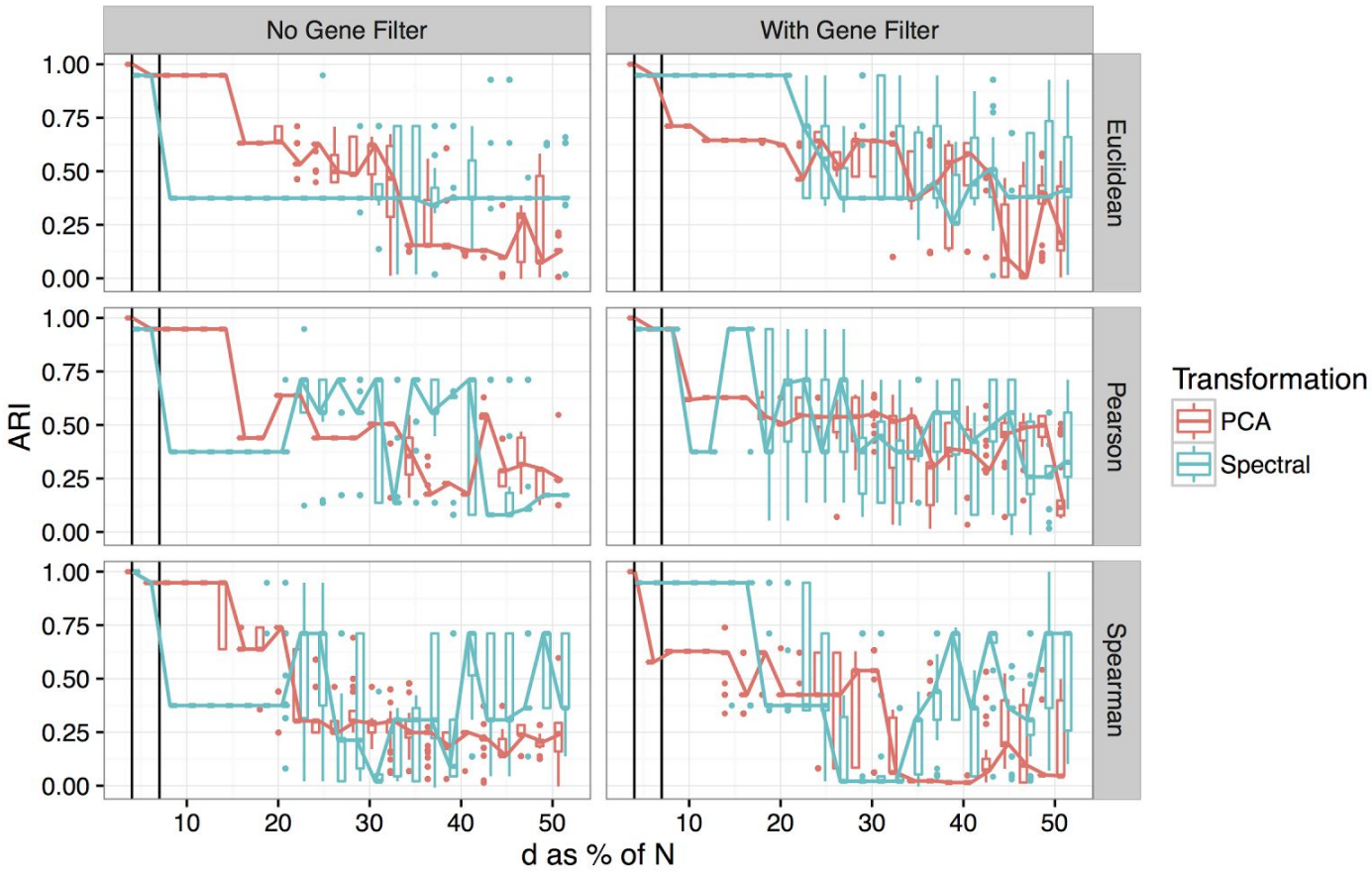

Figure S1. **Boxplots of 100 realizations of the SC3 clustering on the Biase dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 49). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

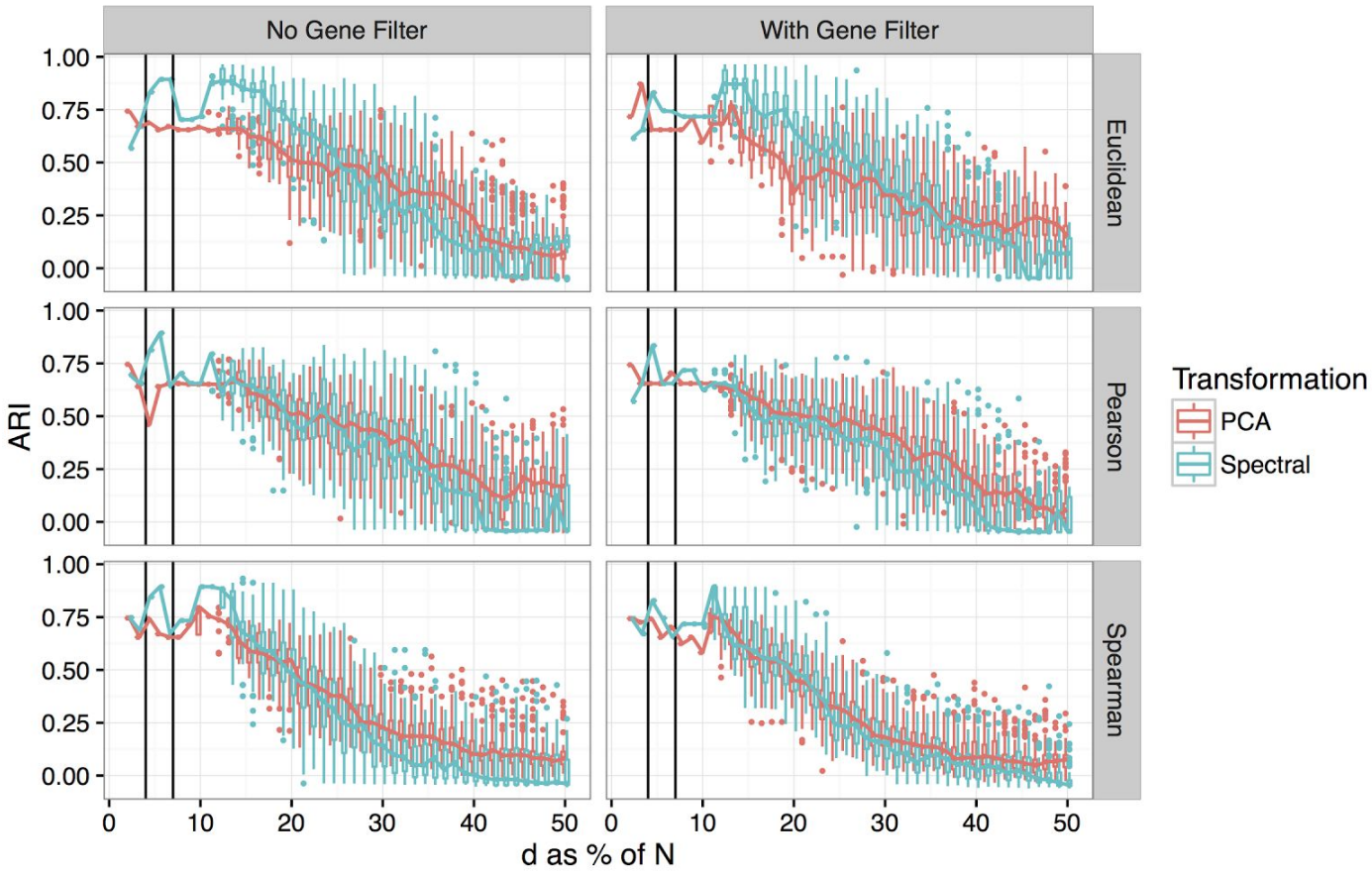

Figure S2. **Boxplots of 100 realizations of the SC3 clustering on the Yan dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 90). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

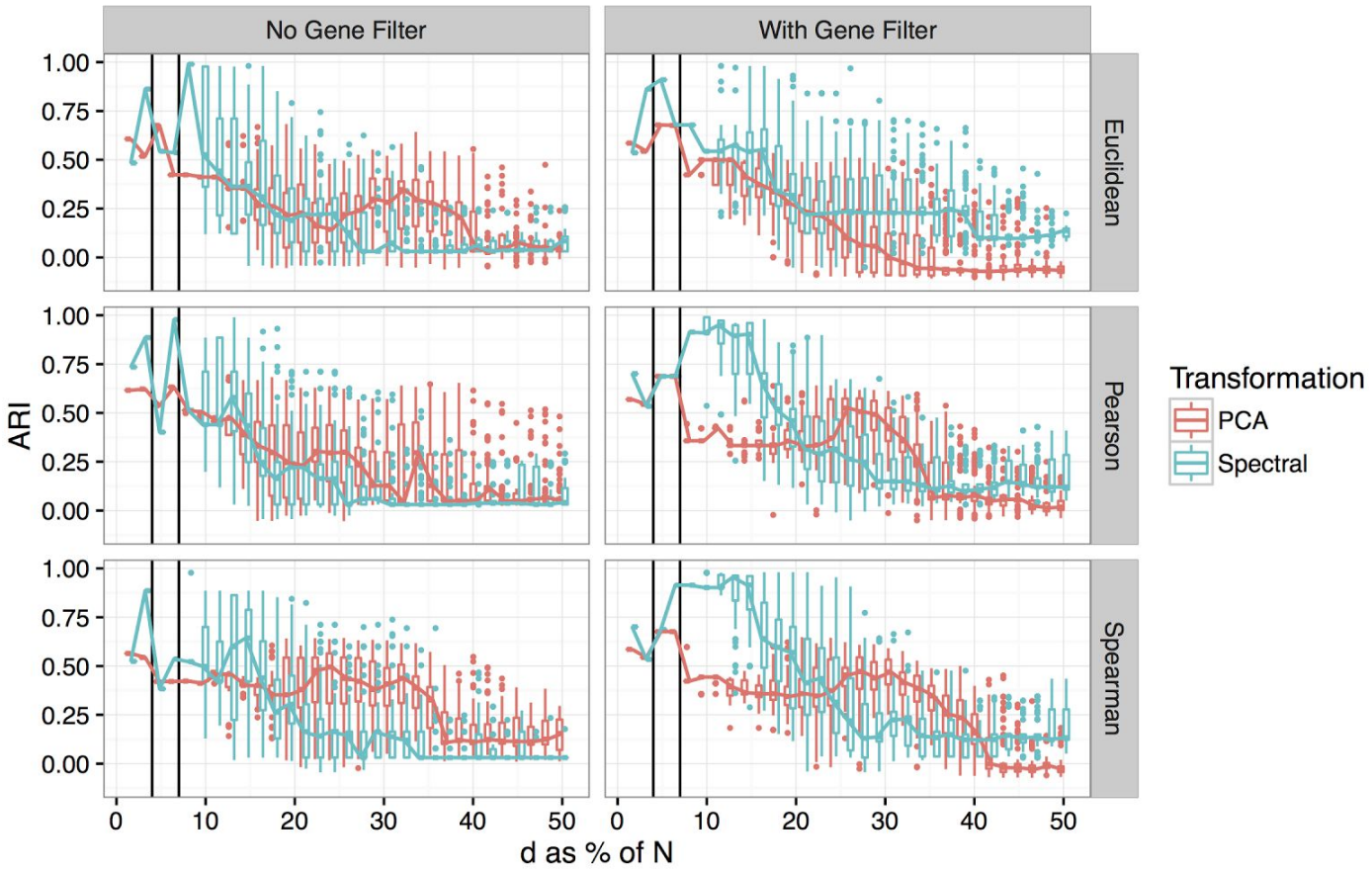

Figure S3. **Boxplots of 100 realizations of the SC3 clustering on the Goolam dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 124). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

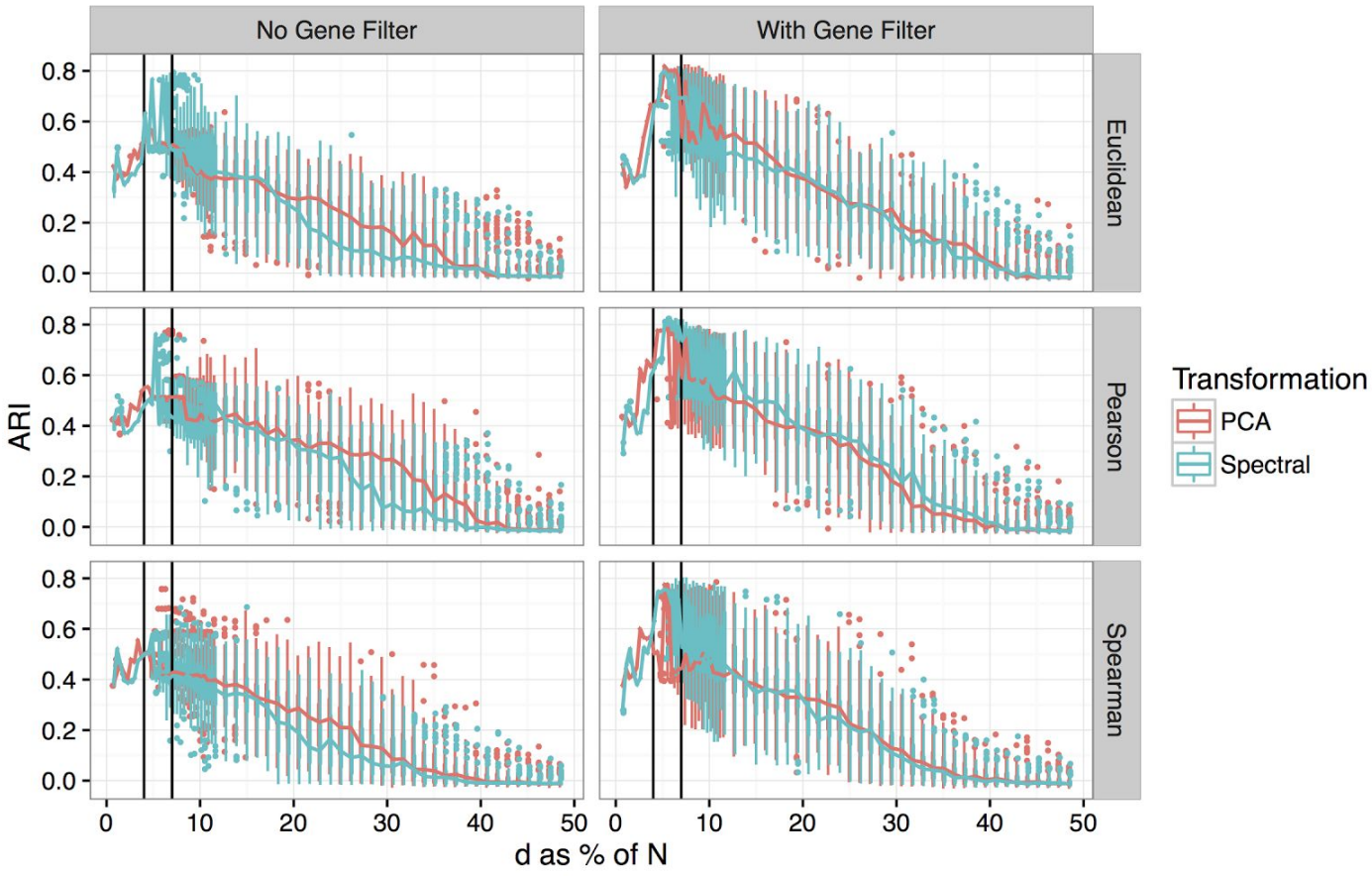

Figure S4. **Boxplots of 100 realizations of the SC3 clustering on the Deng dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 268). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

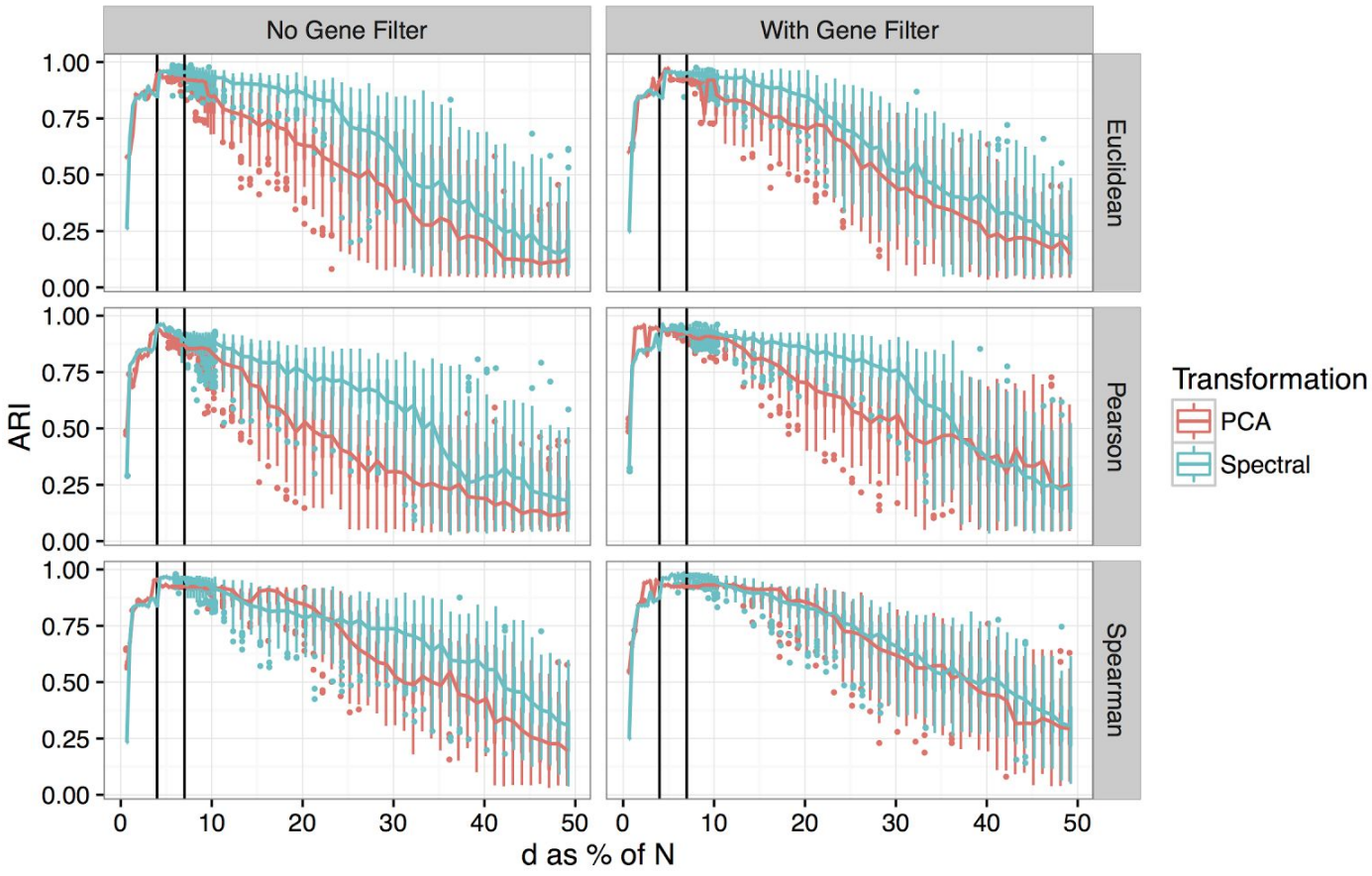

Figure S5. **Boxplots of 100 realizations of the SC3 clustering on the Pollen dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 301). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

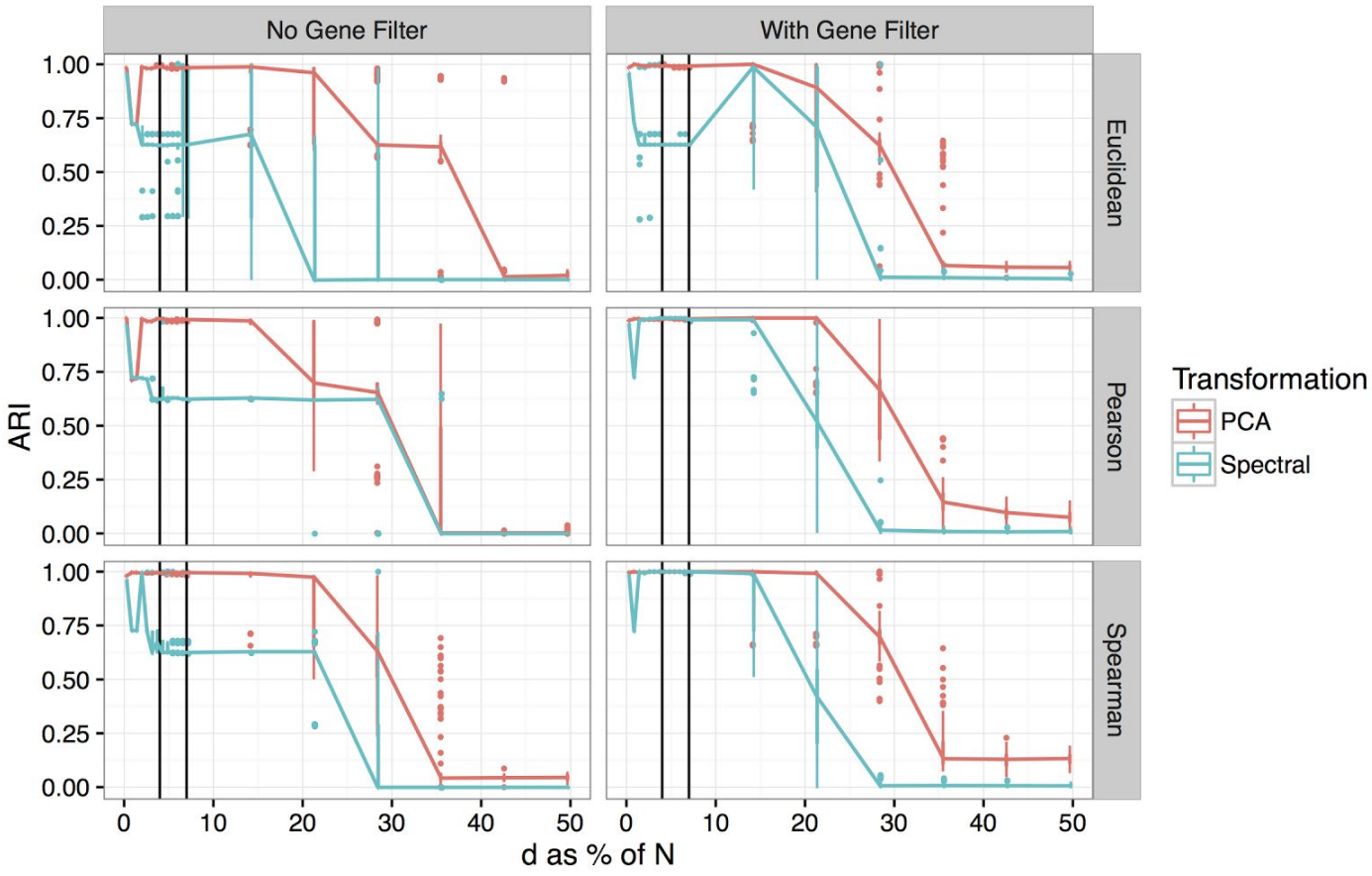

Figure S6. **Boxplots of 100 realizations of the SC3 clustering on the Kolodziejczyk dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 704). Dots represent outliers that are higher than the highest value (or lower than the lowest value) within 1.5 * IQR, where IQR is the inter-quartile range, or distance between the first and third quartiles.

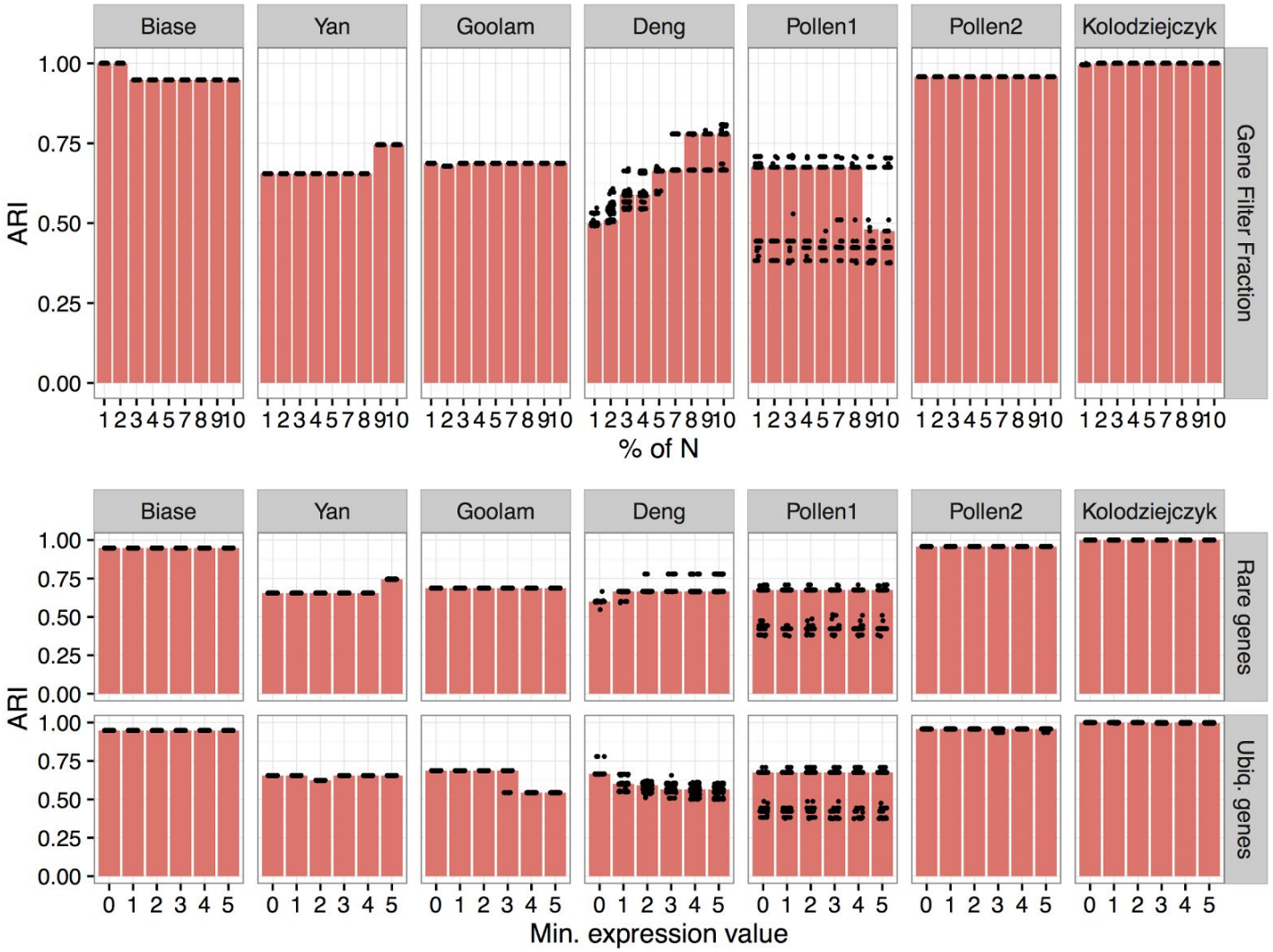

Figure S7. Exploration of the gene filter parameters (see Methods).

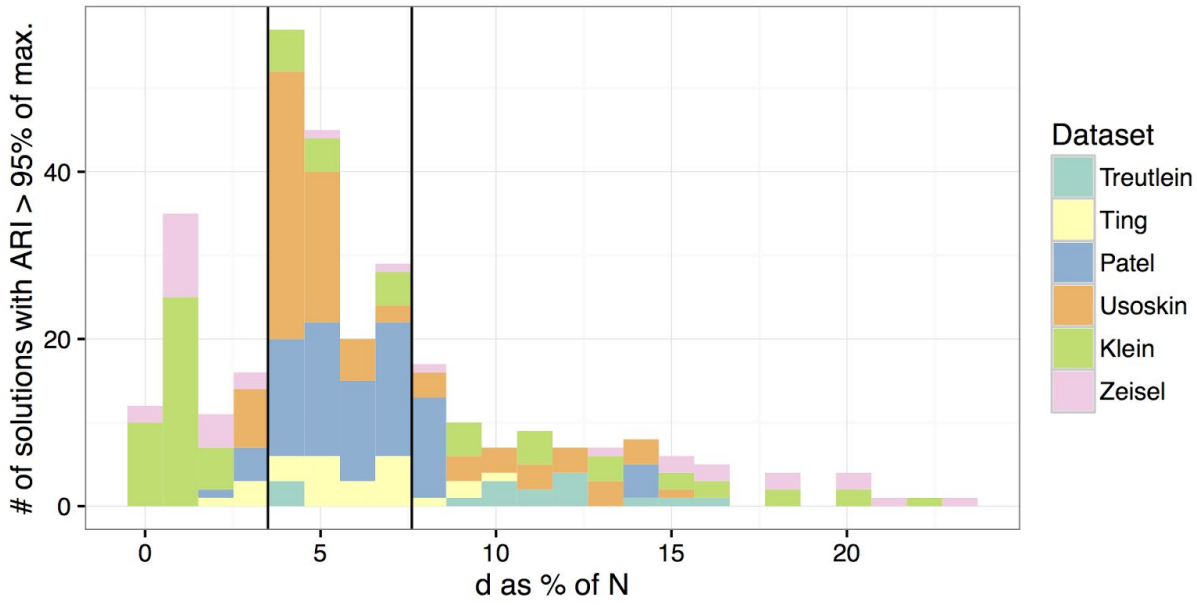

Figure S8. Histogram of the *d* values where ARI>.95 is achieved for the silver standard datasets from Fig. 1b. The black vertical lines indicate the interval *d* = 4-7% of the total number of cells *N,* showing high accuracy in the classification.

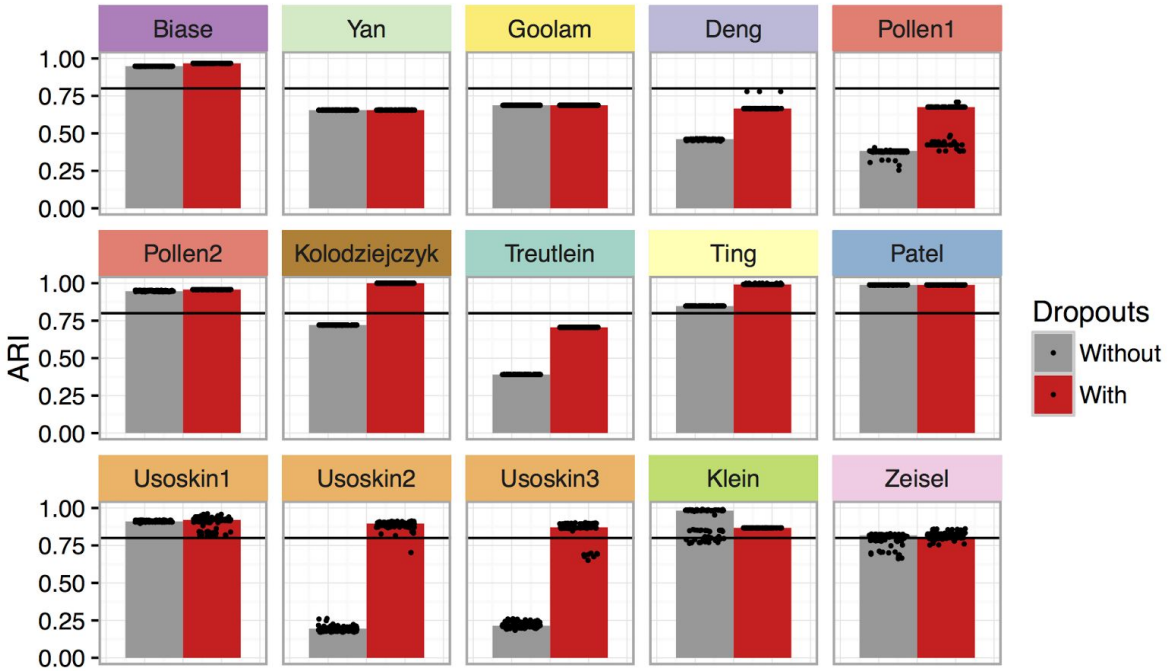

Figure S9. The effect of dropouts in the distance calculations step on the accuracy of SC3 clustering (Methods).

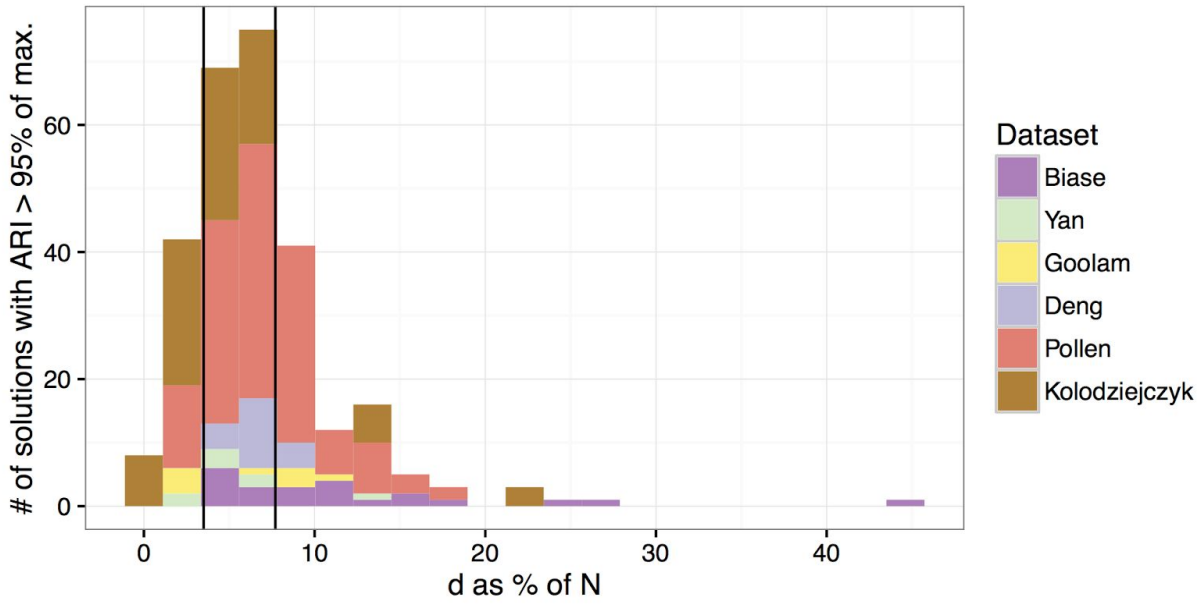

Figure S10. Histogram of the *d* values where ARI>.95 is achieved for the downsampled (by a factor of ten) gold standard datasets. The black vertical lines indicate the interval *d* = 4-7% of the total number of cells *N*, showing high accuracy in the classification.

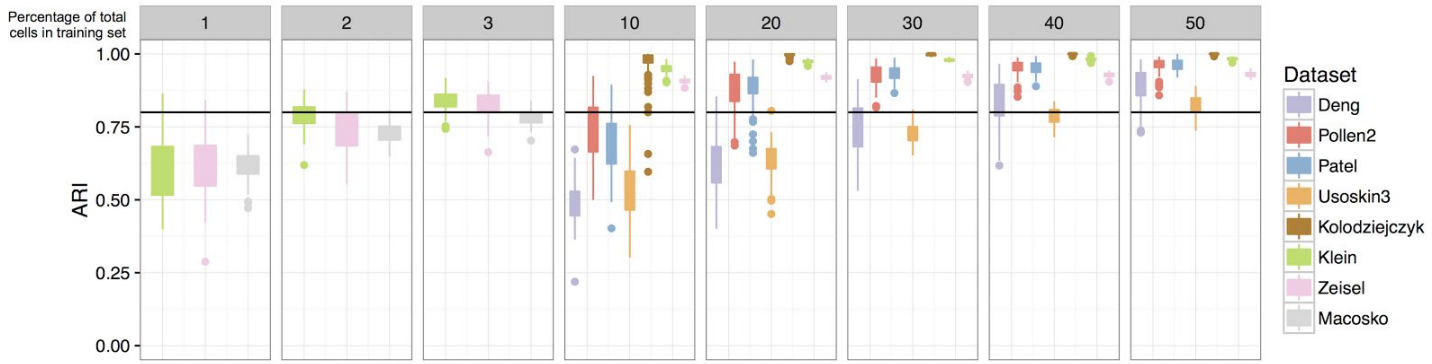

Fig. S11. **Using the hybrid SC3 based on reference labels provided by the authors.** Same as Figure 3a in the main text, but using the reference labels provided by the authors as inputs to the SVM. Dots represent outliers higher (lower) than the highest (lowest) value within 1.5 x IQR, where IQR is the interquartile range. The black line indicates ARI = 0.8.

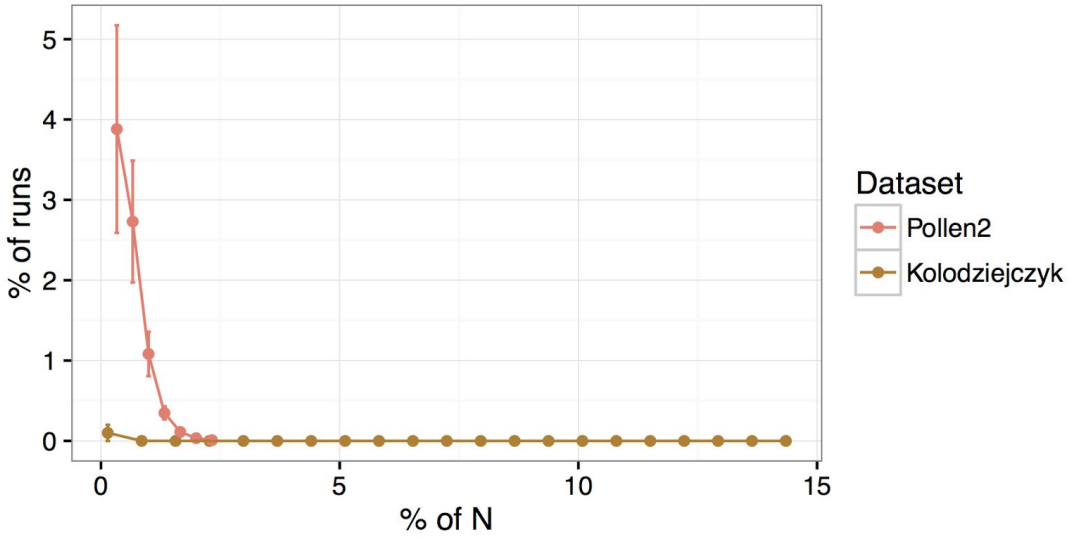

Figure S12. Sensitivity of SC3 for identifying rare cell-types when the hybrid SC3 approach is used with 30% of cells to train the SVM. This figure was derived from Fig. 3b by correcting the mean fraction of times that the rare cells were located in the same cluster using the probability of drawing rare cells within the 30% of all cells (Methods).

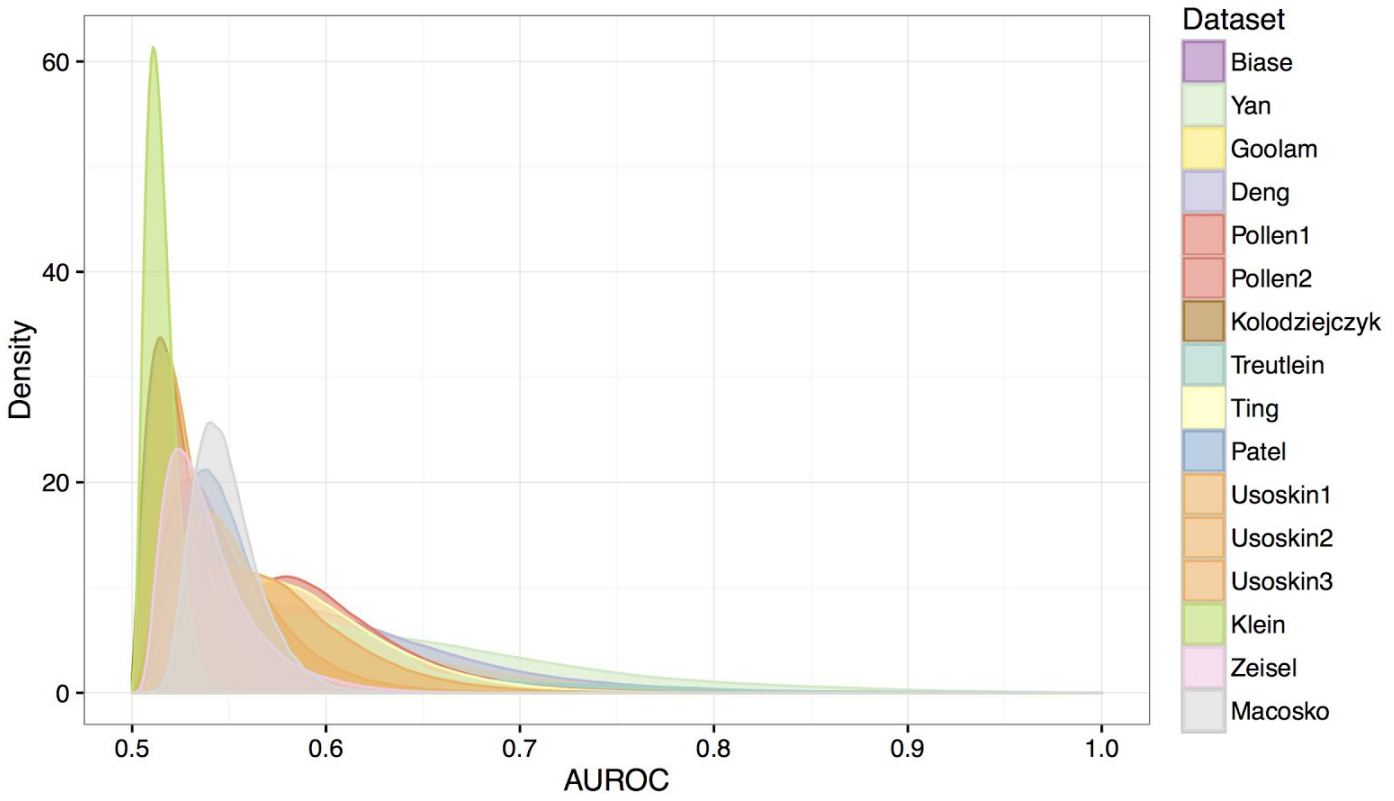

Figure S13. Density of distributions of AUROC obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods).

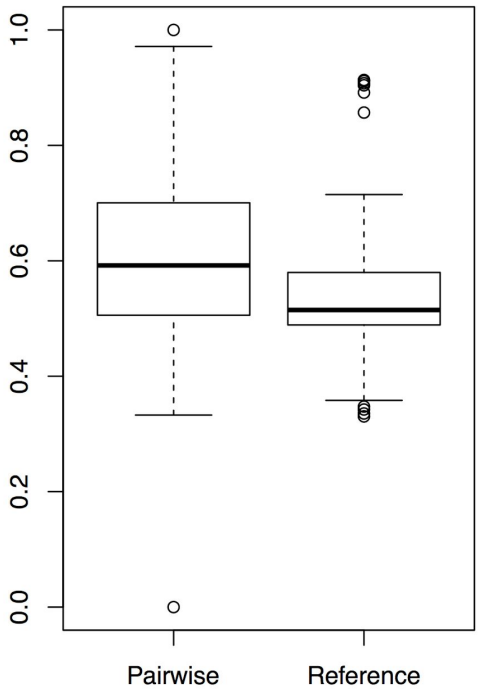

Fig. S14 The cells from the Drop-Seq dataset were clustered 100 times using SC3. “Pairwise” indicates the ARIs between the different solutions obtained and “Reference” indicates the ARI as compared to the labels obtained by Macosko et al.

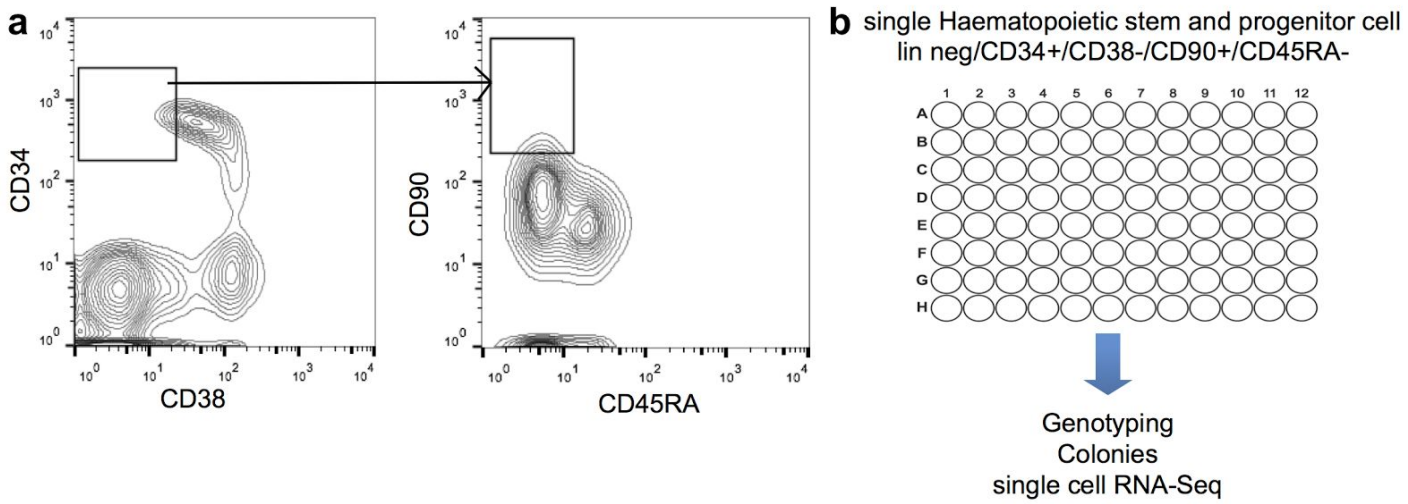

Figure S15. **Cell sorting procedure for patients.** (**a**) Contour plots describing the sorting strategy for isolating HSCs in patient 2 (the same was done for patient 1). CD34, CD38, CD90 and CD45RA expression is displayed using a log scale. (**b**) Lineage negative, CD34+/CD38-/CD90+/CD45RA- single cells were sorted into individual wells for scRNA-Seq or colony growth in cytokine cocktail allowing progenitor cell expansion.

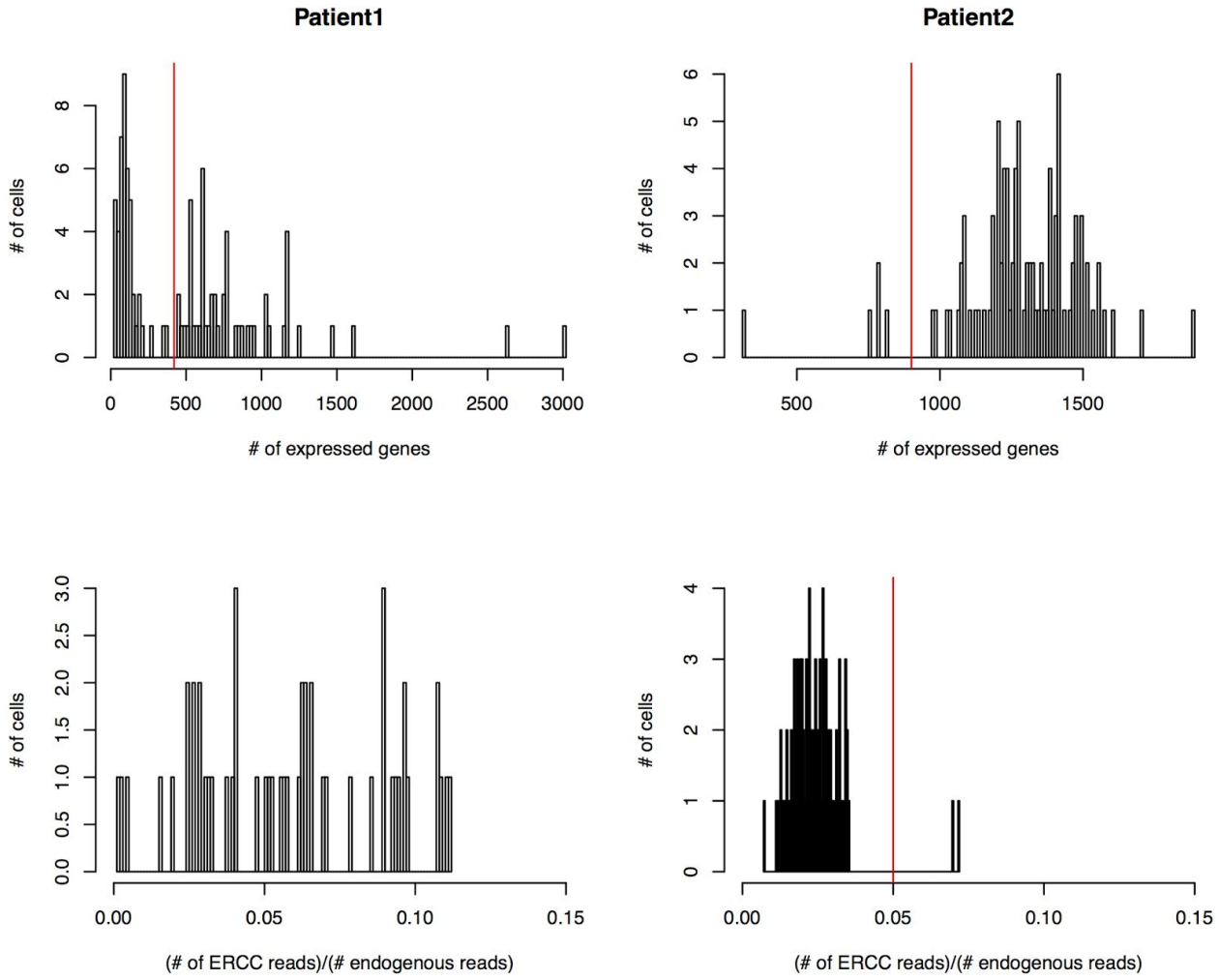

Figure S16. **Quality control of cells in the patient data.** (**a**) Number of cells with a given number of expressed genes in each patient. Cells on the left side of the red line were removed from further analysis as lowly expressed. (**b**) Number of cells with a given (# of ERCC reads)/(# endogenous reads) ratio in each patient. Cells on the right side of the red line were removed from further analysis as outliers.

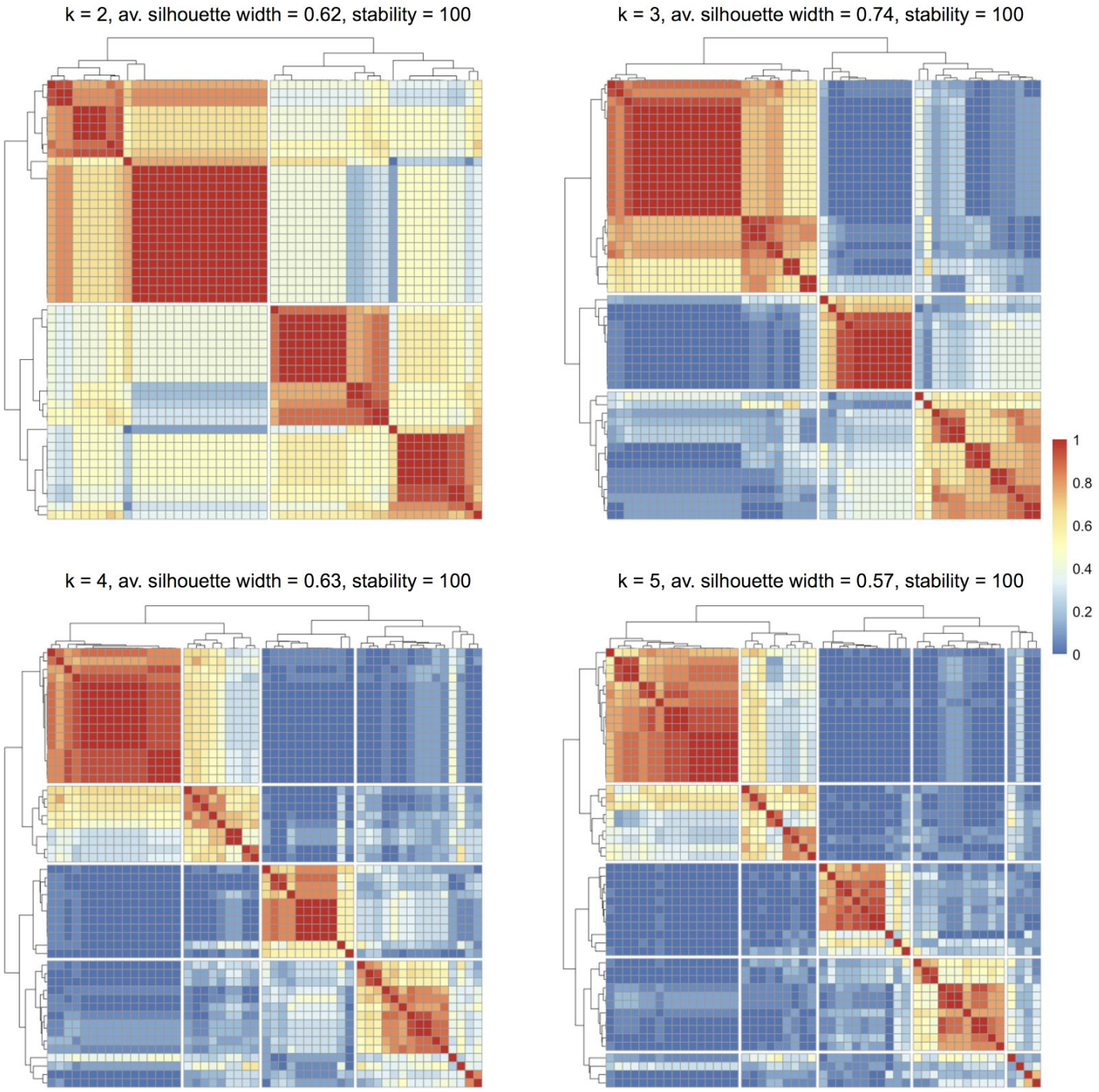

Figure S17. **Clustering of scRNA-seq data from patient 1.** Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.

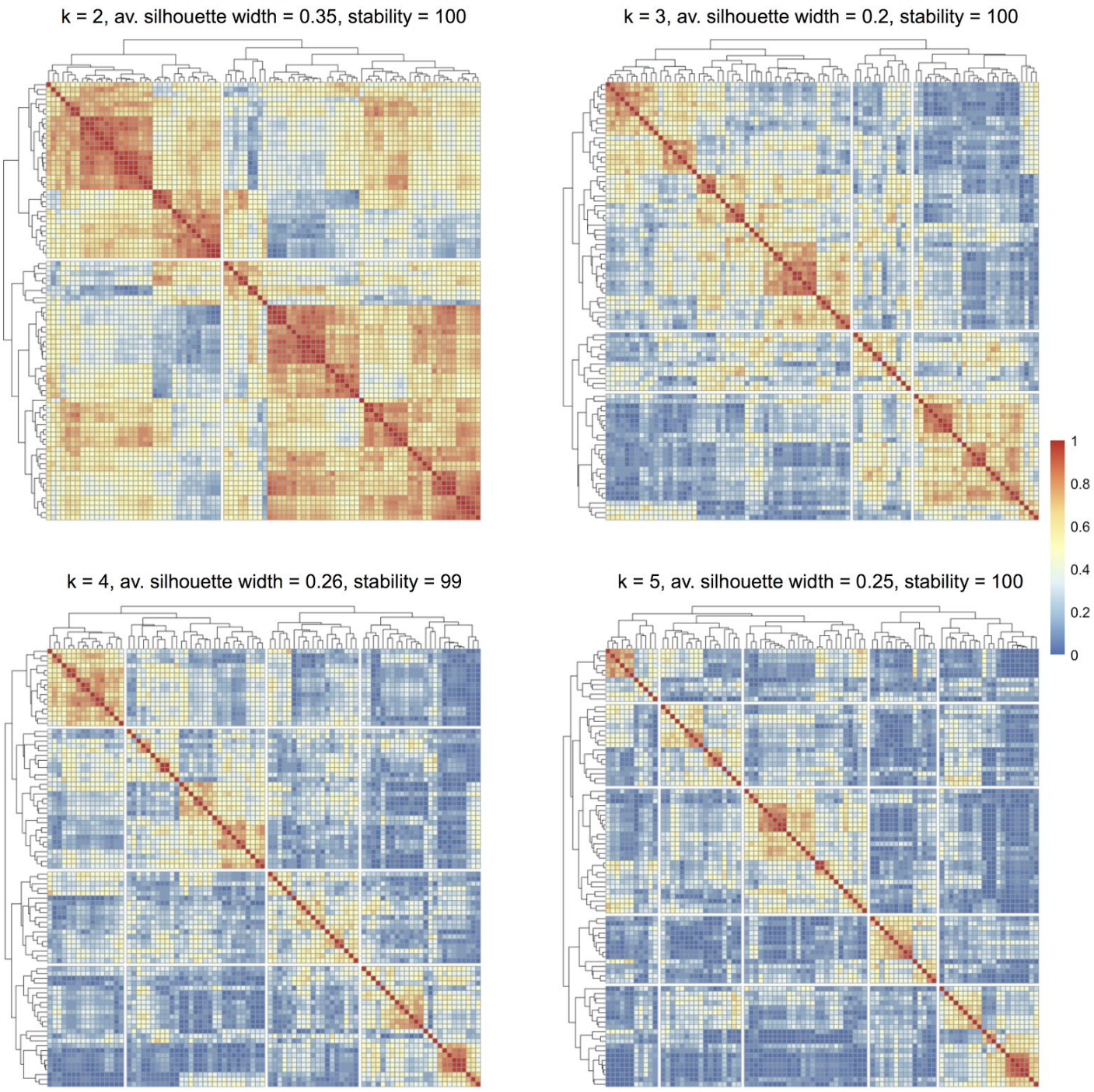

Figure S18. **Clustering of scRNA-seq data from patient 2.** Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.

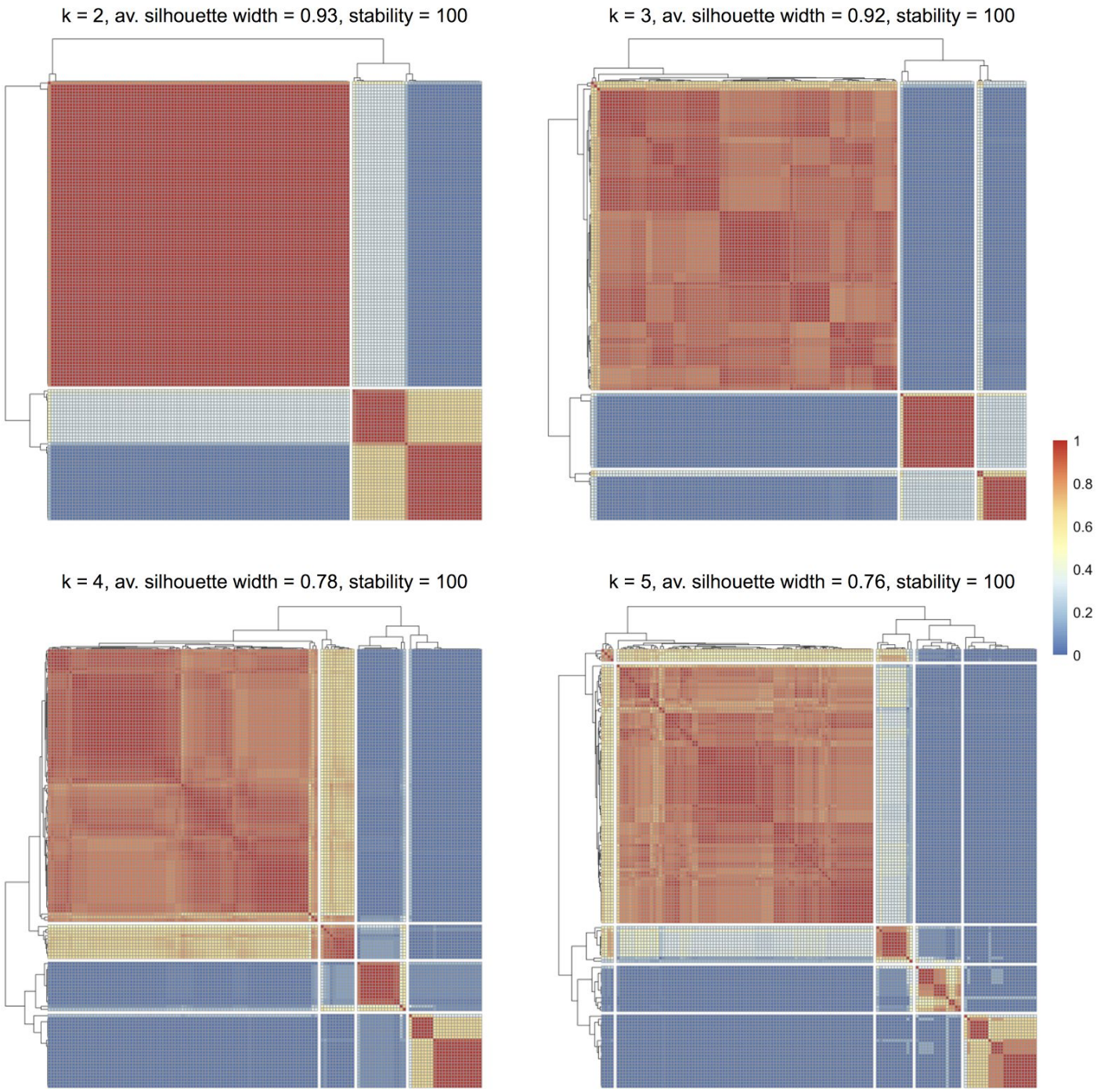

Figure S19. **Clustering of scRNA-seq data from combined patient 1 and patient 2 datasets.** Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.

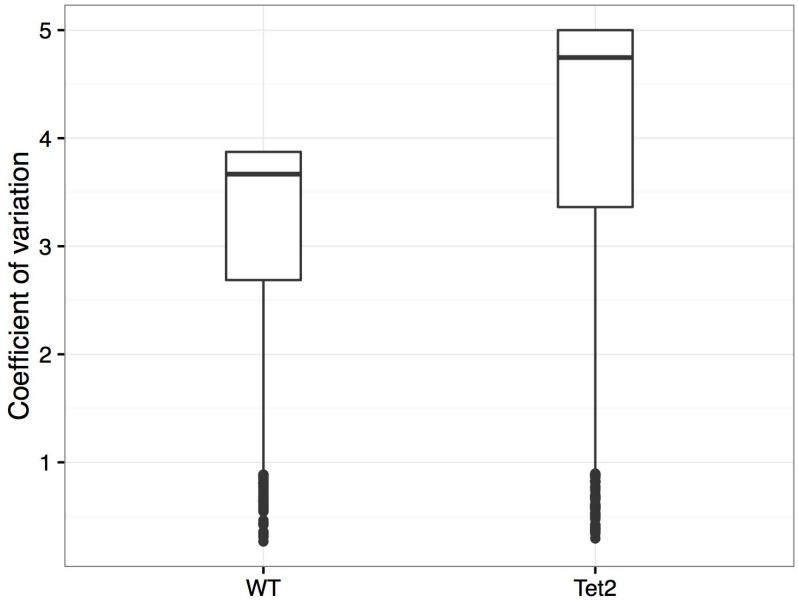

Figure S20. Comparison of the coefficient of variation of gene expression in Tet2 and WT subclones of patient 1.

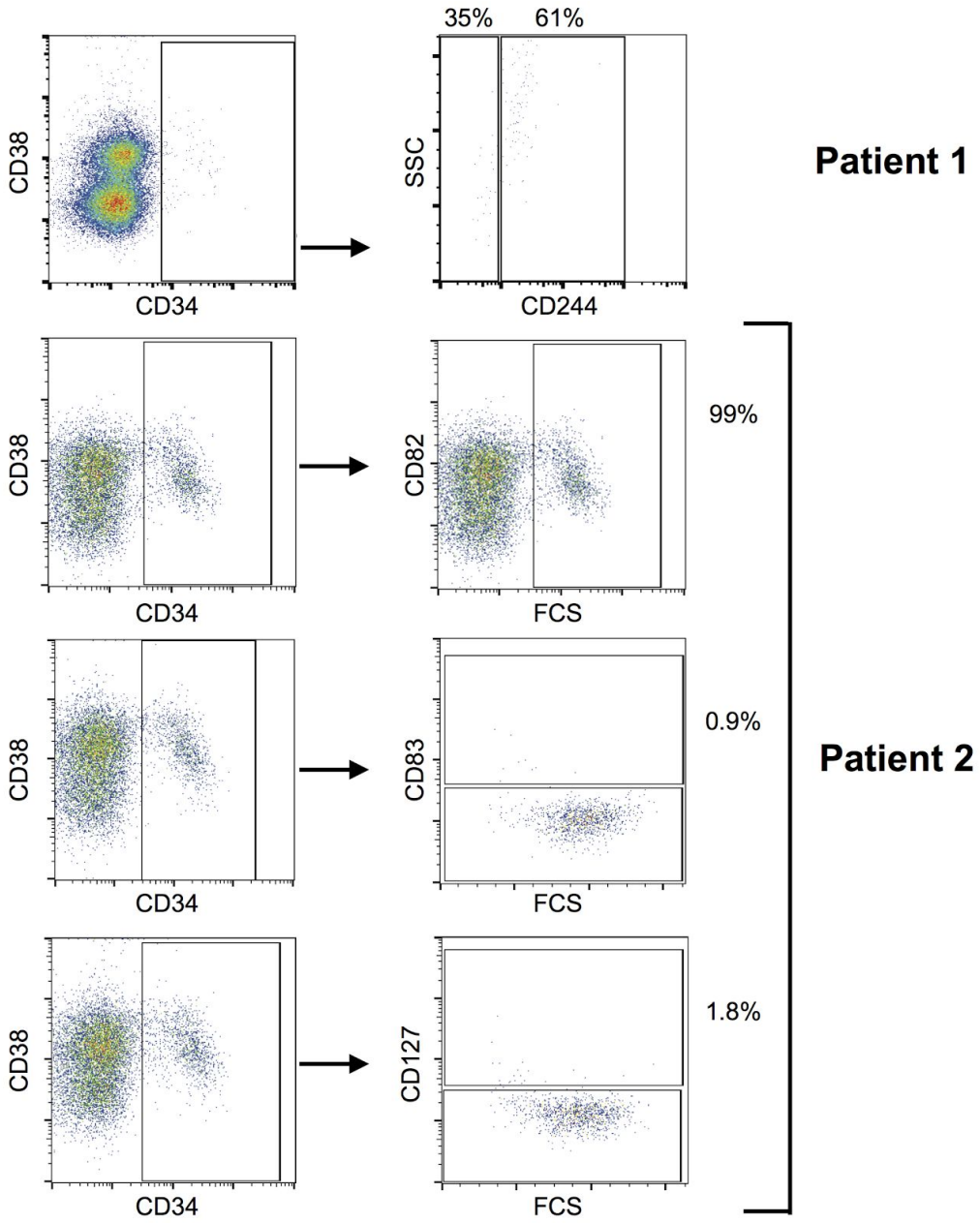

Figure S21. Sorting of haematopoietic stem and progenitor cells from patient 1 and 2 using antibodies that target surface markers identified using SC3. Our analysis suggests that CD83 should be specific for WT clones, CD127 and CD244 for the Tet2 only mutant clones, while CD82 is specific to double mutant clones. Percentages account for $CD38^{+}CD34^{+}$ cells positive for the indicated surface marker.

Table S1. 99% quantiles of AUROC density distributions (Fig. S13) obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods).

Table S2. SC3 output file containing all 3,500 identified marker genes from the Deng dataset.

Table S3. SC3 clustering, marker genes, DE genes (from clusters 4 and 8) and gene ontology and pathway enrichment analysis (of DE genes from clusters 4 and 8) results of the Macosko *et al* dataset.

Table S4. A summary of the patient information. ET, essential thrombocytosis; MF, myelofibrosis

Table S5. Marker genes for the comparison of patient 1 & 2, gene ontology and pathway enrichment analysis results of marker genes for patient 1 & 2