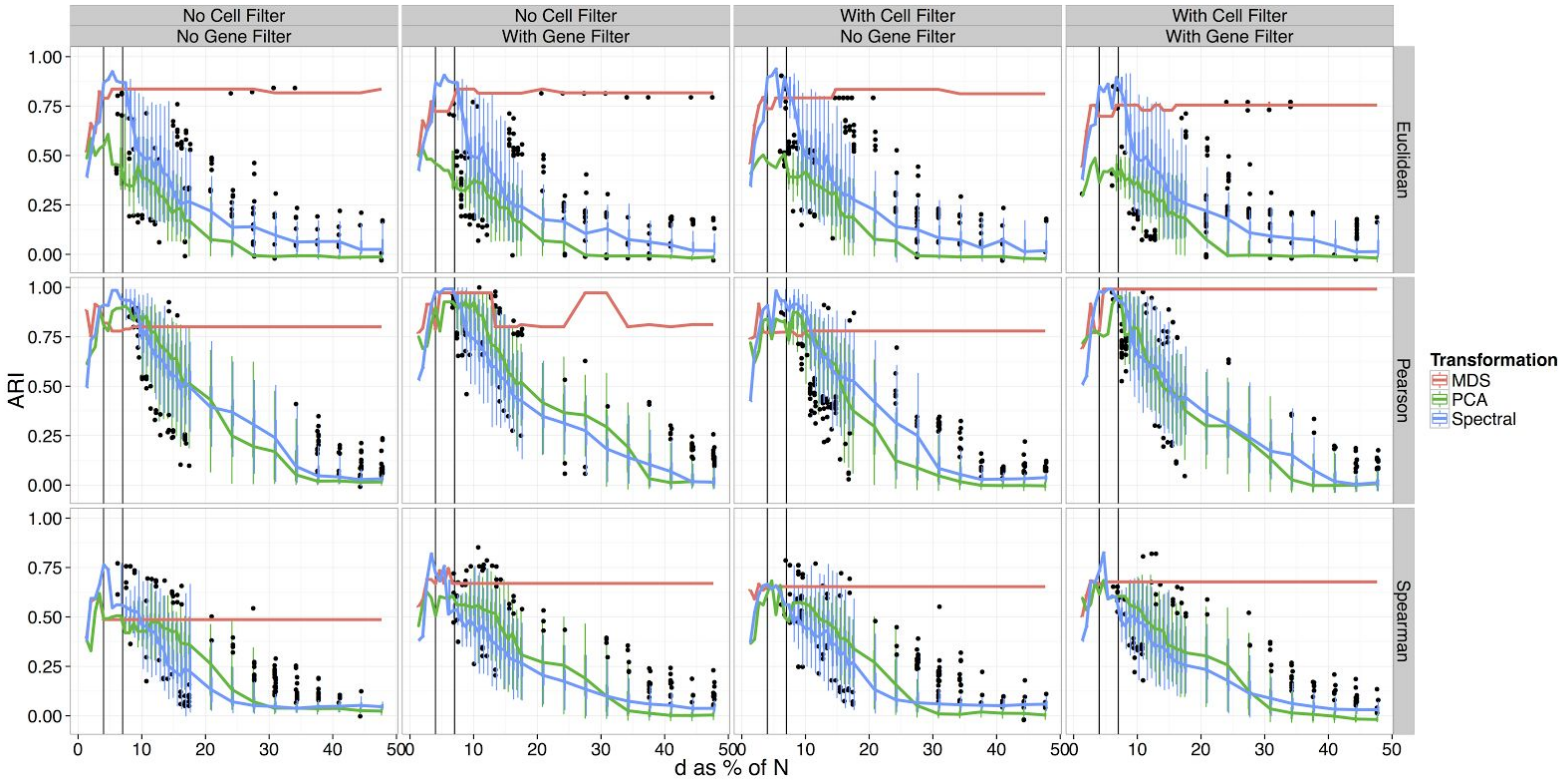

Figure S1. **Boxplots of 100 realizations of the SC3 clustering on the Ting dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 149). 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.

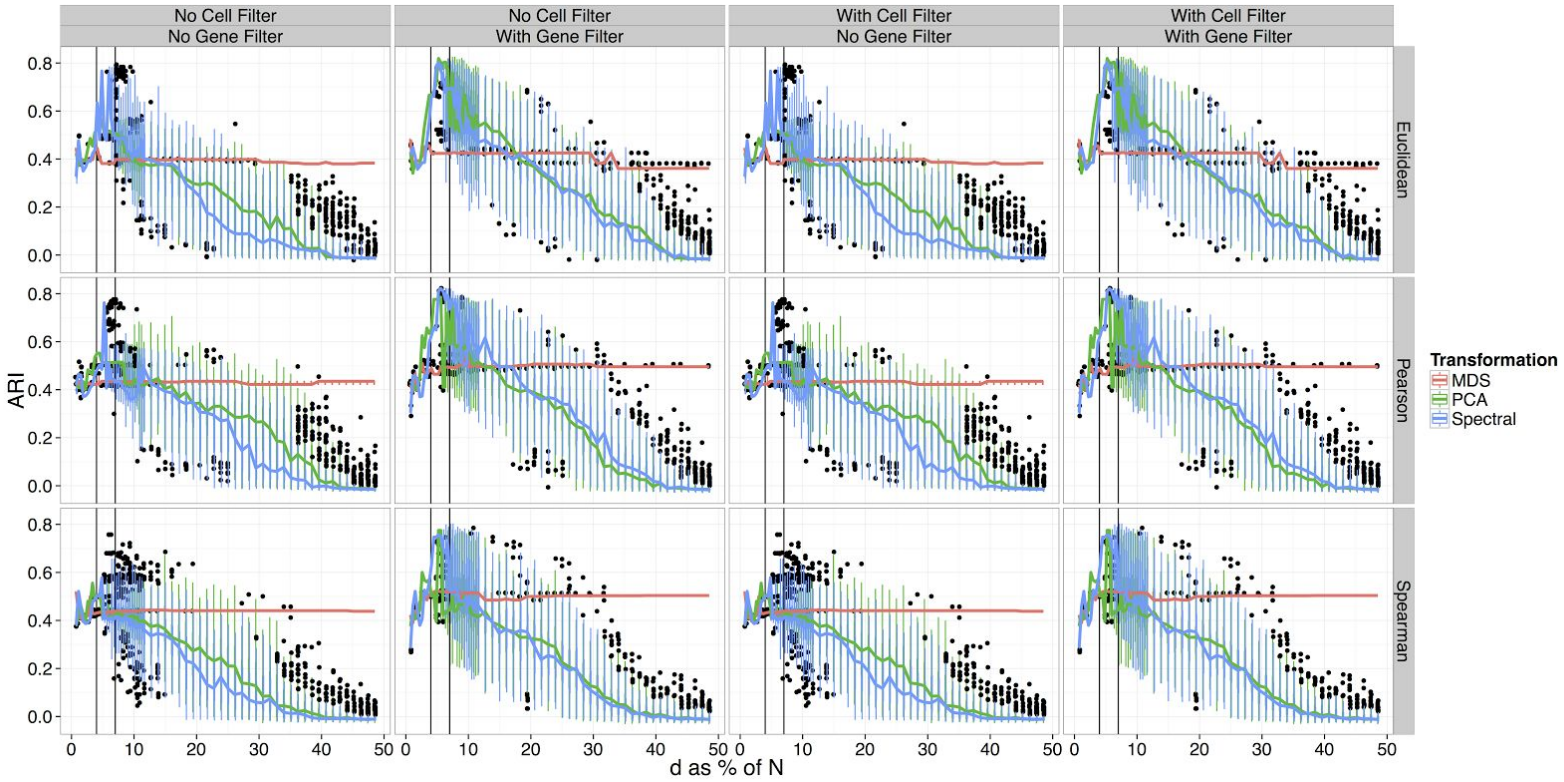

Figure S2. **Boxplots of 100 realizations of the SC3 clustering on the Deng dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. 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.

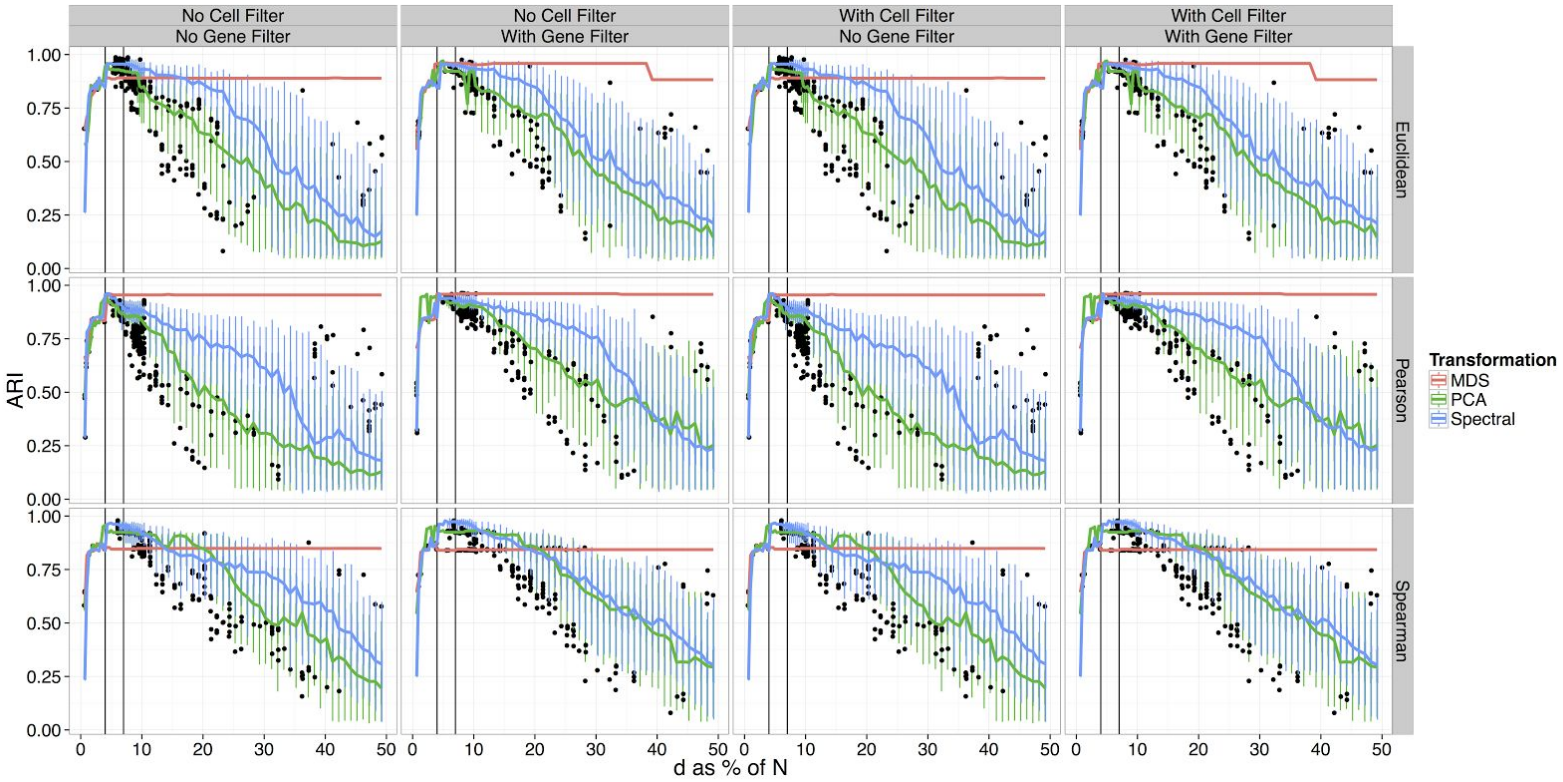

Figure S3. **Boxplots of 100 realizations of the SC3 clustering on the Pollen dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. 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.

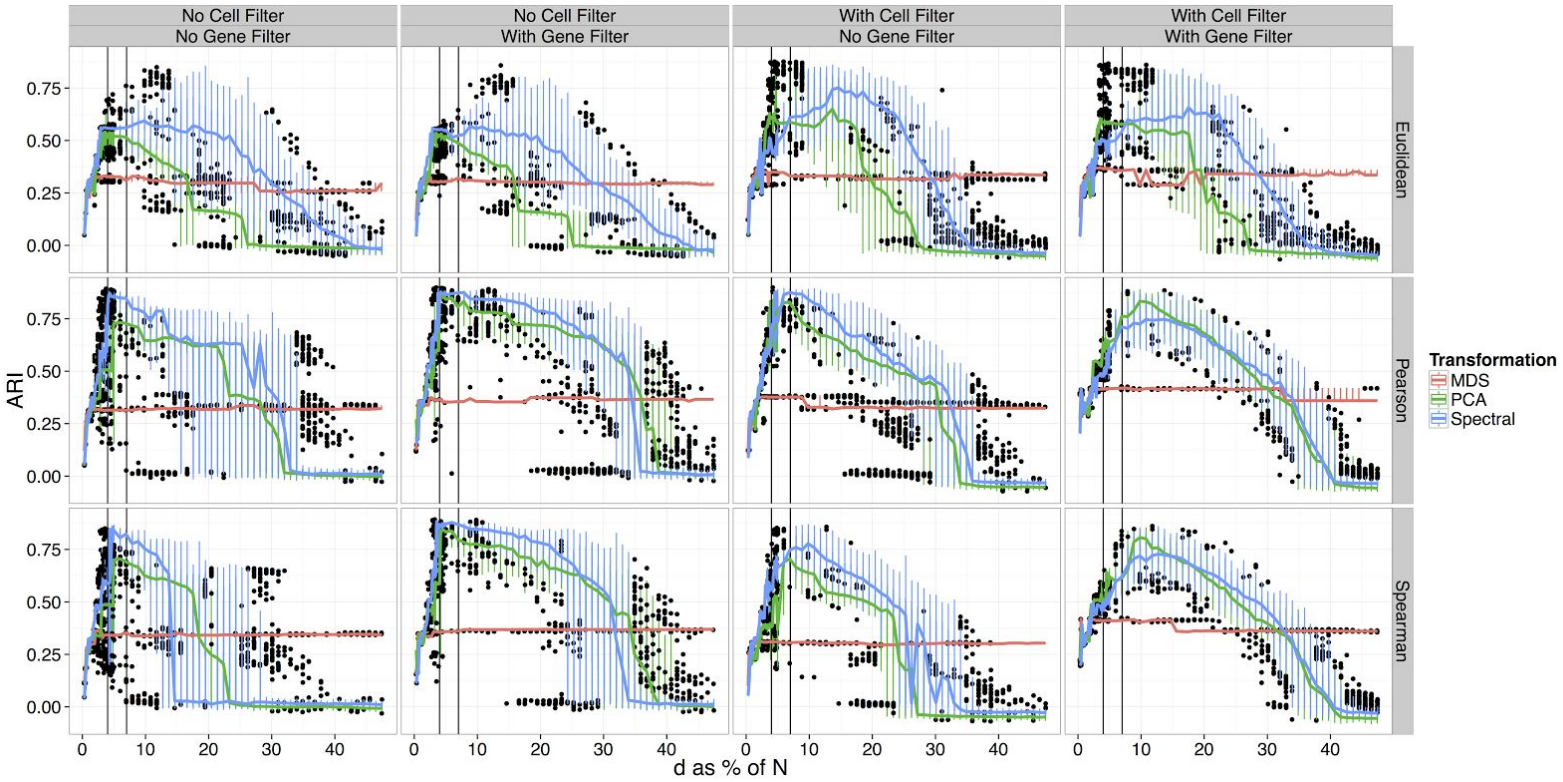

Figure S4. **Boxplots of 100 realizations of the SC3 clustering on the Usoskin dataset.** For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N* (*N* = 622). 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.

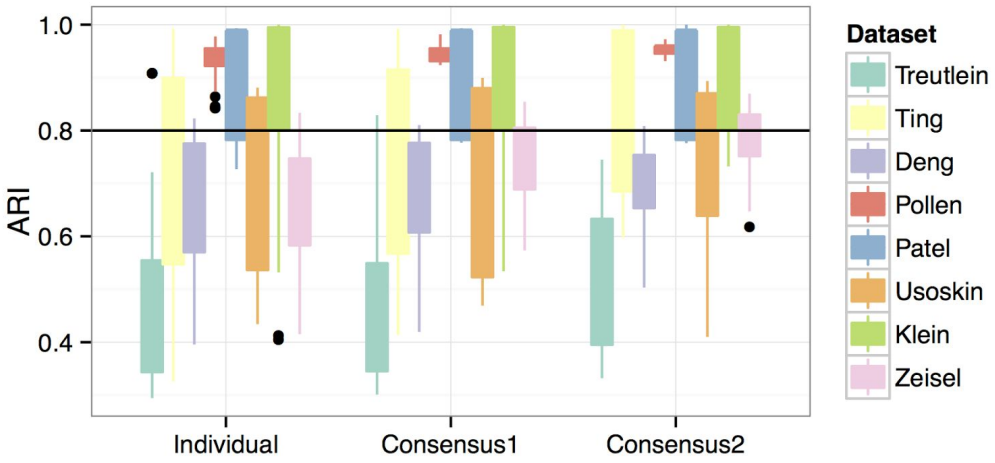

Figure S5. **Effect of consensus clustering on ARI.** Boxplots of 100 realizations of the SC3 clustering of all validation datasets (Fig. 1b). *Individual* corresponds to clustering without consensus approach. *Consensus1* corresponds to the consensus clustering over the *d* range. *Consensus2* corresponds to the consensus clustering over the parameter set (see Methods for more details). The black line corresponds to ARI=0.8. 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.

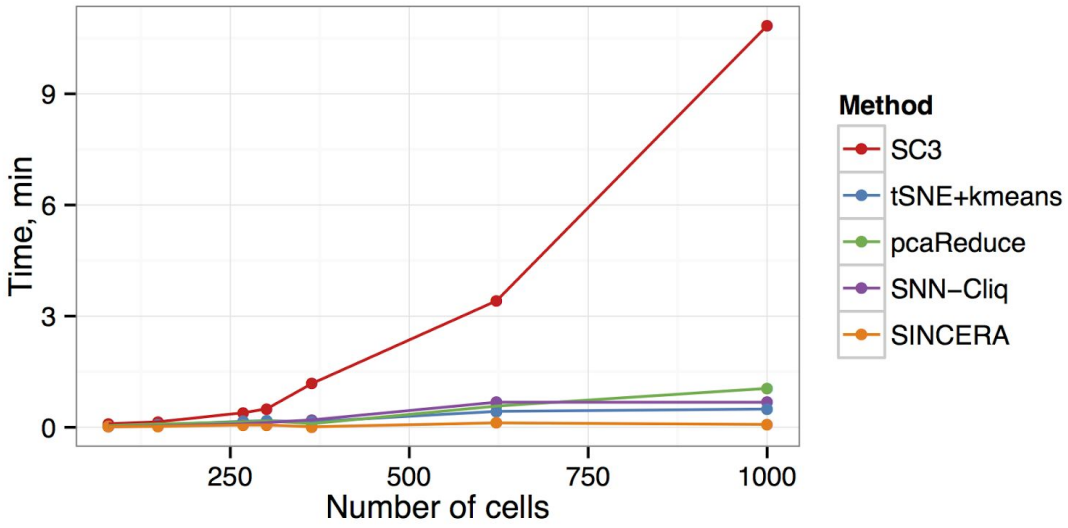

Figure S6. **Comparison of running times of SC3 with existing methods for different number of cells in an input expression matrix.** These measurements were performed on a MacBook Pro (Mid 2014), OS X Yosemite 10.10.5 with 2.8 GHz Intel Core i7 procession, 16 GB 1600 MHz DDR3 of RAM.

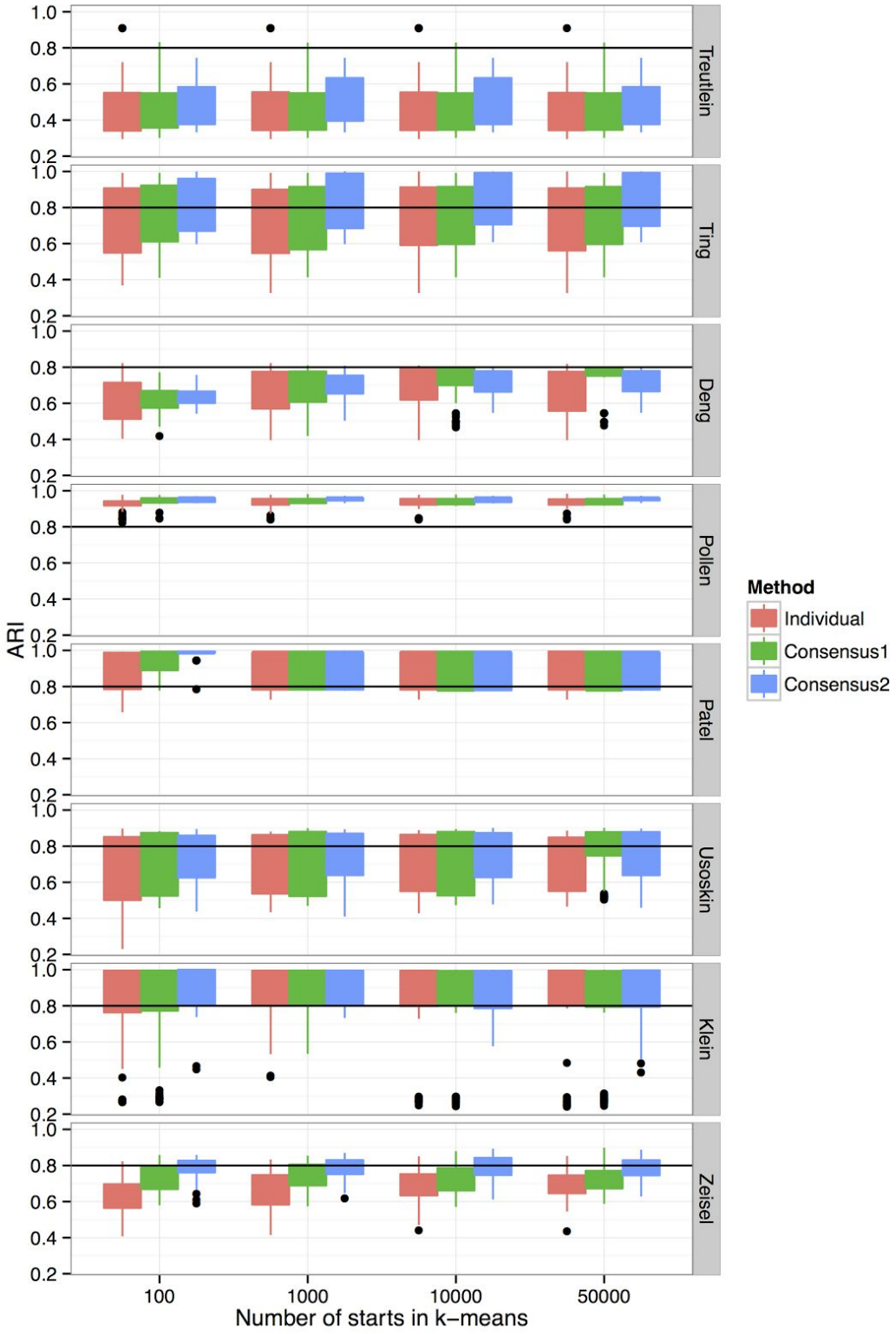

Figure S7. **Effect of the number of starts on ARI.** Boxplots of 100 realizations of the SC3 clustering of all validation datasets (Fig. 1b). Note that for Klein and Zeisel datasets a random sample of 1000 cells was used when the number of starts were 10000 and 50000 because it was unfeasible to calculate the ARI in a reasonable time. 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.

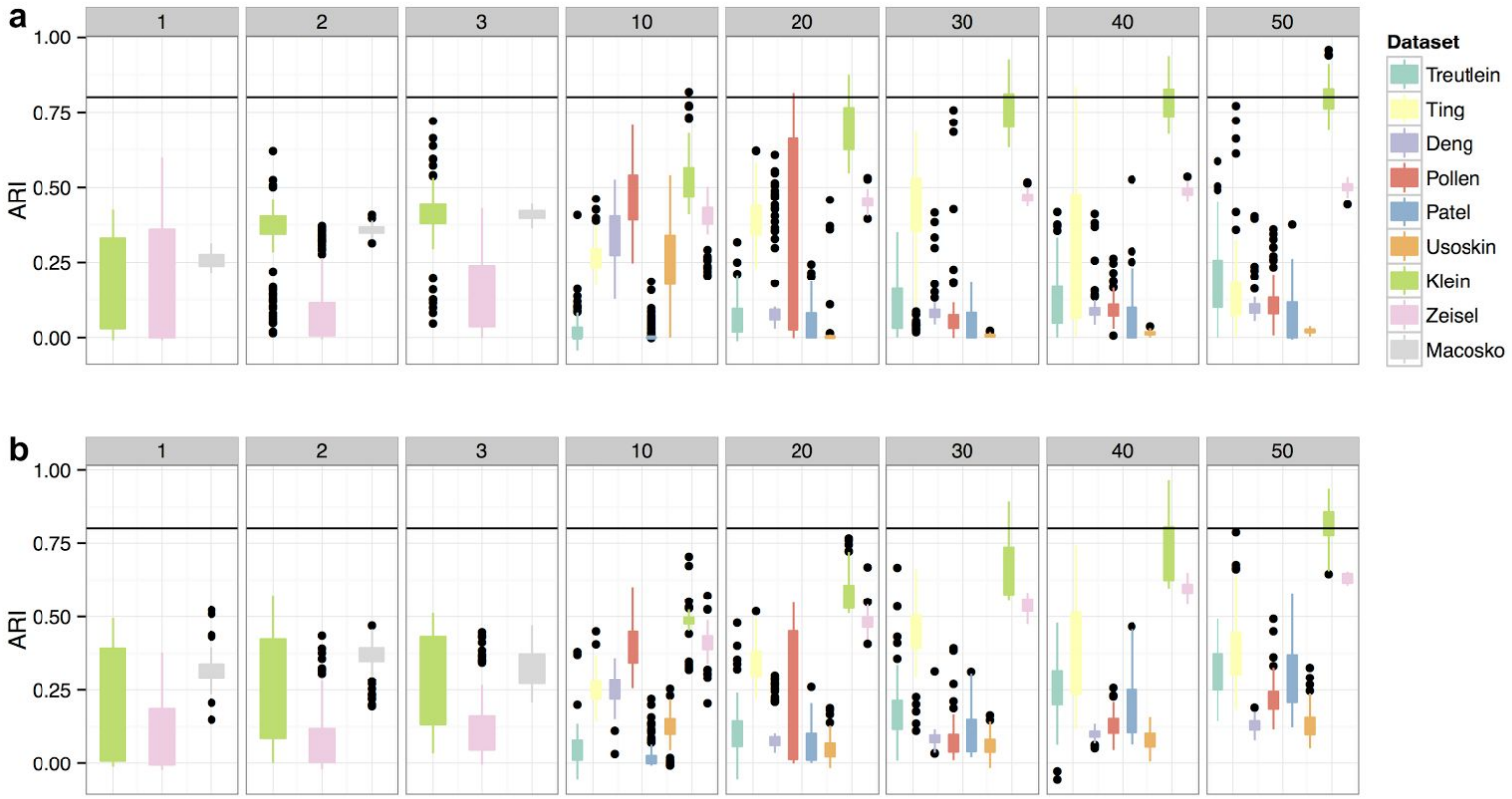

Figure S8. **Results for the hybrid clustering approach using a polynomial kernel.** The black line corresponds to ARI=0.8. Numbers in grey boxes correspond to the number of training cells as % of *N.* (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**b**) ARI levels for SVM prediction when labels calculated by SC3 are used for training. 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.

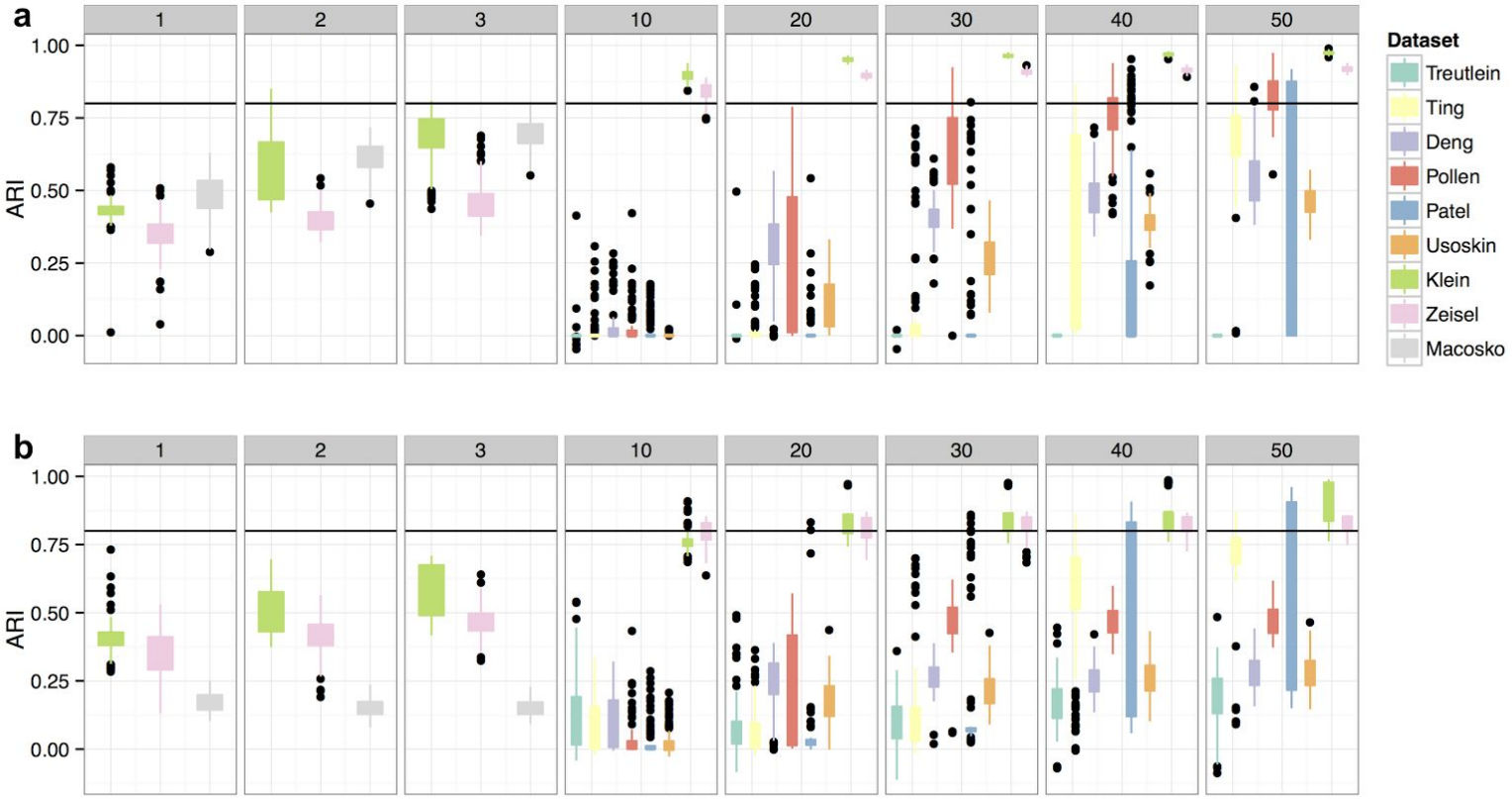

Figure S9. **Results for the hybrid clustering approach using a radial kernel.** The black line corresponds to ARI=0.8. Numbers in grey boxes correspond to the number of training cells as % of *N*. (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**b**) ARI levels for SVM prediction when labels calculated by SC3 are used for training. 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.

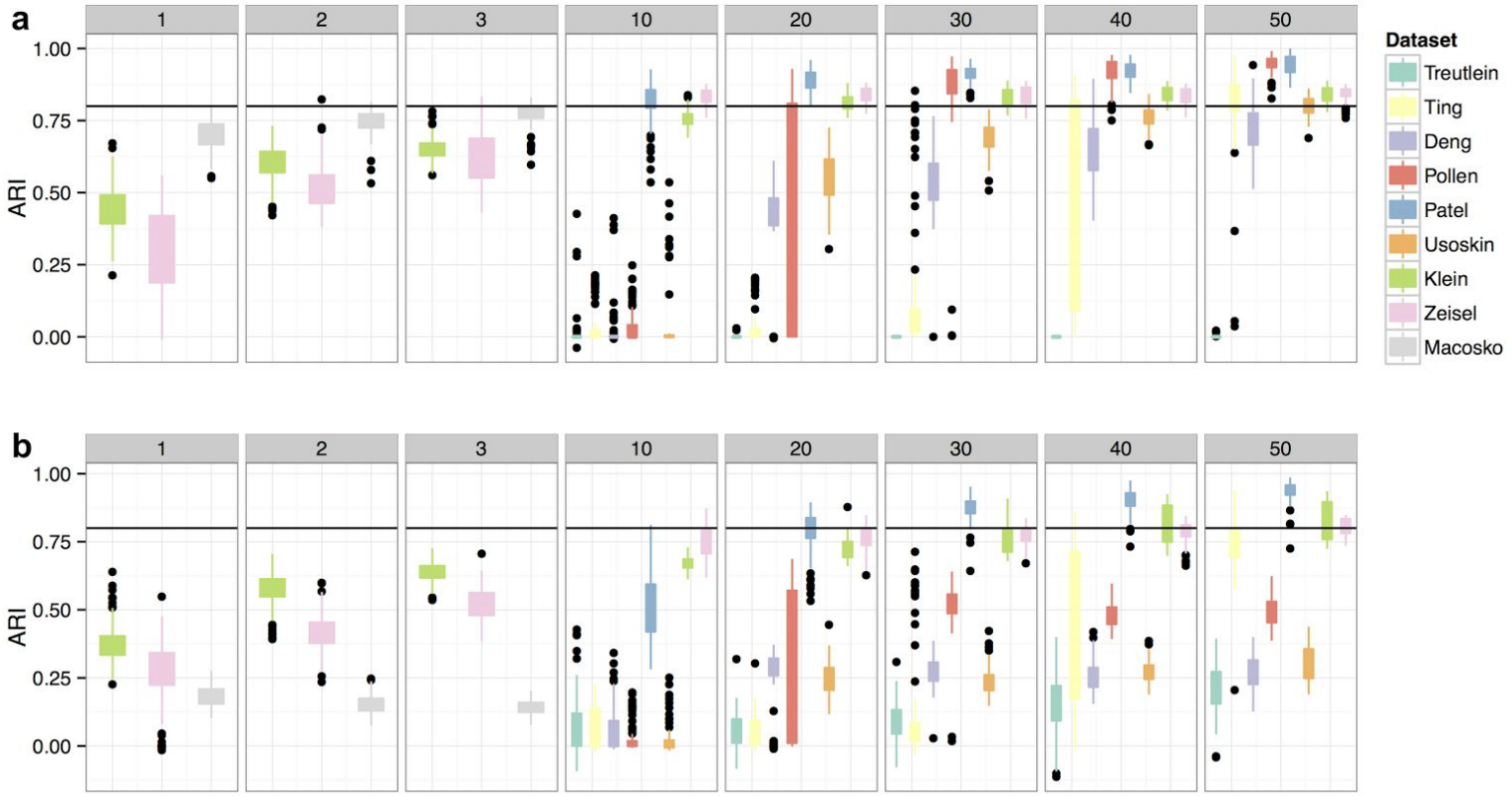

Figure S10. **Results for the hybrid clustering approach using a sigmoid kernel.** The black line corresponds to ARI=0.8. Numbers in grey boxes correspond to the number of training cells as % of *N.* (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**b**) ARI levels for SVM prediction when labels calculated by SC3 are used for training. 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.

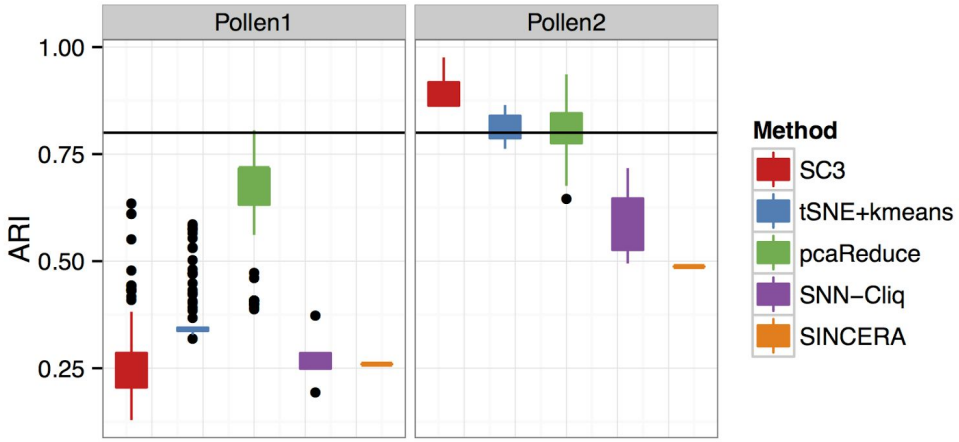

Figure S11. **Benchmarking of SC3 against existing methods using the alignment provided by Zurauskiene and Yau(Zurauskiene and Yau 2015).** Boxplots are the results of 100 realisations of a given method. Pollen1, Pollen2 correspond to different levels of hierarchies as described by Zurauskiene and Yau(Zurauskiene and Yau 2015). The black line corresponds to ARI=0.8. 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.

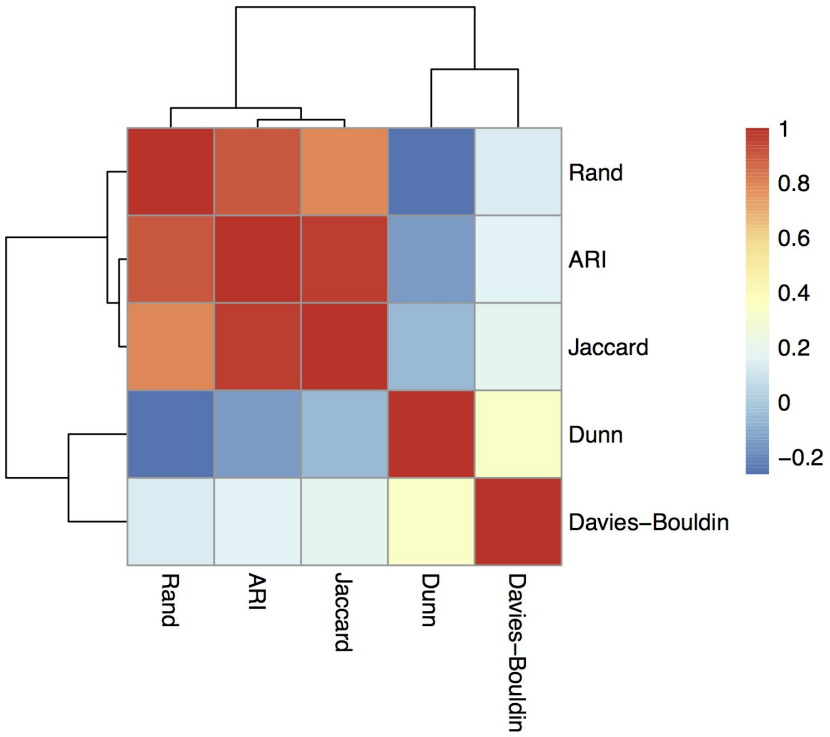

Figure S12. **Correlations between different external and internal measures of clustering.** Correlations are based on all results of SC3 clustering presented in Figs. S1-S8.

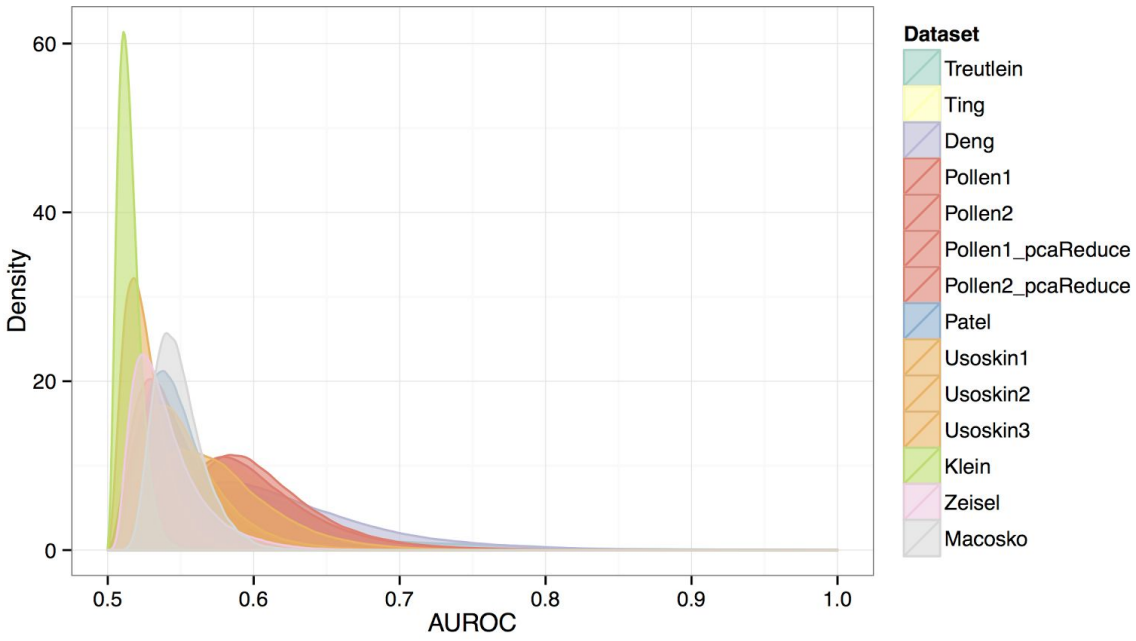

Figure S13. Density of distributions of AUROC obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods).

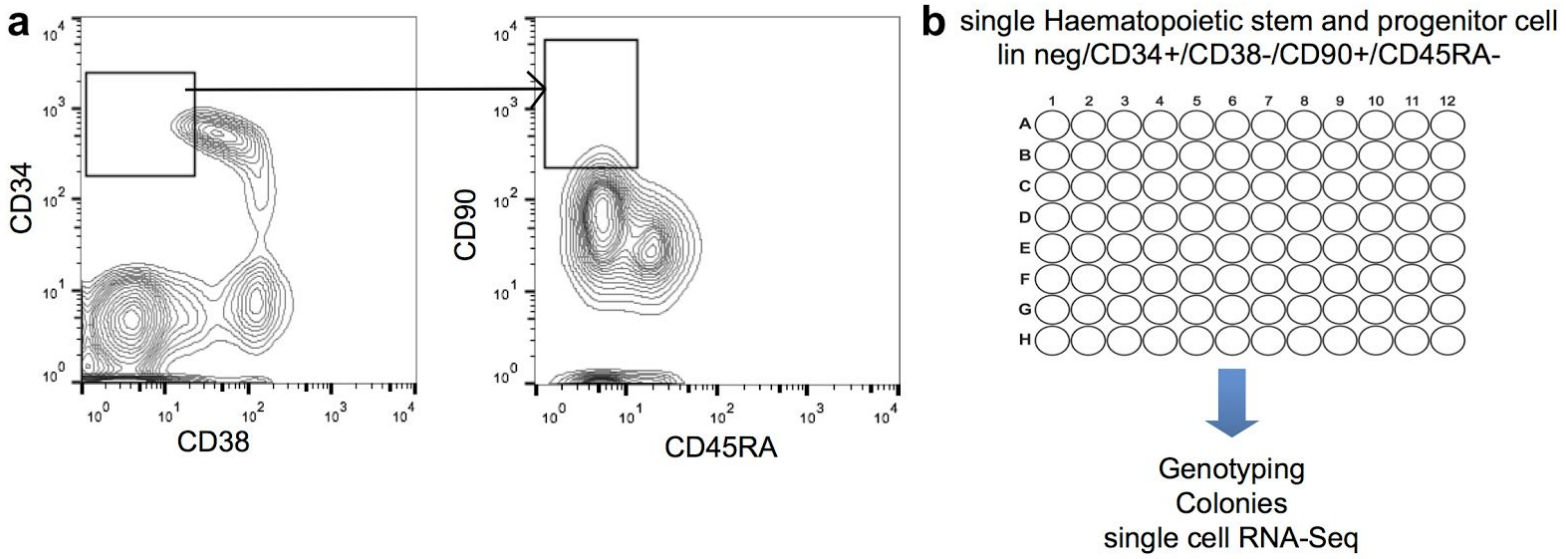

Figure S14. **Cell sorting procedure for patients.** (**a**) Contour plots describing the sorting strategy for isolating HSCs in patient 2 (the same was done for patient 1). CD34, CD38, CD90 and CD45RA expression is displayed using a log scale. (**b**) Lineage negative, CD34+/CD38-/CD90+/CD45RA- single cells were sorted into individual wells for scRNA-Seq or colony growth in cytokine cocktail allowing progenitor cell expansion.

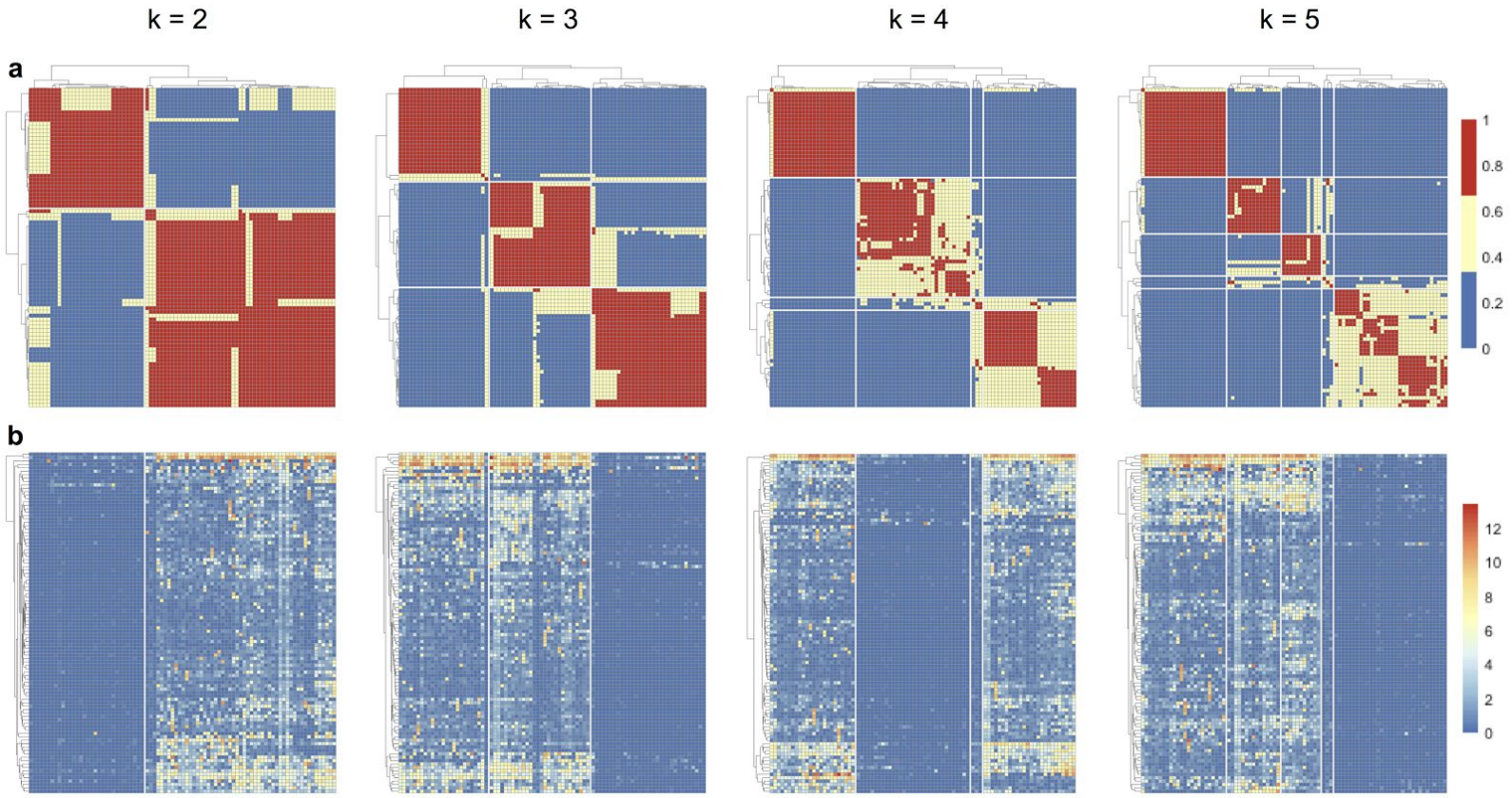

Figure S15. **Clustering of scRNA-seq data from patient 1.** (**a**) Consensus matrices corresponding to different values of *k*. (**b**) Heatmaps of the expression matrix (after Gene Filter and Log-transformation, Methods) corresponding to different values of *k*. Genes are clustered by *k*-means with *k* = 100 and the heatmap represents the expression levels of the cluster centers.

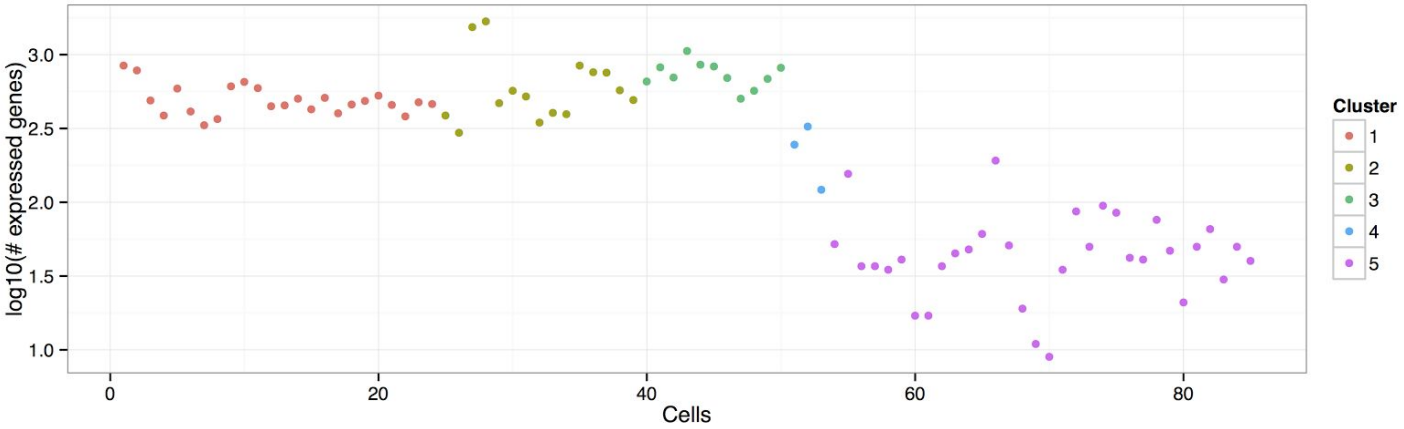

Figure S16. **Number of expressed genes in all cells from patient 1.**

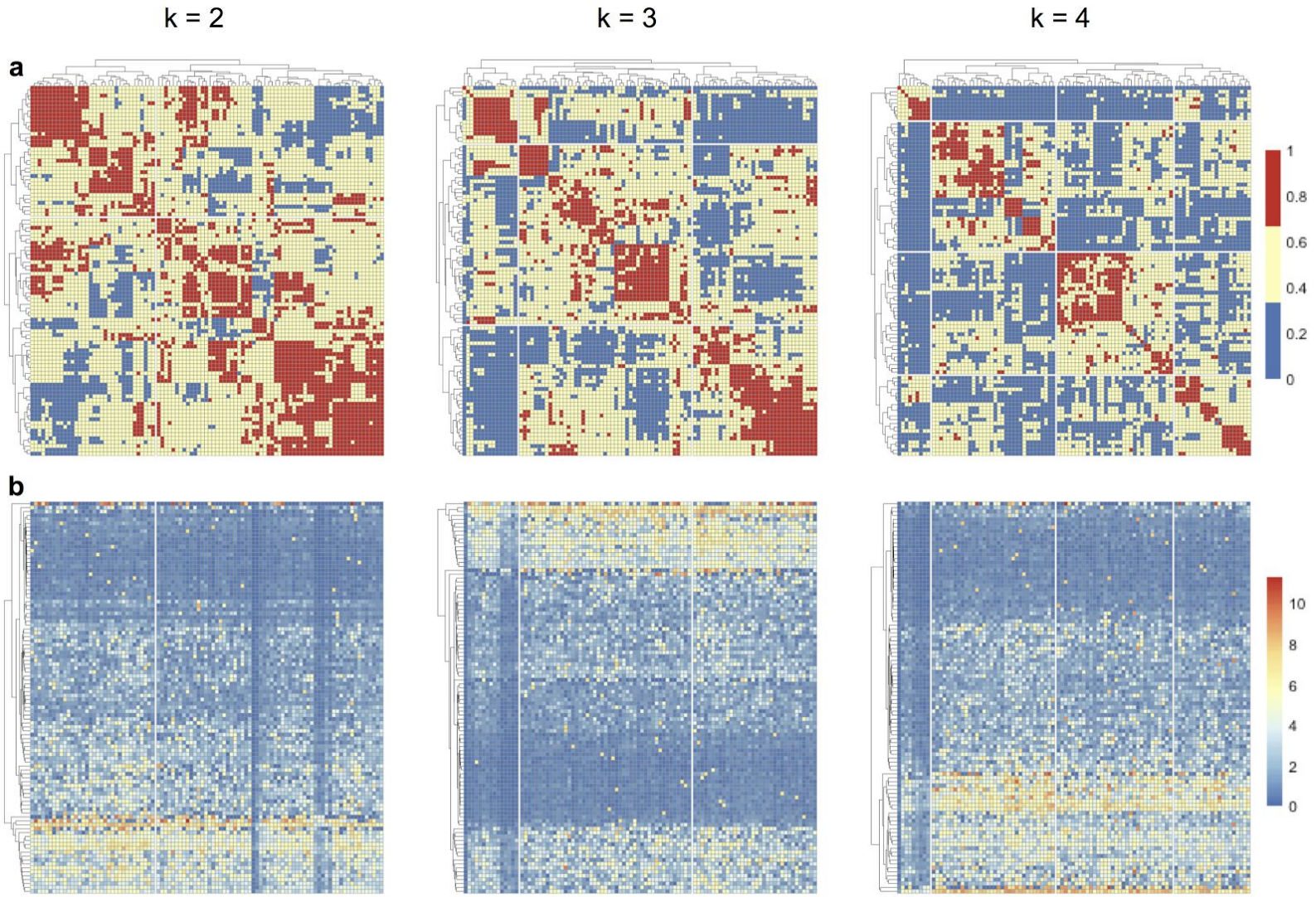

Figure S17. **Clustering of scRNA-seq data from patient 2.** (**a**) Consensus matrices corresponding to different values of *k*. (**b**) Heatmaps of the expression matrix (after Gene Filter and Log-transformation, Methods) corresponding to different values of *k*. Genes are clustered by *k*-means with *k* = 100 and the heatmap represents the expression levels of the cluster centers.

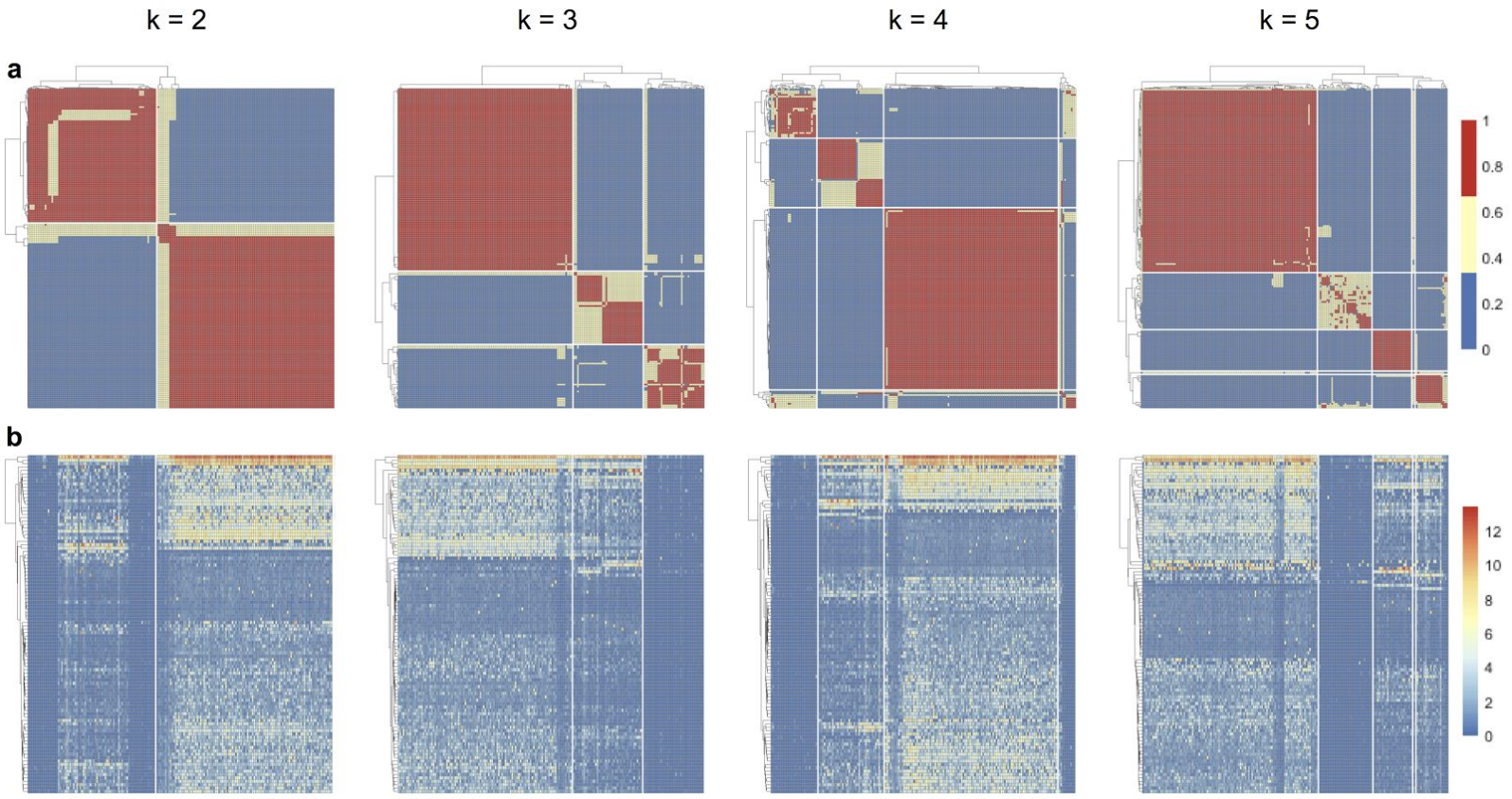

Figure S18. **Clustering of the combined scRNA-seq data from patient 1 and patient 2.** (**a**) Consensus matrices corresponding to different values of *k*. (**b**) Heatmaps of the expression matrix (after Gene Filter and Log-transformation, Methods) corresponding to different values of *k*. Genes are clustered by *k*-means with *k* = 100 and the heatmap represents the expression levels of the cluster centers.

Dataset	99% quantile of AUROC density distribution
Treutlein	0.83
Deng	0.82
Pollen2	0.74
Pollen2_pcaReduce	0.73
Ting	0.72
Usoskin3	0.70
Usoskin2	0.65
Pollen1_pcaReduce	0.64
Zeisel	0.62
Pollen1	0.61
Patel	0.60
Macosko	0.60
Usoskin1	0.57
Klein	0.54

Table S1. 99% quantiles of AUROC density distributions (Fig. S18) 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.

Driver Mutations	patient ID	Gender	Diagnosis	Age at diagnosis	Disease duration at assay (years)	Therapy at assay
Tet2 c.3120_3121het _insA Jak2V617F	1	M	ET	75	12	hydroxycarbamide
Tet2 c.5447 T>A p.L1816X Jak2V617F	2	F	post-ET MF	78	14	pacritinib

Table S3. A summary of the patient information. ET, Essential thormocytosis; MF, myelofibrosis