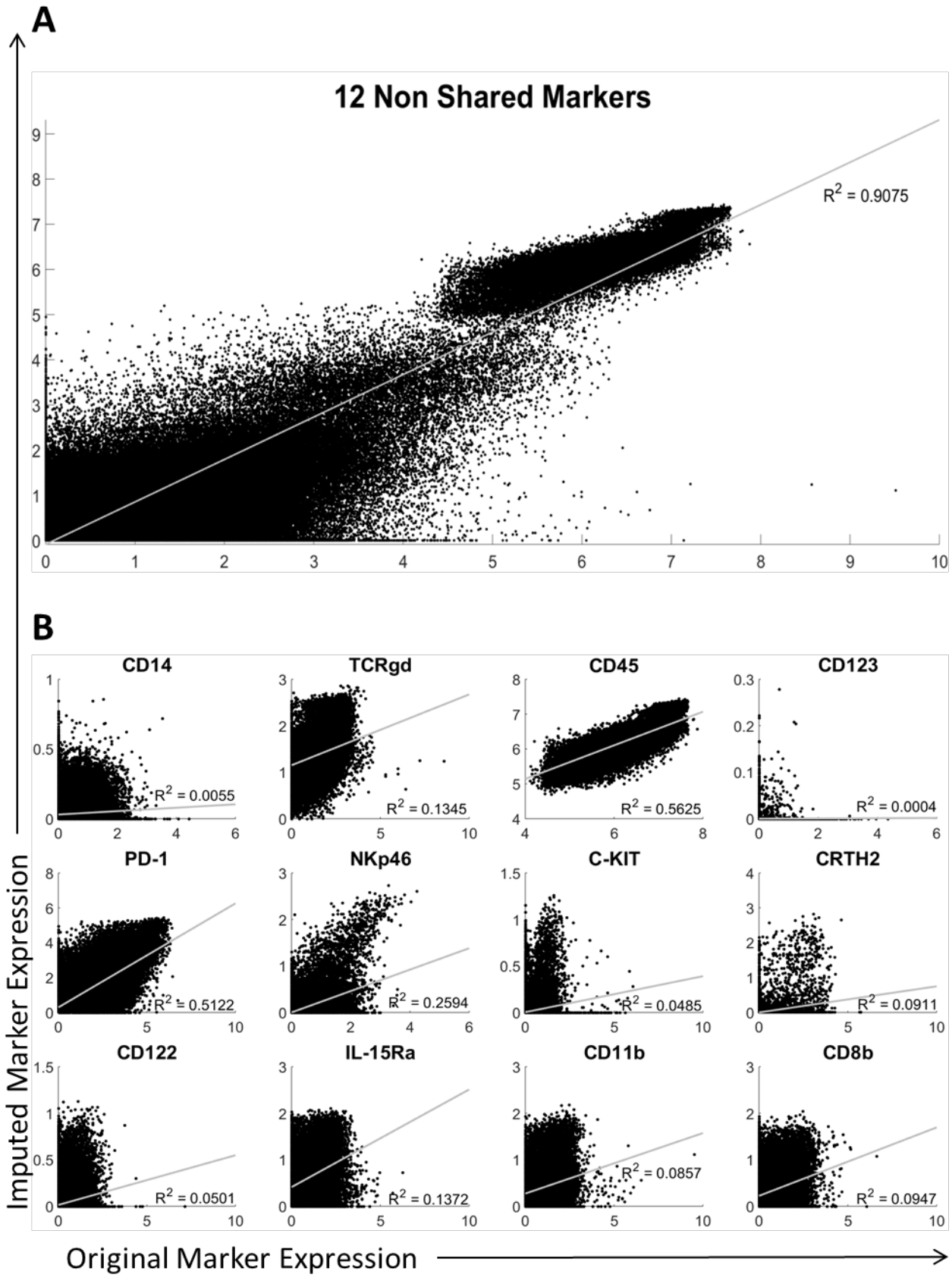



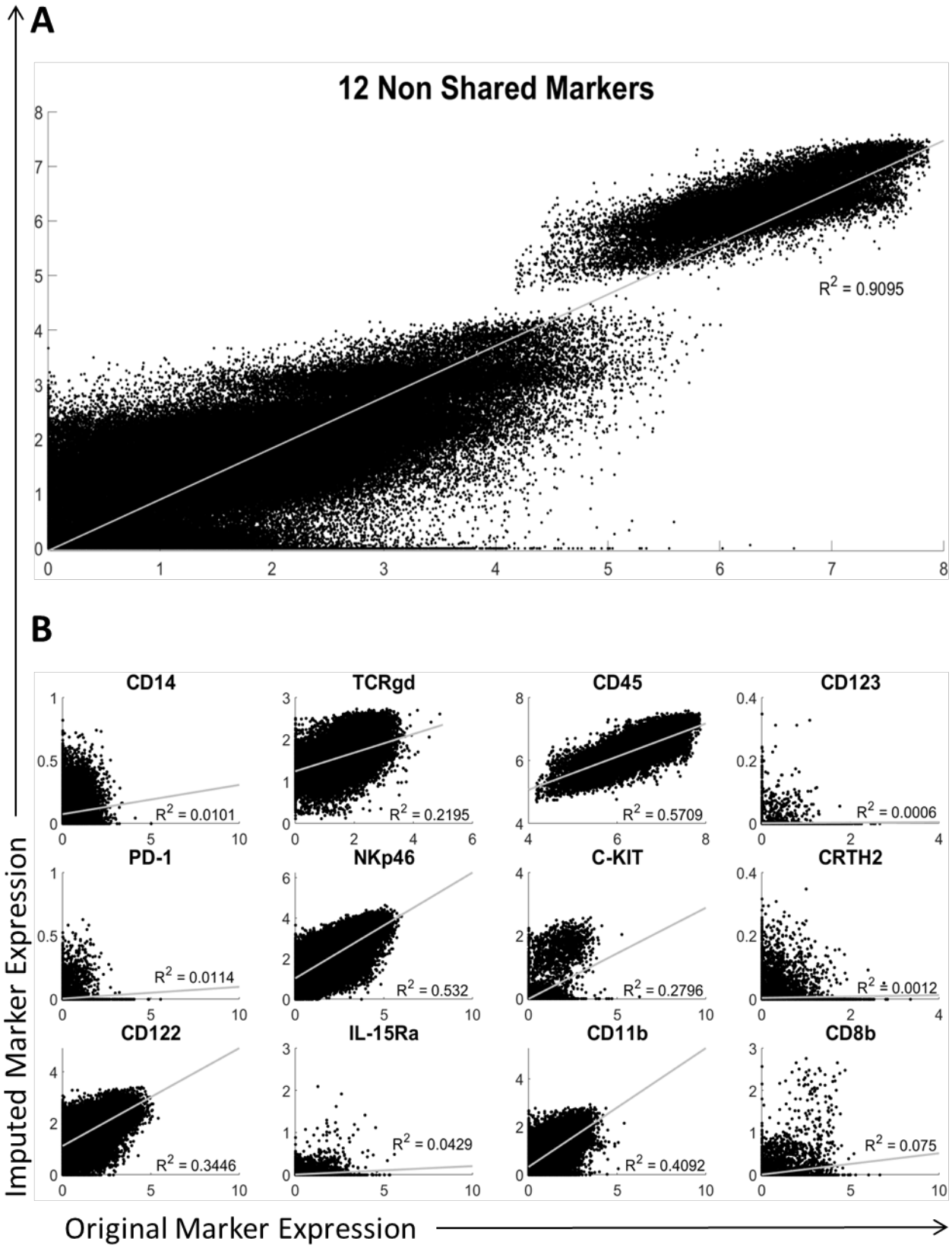

**Supplementary Fig. S3 Evaluation scores for imputed CyTOF data (A-C)** Evaluation scores for the HMIS dataset as a function of shared markers for different lineages: **(A)** Approximate cluster score. **(B)** Distance score. **(C)** Nearest Neighbor score. These performance scores are calculated per lineage and for one dataset having all six lineages together (last row). **(D)** Evaluation scores for the Vortex dataset as a function of shared markers.



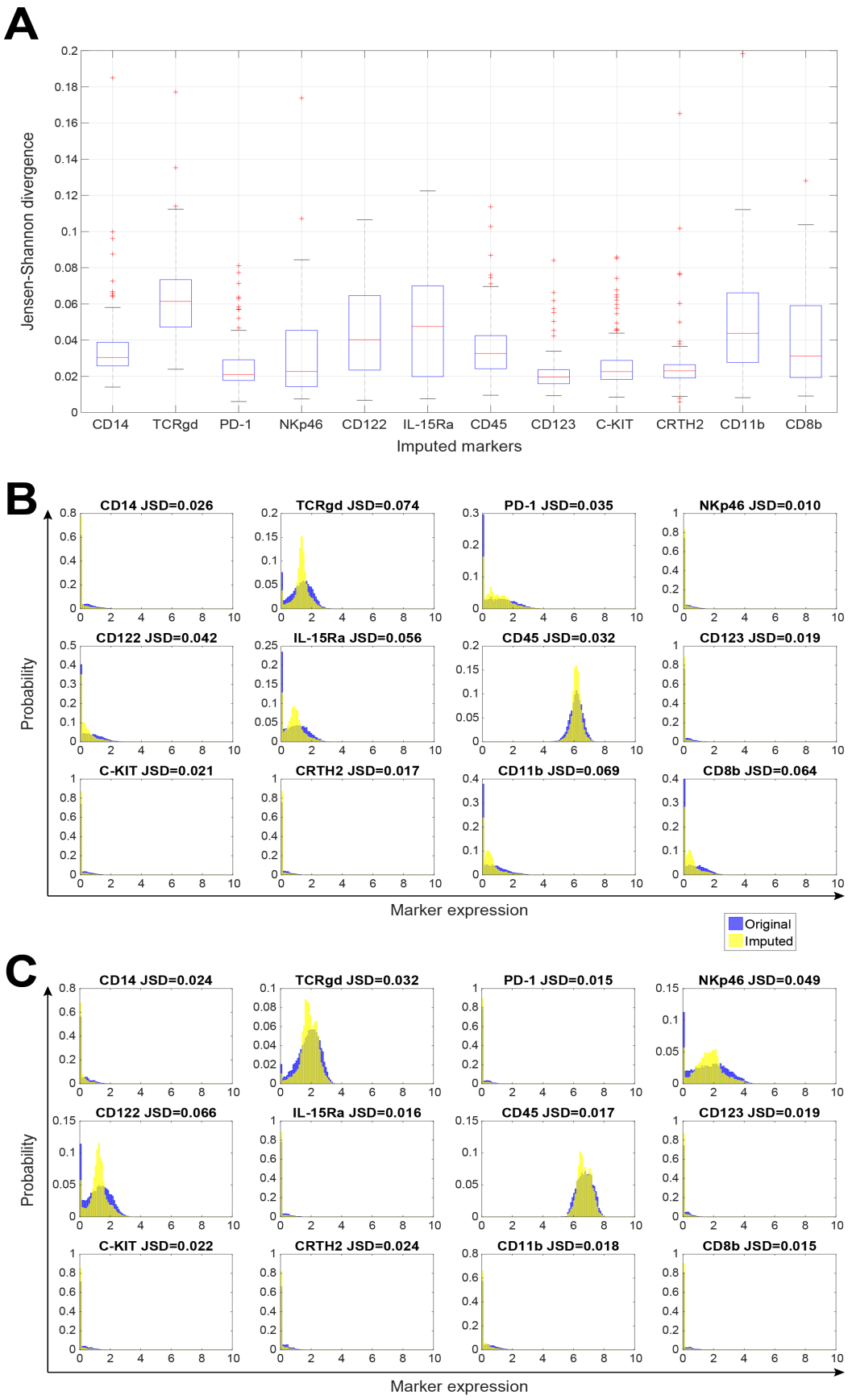

**Supplementary Fig. S4** ***Approximate Cluster Score*** **vs the number of shared markers (*m*)**, combination was performed using the k-nearest neighbor algorithm for different values of *k* represented by the separate lines in each plot.



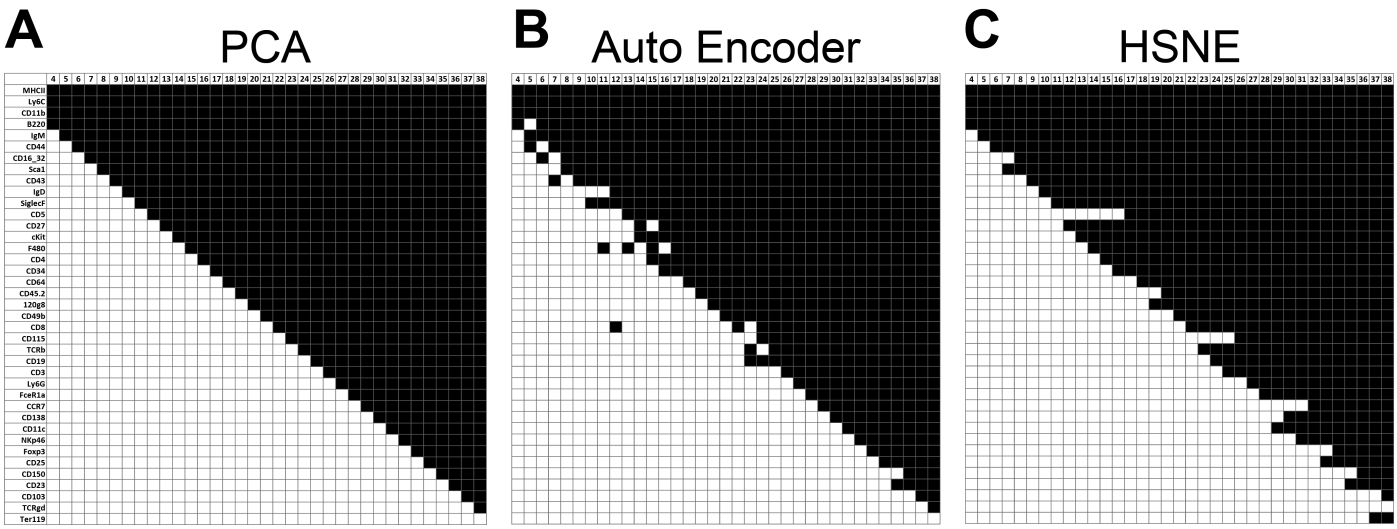

**Supplementary Fig. S5** Scatter plots showing the correlation between the original and the imputed expression values for the 12 non-shared markers of both panels A & B for the CD4+ T cells lineage: **(A)** all non-shared markers concatenated in one vector, showing a global high correlation, **(B)** separate scatter plots per marker, as shown within a specific lineage most of the markers are not expressed (≈ 0) resulting in a low correlation with the imputed values.



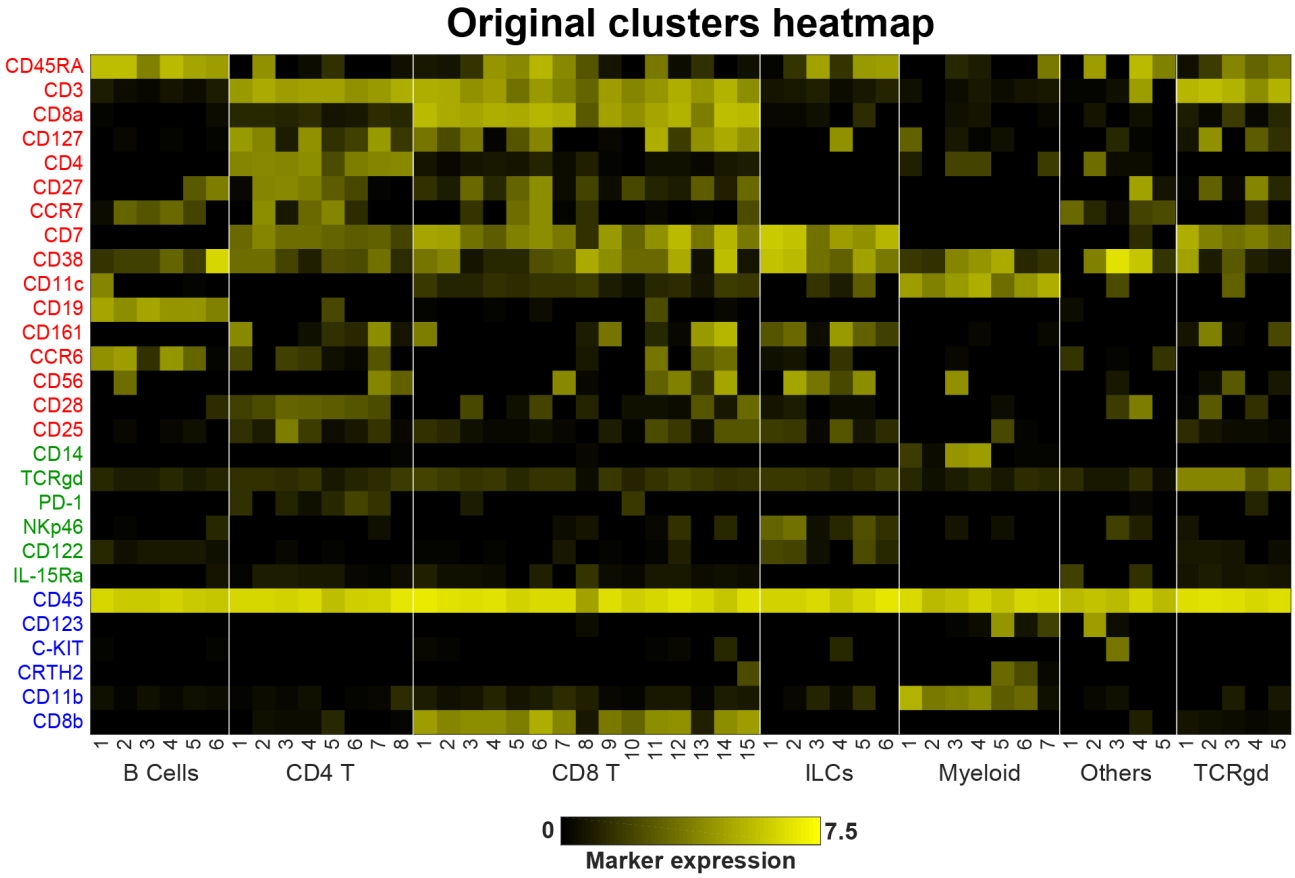

**Supplementary Fig. S6** Scatter plots showing the correlation between the original and the imputed expression values for the 12 non-shared markers of both panels A & B for the CD3-CD7+ cells lineage: **(A)** all non-shared markers concatenated in one vector, showing a global high correlation, **(B)** separate scatter plots per marker, as shown within a specific lineage most of the markers are not expressed (≈ 0) resulting in a low correlation with the imputed values.



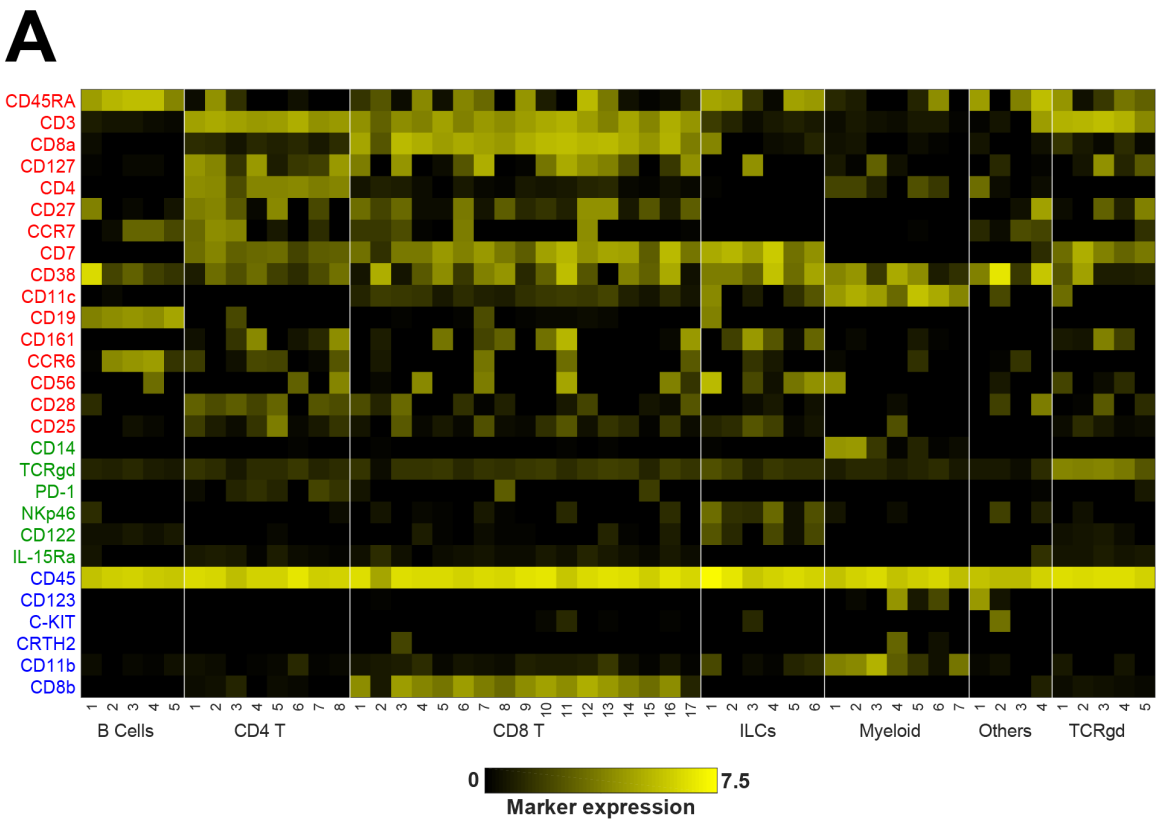

**Supplementary Fig. S7 (A)** Box plots showing the Jensen-Shannon divergence (JSD) values for each of the 12 imputed markers across all 121 cell populations in the HMIS dataset. The JSD value measures the similarity between the original and the imputed distribution of one marker within one population. **(B-C)** Histograms showing the original and the imputed distributions of the 12 imputed markers for **(B)** population CD4+ T cells 01, and **(C)** population CD3-CD7+ cells 01. For each marker, the JSD value is indicated.

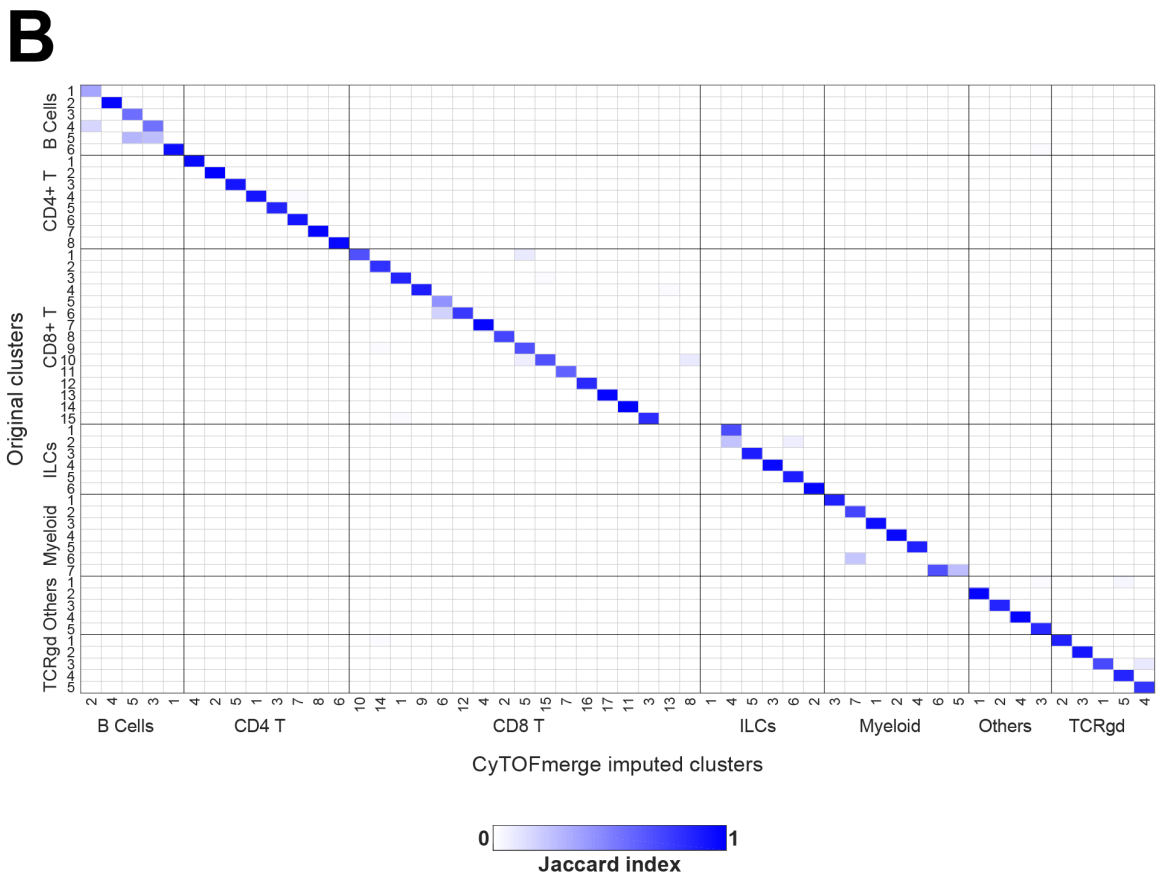



**Supplementary Fig. S8 Selected shared markers for the Vortex dataset.** The selected markers that can best represent the dataset using **(A)** PCA, **(B)** Auto Encoder and **(C)** HSNE. (Marker ordering is based on the PCA selection profile, black is selected, white is not selected). No markers are removed during preprocessing.

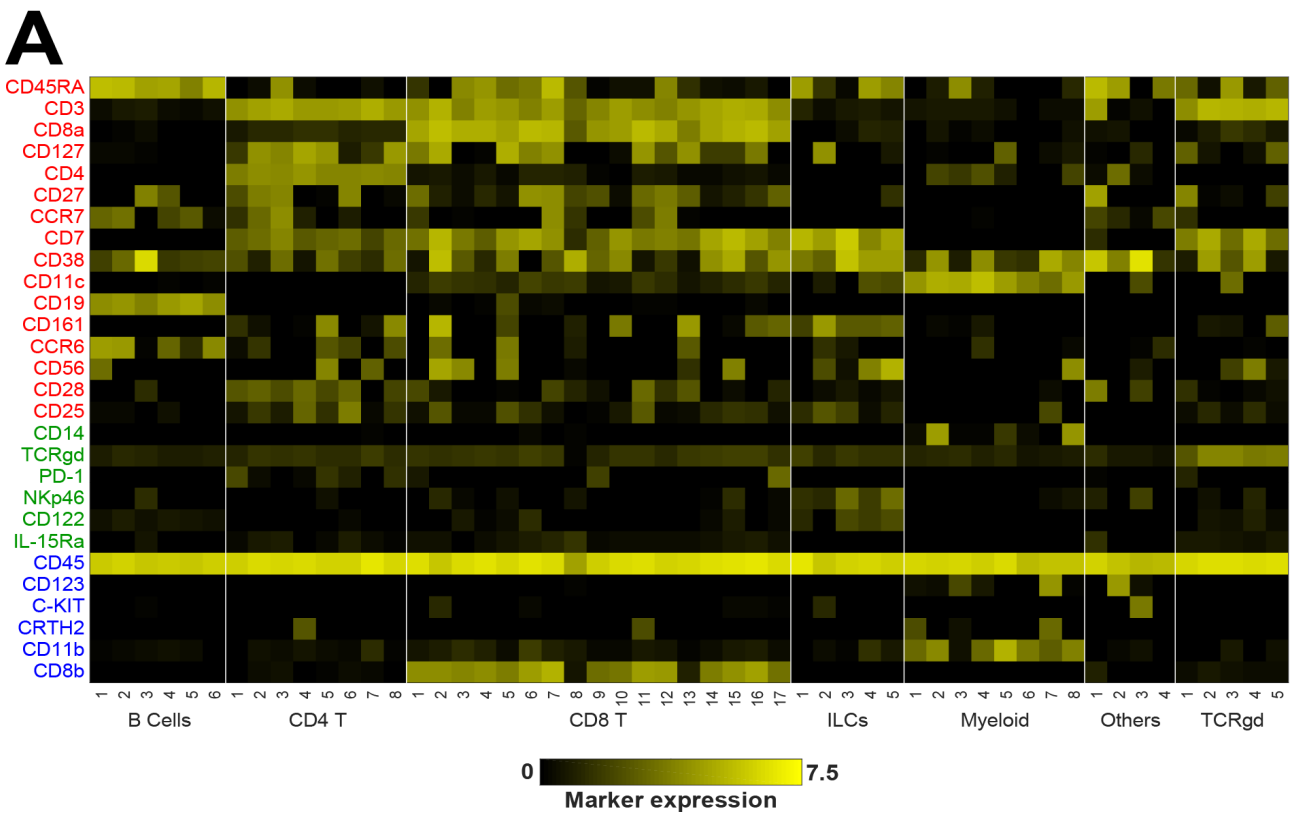


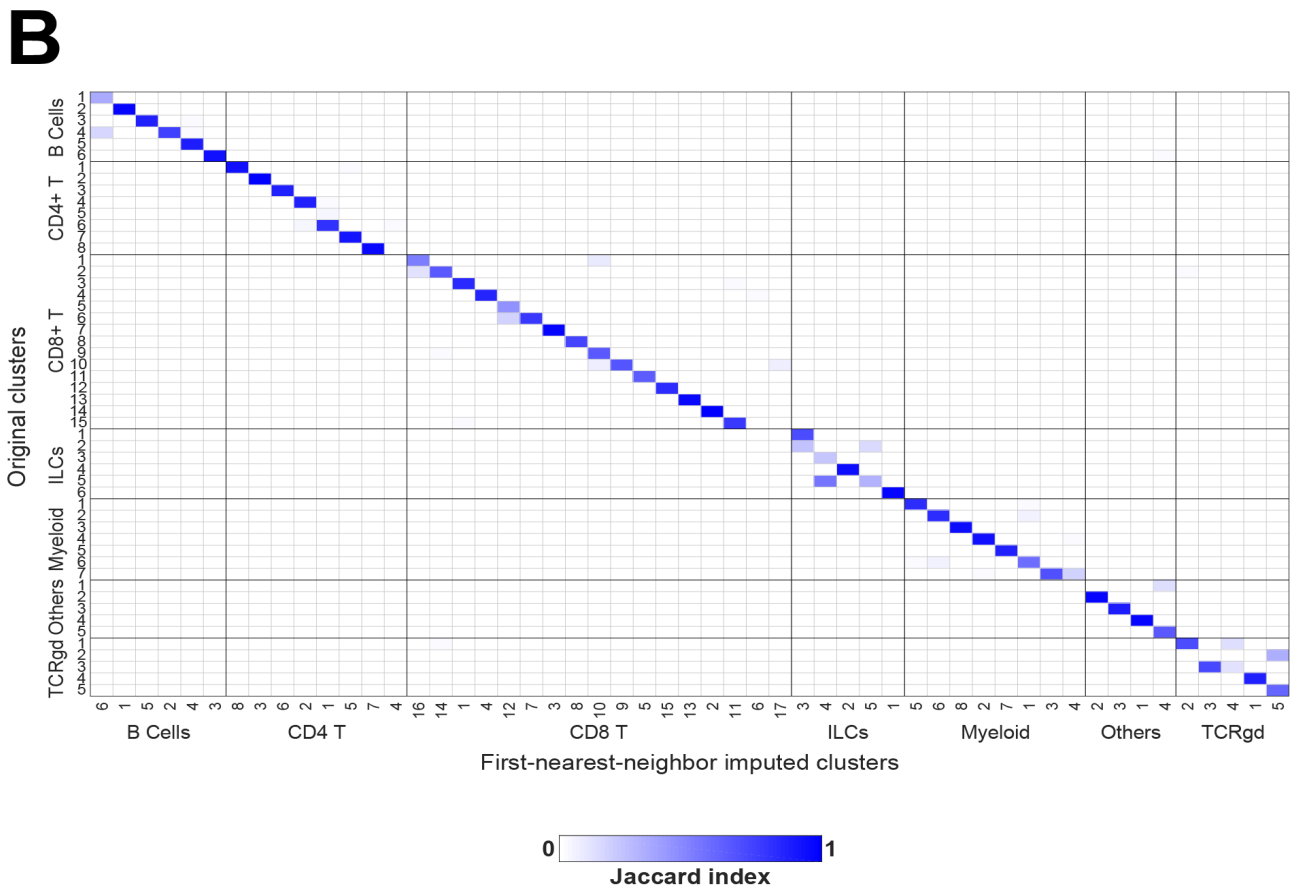

**Supplementary Fig. S9 HMIS original dataset clusters.** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 52 cell clusters obtained by clustering the original HMIS data using Phenograph. Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B).



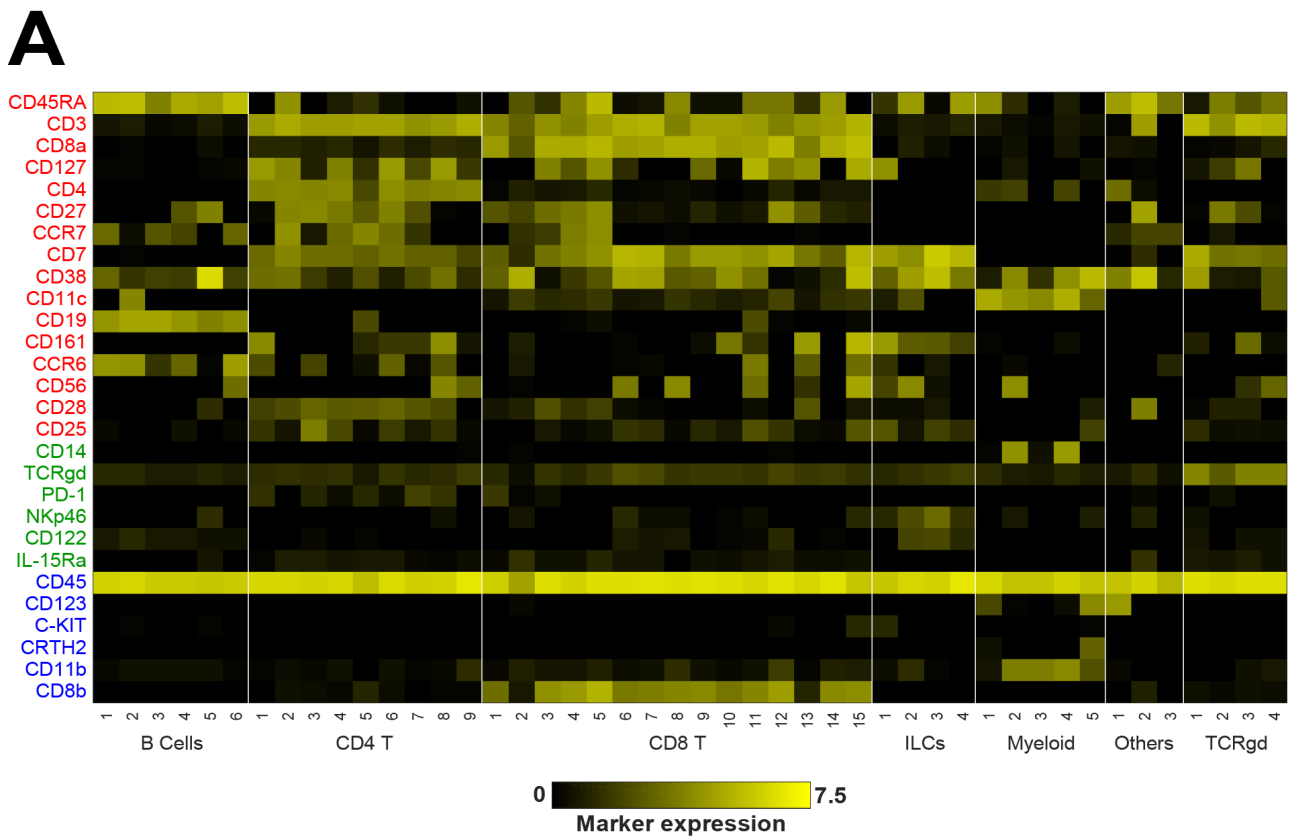



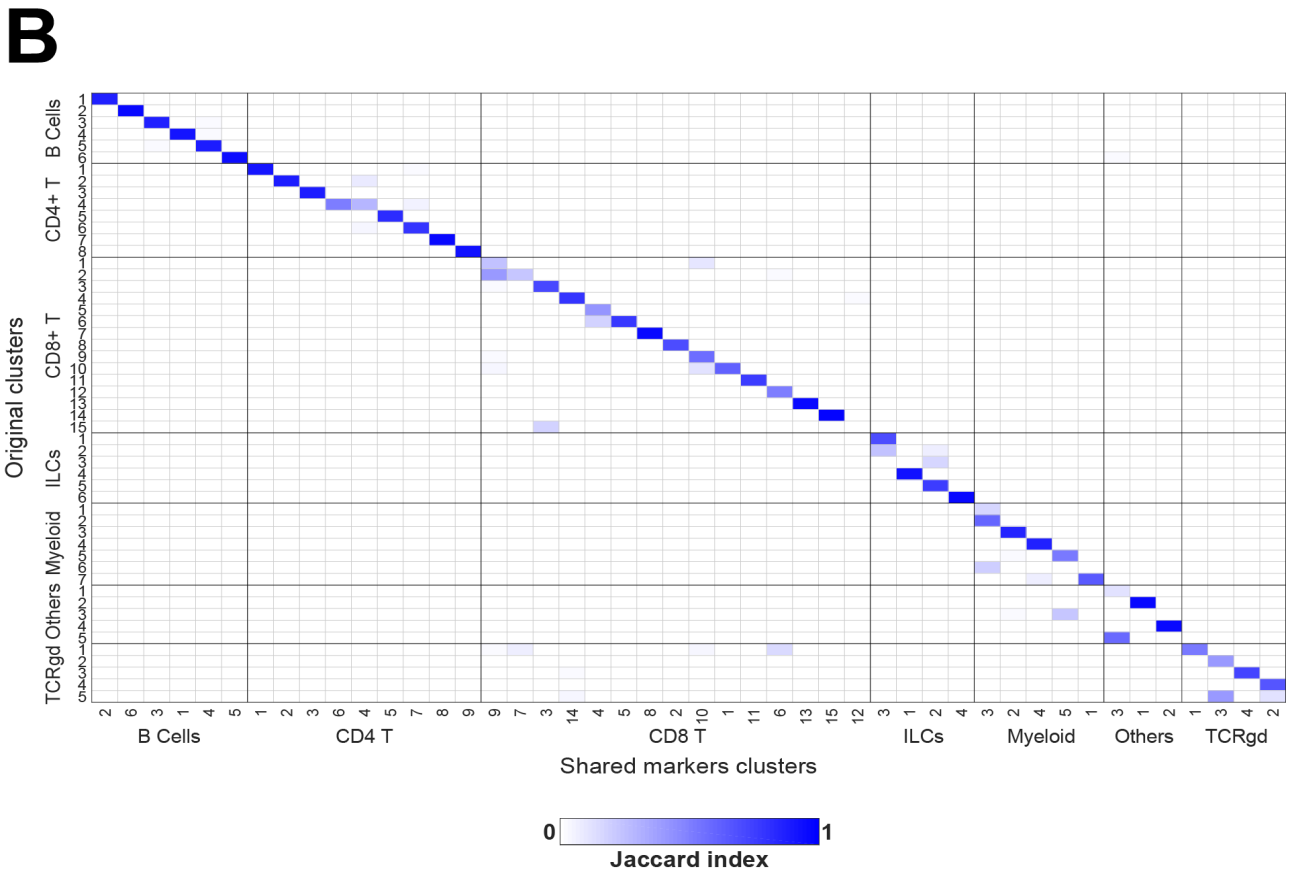

**Supplementary Fig. S10 HMIS imputed data by CyTOFmerge. (A)** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 52 cell clusters obtained by clustering the imputed HMIS data using Phenograph (with $m$ = 16 and $k$ = 50). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset.



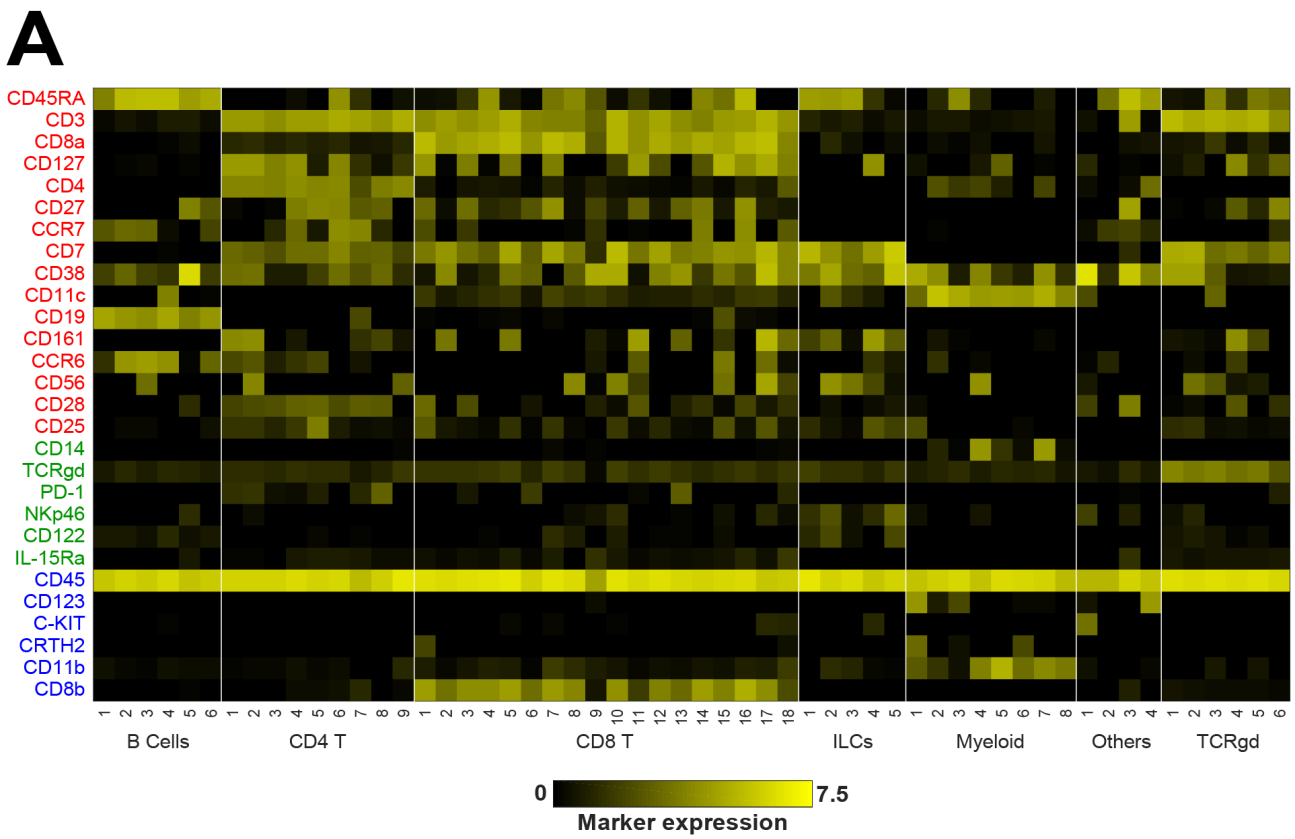



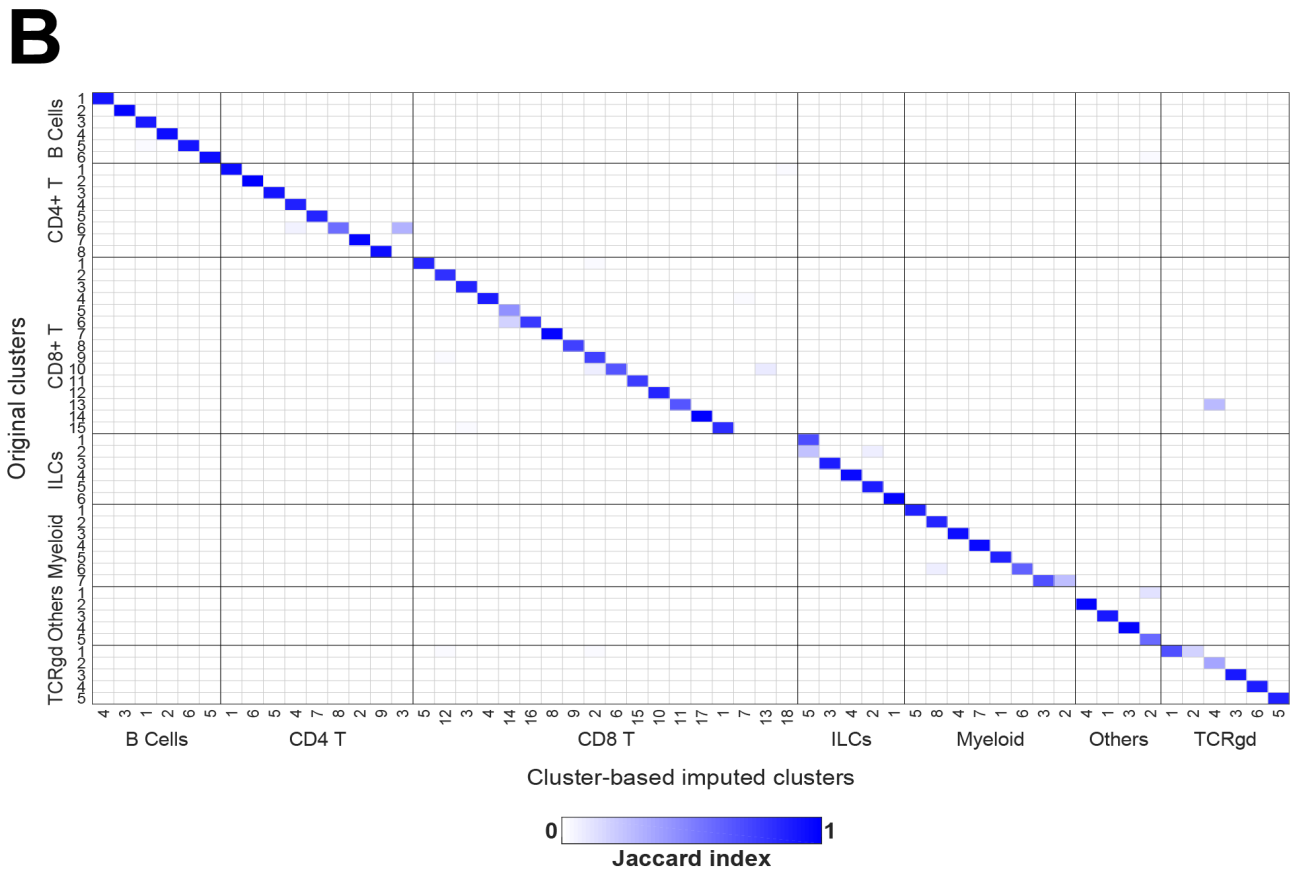

**Supplementary Fig. S11 HMIS imputed data by first-nearest-neighbor. (A)** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 53 cell clusters obtained by clustering the imputed HMIS data using Phenograph (with $m = 16$ and $k = 1$). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset.



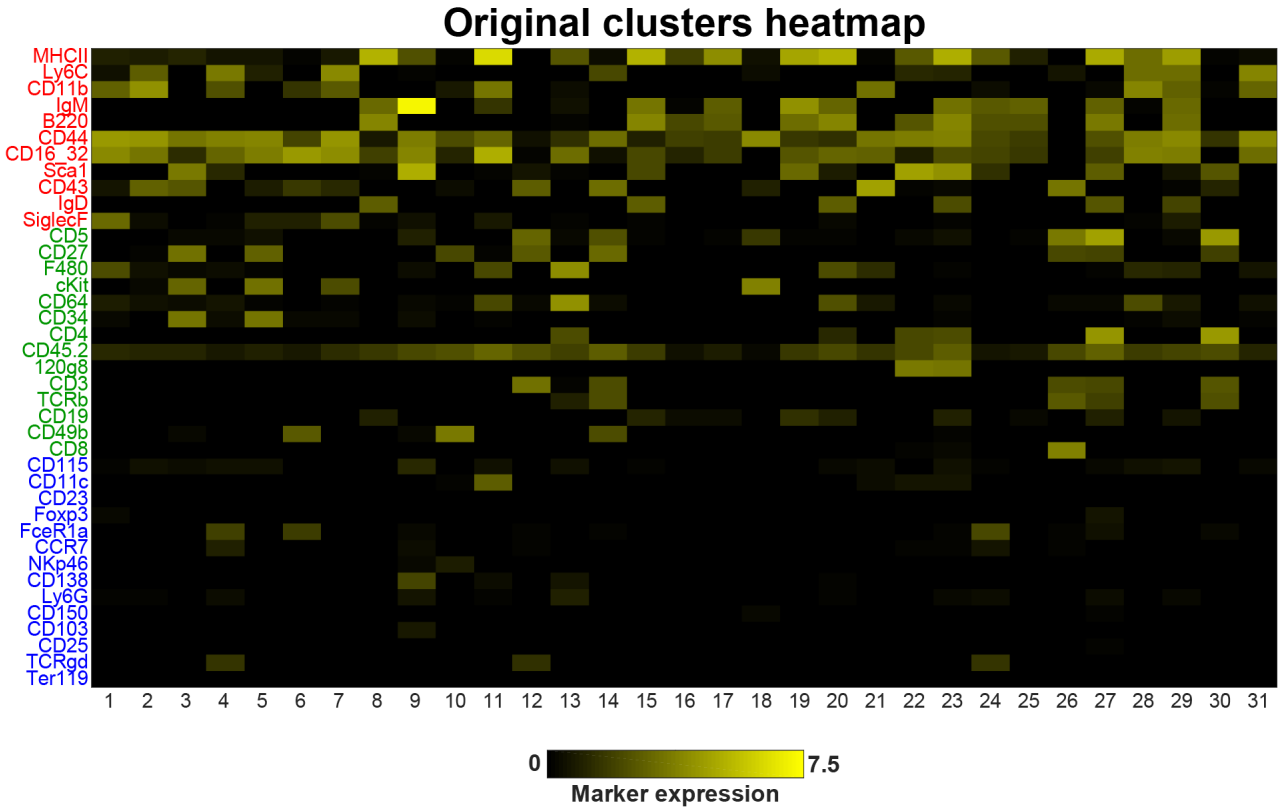




**Supplementary Fig. S12 HMIS shared markers clusters. (A)** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 42 cell clusters obtained by clustering the shared markers of the original HMIS data using Phenograph ($m = 16$). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the shared markers clusters of the HMIS dataset.

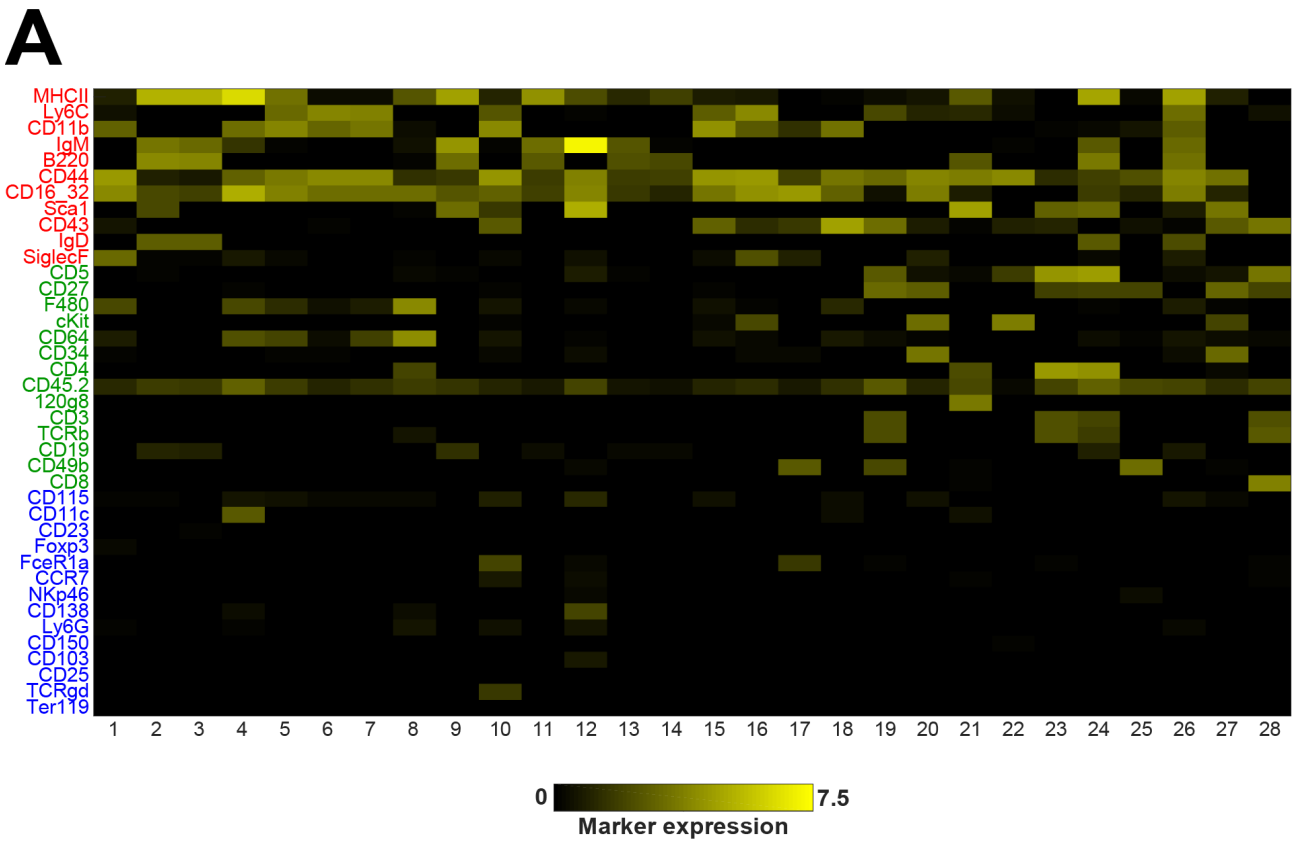


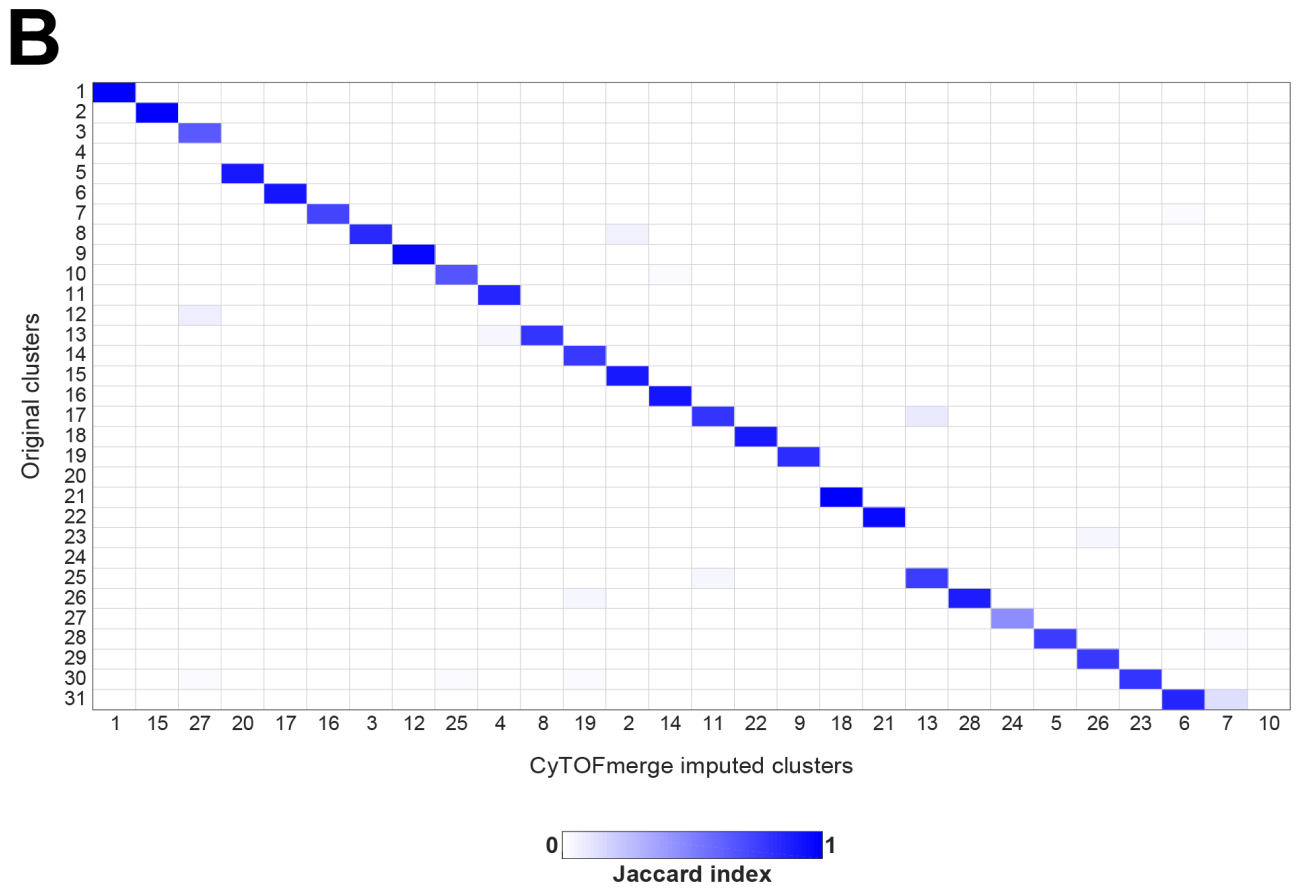




**Supplementary Fig. S13 HMIS imputed data by Cluster-based imputation. (A)** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 56 cell clusters obtained by clustering the imputed HMIS data using Phenograph (with $m = 16$ and $k = 50$, imputation performed within the same cluster found based on the shared markers space). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset.

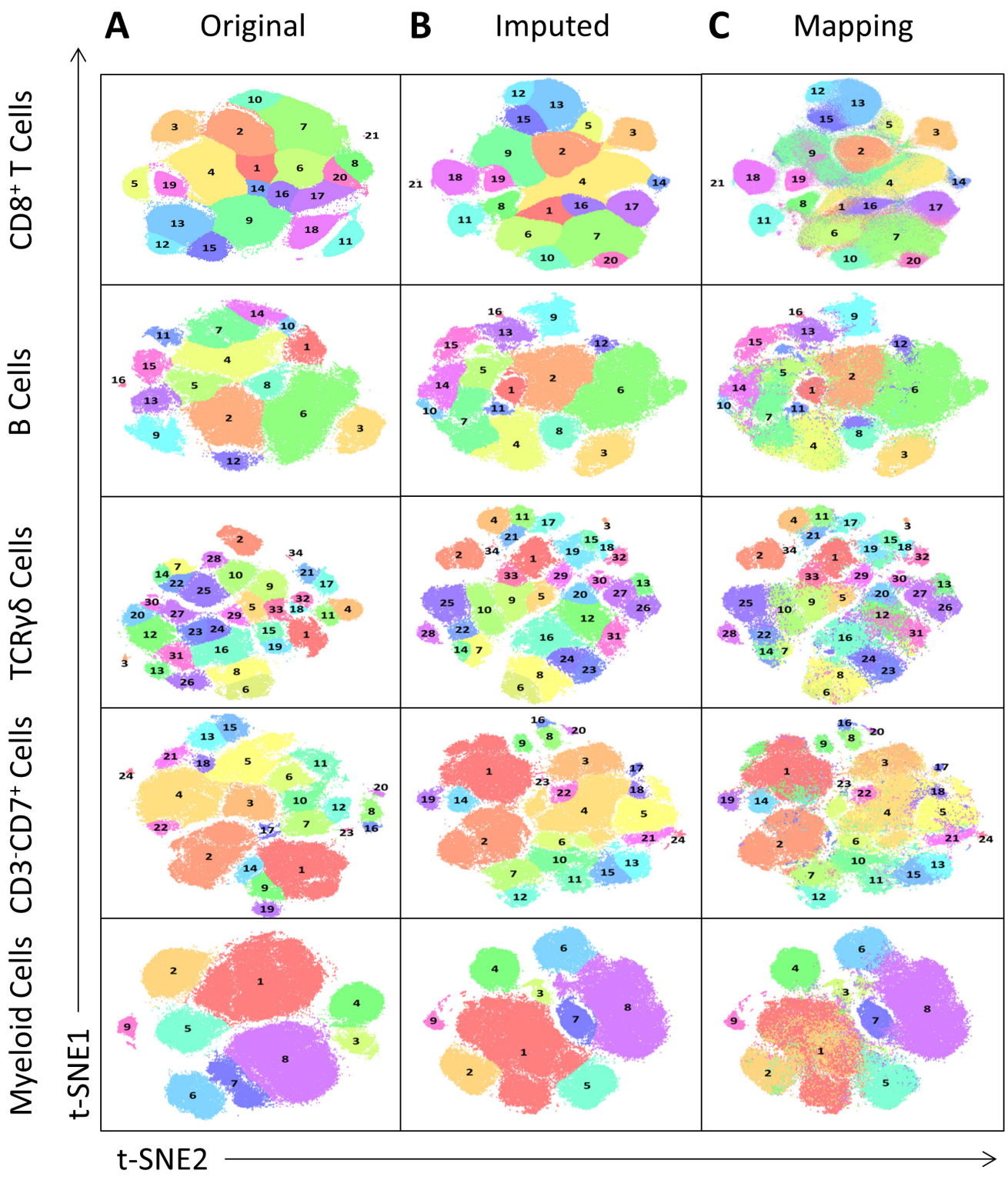



**Supplementary Fig. S14 Vortex original dataset clusters.** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 31 cell clusters obtained by clustering the original Vortex data using Phenograph. Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B).

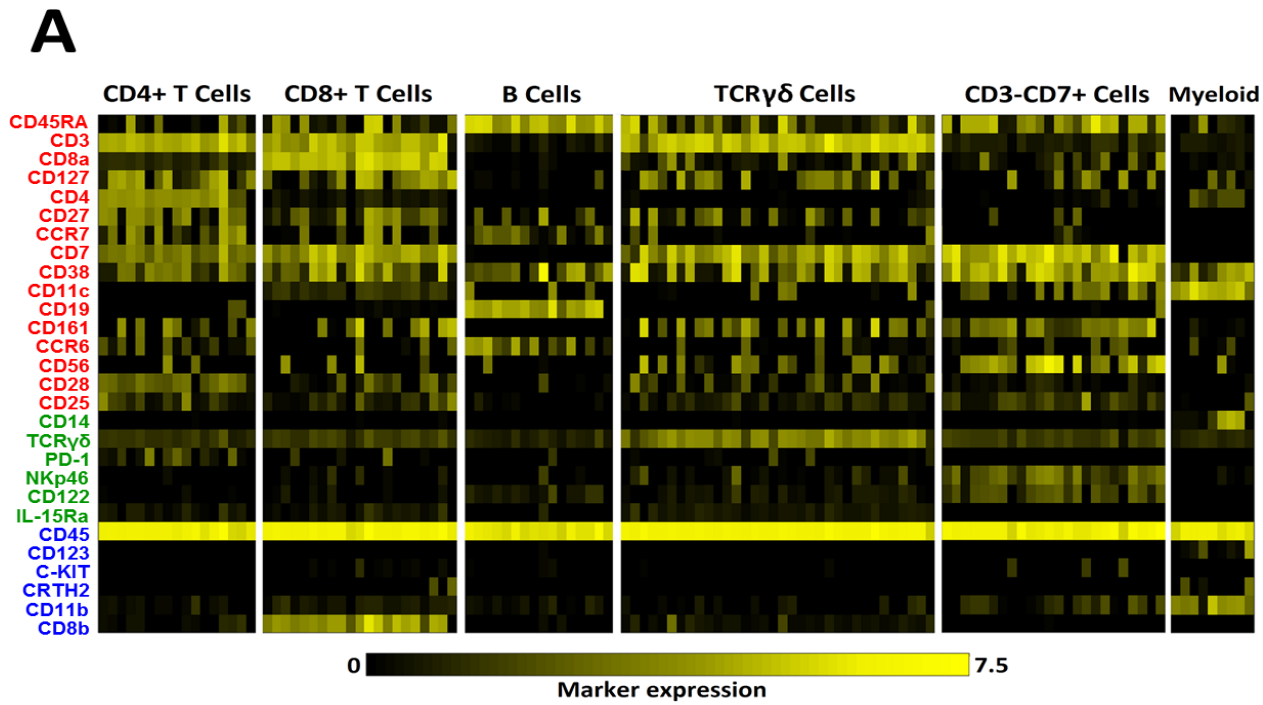


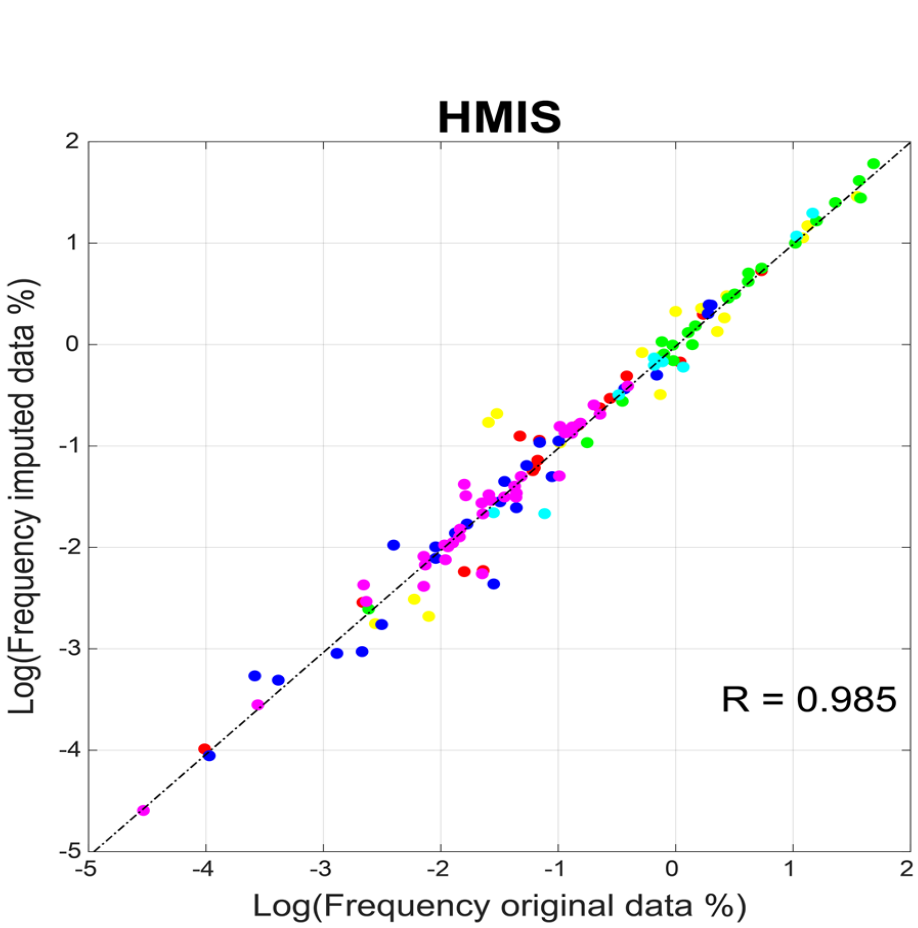




**Supplementary Fig. S15 Vortex imputed data by CyTOFmerge. (A)** Heatmap showing the median arcsinh5-transformed marker expression values (black-to-yellow scale) for the total 28 cell clusters obtained by clustering the imputed Vortex data using Phenograph (with $m$ = 11 and $k$ = 50). Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Pairwise Jaccard index map between the original and the imputed clusters of the Vortex dataset.

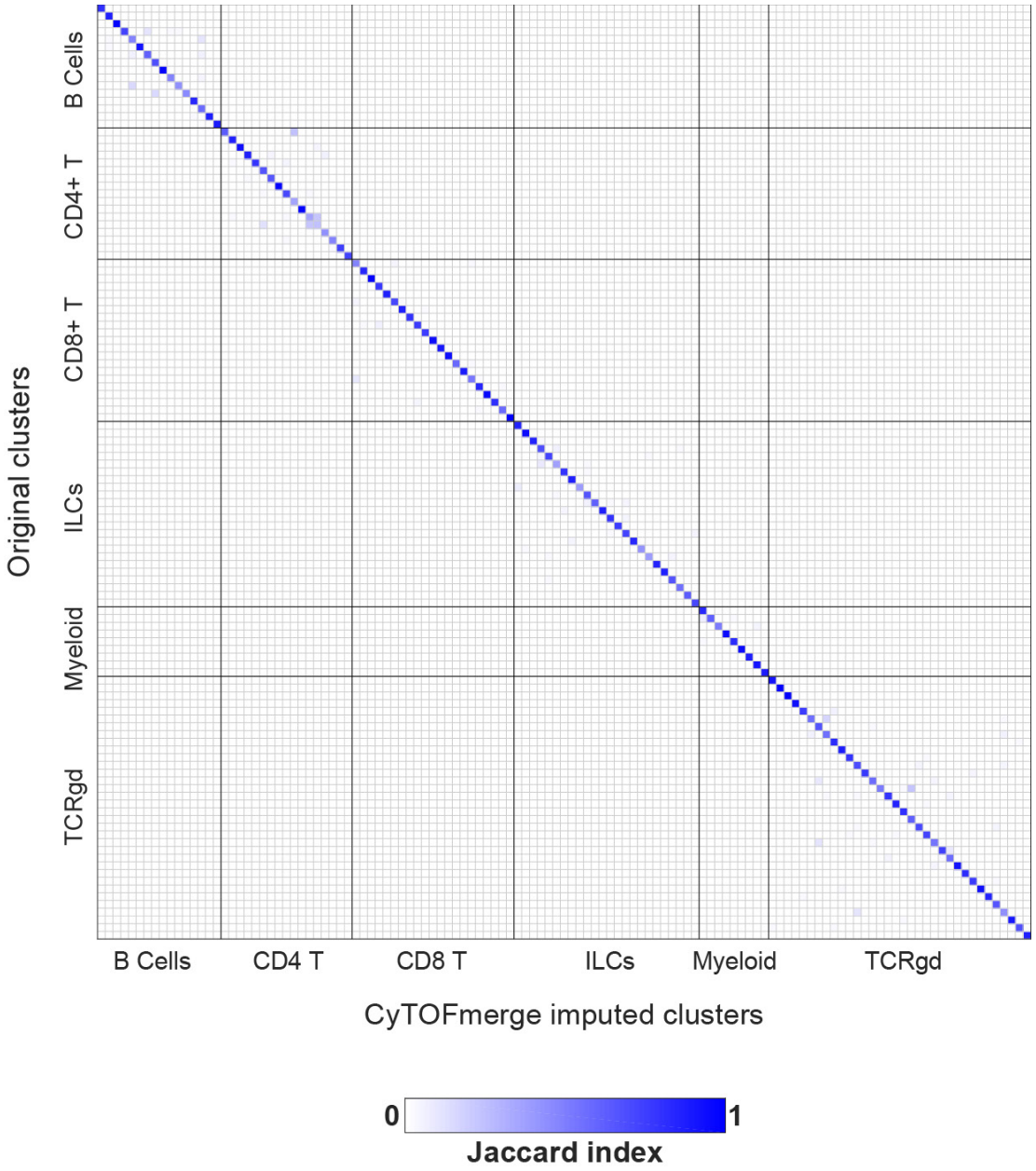



**Supplementary Fig. S16 Clustering of the original and the imputed datasets:** t-SNE maps showing the different identified populations in each immune lineage, each row represent a separate lineage, column **(A)** shows the populations of the original data, column **(B)** shows the populations of the imputed data (for $m$=16, $L1$=6 and $L2$=6) and column **(C)** is the mapping of the original clusters labels on the t-SNE map of the imputed data.

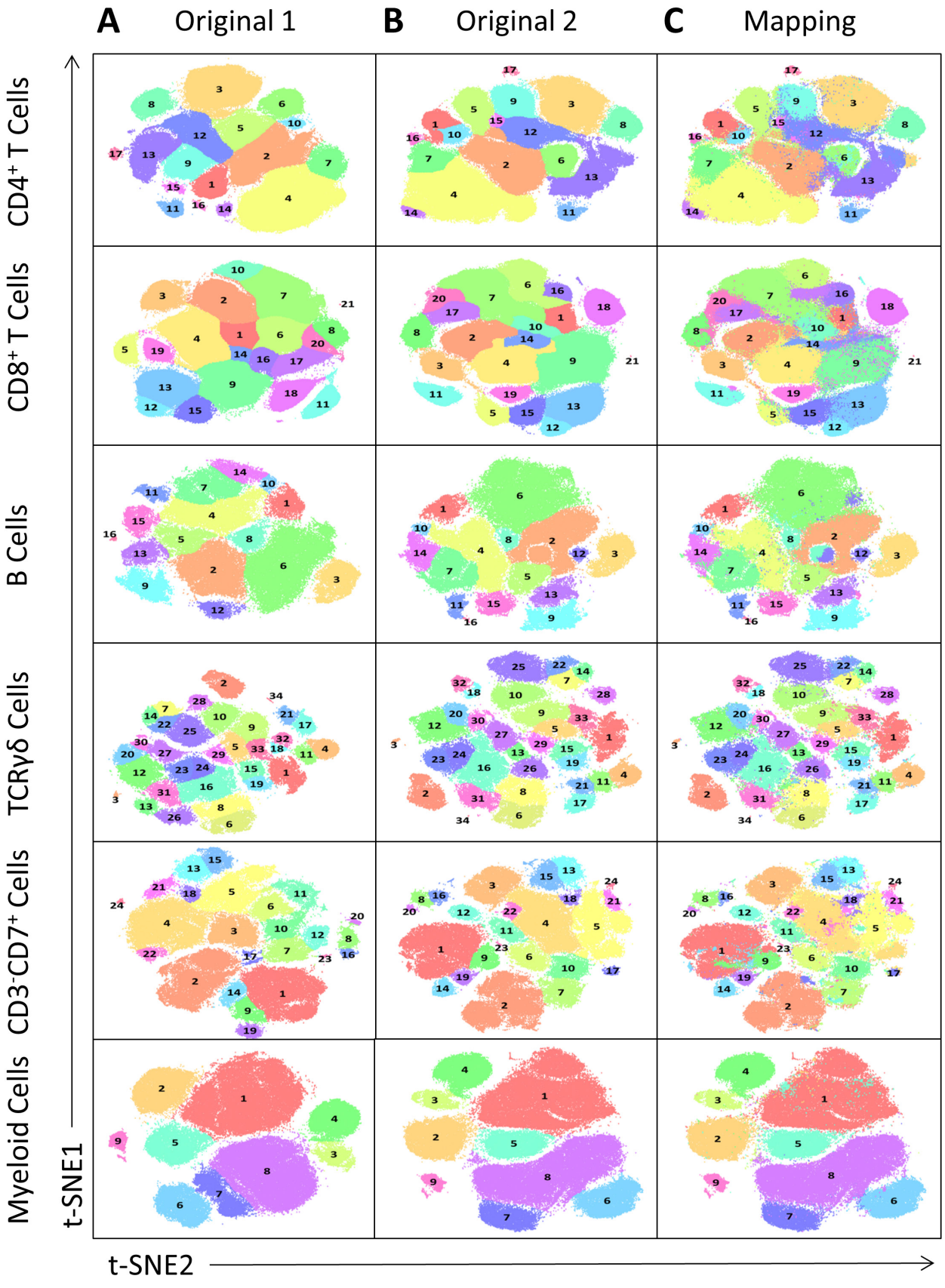


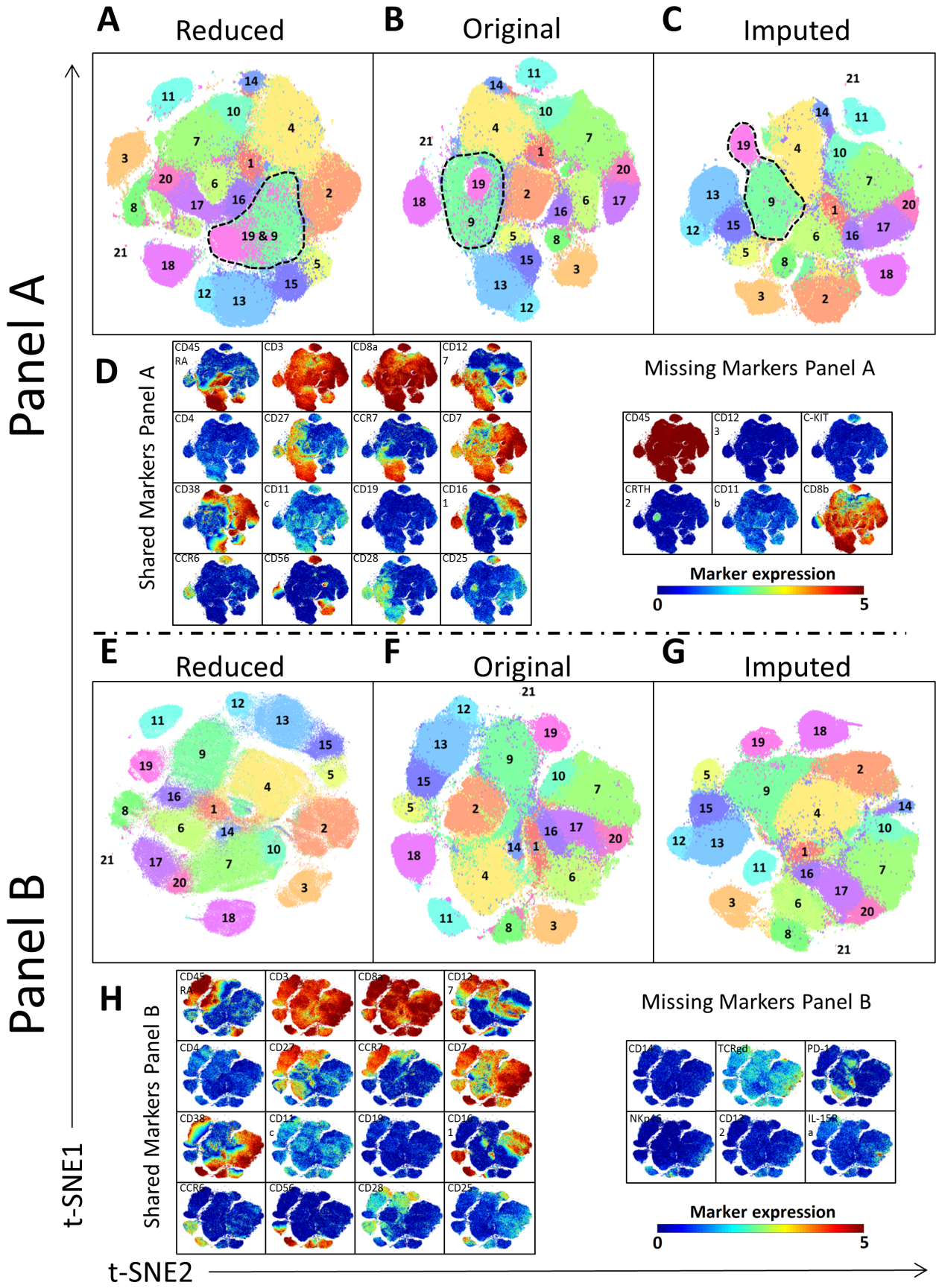

**Supplementary Fig. S17 (A)** Heatmap of markers expression for the 121 characterized immune cells populations of the imputed dataset for $m = 16$. Black-to-yellow scale shows the median arcsinh-5 transformed values for the markers expression. Marker colors indicate whether a marker is shared between panels or unique to a single panel, during panels combination (red is shared, green is unique to panel A, blue is unique to panel B). **(B)** Scatter plots between original and imputed data population frequencies, the dashed line shows the least-squares fit error line, and the R value represents Pearson correlation coefficient between original and imputed frequencies.



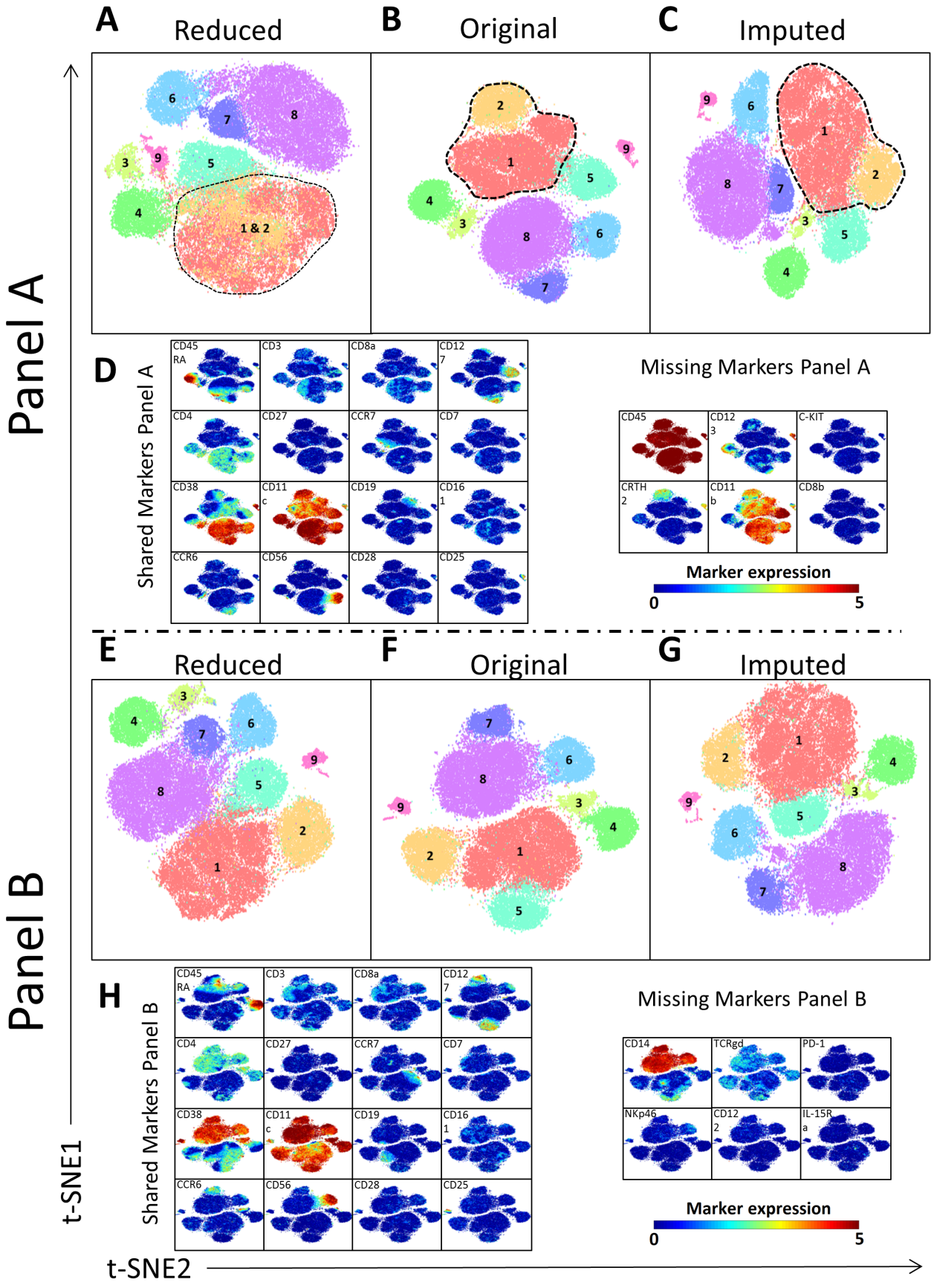

**Supplementary Fig. S18** Pairwise Jaccard index map between the original and the imputed clusters of the HMIS dataset, clustered using Cytosplore.



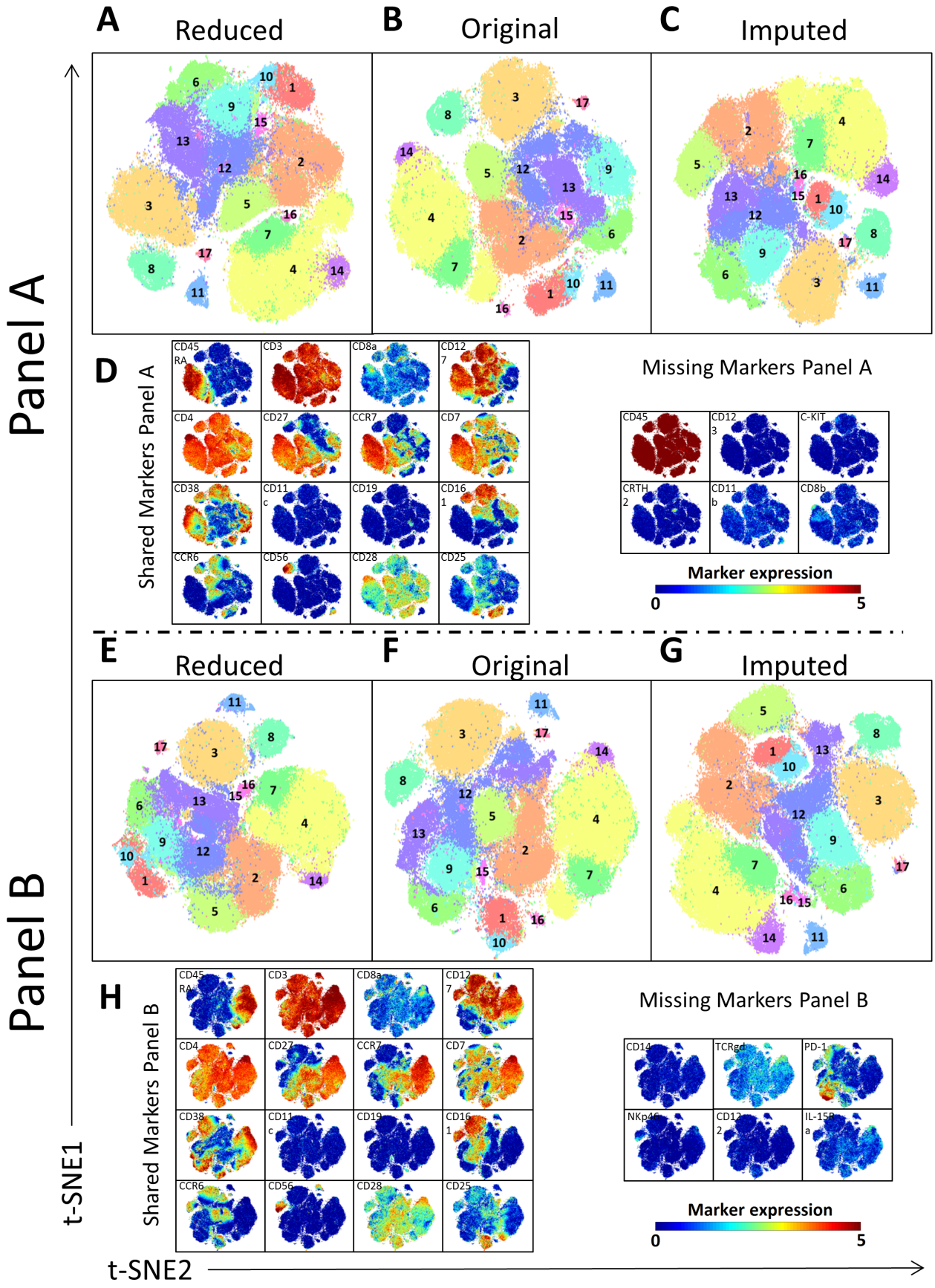

**Supplementary Fig. S19 Evaluation of t-SNE rerun:** t-SNE maps showing the different identified populations in each immune lineage by running the t-SNE twice to the original data, each row represent a separate lineage, column **(A)** shows the populations of the original data for the first t-SNE map (Original 1), column **(B)** shows the populations of the original data for the second t-SNE map (Original 2) and column **(C)** is the mapping of the Original 1 clusters labels on the Original 2 t-SNE map.



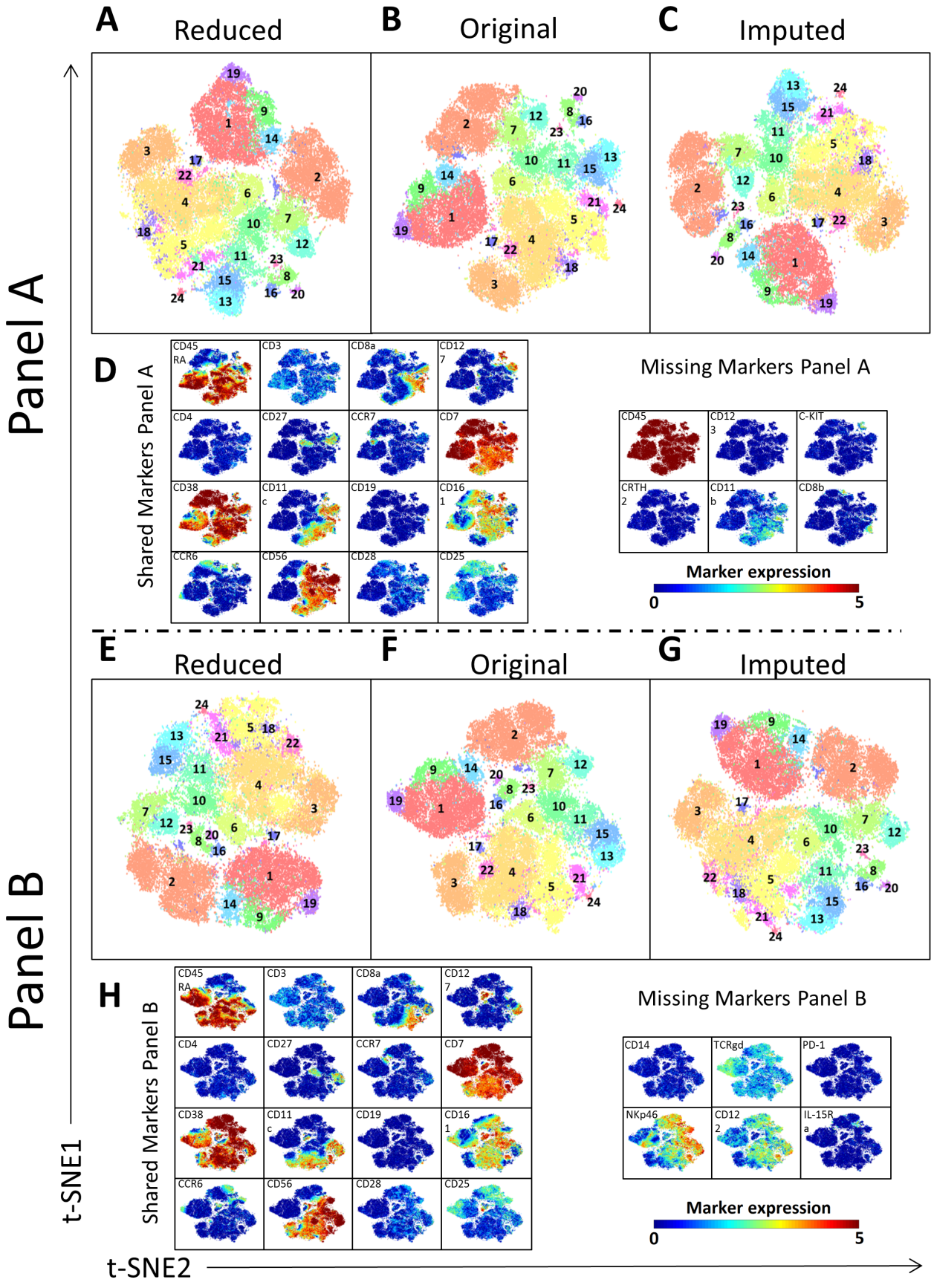

**Supplementary Fig. S20 Marker extension impact on identification of distinct populations in the CD8+ T Cells immune lineage**: The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



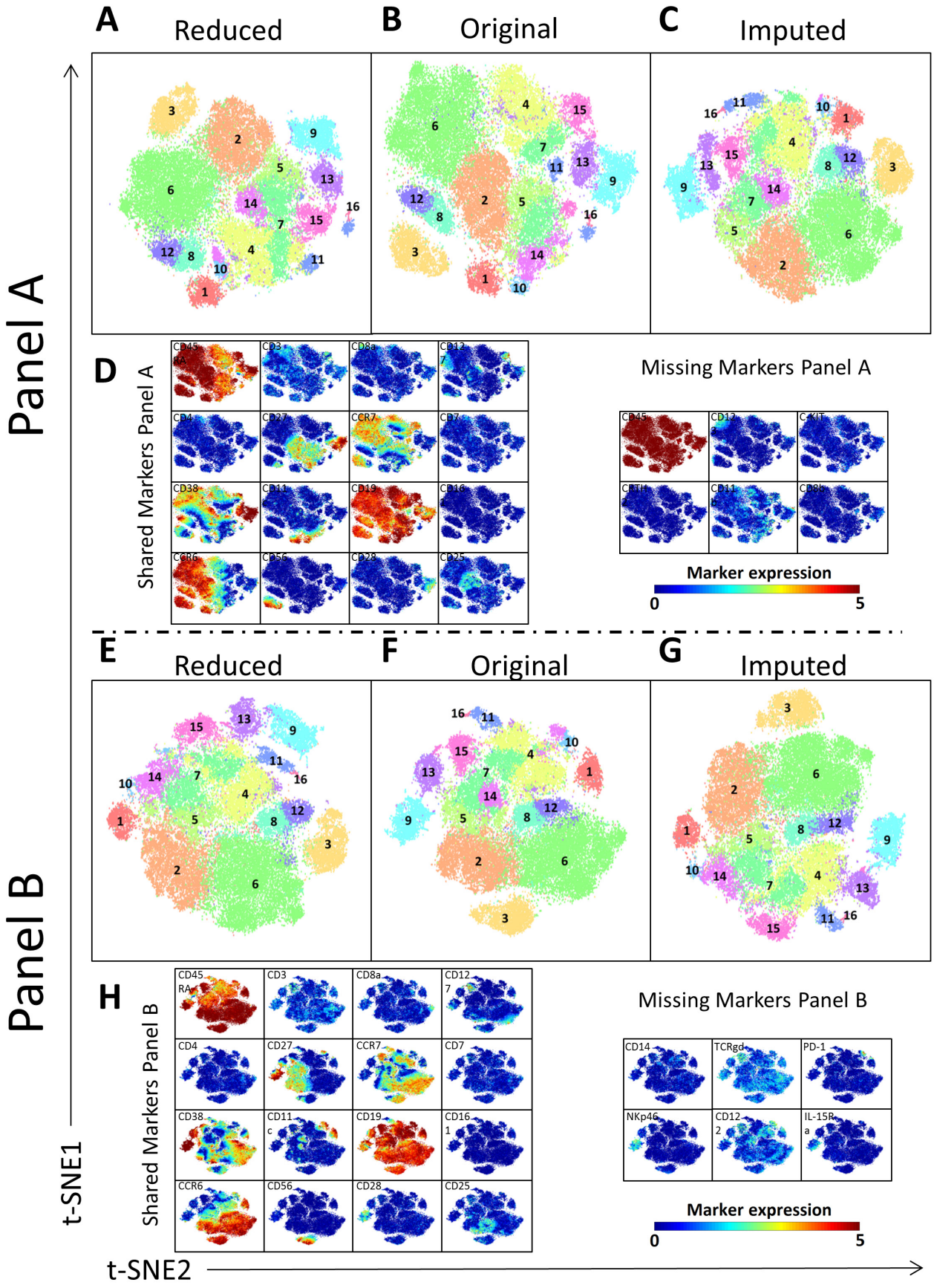

**Supplementary Fig. S21 Marker extension impact on identification of distinct populations in the Myeloid Cells immune lineage:** The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.



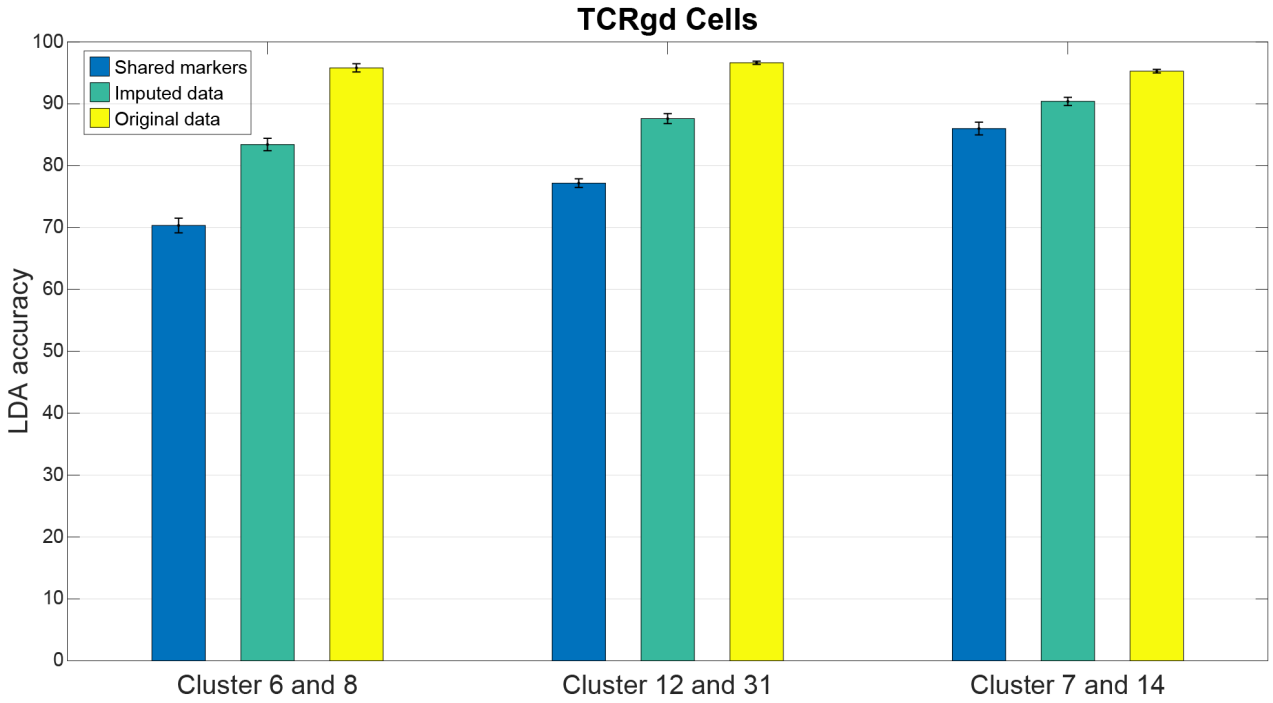

**Supplementary Fig. S22 Marker extension impact on identification of distinct populations in the CD4+ T Cells immune lineage:** The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.




**Supplementary Fig. S23 Marker extension impact on identification of distinct populations in the CD3-CD7+ Cells immune lineage:** The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.




**Supplementary Fig. S24 Marker extension impact on identification of distinct populations in the B Cells immune lineage:** The upper half presents Panel A and the lower part presents Panel B. For each panel, the Reduced **(A, E)**, Original **(B, F)** and Imputed **(C, G)** t-SNE map are shown colored with the populations labels. Shared and missing markers expression profiles are shown on the Original t-SNE map **(D, H)**. The color scale bar shows the arcsinh-5 transformed values for the markers expression.




**Supplementary Fig. S25** LDA classification accuracy for clusters 6-8, 12-3 and 7-14, in the TCRgd cells from the HMIS dataset. Classification is applied using the 16 shared markers only, all 28 markers from the imputed dataset, and all 28 markers from the original dataset. Error bar shows the performance variation across the 5-folds of the cross validation.