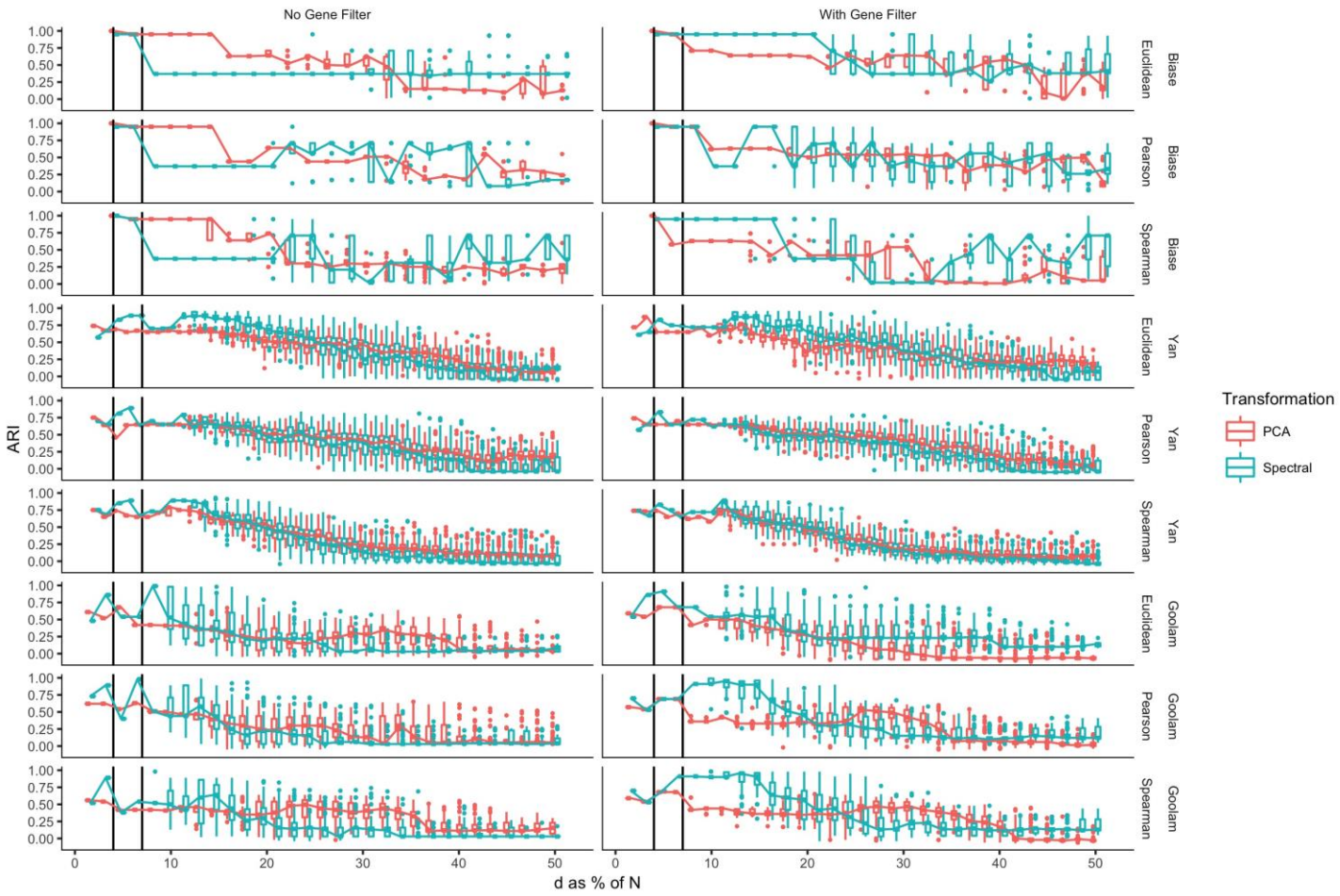
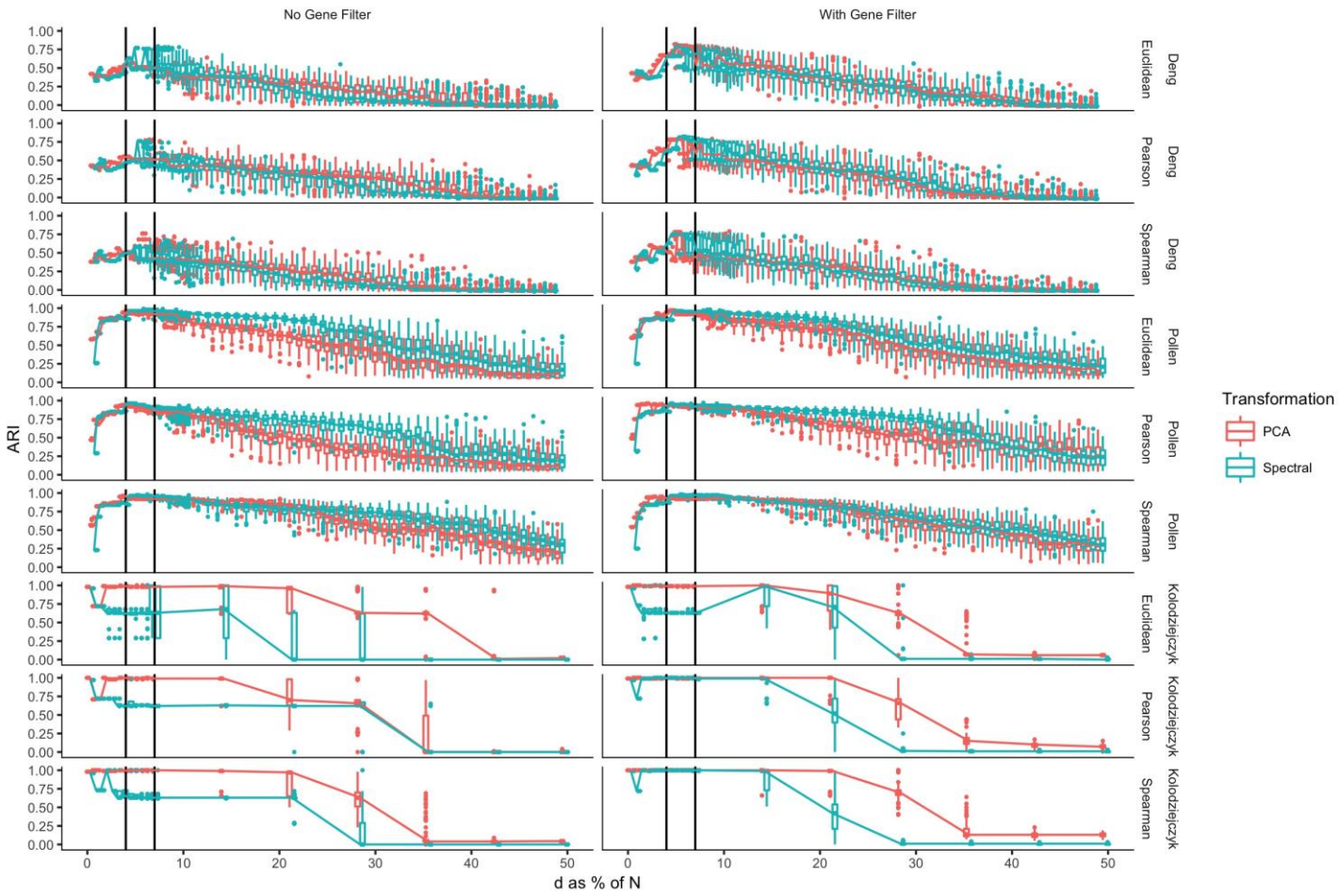


Supplementary Figure 1

Boxplots of 100 realizations of the SC3 clustering on the Biase, Yan and Goolam datasets.

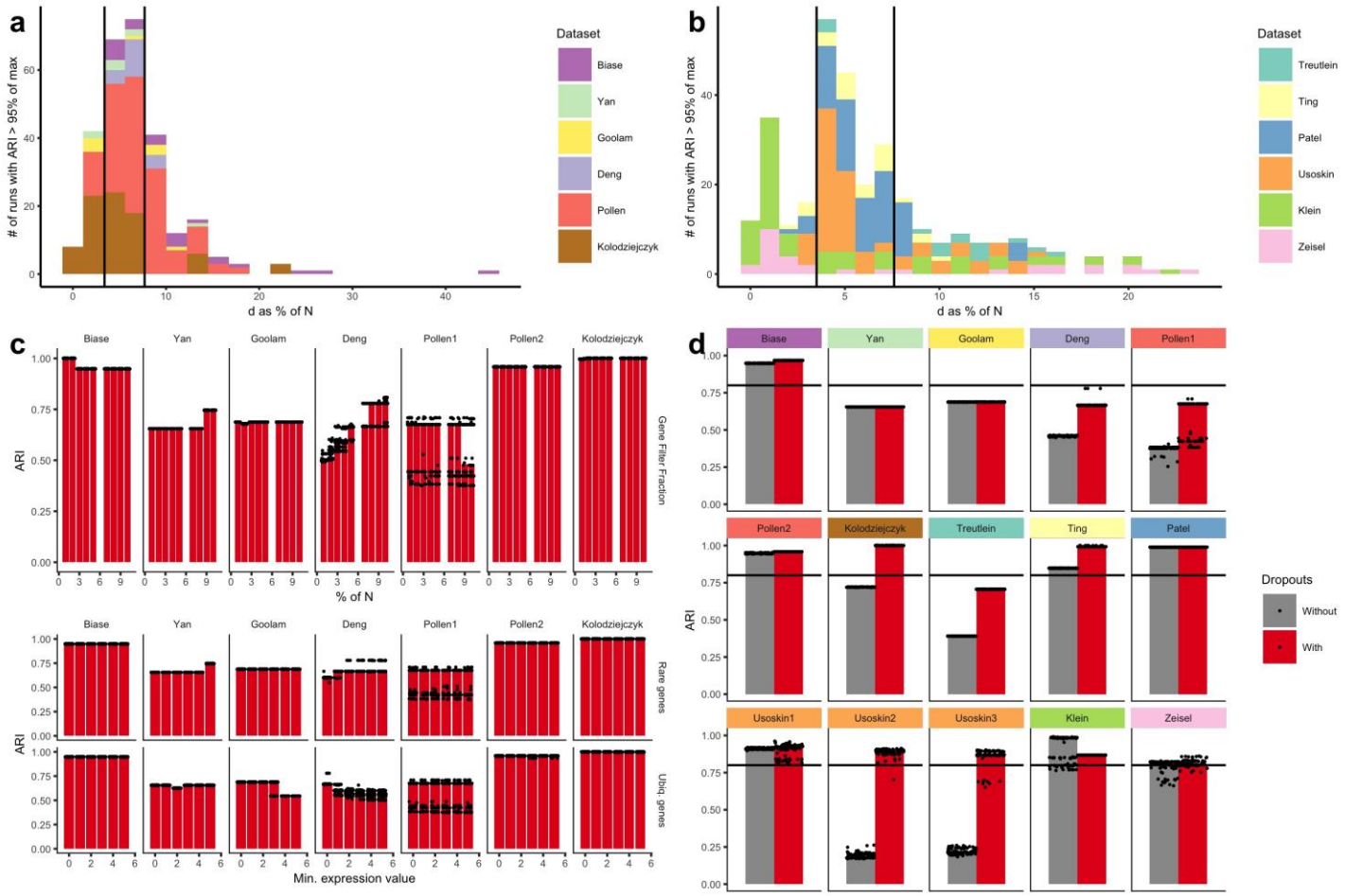
For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N.* Dots represent 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.



Supplementary Figure 2

Boxplots of 100 realizations of the SC3 clustering on the Deng, Pollen and Kolodziejczyk datasets.

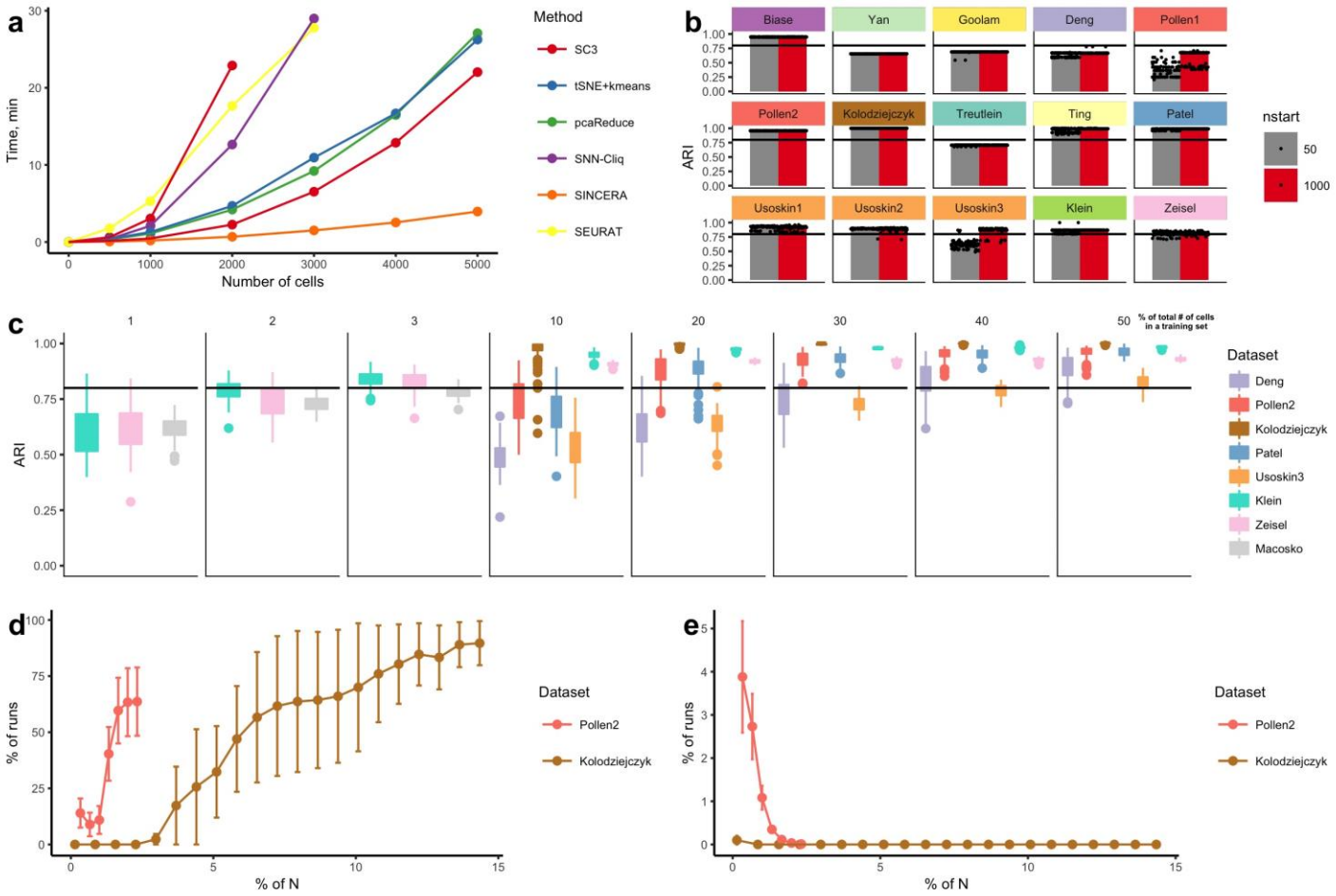
For clarity, lines are drawn through the medians of the boxplots. The x-axis shows the number of eigenvectors *d* of the transformed distance matrix as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% of *N* and *d* = 7% of *N.* Dots represent 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.



Supplementary Figure 3

Exploration of SC3 pipeline parameters

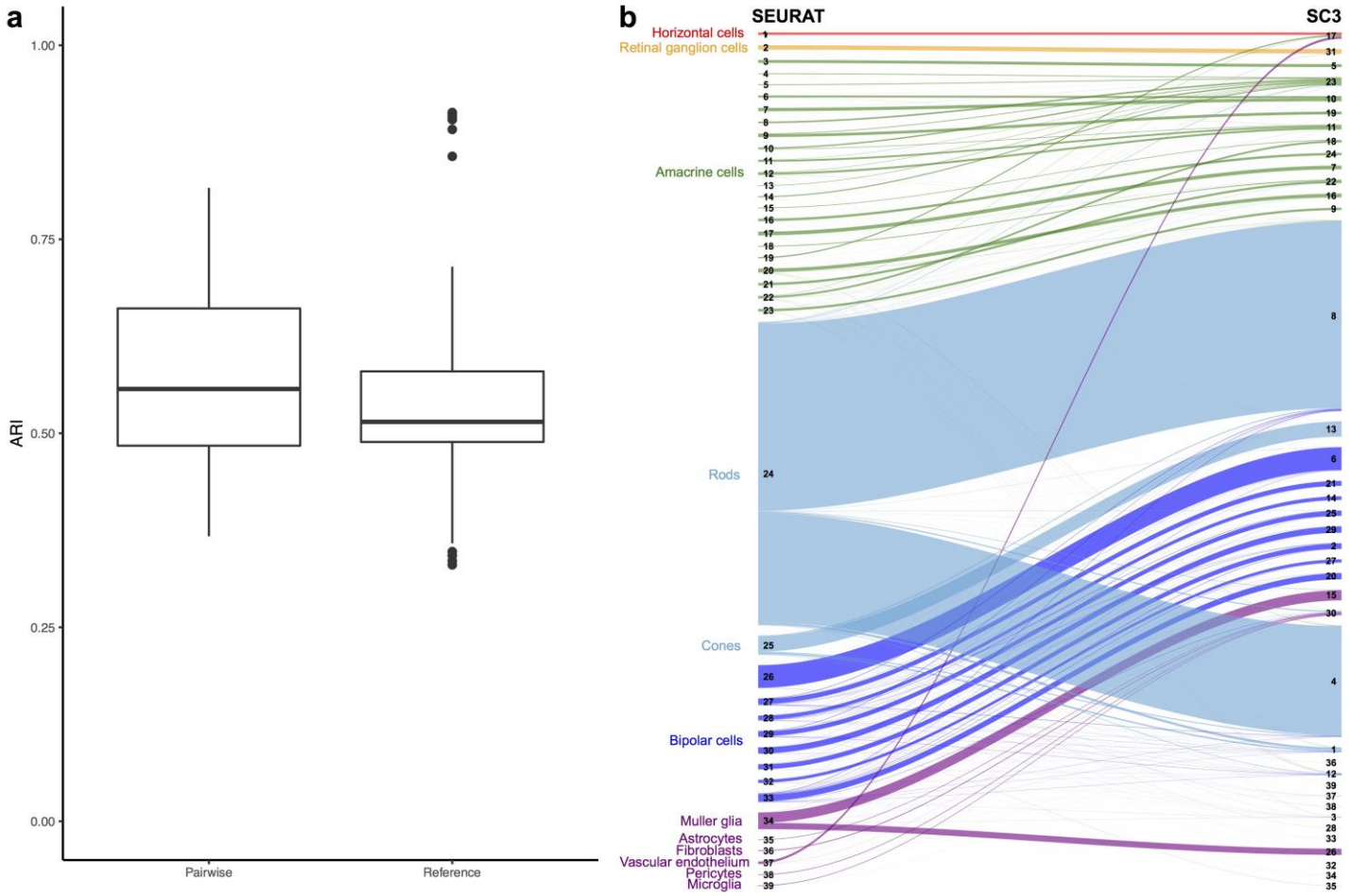
(a) Histogram of the *d* values where ARI>.95 is achieved for the downsampled (by a factor of 10, Methods) gold standard datasets from Fig. 1b. The black vertical lines indicate the interval *d* = 4-7% of the total number of cells *N*, showing high accuracy in the classification; (b) Histogram of the *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; (c) Exploration of the gene filter parameters (see Methods for more details). Dots represent individual clustering runs. Bars correspond to the median of the dots; (d) The effect of dropouts in the distance calculations step on the accuracy of SC3 clustering (Methods for more details). Dots represent individual clustering runs. Bars correspond to the median of the dots. Red and grey colours correspond to clustering with and without dropouts. The black line corresponds to ARI=0.8.



Supplementary Figure 4

Scalability, accuracy and rare cell-type detection rate of SC3 and benchmarking of the hybrid SC3

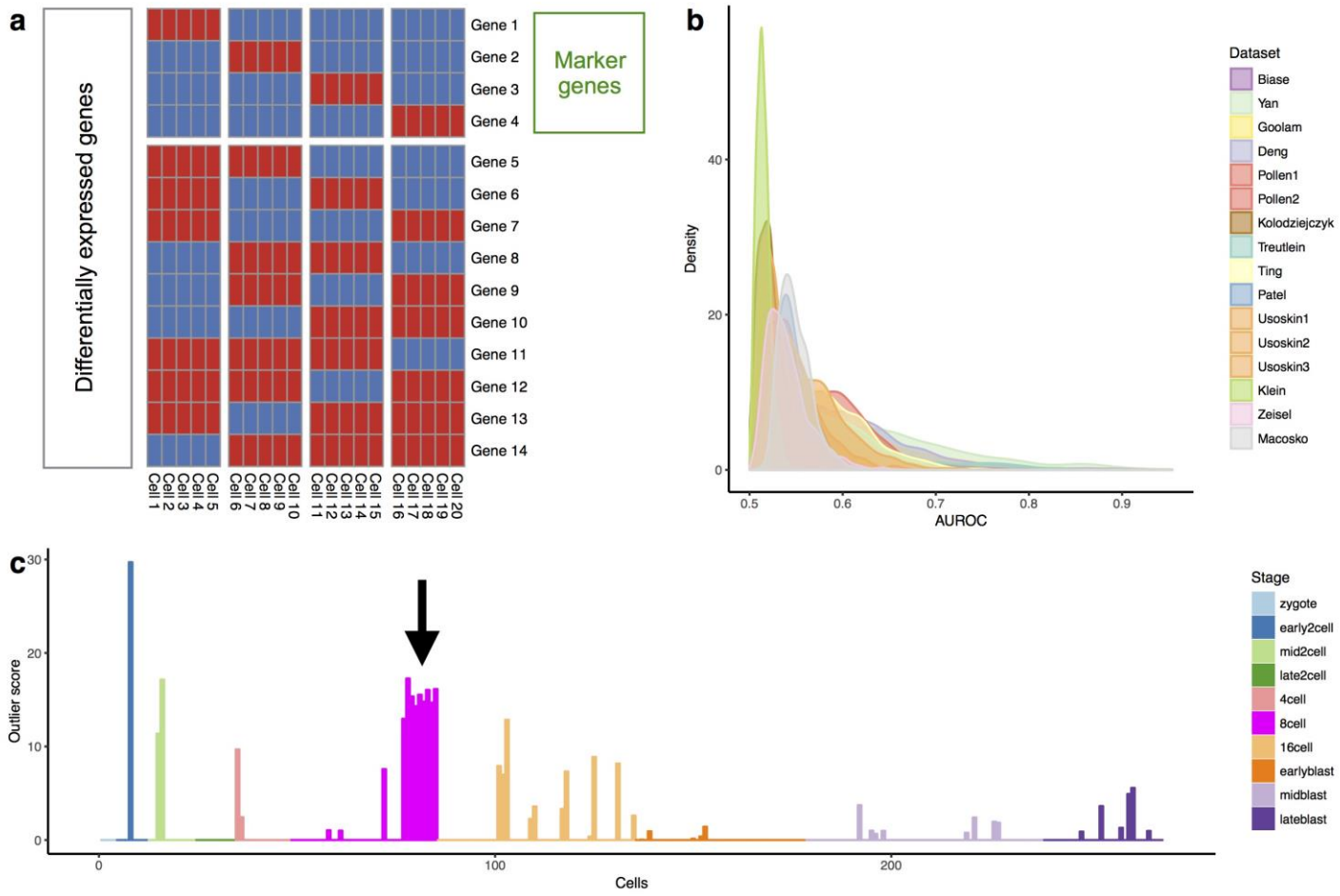
(a) Run times for different clustering methods as a function of the number of cells (*N*). All methods were run on a MacBook Pro (Mid 2014), OS X Yosemite 10.10.5 with 2.8 GHz Intel Core i7 processor, 16 GB 1600 MHz DDR3 of RAM. Two results shown for SC3 correspond to nstart=1000 and nstart=50, where nstart is the number of starting points for *k*-means clustering; (b) Reducing the number of k-means runs (nstart) from 1,000 to 50 results only in a slightly worse performance for SC3, yet with significant computational savings, as shown in (a). The black line indicates ARI = 0.8; (c) Using the hybrid SC3 based on reference labels provided by the authors. Same as Fig. 2c 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; (d) Robustness of SC3 for the detection of rare cell-types. For two of the datasets, we remove different percentages of the cells in the rare cell-types. The figure shows the mean fraction of SC3 runs in which all the rare cells were clustered together as a function of the total number of cells in the rare cell-type; (e) 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 (d) 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).



Supplementary Figure 5

Analysis of SC3 clustering of the Macosko dataset

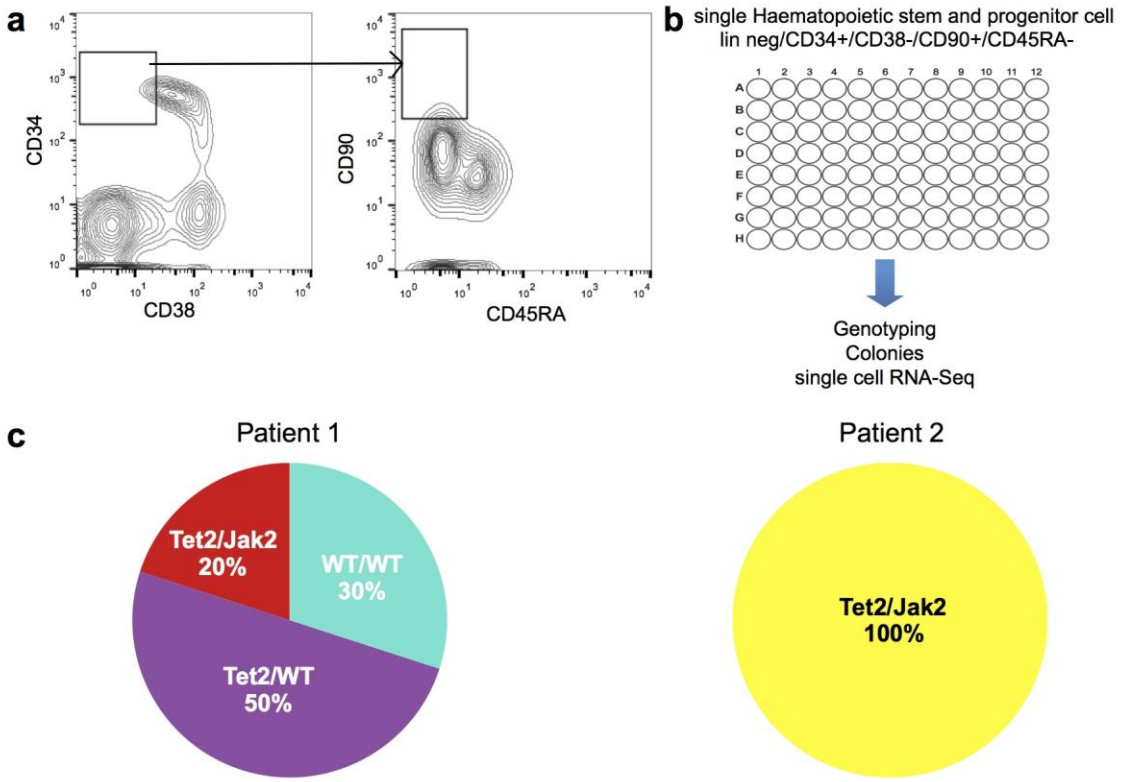
(a) The cells from the Macosko dataset were clustered 100 times using SC3. “Pairwise” indicates the ARIs between the different solutions (a sample of 100 ARIs was taken) obtained and “Reference” indicates the ARI as compared to the labels obtained by Macosko et al.; (b) Sankey diagram comparing the 39 clusters reported by Macosko et al (left) and the 39 clusters obtained with SC3 (right). The widths of the lines linking both sets of clusters correspond to the number of cells they have in common. Colors and cell types as in Macosko et al.



Supplementary Figure 6

Explanation of biological insights provided by SC3

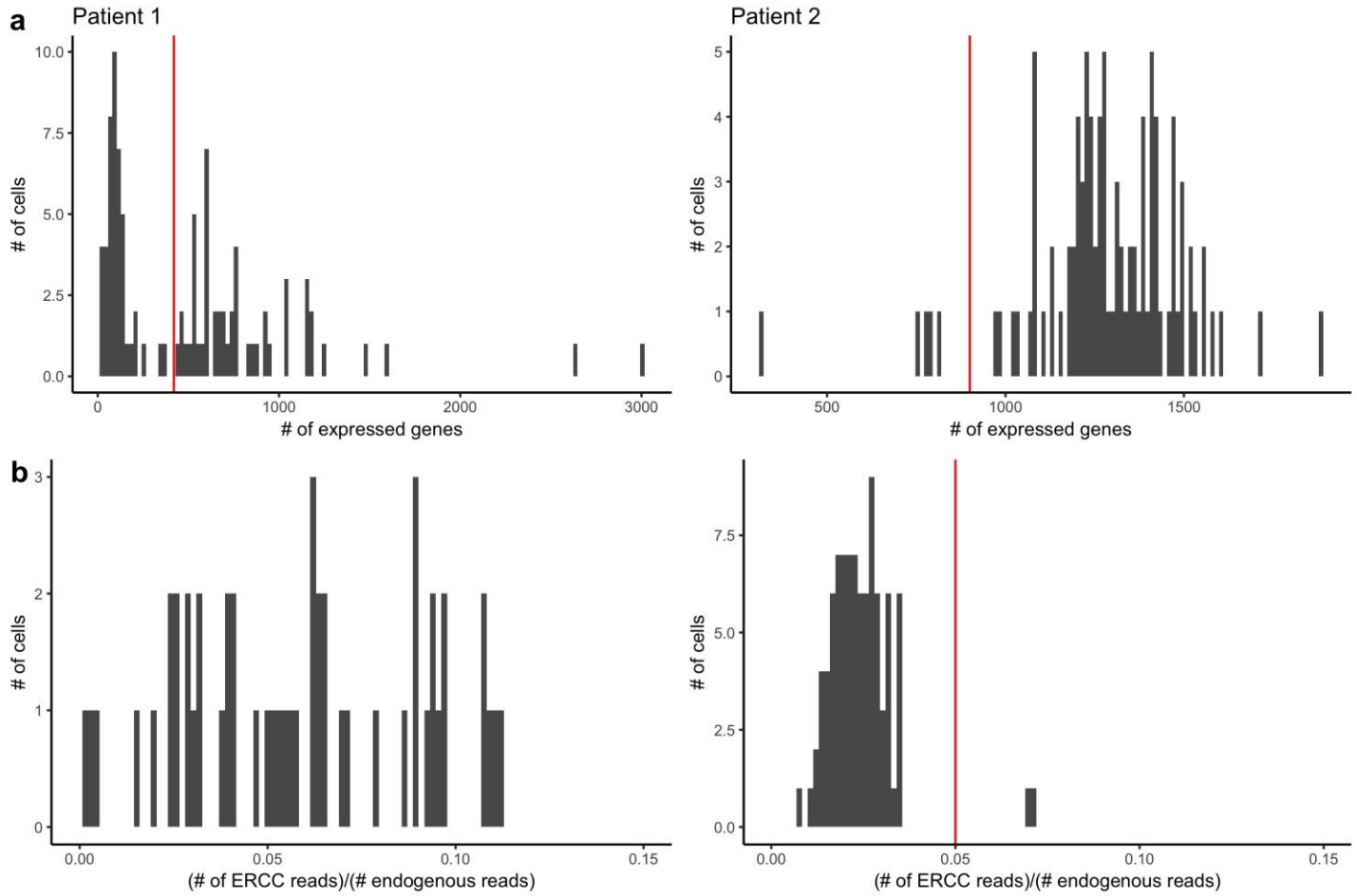
(a) Illustration of the difference between marker genes and differentially expressed genes. In this small example, 20 cells containing 14 genes with binary expression values (blue for 'off', red for 'on') are clustered. Only genes 1-4 can be considered as marker genes, whereas all 14 genes are differentially expressed; (b) Density of distributions of AUROC (sample of 1000 values for each dataset) obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods); (c) Outlier scores for all *N*= 268 cells of the Deng dataset as generated by SC3 (colors correspond to the 10 reference clusters provided by the authors – same as Stage in Fig. 2d). The nine cells with high outlier score in the red cluster (black arrow) were prepared using a different protocol (see text for details), and are thus assigned to a technical artifact.



Supplementary Figure 7

Cell sorting and genotyping procedures for patients

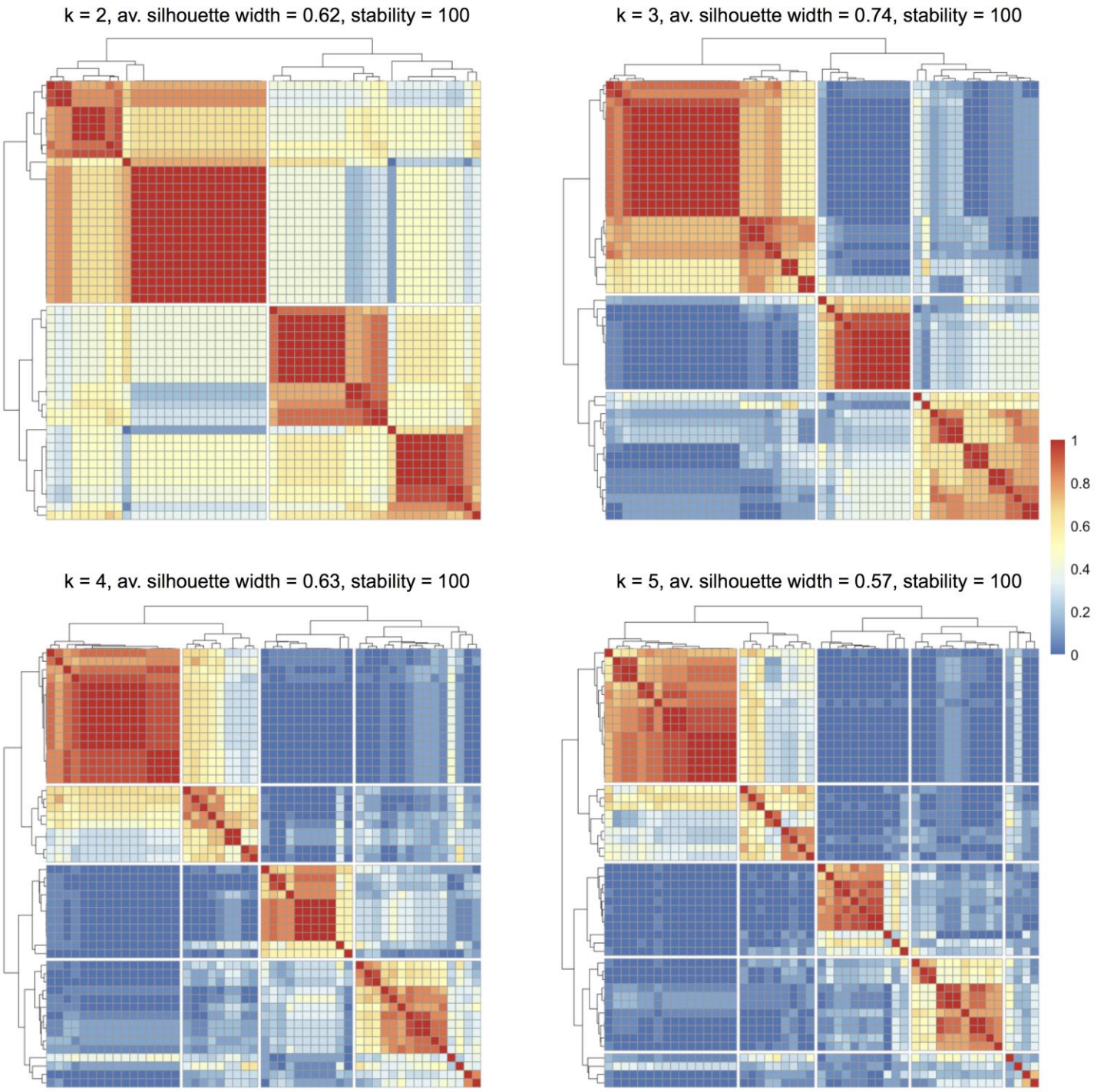
(a) Contour plots describing the sorting strategy for isolating HSCs in patient 2 (the same was done for patient 1). CD34, CD38, CD90 and CD45RA expression is displayed using a log scale; (b) Lineage negative, CD34+/CD38-/CD90+/CD45RA- single cells were sorted into individual wells for scRNA-Seq or colony growth in cytokine cocktail allowing progenitor cell expansion. For genotyping the JAK2V617F and the TET2 loci were characterised using Sanger sequencing. (c) Clonal composition of patients 1, 2 obtained by Sanger sequencing experiments as described in (b) of the JAK2V617F and the TET2 loci (Methods). Colors are the same as Cluster colors in Fig. 3.



Supplementary Figure 8

Quality control of cells in the patient data

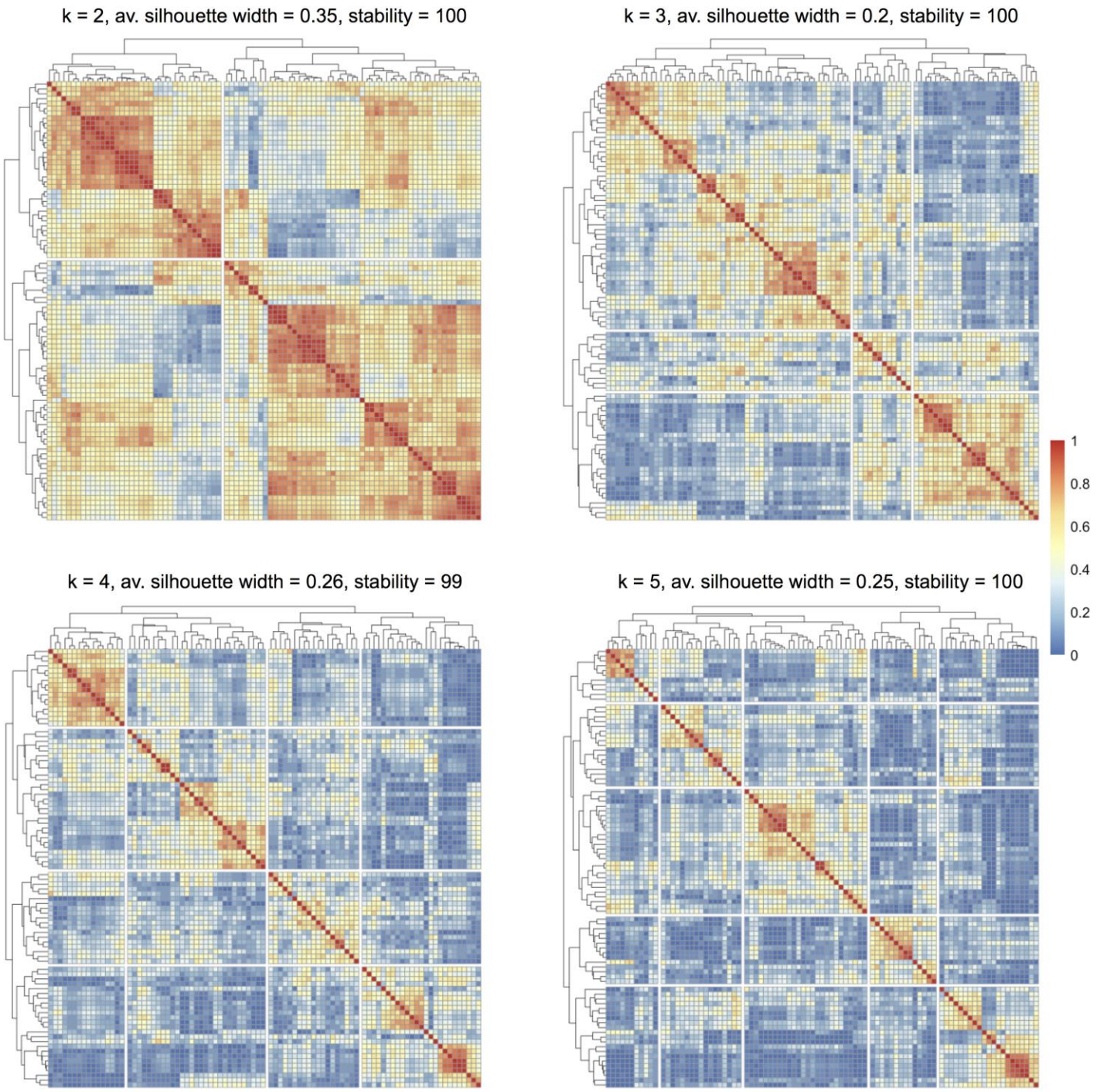
(a) Number of cells with a given number of expressed genes in each patient. Cells on the left side of the red line were removed from further analysis as lowly expressed; (b) Number of cells with a given (# of ERCC reads)/(# endogenous reads) ratio in each patient. Cells on the right side of the red line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



Supplementary Figure 9

Clustering of scRNA-seq data from patient 1

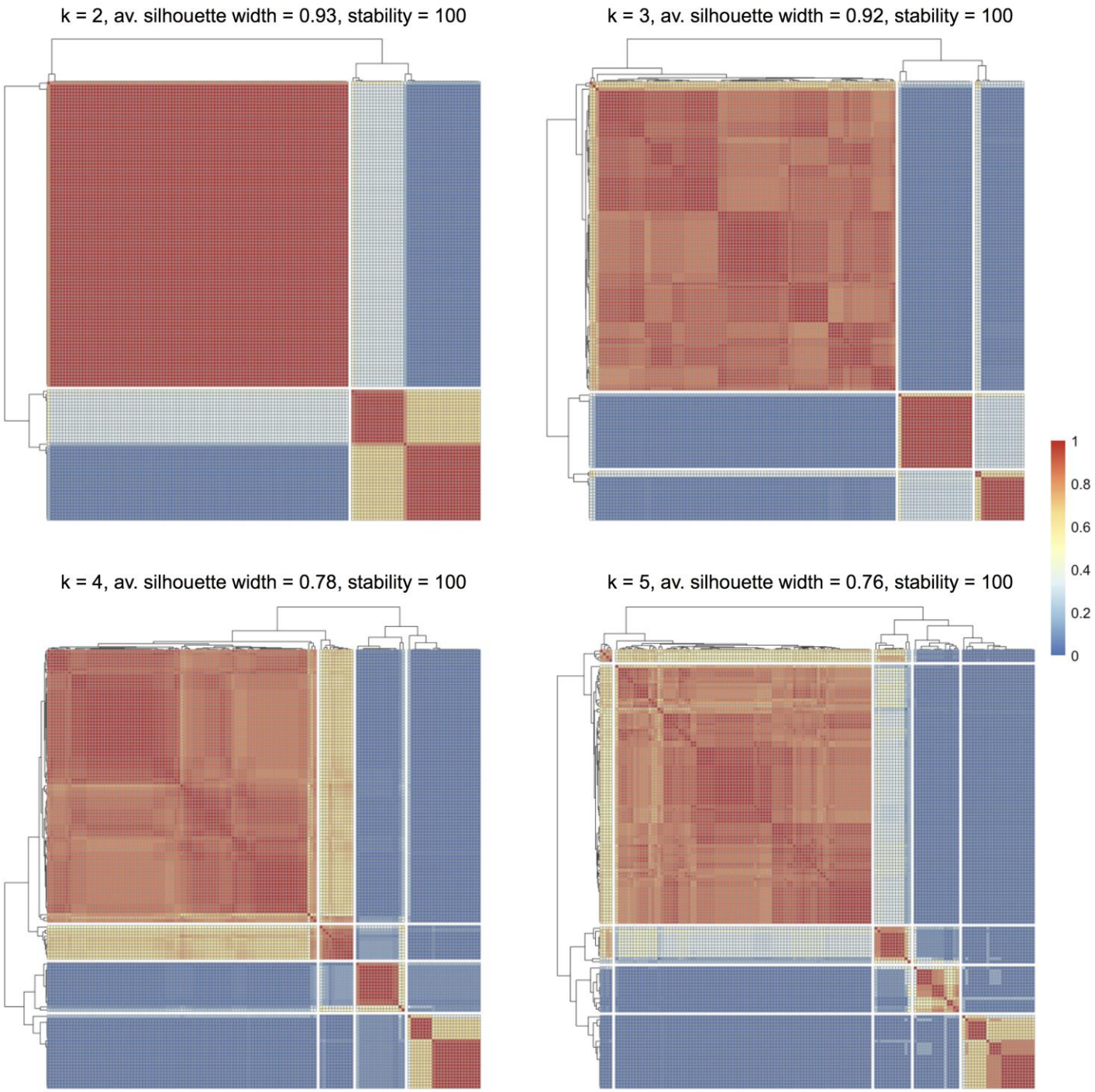
Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.



Supplementary Figure 10

Clustering of scRNA-seq data from patient 2

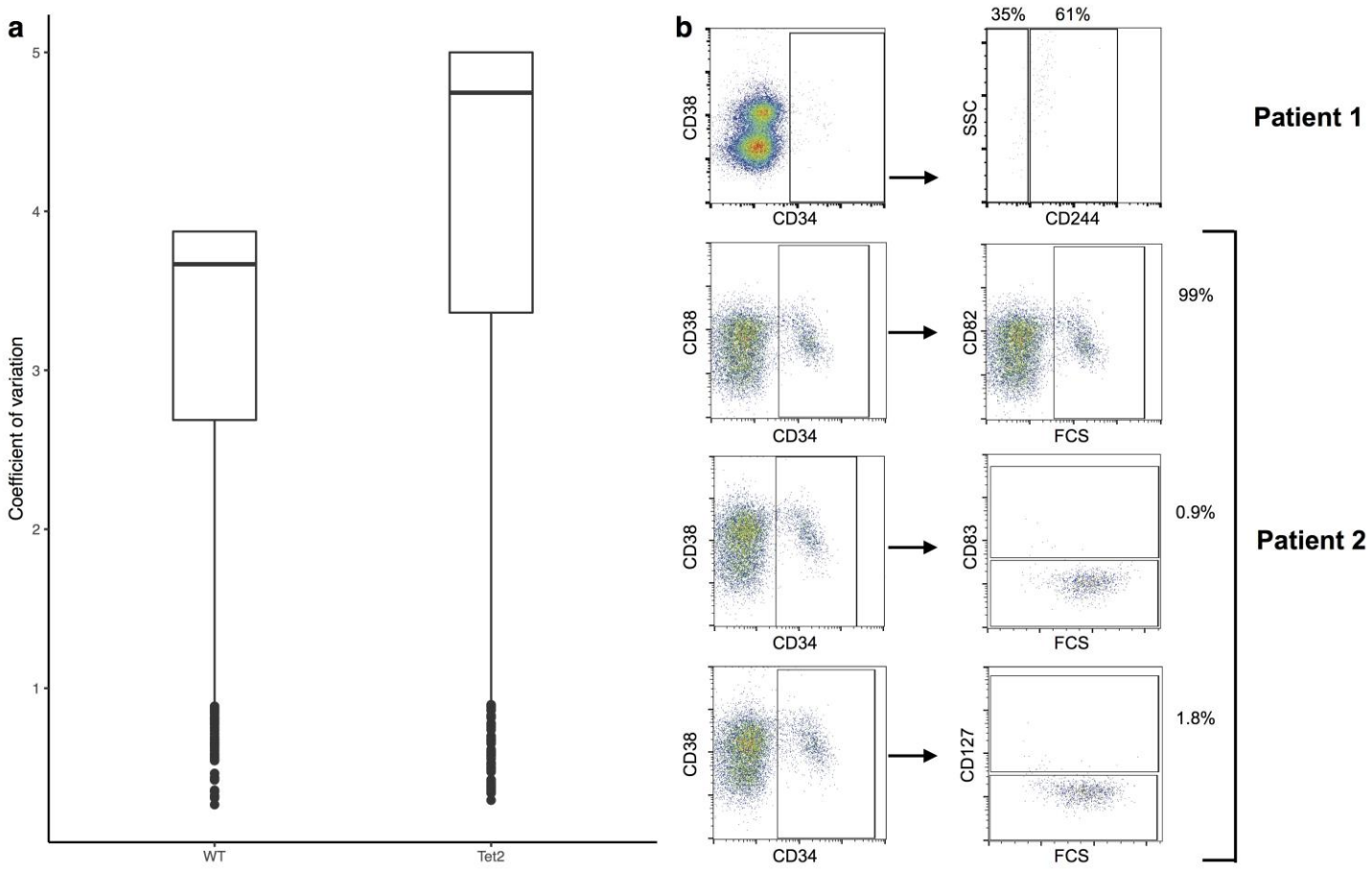
Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.



Supplementary Figure 11

Clustering of scRNA-seq data using combined patient 1 and patient 2 datasets

Consensus matrices corresponding to different values of *k*. For average silhouette width and stability see Methods.



Supplementary Figure 12

Additional lines of evidence that SC3 can help to define subclonal composition

(a) Comparison of the coefficient of variation of gene expression in Tet2 and WT subclones of patient 1; (b) 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.

Supplementary Results 1

Clustering of the Macosko dataset

Using the hybrid approach, we are able to analyse a large Drop-Seq dataset with N = 44,808 cells and k = 39 clusters[1] (Fig. S5, Methods). The ARI between the SC3 clustering and the computationally-derived labels obtained by the original authors is 0.52. This result is largely driven by the fact that Macosko *et al*. lumped a large number of cells into a single "Rods" cluster. This Rods cluster contains 29,400 cells, but using SC3 a finer split of the Rods cluster is revealed with the majority of cells being assigned to 2 large clusters (clusters 4 and 8 on Fig. S5b). Interestingly, several genes related to photoreceptors (e.g. Gngt1, Pde6g, Rho, Rcvrn, Pdc, Gnat1, Nrl, Slc24a1, Rs1 and Sag for cluster 4; Rpgrip1 and Rp1 for cluster 8) are identified as markers distinguishing the two subclusters (Table S3), implying that there is likely a higher degree of heterogeneity amongst those cells than originally reported. We note that 94% of the 29,400 rod cells were lowly expressed (<900 genes detected), and this explains why so few marker genes were identified by SC3. Moreover, 31 of the clusters that were identified by SC3 can be matched with clusters identified by Macosko *et al*. (Fig. S5b and Methods), suggesting that the subsampling employed in our hybrid strategy works well for larger datasets.

Gene and pathway enrichment analysis of the Macosko dataset

Since the cells from the original large cluster have on average fewer than 500 expressed genes, a total of only 15 marker genes (with AUROC threshold of 0.7) for the two clusters could be detected. This low number of genes was not enough to perform enrichment analysis. To overcome this hurdle we performed additional differential expression analysis (using the default SC3 algorithm - see Methods) for the two subclusters (4 and 8, identified by SC3, Fig. S5b) of the large original cluster (using g:Profiler[2]). The differential expression analysis provided 3620 differentially expressed genes and we were able to identify 'phototransduction' and 'oxidative phosphorylation' pathways and 'photoreceptor cell differentiation', 'response to light stimulus', 'sensory perception of light stimulus' and 'NADH dehydrogenase activity' GO terms (selected with green color in Table S3).

Additional lines of evidence that SC3 can help to define subclonal composition

Three additional lines of evidence support the assumption that SC3 can help to define subclonal composition.

Firstly, we used microarray data from erythroid burst-forming units colonies available for patient 1[3] where the genotype of each clone was linked to a specific transcriptional signature (Methods). When comparing differentially expressed genes for the double mutant clone from erythroid burst-forming unit colonies and the marker genes obtained from the pooled putative TET2/JAK2 mutant clone, we found 13 genes in common. This overlap was significant (p-value=0.048, hypergeometric test) and we also found a weak correlation (Spearman's rho = 0.15, p-value=0.031) between the fold changes from the microarray and the scRNA-seq data.

Secondly, we performed Gene and Pathway Enrichment Analysis using the marker genes (Methods). We found several categories related to haematopoiesis (selected with green color in Table S5). Among the enriched pathways were 'Jak-STAT signalling pathway', 'estrogen signalling pathway'[4] and 'GPVI-mediated activation cascade' (the latter plays a role in activation and aggregation of platelets). Furthermore, Gene Ontology analysis showed enrichment for the 'Regulation of cytokine production' term. Cytokines play an important role in haematopoiesis by initiating intracellular signals that govern cell fate choices such as proliferation and differentiation[5]. This confirms that ligands and receptors involved in JAK/STAT pathway activation are highly enriched in our marker genes for the putative double mutant cluster. For the putative TET2 only mutant subclones, none of the above pathways were specifically misregulated. Instead, we hypothesized that since *TET2* is involved in DNA de-methylation there would be a global impact on the transcriptome. Loss of TET enzymes has been reported to impact on the variability in gene expression in mouse embryos[37]. Comparing the genome-wide distribution of the coefficient of variations revealed that the putative TET2 mutants have more variable transcriptomes than putative wild-type cells (Mann-Whitney test *p*-value <2.2e-16, Methods and Fig. S12a).

Thirdly, SC3 identified several surface receptors from the list of marker genes corresponding to the different putative clusters. In particular, CD82 (corresponding to the putative double mutant), CD83 (WT clone) and CD127 or CD244 (Tet2 mutant clone) are surface markers that can be targeted by readily available, well-characterized commercial antibodies. We therefore carried out cell-sorting using such antibodies, and as predicted, the CD82 antigen isolated cells with a double mutant nature, was found for 99% of $CD34^{+}CD38^{+}$cells from patient 2 (Fig. S12b). In contrast, CD127 and CD83 antibodies were unable to isolate populations containing >2% of the cells from the same patient, strengthening the assumption that SC3 can predict clonal composition by providing specific marker genes. Due to limited material availability, we were only able to test one surface marker for patient 1. We chose CD244 since it was highly expressed in the putative Tet2 only mutant clone (Fig. S12b**)**. Again, we were able to isolate a CD244 positive population in a subset of $CD34^{+}CD38^{+}$cells. This result demonstrates that SC3 is capable of characterising clusters defined by mutations rather than by patient batch.

Supplementary References

1. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
2. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44,** W83–9 (2016).
3. Ortmann, C. A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372,** 601–612 (2015).
4. Gupta, N. & Mayer, D. Interaction of JAK with steroid receptor function. *JAKSTAT* **2,** e24911 (2013).
5. Chen, E., Staudt, L. M. & Green, A. R. Janus kinase deregulation in leukemia and lymphoma. *Immunity* **36,** 529–541 (2012).

Supplementary Results 2

Source file: 'Supplementary_Results2.Rmd'

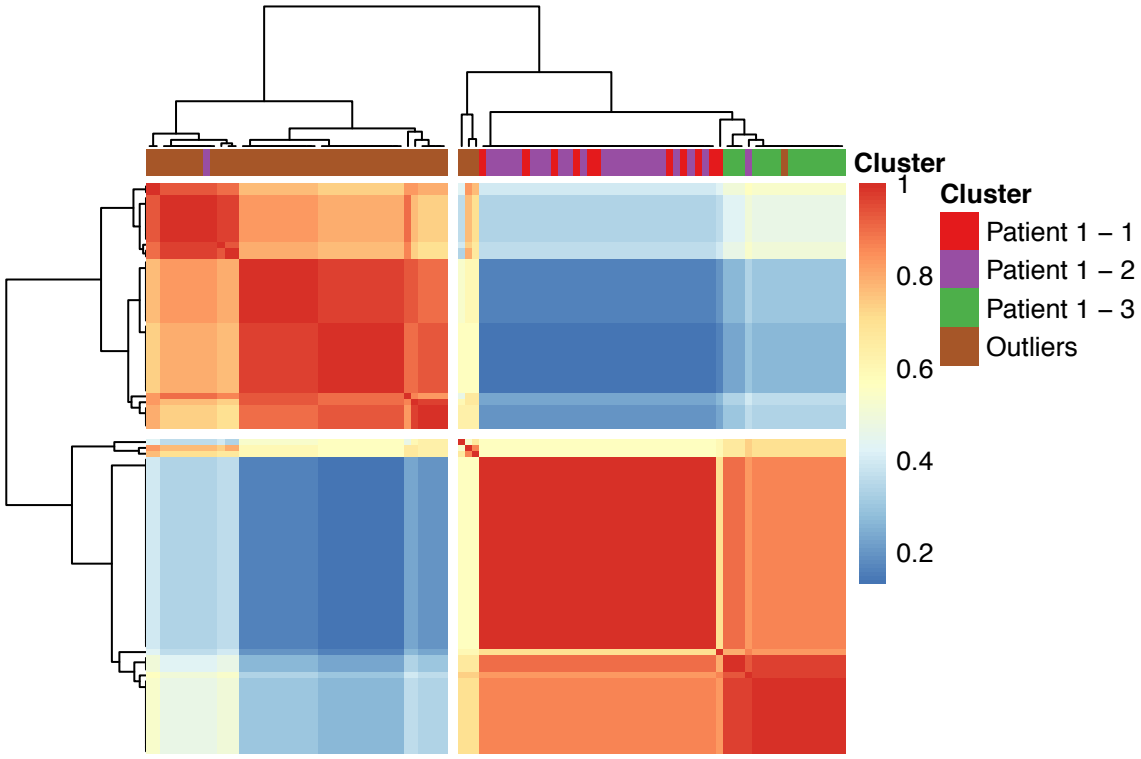
Analysis of cell outliers in patient 1

Here we check whether inclusion of the lower quality cells from patient 1 could shift the proportions of sub-populations.

Fig. S13 shows that we filtered out 45 cells of patient 1 due to either low number of expressed genes or the fraction of ERCC reads. Here we perform clustering of the patient 1 data without quality control keeping all 96 cells in the analysis. However, we still perform the size factor normalisation.

We then run SC3 on the final dataset and look at the results for $\mathbf{k} = 2$, 3, and 4. We highlight the cells corresponding to the cell clusters from Fig. 5b using the same colours as in Fig. 5b. Outlier cells that were excluded from the original analysis are coloured with the brown colour.

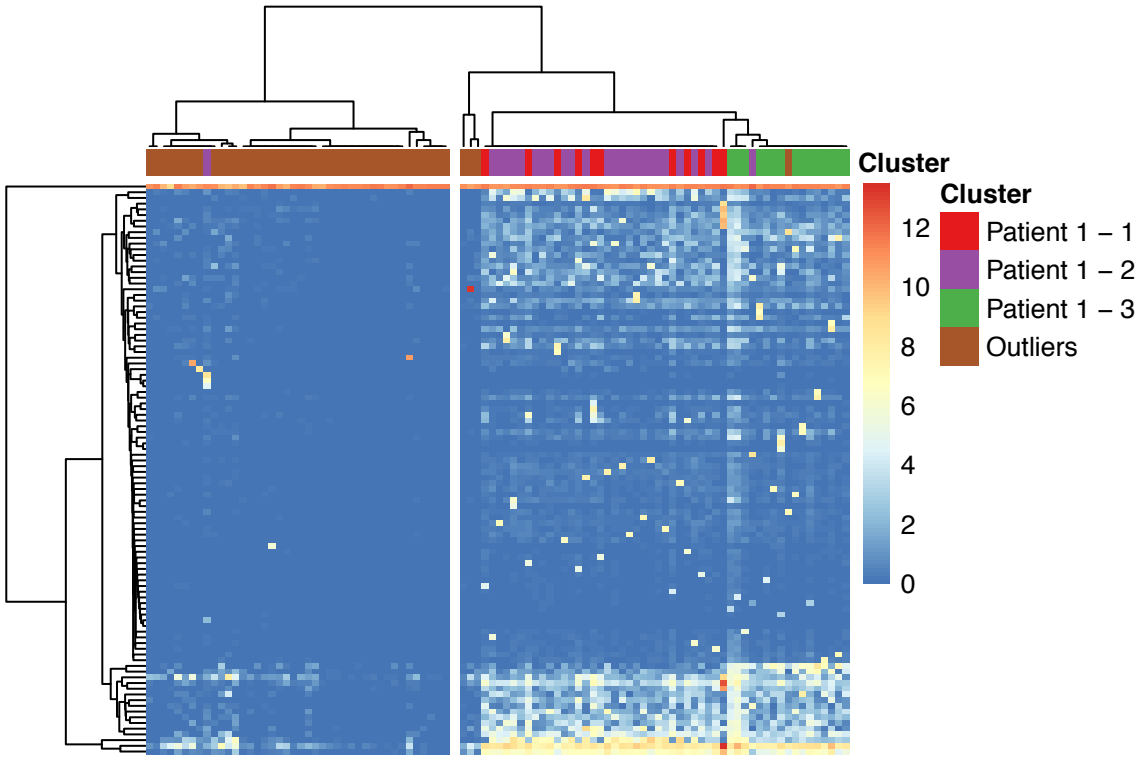
k = 2



When running SC3 with $\mathbf{k} = 2$, one of the clusters consists almost only of the outlier cells and the three previously obtained clusters of the patient 1 are merged into the second large cluster.

The expression plot clearly shows that the separation of the clusters is mainly based on the number of expressed genes in the cells (almost all outlier cells with a low number of expressed genes belong to the left cluster):

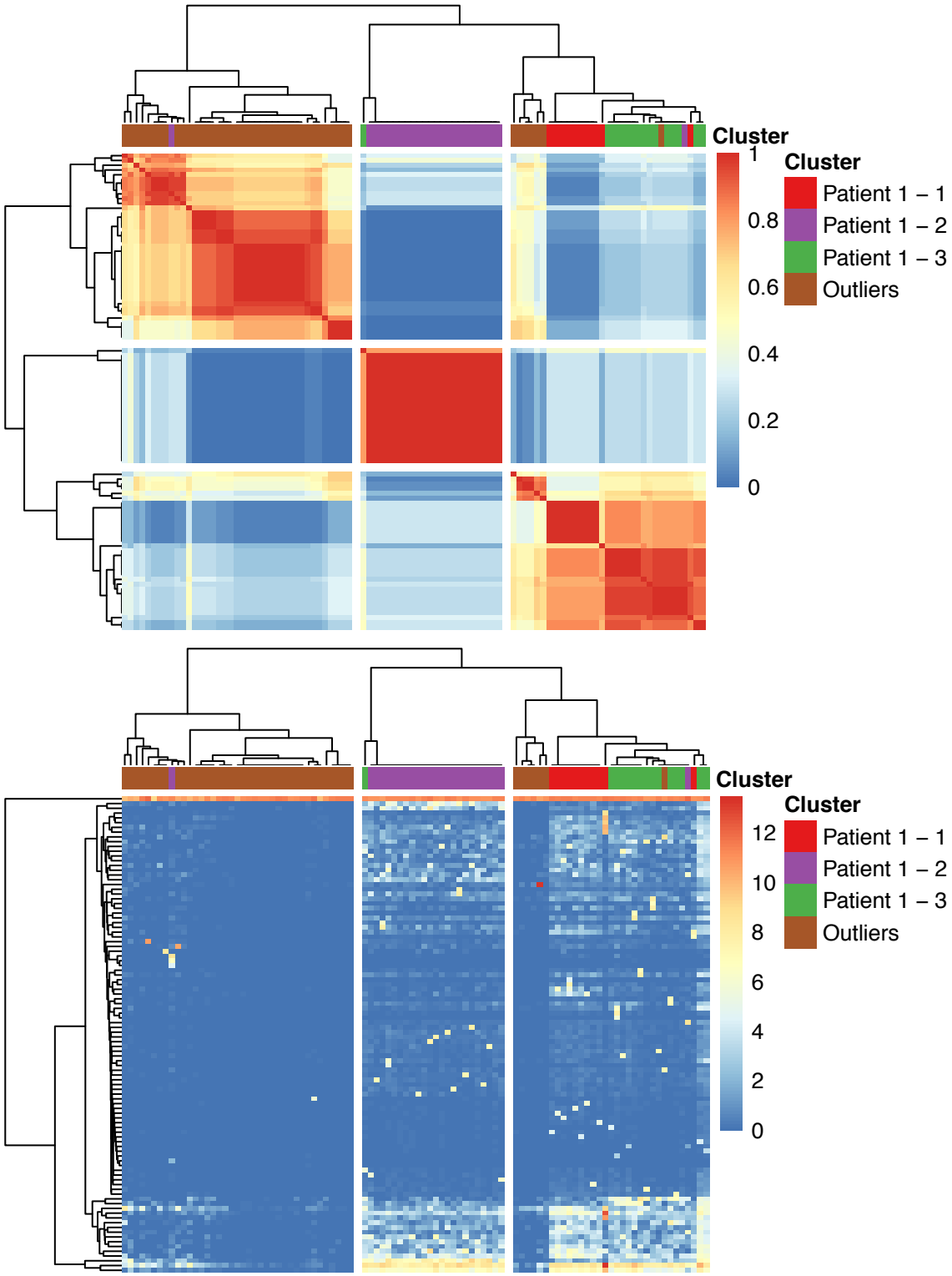
Nature Methods: doi:10.1038/nmeth.4236



$\mathbf{k = 3}$

When $\mathbf{k} = 3$ the purple cluster (supposedly Tet2/WT subclone) gets separated from the the red and the pink cluster. However, the brown outlier cells still mostly stay in one cluster, with only 6 of them mixing with the red and green clusters.

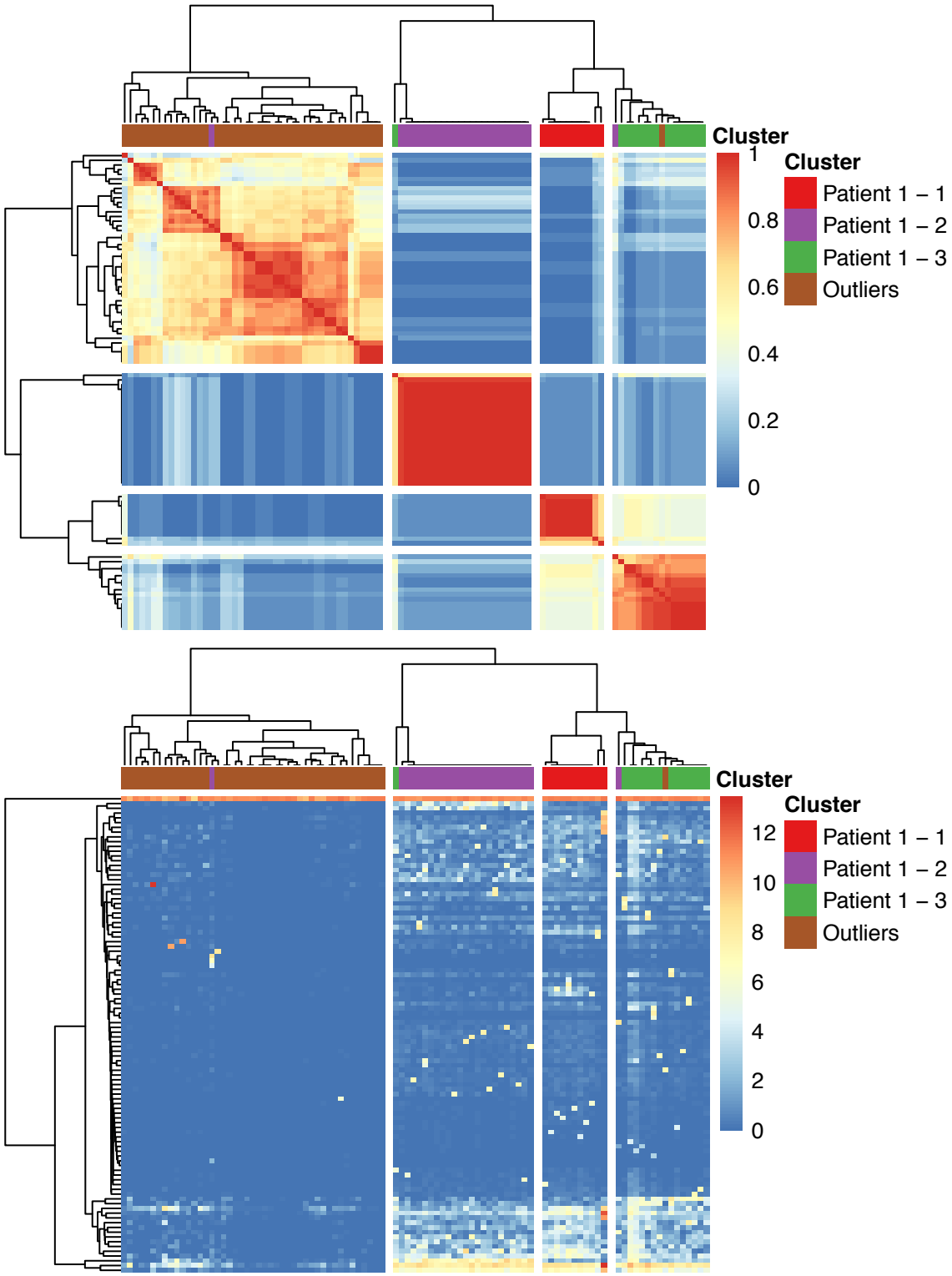
Nature Methods: doi:10.1038/nmeth.4236



$\mathbf{k} = 4$

Finally, when $\mathbf{k} = 4$ the red cluster also separates from the green cluster and all but one outlier cells again stay separately in one cluster.

Nature Methods: doi:10.1038/nmeth.4236



Conclusion

Based on this analysis we conclude that filtering the outlier cells does not affect the subclone composition much. Most of the outlier cells stay always separated from the highly expressed cells in a distinguished cluster

Nature Methods: doi:10.1038/nmeth.4236

and we think tha the main reason for this is that the number of expressed genes in the outlier cells is very low. Moreover, if we assume that for k = 4 the cluster containing the outlier cells constitutes any biological properties, we are still unable to find any marker gene in this cluster. Therefore, we believe one should not include the outlier cells in the analysis.

Nature Methods: doi:10.1038/nmeth.4236

Supplementary Results 3

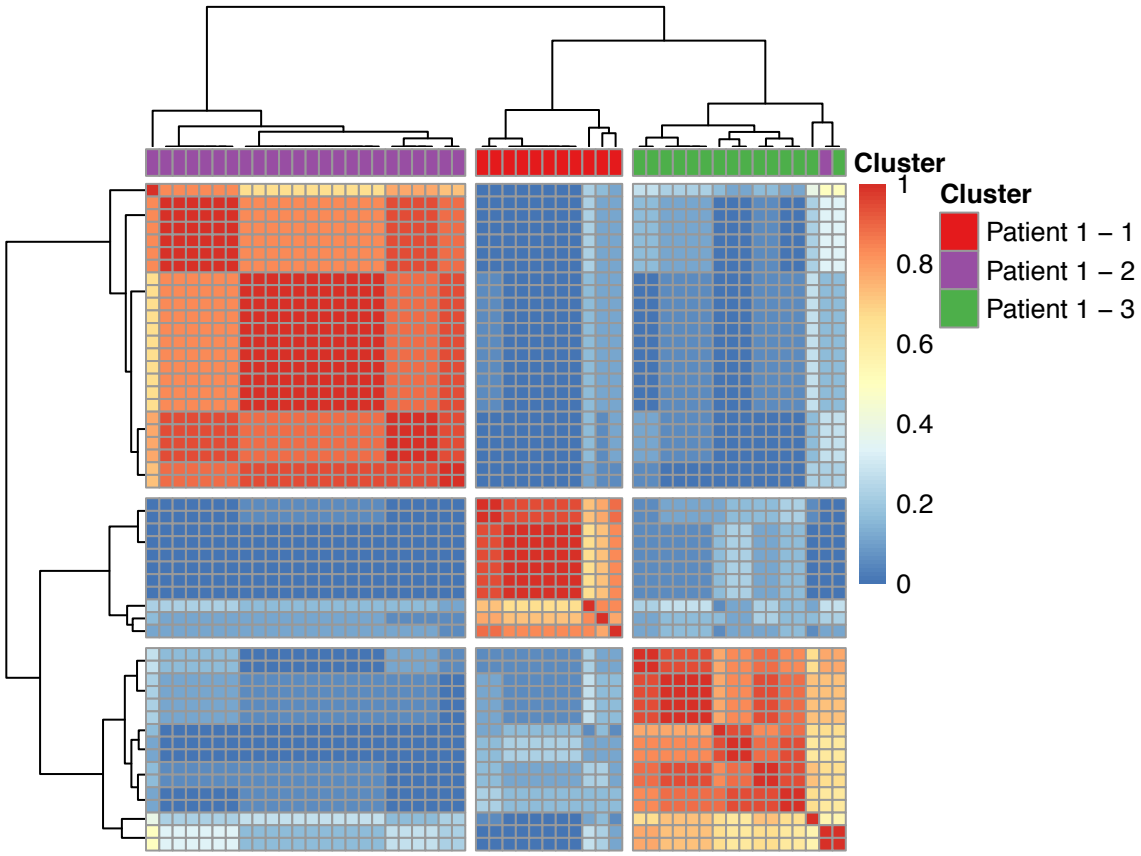
Source file: 'Supplementary_Results3.Rmd'

Analysis of the effect of normalisation (SF, RUV) on clustering of patient 1 data

Here we show that clustering of patient 1 data is not strongly affected by normalisation procedures. We carried out the SC3 clustering of patient 1 data for three scenarios: no normalisation, size-factor normalisation and size-factor+RUV normalisation.

No normalisation

Here we perform clustering of the patient 1 data without any normalisation, however we still perform the quality control and remove cells with a low number of the expressed genes. We also highlight the cells corresponding to the cell clusters from Fig. 6b using the same colours as in Fig. 6b.

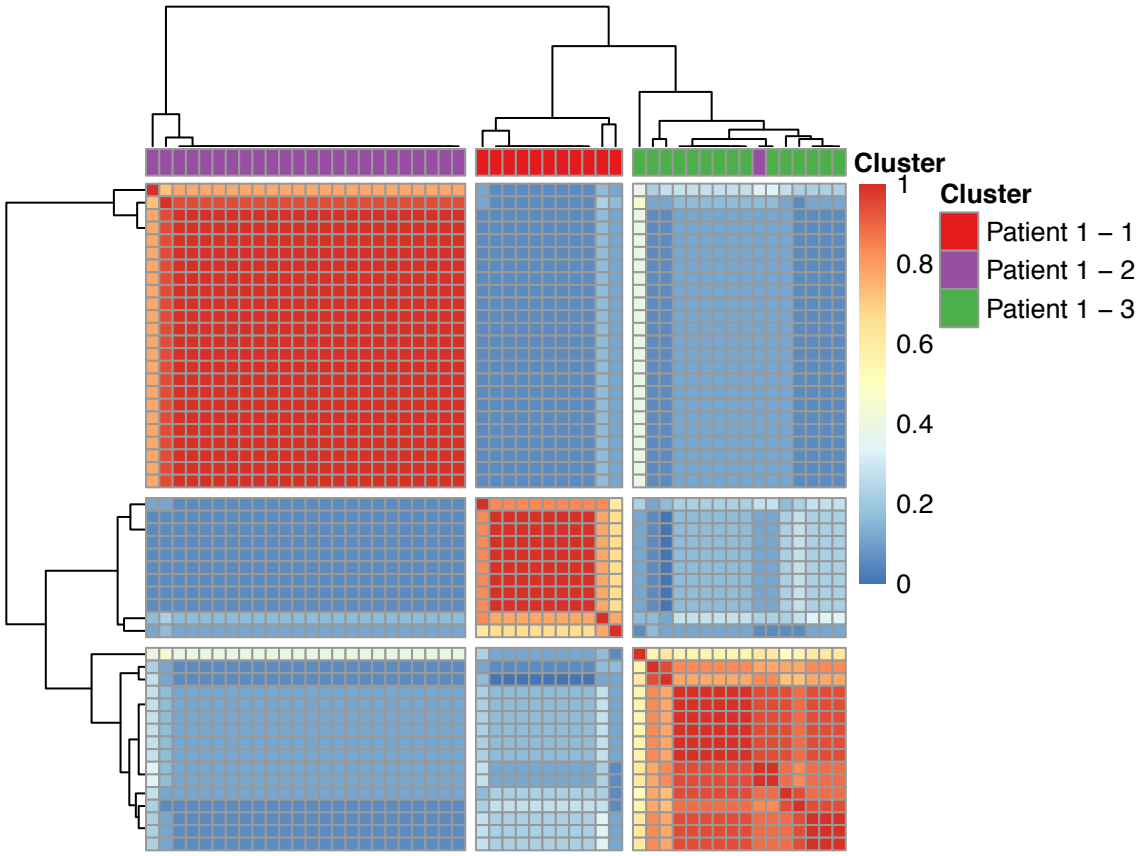


Clearly the clustering result did not change much. There is only one cell that change its label compared to the the clustering results in Fig. S14 (k = 3).

Size-factor normalisation

Here we perform clustering after only SF normalisation.

Nature Methods: doi:10.1038/nmeth.4236

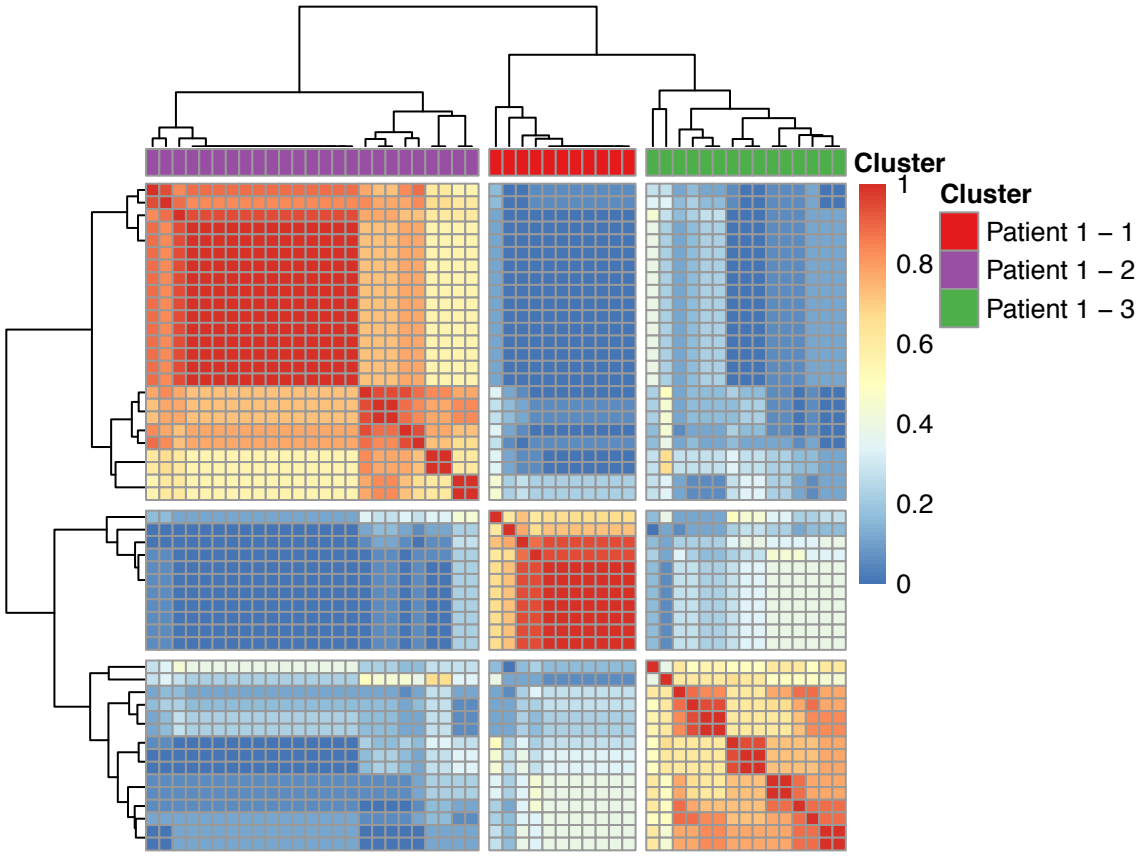


Again there is only one cell that change its label compared to the the clustering results in Fig. S14 (k = 3).

Size-factor + RUV normalisation

Here we perform both SF and RUV normalisations (this case is described in the main text of the manuscript) and obtain a figure identical to Fig. S14 (k = 3).

Nature Methods: doi:10.1038/nmeth.4236



Conclusion

We detected no significant difference in the clustering outcome when different type of normalisations (or no normalisation) were used.

Nature Methods: doi:10.1038/nmeth.4236

Supplementary Results 4

Source file: 'Supplementary_Results4.Rmd'

Additional comparison of the clustering methods

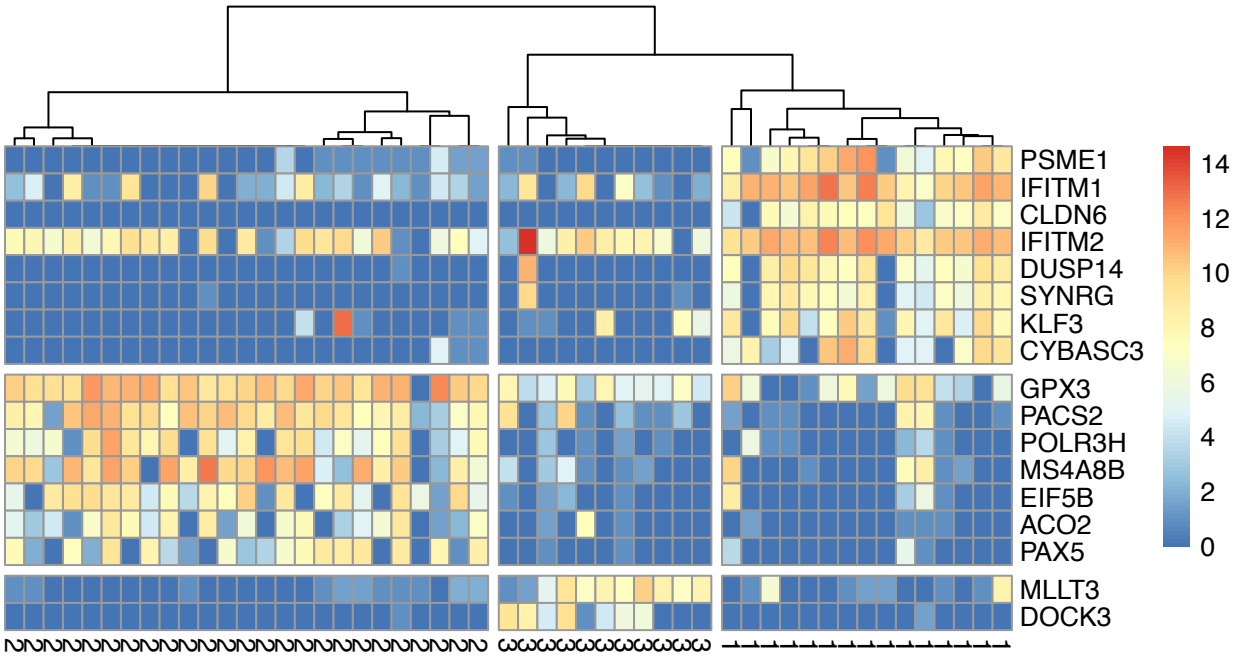
Here we show how we used various clustering methods to identify marker genes and other biologically meaningful information from the HSC data collected from patient 1. Note that some of the methods are stochastic and one could get different results by running the same analysis with a different random seed.

Patient 1 dataset

We use the patient 1 dataset after the quality control and normalisation (SF + RUV). The dataset contains 51 cells and 8710 genes. From the SC3, RMT algorithm and genotyping (see main text), we expect three clusters from patient 1 with 50%, 30% and 20% of the cells. This corresponds to cluster sizes of 26, 15 and 10.

SC3

First, we run SC3 with $\mathbf{k} = 3$ and calculate marker genes. The heatmap shows the clusters and marker genes identified by SC3 for patient 1. The clusters include 15, 25 and 11 cells, in excellent agreement with the genotype data.



Note that because of the small number of cells SC3 was only able to identify 2 marker genes corresponding to cluster 3. In the further analysis presented in the paper, the cells from cluster 3 were then clustered together with patient 2 cells (red cells in Fig. 5c). This allowed us to find more marker genes corresponding to the possible double mutant cells (Tet2/Jak2). MLLT3 gene is present in both figures.

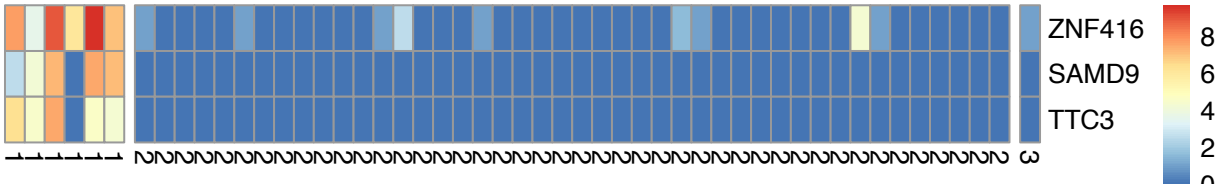
To compare how well the clusters identified by SC3 correspond to the ones identified by the genotyping, we calculate the Kullback-Leibler divergence between the distribution of the cluster sizes. For the genotype data we have that p = [26/51, 15/51, 10/51] and for the SC3 clustering we have that q = [25/51, 15/51, 11/51] which results in a divergence of 0.0013066 nats.

Nature Methods: doi:10.1038/nmeth.4236

pcaReduce

Next, we use pcaReduce to cluster the cells with $\mathbf{k} = 3$. The clusters identified by pcaReduce are of size 6, 44 and 1. Comparing to the cluster sizes predicted by the genotyping, we find that the divergence is 0.4527802 nats.

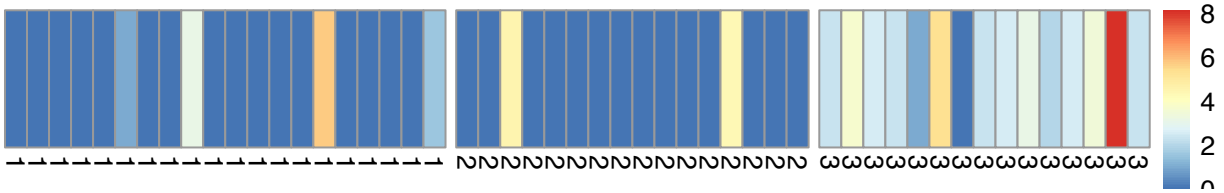
Since pcaReduce does not identify marker genes, we use the SC3 procedure and only three marker genes for one of the clusters are found.



tSNE + kmeans

The tSNE+k-means strategy performs a little bit better in terms of the cluster sizes as it reports 20, 16 and 15 cells in the three clusters. Comparing the cluster sizes to the one obtained from the genotyping, we find that the divergence is 0.0352695 nats.

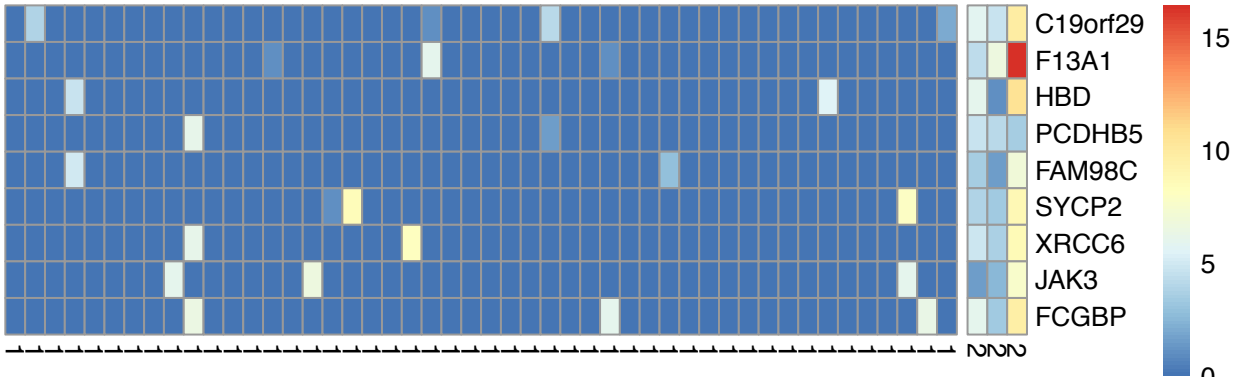
However, the marker gene analysis is only able to identify a single gene, making it difficult to draw any conclusions about the nature of the clusters.



SNN-Cliq

We run SNN-cliq with the default parameters provided in the authors' example.

SNN-cliq reports two clusters, containing 48 and 3 cells. This solution is clearly incompatible with the result suggested by the genotyping.

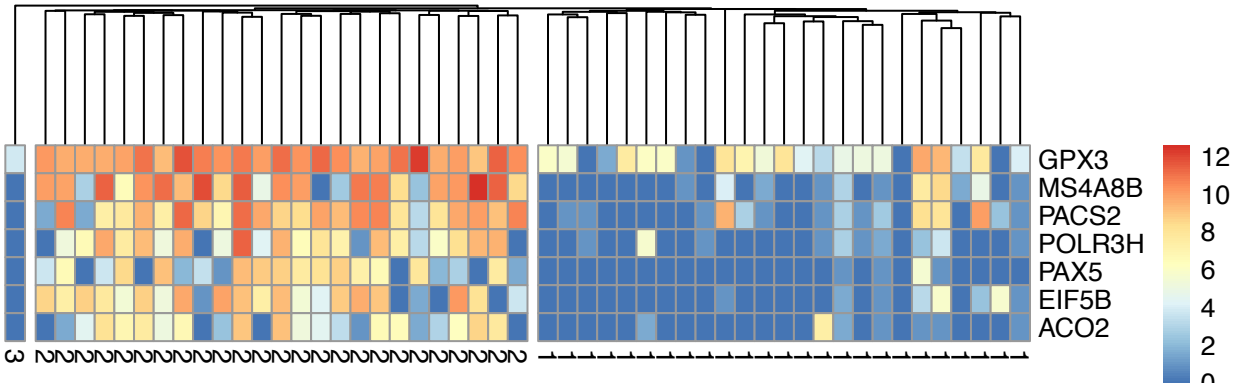


Although there are nine marker genes, the heatmap shows that they are all found in the smaller population, with no positive markers for the larger population.

Nature Methods: doi:10.1038/nmeth.4236

SINCERA

When asking SINCERA for three clusters, their sizes are 25, 25 and 1. Compared to the genotyping clusters, the divergence is 0.3212393 nats. The marker gene analysis reports seven genes and they mostly correspond to the genes of cluster 2 reported by SC3. However, SINCERA was not able to identify the other two SC3 clusters and their marker genes.



Note that SINCERA's own estimate of **k** provided only 1 cluster.

SEURAT

We followed an example provided by the authors. We had to introduce some modifications due to the errors produced by the original code.

It is known that density clustering is sensitive to the density of the data points. In SEURAT this is controlled by the density parameter G. We checked clusterings corresponding to a large range of G, however were not able to infer more than one cluster in any case.

G = 8:

```
##
##  1
## 51
```

G = 1:

```
##
##  1
## 51
```

G = 80:

```
##
##  2
## 51
```

G = 0.0008

```
##
##  1
## 51
```

This poor performance is consistent with what we observed for the other small datasets in Fig. 2a and it is likely to reflect the difficulties of estimating densities when the number of points is low.

Nature Methods: doi:10.1038/nmeth.4236

Conclusions on clustering results

We conclude that SC3 performed better than other clustering methods when clustered patient 1 data. It follows from both the marker gene analysis (SC3 found biologically relevant genes from all three obtained clusters) and from the Kullback-Leibler divergence analysis (the distribution of the cell cluster sizes obtained by the SC3 was the least diverged from the distribution of the cell cluster sizes defined by genotyping).

Nature Methods: doi:10.1038/nmeth.4236

Supplementary Results 5

Source file: 'Supplementary_Results5.Rmd'

Stability analysis of clustering

In addition to the comparison of the actual clusterings of patient 1 data by different methods we also look at how stable these clustering solutions were. This could only be applied to stochastic methods (SC3, pcaReduce, tSNE+kmeans and SEURAT), however we will not consider the stability of SEURAT, since it failed to find more than 1 cluster in the data.

We will run each of the three methods (SC3, pcaReduce, tSNE+kmeans) 10 times (with different random seeds) and look at how stable the solutions are by taking the mean of the ARIs calculated by pairwise comparisons of all different combinations of the obtained solutions (the same way as in Fig. 2b).

SC3

Stability of SC3 is 1.

pcaReduce

Stability of pcaReduce is 0.5293317.

tSNE+kmeans

Stability of tSNE+kmeans is 0.3090761.

Conclusion

SC3 provides a single stable solution, whereas other stochastic methods are less stable.

Nature Methods: doi:10.1038/nmeth.4236

Supplementary Tables

Table S1. 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 dataset.

Dataset	99% quantile of AUROC density distribution
Yan	0.9
Treutlein	0.83
Deng	0.82
Goolam	0.79
Pollen2	0.74
Biase	0.73
Ting	0.72
Usoskin3	0.7
Usoskin2	0.65
Zeisel	0.62
Pollen1	0.61
Patel	0.6
Macosko	0.6
Usoskin1	0.57
Kolodziejczyk	0.56
Klein	0.54

Table S2. 99% quantiles of AUROC density distributions (Fig. S6b) obtained from merging of 100 calculations of marker genes using randomly shuffled assignments of reference labels (provided by the authors, see Methods).

Table S3. 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 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