SC3 - consensus clustering of single-cell RNA-Seq data

Vladimir Yu. Kiselev[1], Kristina Kirschner[2], Michael T. Schaub[3,4], Tallulah Andrews[1], Tamir Chandra[1,5], Kedar N Natarajan[1,6], Wolf Reik[1,5,7], Mauricio Barahona[8], Anthony R Green[2], Martin Hemberg[1]

[1] Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
[2] Cambridge Institute for Medical Research, Wellcome Trust/MRC Stem Cell Institute and Department of Haematology, University of Cambridge, Hills Road, Cambridge, UK
[3] Department of Mathematics and naXys, University of Namur, Belgium
[4] ICTEAM, Université catholique de Louvain, Belgium
[5] Epigenetics Programme, The Babraham Institute, Babraham, Cambridge, UK
[6] EMBL-European Bioinformatics Institute, Hinxton, Cambridge, UK
[7] Centre for Throphoblast Research, University of Cambridge, Cambridge, UK
[8] Department of Mathematics, Imperial College London, London, UK

Abstract

Using single-cell RNA-seq (scRNA-seq), the full transcriptome of individual cells can be acquired, enabling a quantitative characterisation of cell-type based on expression profiles. Due to the large variability in gene expression, assigning cells into groups based on the transcriptome remains challenging. We present Single-Cell Consensus Clustering (SC3), a tool for unsupervised clustering of scRNA-seq data. SC3 integrates many different clustering solutions through a consensus approach, thereby increasing its accuracy and robustness against noise. Tests on nine published datasets show that SC3 outperforms existing methods, yet SC3 remains scalable for large datasets, as shown by the analysis of a dataset containing ~45,000 cells. To enhance the accessibility to users with limited bioinformatics expertise, SC3 features an interactive graphical implementation, which aids the biological interpretation by identifying marker genes, differentially expressed genes and outlier cells. Finally, we apply SC3 to identify different subclones of neoplastic cells in data collected from patients.

Introduction

With the recent advent of single cell RNA-seq (scRNA-seq) technology, researchers are now able to quantify the entire transcriptome of individual cells, opening up a wide spectrum of biological applications. A key application of scRNA-seq is the ability to determine cell types based on the transcriptome profile alone[1–3]. The diversity of cell-types is a fundamental property of higher eukaryotes. Traditionally, cell type was defined based on shape and morphological properties. More recently, tools from molecular biology have enabled researchers to categorize cells based on surface markers[4,5]. However, morphology or a small number of marker proteins are not sufficient to characterise complex cellular phenotypes. scRNA-seq makes it possible to group cells based on their genome-wide transcriptome profiles, which is likely to provide a better representation of the cellular phenotype. Indeed, several studies have already used scRNA-seq to identify novel cell-types[1–3], demonstrating the potential to unravel and catalogue the full diversity of cells in the human body.

Mathematically, the problem of *de novo* identification of cell-type from data may be seen as an unsupervised clustering problem, i.e., how to separate cells into non-overlapping groups without *a priori* information as to the number of groups or group labels. However, the lack of training data renders this unsupervised clustering a hard problem, and for scRNA-seq the challenge is further compounded by the fact that there are few reliable benchmarks for validation.

Although scRNA-seq is a relatively new technology, several groups have already developed custom clustering methods for single cell data[2,6–10]. However, these clustering methods have various shortcomings: (i) they have not been thoroughly benchmarked against one another or against standard algorithms; (ii) it is not clear how they can be scaled to large datasets; (iii) there is no interactive, user-friendly implementation that includes support to facilitate the biological interpretation of the clusters; and (iv) the number of clusters, k, has to be fixed *a priori* by the user and there is no support to explore different hierarchies of clusters. The last point is particularly relevant when studying complex biological tissues as previous studies have found biologically meaningful cell populations at several levels of granularity[8,11,12]. Thus, a clustering method should enable the user to explore how the clusters change with k.

We present a novel interactive clustering tool for scRNA-seq data, SC3 (**S**ingle **C**ell **C**onsensus **C**lustering). SC3 obtains robust results by combining several well-established techniques using a consensus clustering approach. We demonstrate the performance of SC3 by applying it to nine published datasets and thoroughly benchmarking the method against four other methods. We showcase the scalability of SC3 to very large datasets by analysing a dataset with ~45,000 cells[9]. SC3 is implemented as an R-package with a graphical user interface, thus making it possible for users with limited bioinformatics training to explore different clusterings in an interactive and intuitive manner. In addition to providing cell clusters, SC3 facilitates their biological interpretation by identifying marker genes, differentially expressed genes, outlier cells, and by providing an integrated link to Gene Ontology analysis. To demonstrate how SC3 can be used to provide novel biological insights, we analyse scRNA-seq data from two patients with a myeloproliferative neoplasm. Strikingly, the clusters that we identify can be directly linked to subclonal populations that were characterised through independent experiments.

Results

Consensus clustering as a robust methodology underpinning SC3

The output of a scRNA-seq experiment is typically represented as an expression matrix (***M***) consisting of *g* rows (corresponding to genes or transcripts) and *N* columns (corresponding to cells). SC3 takes this expression matrix as its input. The SC3 algorithm consists of several steps: a cell filter, a gene filter, distance calculations, nonlinear spectral transformations and *k*-means clustering, followed by the consensus step (Fig. 1a, Methods). Note, in particular, that the distance calculation reflects a change of coordinate space, as we go from the expression matrix (*g* x *N*) to a cell-to-cell matrix (*N* x *N)*. The nonlinear spectral transformation step of the distance matrix retains the first *d* eigenvectors for clustering. Each of the above steps requires the specification of a number of parameters, e.g. different metrics to calculate distances between the cells or the type of nonlinear transformation. However, choosing optimal parameter values is difficult. To avoid this problem, SC3 utilizes a parallelisation approach, whereby a significant subset of the parameter space is evaluated simultaneously. The outcome of this strategy is a set of clusterings. Instead of trying to pick one optimal solution from within this set, we combine all the different clustering outcomes into a consensus matrix which summarises how often each pair of cells is located in the same cluster. The final result provided by SC3 is a consensus cluster determined by complete-linkage hierarchical clustering of the consensus matrix into *k* groups (Fig. 1b). Using this approach, we can leverage the power of a plenitude of well-established methods, while additionally gaining robustness against the particularities of any single method.

We first validated SC3 using eight publicly available scRNA-Seq datasets[8,11–17], ranging from *N* = 80 to *N* = 3,005 cells (Fig. 1c). For each dataset, we used the clustering provided by the authors of the original study as *reference labels* against which we compared the clusters obtained by SC3. To quantify the discrepancy of clusterings, we use the Adjusted Rand Index (ARI, see Methods), a normalised similarity index, ranging from 1 when the clusterings are identical to 0 when the similarity is what one would expect by chance. In this study, we take 0.8 to indicate a good score. A suitable set of methods and parameters to include within our consensus approach was obtained by analysing four of the eight datasets (Fig. 1d). Interestingly, we found that clustering performance was largely unaffected by the cell filter, and for different choices of gene filter, distance metrics and spectral transformations, i.e., the ARI values were largely similar (Fig. 1d, S1-S4). In contrast, we found that ARI was most sensitive to the number of eigenvectors, *d*, retained after the spectral transformation: the ARI is low for small values of *d*, and then increases to its maximum before dropping close to 0 as *d* approaches *N*/2. In particular, we find that the best clusterings were achieved when *d* is between 4-7% of the number of cells, *N* (Methods). We hypothesise that this range represents an optimal dimensionality reduction such that the technical noise is minimized, and use this range of parameters within our consensus approach to narrow down the parameter space to be explored. Strikingly, consensus clustering[18] of the solutions obtained in this range further improves the ARI (Fig. S5).

To benchmark SC3, we considered four other methods: tSNE[19] followed by *k*-means clustering (a method similar to those used by Grün et al[1] and Macosko et al[9]), pcaReduce[7], SNN-Cliq[6] and SINCERA[10]. SC3 performs better than all of the other methods across all datasets, with the exception of the Pollen1 dataset (Fig. 2). Furthermore, the higher ARI attained by SC3 comes at a

moderate computational cost (Fig. S6): the run time for N = 1,000 is ~10 min, compared to ~1 min for the other methods.

SC3 can be scaled to large datasets

The main computational bottleneck of SC3 is the calculation of the eigenvectors of the distance matrix (which scales as $O(N^3))$[20]. This scaling makes it impractical to use the original, simpler version of SC3 described above for datasets with >10,000 cells.To allow SC3 to be applied to very large datasets, we have implemented a hybrid approach which combines unsupervised and supervised methodologies. When the number of cells is large (N > 1,000), SC3 selects 1,000 cells uniformly at random, and clusters this subset as described above. Subsequently, the inferred labels are used to train a support vector machine (SVM, Methods), which is used to assign labels to the remaining cells. Training the SVM is fast, thus allowing SC3 to be applied to very large datasets.

To test the SVM in isolation, we used all the validation datasets from Fig. 1c with the original reference labels provided by the authors to train the SVM. Our results demonstrate that using <20% of the cells for training it is possible to accurately predict the labels of the remaining cells (Fig. 3a). Overall, the performance improves with higher N, and for the larger datasets, an ARI of 0.8 can be achieved with <10% of training cells. Using this approach enables us to analyse a large Drop-Seq dataset that would otherwise be intractable (Macosko, N = 44,808, k = 39)[9]. For this dataset we found that an ARI of 0.8 can be achieved by the SVM with only 3% of the cells used for training.

Using the hybrid approach it is still possible to achieve an ARI > 0.8 with only 20% of the training cells for the larger datasets (Klein and Zeisel, Fig. 3b). Interestingly, in the case of the Macosko dataset our consensus clustering produces a result that is drastically different from the authors'. Closer inspection of our results shows that SC3 does not identify cluster 24 (labelled as "Rods") with 29,400 cells[9]. This implies that the original large cluster 24 may be more heterogeneous than suggested by the authors.

SC3 is implemented through an interactive and user-friendly interface

To increase its usability, SC3 features a graphical user interface displayed through a web browser, thus minimising the need for bioinformatics expertise. The user is only required to provide the input expression matrix and a range for the number of clusters, k, and SC3 will calculate the possible clusterings for this range. Since the clustering is performed during startup, the user can thus explore different choices of k in real-time. The outcome is presented graphically as a consensus matrix to facilitate visual assessment of the clustering quality (Fig. 4b). The elements of the consensus matrix represent a score, normalised between 0 and 1, indicating the fraction of SC3 parameter combinations that assigned the two cells to the same cluster. By considering the consensus scores within and between clusters, one may quickly assess the clustering quality by visual inspection. Furthermore, to aid the user in the selection of k, SC3 also calculates the silhouette index[21], a measure of how tightly grouped the cells in the cluster are.

SC3 assists with biological interpretation

A key aspect for evaluating the quality of the clustering, which cannot be captured by traditional mathematical consistency criteria, is the biological interpretation of the clusters. To help the user characterise the clusters, SC3 identifies differentially expressed genes, marker genes, and outlier cells. By definition, differentially expressed genes differ between two or more clusters. To detect such genes, SC3 employs the Kruskal-Wallis[22] test (Methods) and reports the differentially expressed genes across all clusters, sorted by *p*-value. In contrast, marker genes are highly expressed in only one of the clusters and are selected based on their ability to distinguish one cluster from the remaining ones (Fig. 4a). To select marker genes, SC3 uses a binary classifier based on the ranked gene expressions to distinguish one cluster from all the others. For each gene, a receiver operator characteristic (ROC) curve is then calculated to evaluate how well the classifier identifies the cluster, and the genes are ranked by the area under the ROC curve. The most significant candidates are then reported as marker genes. Cell outliers are identified by calculating a score for each cell using the Minimum Covariance Determinant[23]. Cells that fit well into their clusters receive an outlier score of 0, whereas high values indicate that the cell should be considered an outlier (Fig. 4d). The outlier score helps to identify cells that could correspond to, e.g. rare cell-types or technical artifacts. Moreover, SC3 also makes it possible to obtain a gene ontology analysis for each cluster by directly exporting the list of marker genes to WebGestalt[24]. All the results from SC3 can be saved to text files for further downstream analyses.

To illustrate the above features, we analysed the Deng dataset tracing embryonic developmental stages, and clustered it into 10 groups as suggested in the original study[15]. As shown in Fig. 4b, our clusters largely agree with those obtained by the authors. We identified ~3000 marker genes (Fig. 4c, Table S2). Characteristic genes for the zygote and 2-cell stages include Xpr1, Elf3, Cux2, Klf6, Emp2, Krt8, Serpinb6a/6b, Fn1 and Pecam1. From 8-cell to blastocyst developmental transition we found several Zscan subunits (Zscan4a, Zscan4b, Zscan4c Zscan4d, Zscan4f and Zscan5f – 16-cell stage), Mad2l2, Krt28, Anxa2, Cyb5, Dppa1, Egfr, Eomes, Id2, Lcp1, Mbnl3, Snai1, Tcfap2a, Tspan8, Klf2, Hand1, Gdf9, Tcl1, Pabpc1l, Cryl1, Ddx4, Il10rb and Il6st (receptor for Il6, Lif, Osm, Cntf, Il11, Ctf1 for blastocyst stages). In addition, we investigated several other highly ranked marker genes by comparing to public repositories and publications[25–29]. The analysis revealed several genes specific to a developmental stage which had previously not been reported (Table S2). Using the reference labels reported by the authors[15], we identified nine cells with high outlier scores (green cells in Fig. 4d). Further examination revealed that these nine cells were prepared using the Smart-Seq2 protocol instead of the Smart-Seq protocol[6,15], thus demonstrating the ability of SC3 to identify technical outliers.

SC3 provides biological insights to clinically relevant patient sub-clones

Next, we applied SC3 to scRNA-seq data from two patients diagnosed with myeloproliferative neoplasm (MPN), a group of diseases which reflects an early stage of tumorigenesis. MPNs are predisposed to acute myeloid leukemia due to overproduction of cells from the myeloid lineage. This chronic myeloid nature of MPNs is a result of mutations in haematopoietic stem cells (HSC). MPNs are a clonal disease where several clones are able to coexist for long periods of time[30].

Each patient typically harbours multiple neoplastic subclones intermingled with normal blood cells, thus presenting a severe challenge for data analysis.

We collected HSCs from both patients and for patient 1 we obtained scRNA-seq data for $N = 85$ cells and for patient 2 for $N = 96$. To characterise the samples, we first performed a separate clustering of the single HSCs from the two patients. For patient 1, 35 of the cells showed little or no expression for all genes and we thus hypothesise that they represent technical artifacts. Even though these cells had enough reads to pass the quality control (Methods), closer inspection revealed low diversity evidenced by the number of expressed genes (Fig. S16). The remaining 50 HSCs separated into three biologically meaningful distinct clusters (clusters 1-3 in Figs. 5a and Methods). In contrast, SC3 found only one cluster for the 96 cells from patient 2 (Methods).

To gain further insight into the clusters, we used previously published information derived from exome sequencing data[31] to determine driver mutations for both patients. This analysis identified one driver mutation in the Tet2 gene and one driver mutation in the Jak2V617F gene (Table S3). To assess the clonal composition of each patient, single HSCs were cultured to give rise to individual granulocyte/macrophage colonies, to which Sanger sequencing was applied to quantify the prevalence of both driver mutations. Interestingly, we identified three genotypes for patient 1: 30% of the clones had no mutation, 50% of the clones harboured a Tet2 mutation, and 20% of the clones carried both a Tet2 and a Jak2V617F mutation (Fig. 5b). This ratio is reflected in the number of cells contributing to each of the three clusters identified by SC3 for patient 1 (Fig. 5a). Moreover, all clones grown up for patient 2 carried both a Tet2 and a Jak2V617F mutation (Fig. 5b). Again, this is reflected in the single cluster obtained for this patient from the scRNA-seq data. To compare both patients, we clustered all cells simultaneously (Fig. 5c). Strikingly, 10/11 cells from cluster 3 in patient 1 are grouped with the cells from patient 2, whereas clusters 1 and 2 from patient 1 remain separate. Thus, we hypothesise that cells in cluster 3 from patient 1 correspond to the double mutant clone. Taken together, these results demonstrate that SC3 is capable of extracting biologically relevant data from primary patient samples.

Discussion

We have presented SC3, an interactive tool for unsupervised clustering of scRNA-seq data. By comparing to existing methods, we demonstrate that SC3 provides a highly accurate clustering for published datasets. For large datasets, SC3 employs a hybrid approach, which makes it possible to apply the method to very large experiments, e.g. Drop-seq[9]. Importantly, SC3 features a graphical user interface, making it interactive and user-friendly. SC3 also aids biological interpretation by identifying differentially expressed genes, marker genes, and outlier cells.

A major challenge when developing unsupervised clustering algorithms for scRNA-seq data is the lack of good mathematical models that can be used to generate realistic, synthetic surrogate datasets to benchmark the methods. Instead, we must rely on published datasets where the labels have been provided by the original authors. For some of the datasets (Ting, Patel and Klein), the labels are likely to be accurate since they correspond to cells taken from different tissues, patients or time-points. For the others, however, the labelling was based on a combination of the authors' clustering methods together with their biological knowledge. In these latter cases, the labelling is less reliable, and we cannot be certain that the original clustering is more meaningful than ours. Reassuringly, the degree of confidence in the original labels correlates well with the accuracy of SC3; for the datasets where the labels are almost certainly correct (Ting, Patel and Klein) SC3 obtains ARIs close to 1, whereas for the remaining, less certain datasets, SC3 reaches ARIs of ~0.8.

Another central problem of unsupervised clustering relates to the choice of the number of clusters, *k*. We investigated whether the internal measures of clustering quality, such as the Dunn[32] and Davies-Bouldin[33] indexes, can be used to choose *k*, but we did not find any correlation between them and the external indexes (Rand, ARI and Jaccard[34]) (Fig. S12). In addition, for many datasets (e.g. Pollen, Usoskin and Zeisel) there are at least two hierarchies present, and this is likely to be the case for most samples from complex tissues. Rather than identifying a single *k*, SC3 allows the user to interactively explore different options.

We applied SC3 to scRNA-seq data from two patients diagnosed with MPN. We found strong evidence in support of the hypothesis that the clusters identified by SC3 directly correspond to subclones that were previously characterised by independent experiments[35]. Importantly, SC3 can identify marker genes (Fig. 5a), thus making it possible to determine the impact of different mutations on the transcriptome. These analyses could aid in the development of novel therapies.

As sequencing costs decrease, larger scRNA-seq datasets will become increasingly available, furthering their potential to advance our understanding of biology. An exciting aspect of scRNA-seq is the possibility to address fundamental questions that were previously inaccessible, e.g. *de novo* identification of cell-types. However, the current lack of computational methods for analyzing scRNA-seq has made it difficult to exploit fully the information contained in such datasets. We have shown that SC3 is a versatile, accurate and user-friendly tool, which will facilitate the analysis of complex scRNA-seq datasets. We believe that SC3 can provide experimentalists with a hands-on tool that will help extract novel biological insights from such rich datasets.

Methods

Validation datasets

All validation datasets (Fig. 1c), except the Pollen dataset, were acquired from the accessions provided in the original publications. The Pollen dataset[11] was acquired from personal communication with Alex A Pollen. When benchmarking against pcaReduce method (Fig. 2) we used this version of the Pollen data. However, in the original pcaReduce paper[7] an independent alignment of the data was performed. SC3 benchmarking of this version of the Pollen data is shown in Fig. S11.

SC3 clustering

SC3 takes as input an expression matrix ***M*** where columns correspond to cells and rows correspond to genes. SC3 does not carry out any form of normalization or correction for batch effects. SC3 is based on seven elementary steps. These parameters can be easily adjusted by the user, but are set to sensible default values, determined via cross-validation on the Ting, Deng, Pollen and Usoskin datasets.

1. Cell Filter
The cell filter should be used if the quality of data is low, i.e. if one suspects that some of the cells may be technical outliers with poor coverage. The cell filter removes cells containing fewer than a specified number of non-zero genes. In our calculations the minimum number of expressed genes was set to 2,000. The Cell Filter was removed from the final algorithm since it only affected the results for the Treutlein dataset.

2. Gene Filter
The gene filter removes genes that are either expressed or absent (expression value is less than 2) in at least 94% of cells. The motivation for the gene filter is that ubiquitous and rare genes most often are not informative for the clustering.

3. Log-transformation
For further analysis the filtered expression matrix ***M*** is log-transformed after adding a pseudo-count of 1: $\boldsymbol{M}' = \log2(\boldsymbol{M} + 1)$.

4. Distance calculations
Distance between the cells, i.e. columns, in ***M'*** are calculated using the Euclidean, Pearson and Spearman metrics to construct distance matrices.

5. Transformations
All distance matrices are transformed using either principal component analysis (PCA), multidimensional scaling (MDS) or by calculating the eigenvectors of the graph Laplacian (Spectral). The columns of the resulting matrices are then sorted in descending order by their corresponding eigenvalues. MDS method was removed from the final algorithm because of poor performance.

6. k-means

k-means clustering is performed on the first *d* eigenvectors of the transformed distance matrices (Fig. 1a) by using the default kmeans() R function with the Hartigan and Wong algorithm[36]. The maximum number of iterations was set to 10^9 and the number of starts was set to 1,000 (see Fig. S7 for details on different parameter settings).

7. Consensus clustering

SC3 computes a consensus matrix using the Cluster-based Similarity Partitioning Algorithm (CSPA)[18]. For each clustering result (e.g. Labels(i,d_3) in Fig. 1a) a binary similarity matrix is constructed from the corresponding cell labels: if two cells belong to the same cluster, their similarity is 1, otherwise the similarity is 0. A consensus matrix is calculated by averaging all similarity matrices. This can be split in 2 steps:

7a. Consensus over the d range (Consensus1)

SC3 calculates the consensus matrix over a range of *d* from 4% to 7% of *N*. Consensus over the *d* range is performed for each combination of distance measures and transformations (Fig. 1a). To reduce computational times, if the length of the *d* range (*D* on Fig. 1a) is more than 15, a random subset of 15 values selected uniformly from the *d* range is used. Note that consensus over the *d* range does not provide a single solution and further averaging is required.

7b. Consensus over the parameter set (Consensus2)

Consensus over the parameter set takes multiple results of Consensus1 and includes additional averaging of similarity matrices over the distance measures and transformations.

7c. Consensus clustering

The resulting consensus matrix (after Consensus1 and Consensus2) is clustered using hierarchical clustering with complete agglomeration and the clusters are inferred at the *k* level of hierarchy, where *k* is defined by a user (Fig. 1b).

Fig. S5 shows that the quality of clustering after Consensus2 is generally better than only after Consensus1.

Adjusted Rand Index

If cell-labels are available (e.g. from a published dataset) the Adjusted Rand Index (ARI)[37] can be used to calculate similarity between the SC3 clustering and the published clustering. Since the reference labels are known for all validation datasets, ARI is used for all comparisons throughout the paper. We also investigated the correlation between external (Rand index and Jaccard index[34]) and internal (Dunn index[32] and Davies-Bouldin index[33]) (Fig. S12) measures of clustering. An external index evaluates the clustering based on an external reference, whereas an internal index does not require any additional information. Even though external indexes were in a good agreement with each other, there was little or no correlation between external and internal indexes.

Benchmarking

Before benchmarking we applied the Gene Filter to all the datasets for tSNE+k-means, SNN-Cliq and SINCERA. For pcaReduce no Gene Filter was used. For tSNE+k-means, SNN-Cliq and pcaReduce we additionally applied Log-transformation as described above ($\boldsymbol{M}' = \log2(\boldsymbol{M} + 1)$). For SINCERA we used the original z-score normalisation[10] instead of the Log-transformation.

Biological insights

Identification of differential expression

Differential expression is calculated using the non-parametric Kruskal-Wallis test[22], an extension of the Mann-Whitney test for the scenario when there are more than two groups. A significant *p*-value indicates that gene expression in at least one cluster stochastically dominates one other cluster. SC3 provides a list of all differentially expressed genes with adjusted (using the default "holm" method of p.adjust() R function) *p*-values < 0.01 and plots gene expression profiles of the 70 most significant differentially expressed genes. Note that the calculation of differential expression after clustering can introduce a bias in the distribution of *p*-values, and thus we advice to use the *p*-values for ranking genes only.

Identification of marker genes

For each gene a binary classifier is constructed based on the mean cluster expression values. A classifier prediction is then calculated using the gene expression ranks and its ability to distinguish the cluster with the most highly ranked cells from the remaining clusters is assessed. The area under the receiver operating characteristic (ROC) curve is used to quantify the accuracy of the prediction. A *p*-value is assigned to each gene by using the Wilcoxon signed rank test comparing gene ranks in the cluster with the highest mean expression with all others (*p*-values are adjusted by using the default "holm" method of p.adjust() R function). The genes with the area under the ROC curve (AUROC) > 0.85 and with the *p*-value < 0.01 are defined as marker genes. The AUROC threshold was derived from 99% quantile of the AUROC distributions obtained from 100 random permutations of cluster labels for all datasets (Table S1 and Fig. S13). SC3 provides a visualization of the gene expression profiles for the top 10 marker genes of each obtained cluster.

Cell outlier detection

Outlier cells in each cluster are detected by first reducing the dimensionality of the cluster using the robust method for principal component analysis[38]. Second, robust distances are calculated using the minimum covariance determinant (MCD)[23]. We then used a threshold based on 99.99% quantile of the chi-squared distribution to define outliers. Finally we define an outlier score as the differences between the square root of the robust distance and the square root of the 99.99% quantile of the chi-squared distribution. The outlier score is plotted as a barplot (Figs. 4d).

Support Vector Machines (SVM)

When a dataset contains more than 1,000 cells, SC3 randomly selects and clusters 1,000 cells. Next, a support vector machine (SVM)[39] model with a linear kernel (this kernel provided the best clustering predictions, results from using other kernels - polynomial, radial and sigmoid - are shown in Figs. S8-S10) is constructed based on the obtained clustering. We used the *svm* function of the *e1071* R-package with default parameters. The cluster IDs for the remaining cells are then predicted by the SVM model.

Patients

Both patients provided written informed consent. Diagnoses were made in accordance with the guidelines of the British Committee for Standards in Haematology.

Isolation of haematopoietic stem and progenitor cells

Cell populations were derived from peripheral blood enriched for haematopoietic stem and progenitor cells (CD34+, CD38-, CD45RA-, CD90+), hereafter referred to as HSCs. For single cell cultures, individual HSCs were sorted into 96-well plates (Fig. S14) and grown in a cytokine cocktail designed to promote progenitor expansion as previously described[40]. For scRNA-seq studies, single HSCs were directly sorted into lysis buffer as described in Picelli et al[41].

Determination of mutation load

Colonies of granulocyte/macrophage composition were picked and DNA isolated for Sanger sequencing for Jak2V617F and Tet2 mutations as previously described Ortmann et al[35].

Single cell RNA-Sequencing

Single HSCs were sorted into 96-well plates and cDNA generated as described[41]. Nextera XT library making kit was used for library generation as described in Picelli et al[41].

Processing of scRNA-seq data from HSCs

96 single cell samples per patient with 2 sequencing lanes per sample were sequenced yielding a variable number of reads (*mean*=2,180,357, *std dev*=1,342,541). FastQC[42] was used to assess the sequence quality. Foreign sequences from the Nextera Transposase agent were discovered and subsequently removed with Trimmomatic[43]. The reads were trimmed to 90 bases to remove low quality positions before being mapped with TopHat[44] to the Ensembl[45] reference genome version GRCh38.77 augmented with the spike-in controls downloaded from the ERCC consortium[46]. Quality control of the cells was performed similarly to Roshan M et al.[47]. TPM (Transcripts per Million) values per gene per cell were calculated with RSEM[48] and compared with the percentage of aligned reads in order to detect and remove outlier cells from the analysis. Counts of the uniqely mapped reads in each gene were calculated using SeqMonk (http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk) and were used for further downstream analysis (clustering).

SC3 clustering of patient scRNA-seq data

We clustered scRNA-seq data from patient 1, patient 2 separately as well as a merged dataset containing data from both patients. For patient 1 the best clustering was achieved for k = 5 (Fig. S15a). In this case there are three meaningful clusters and two other clusters containing lowly expressed cells (the two clusters to the right Fig. S15b, k = 5). Data from patient 2 was homogeneous and we were not able to identify more than one meaningful cluster (Fig. S17). The combined dataset (patient 1 + patient 2) was clustered for k in the range [2:5] (Fig. S18). For all values of k the majority of the cells from cluster 3 in patient 1 clustered together with the cells from patient 2. We selected k = 5 as a representative example (Fig. 5c) because in this case two other clusters (1 and 2) from patient 1 were also retrieved by SC3.

Contributions

MH conceived the study; VK, MH, MS and TA contributed to the computational framework; KK and TC performed the experiments for the patient data; KN helped with the analysis of embryonic mouse data; MB, WR, AG and MH supervised the research; VK and MH led the writing of the manuscript with input from the other authors.

Accession Numbers

scRNA-seq data for patient 1 and 2 is available from GEO accession GSEXYZ (upload to be completed)

Software availability

The source code for SC3 is available under the GPL-3 licence at https://github.com/hemberg-lab/SC3 and the package "SC3" is currently under review at Bioconductor.

Acknowledgements

We would like to thank Borislav Vangelov, Jean-Charles Delvenne and Renaud Lambiotte for fruitful discussions and their help with computational methods. We would also like to thank David Flores Santa Cruz, Danai Dimitropolou and Jacob Grinfeld for technical assistance with experiments. We thank Ignacio Vasquez-Garcia, David Harmin, Michal Kosicki for helpful comments on the manuscript.

Funding

Work in the Green lab is supported by Bloodwise (grant ref. 13003), the Wellcome Trust (grant ref. 104710/Z/14/Z), the Medical Research Council, the Kay Kendall Leukaemia Fund, the Cambridge NIHR Biomedical Research Center, the Cambridge Experimental Cancer Medicine Centre, the Leukemia and Lymphoma Society of America (grant ref. 07037), and a core support grant from the Wellcome Trust and MRC to the Wellcome Trust-Medical Research Council Cambridge Stem Cell Institute.

References

1. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (2015).
2. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).
3. Mahata, B. *et al.* Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep.* **7,** 1130–1142 (2014).
4. Coons, A. H., Creech, H. J. & Jones, R. N. Immunological Properties of an Antibody Containing a Fluorescent Group. *Exp. Biol. Med.* **47,** 200–202 (1941).
5. Fulwyler, M. J. Electronic separation of biological cells by volume. *Science* **150,** 910–911 (1965).
6. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv088
7. Zurauskiene, J. & Yau, C. pcaReduce: Hierarchical Clustering of Single Cell Transcriptional Profiles. *bioRxiv* 026385 (2015). doi:10.1101/026385
8. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).
9. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
10. Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput. Biol.* **11,** e1004575 (2015).
11. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–1058 (2014).
12. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18,** 145–153 (2015).

13. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).

14. Ting, D. T. *et al.* Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **8,** 1905–1918 (2014).

15. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343,** 193–196 (2014).

16. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).

17. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).

18. Strehl, A. & Ghosh, J. Cluster Ensembles --- a Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **3,** 583–617 (2003).

19. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

20. Pan, V. Y. & Chen, Z. Q. The Complexity of the Matrix Eigenproblem. in *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing* 507–516 (ACM, 1999).

21. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).

22. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47,** 583–621 (1952).

23. Hubert, M. & Debruyne, M. Minimum covariance determinant. *WIREs Comp Stat* **2,** 36–43 (2010).

24. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33,** W741–8 (2005).

25. Guo, G. *et al.* Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell* **18,** 675–685 (2010).

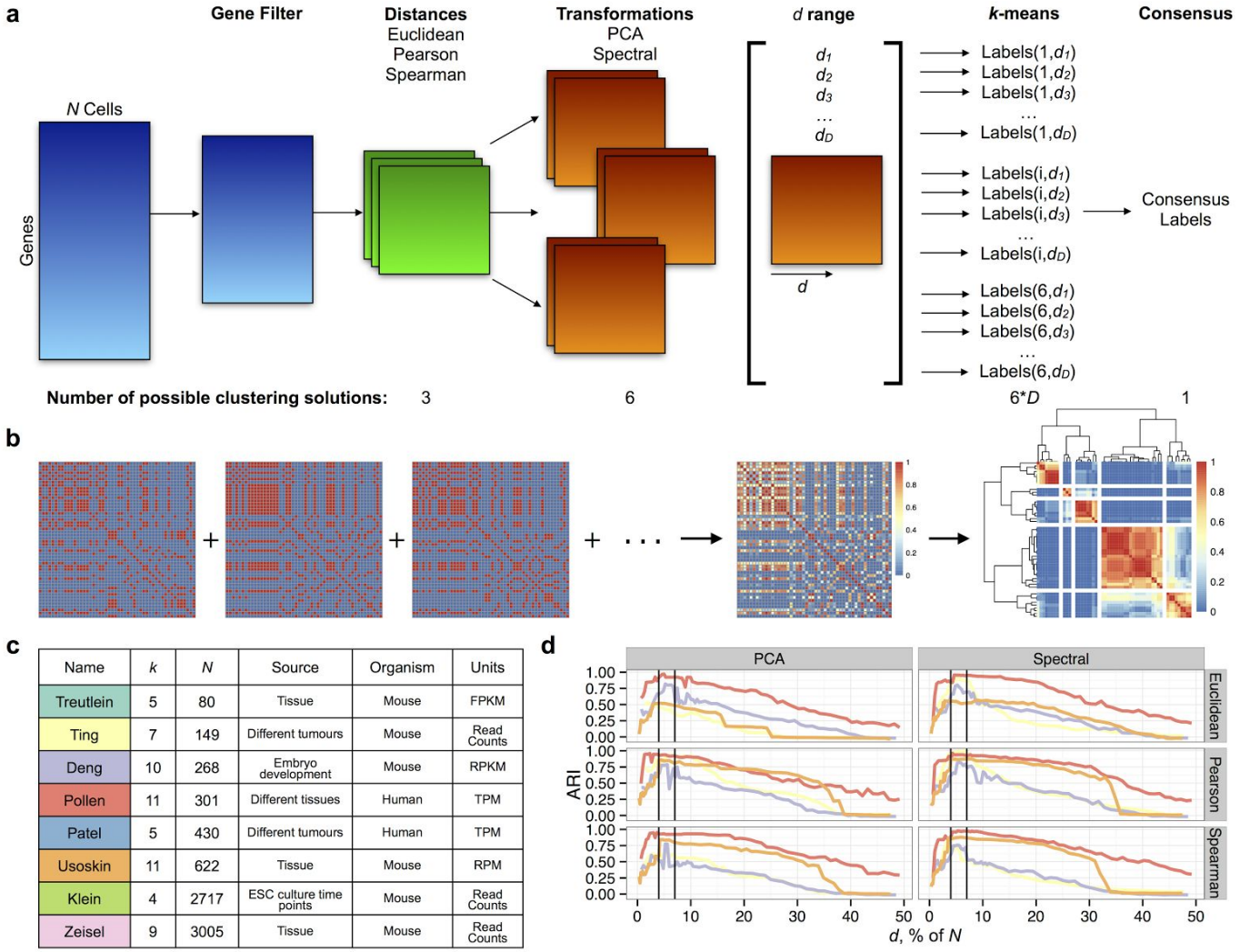
26. Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2014 update. *Nucleic Acids*

Res. **42,** D818–24 (2014).

27. Wiekowski, M., Miranda, M., Nothias, J. Y. & DePamphilis, M. L. Changes in histone synthesis and modification at the beginning of mouse development correlate with the establishment of chromatin mediated repression of transcription. *J. Cell Sci.* **110 (Pt 10),** 1147–1158 (1997).
28. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev. Cell* **35,** 366–382 (2015).
29. Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142,** 3151–3165 (2015).
30. Chen, E., Staudt, L. M. & Green, A. R. Janus kinase deregulation in leukemia and lymphoma. *Immunity* **36,** 529–541 (2012).
31. Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* **369,** 2391–2405 (2013).
32. Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* **3,** 32–57 (1973).
33. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1,** 224–227 (1979).
34. Jaccard, P. Distribution de la Flore Alpine dans le Bassin des Dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles* **37,** 241–272 (1901).
35. Ortmann, C. A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372,** 601–612 (2015).
36. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28,** 100–108 (1979).
37. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66,** 846–850 (1971).
38. Hubert, M., Rousseeuw, P. J. & Branden, K. V. ROBPCA: A New Approach to Robust

Principal Component Analysis. *Technometrics* **47,** 64–79 (2005).

39. Ben-Hur, A., Horn, D., Siegelmann, H. T. & Vapnik, V. Support Vector Clustering. *J. Mach. Learn. Res.* **2,** 125–137 (2002).
40. Petzer, A. L., Zandstra, P. W., Piret, J. M. & Eaves, C. J. Differential cytokine effects on primitive (CD34+CD38-) human hematopoietic cells: novel responses to Flt3-ligand and thrombopoietin. *J. Exp. Med.* **183,** 2551–2558 (1996).
41. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).
42. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Reference Source* (2010).
43. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
44. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
45. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–9 (2015).
46. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat. Methods* **2,** 731–734 (2005).
47. Kumar, R. M. *et al.* Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* **516,** 56–61 (2014).
48. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).



Name	k	N	Source	Organism	Units
Treutlein	5	80	Tissue	Mouse	FPKM
Ting	7	149	Different tumours	Mouse	Read Counts
Deng	10	268	Embryo development	Mouse	RPKM
Pollen	11	301	Different tissues	Human	TPM
Patel	5	430	Different tumours	Human	TPM
Usoskin	11	622	Tissue	Mouse	RPM
Klein	4	2717	ESC culture time points	Mouse	Read Counts
Zeisel	9	3005	Tissue	Mouse	Read Counts

Figure 1. **Validation of the SC3 clustering.** (**a**) Overview of SC3 clustering method (see Methods for the details). *D* is the length of the *d* range. (**b**) Diagram of the consensus matrix construction for the Treutlein data. Binary matrices (Methods) corresponding to each clustering solution in (**a**) are averaged and the resulting matrix is then clustered by hierarchical clustering with *k* level of hierarchy (*k* = 5 in this example). (**c**) Published datasets used to validate SC3. *N* is the number of cells in a dataset, *k* is the number of clusters identified by the authors[8,11–17]. RPKM is Reads Per Kilobase of transcript per Million mapped reads, RPM is Reads Per Million mapped reads, FPKM is Fragments Per Kilobase of transcript per Million mapped reads and TPM is Transcripts Per Million mapped reads. (**d**) Median of ARI over 100 realizations of the SC3 clustering for four validation datasets - Ting, Deng, Pollen and Usoskin (colors correspond to the colors in (**c**)). The x-axis shows the number of eigenvectors *d* (see (**a**)) as a percentage of the total number of cells *N* in each dataset. The black vertical lines correspond to *d* = 4% and *d* = 7% of the total number of cells *N*.

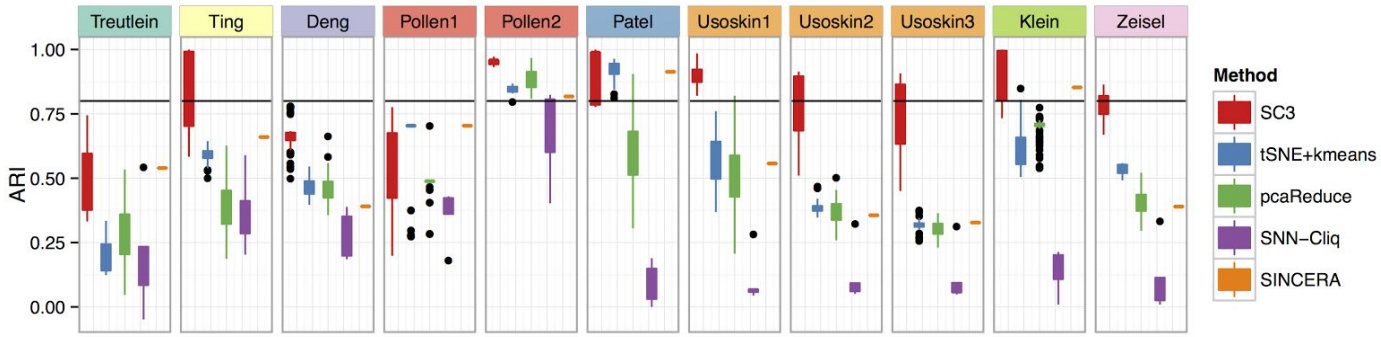


Figure 2. **Benchmarking of SC3 against existing methods.** Each boxplot contains results of 100 realizations of a given method except for SINCERA which is based on hierarchical clustering and provides a single solution. Pollen1, Pollen2, Usoskin1, Usoskin2, Usoskin3 correspond to different levels of hierarchies as described by Zurauskiene and Yau[7]. The black horizontal 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.

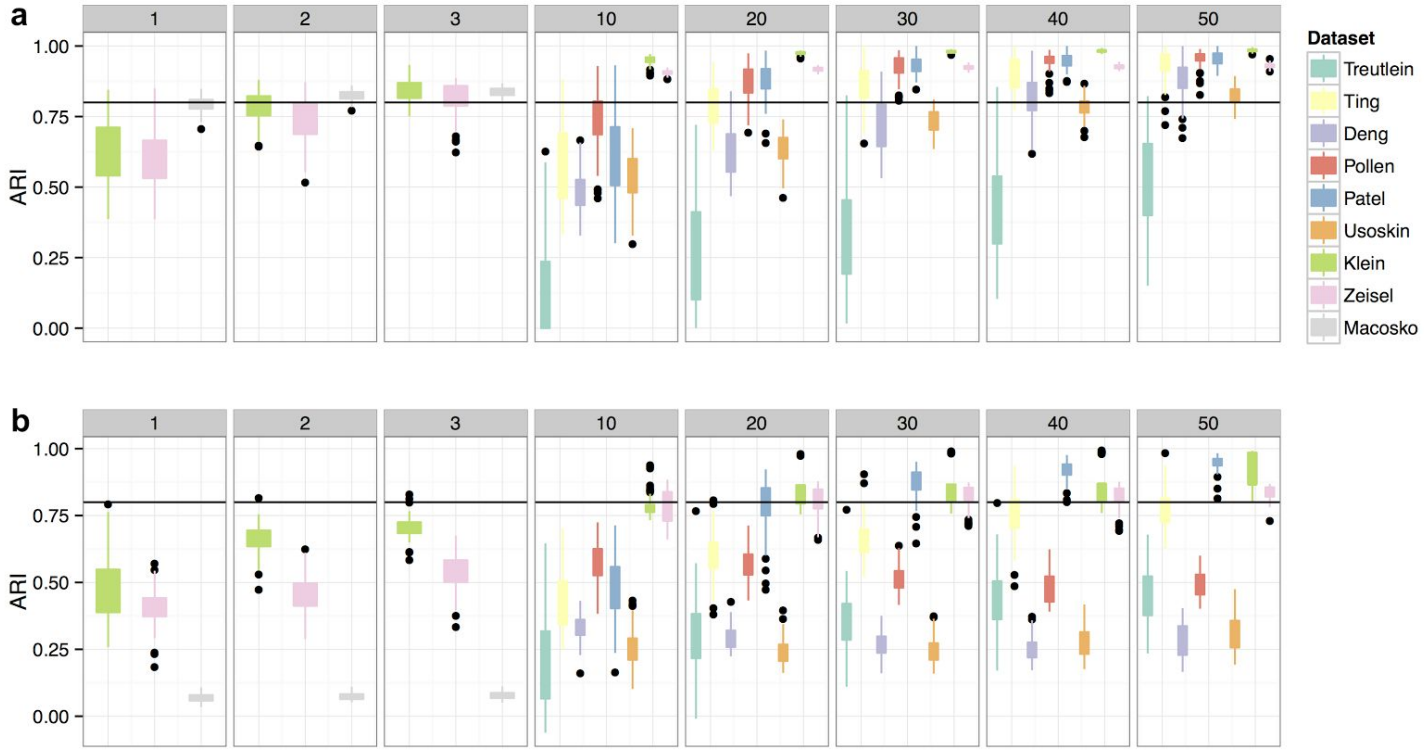


Figure 3. **Results for the hybrid clustering approach using a linear kernel.** The black line corresponds to ARI = 0.8. Numbers in grey boxes correspond to the number of training cells as % of *N.* (**a**) ARI levels for SVM prediction when reference labels (provided by the authors) are used for training. (**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.

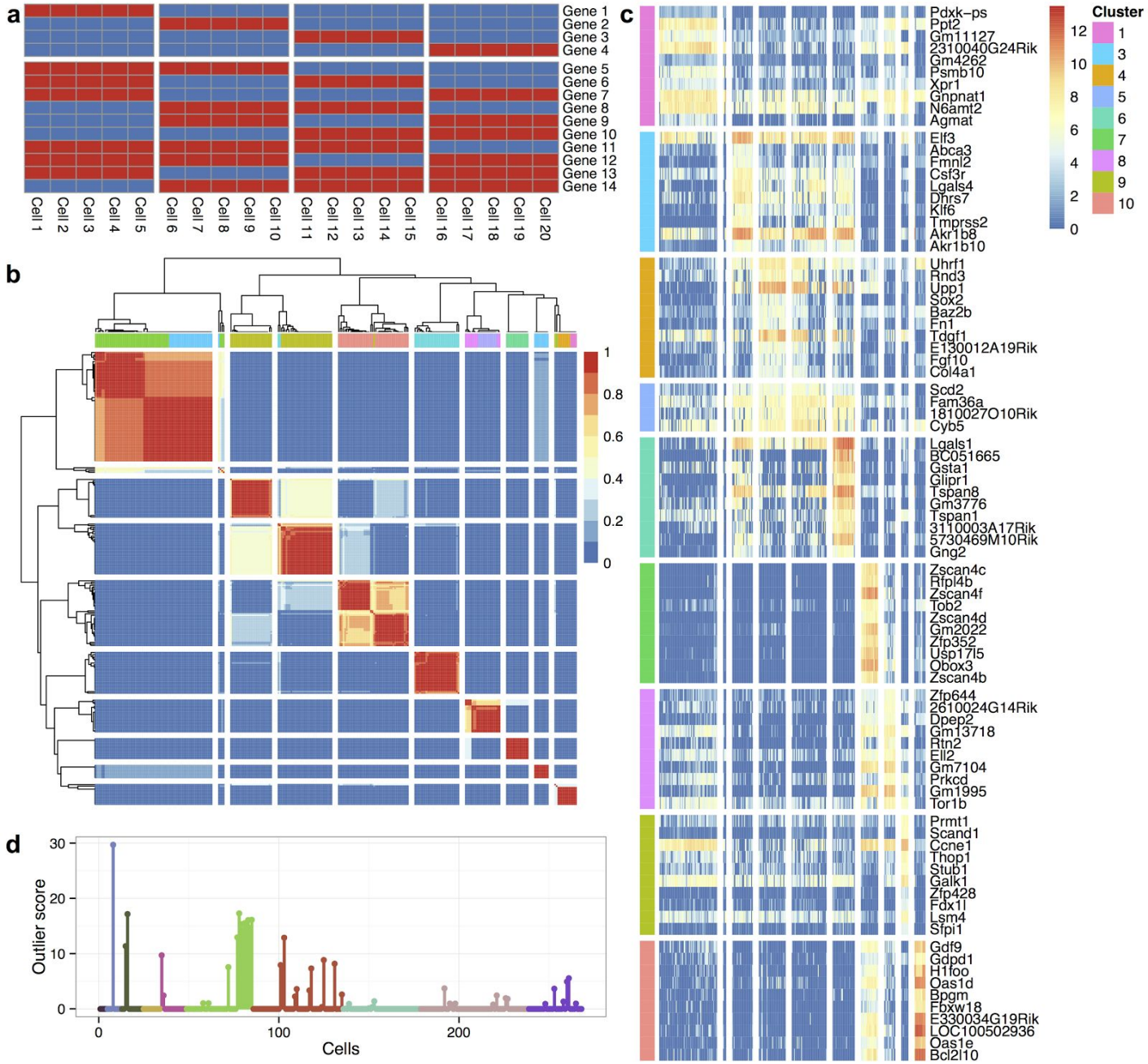


Figure 4. **Biological interpretation of the Deng dataset.** (**a**) Toy example illustrating the difference between marker genes and differentially expressed genes. 20 cells containing only 14 genes with binary expression values (blue for off and red for on) are clustered. Only genes 1-4 can be considered as marker genes, whereas all genes 1-14 are differentially expressed. (**b**) The consensus matrix panel from SC3. The matrix indicates how often each pair of cells was assigned to the same cluster by the different parameter combinations. Dark red indicates that the cells were often assigned to the same cluster whereas dark blue indicates that they were assigned to different clusters most of the time. The colors at the top represent the labels assigned by the authors and the white lines are visual guides to separate the clusters. (**c**) The marker genes panel from SC3. The heatmap shows expression profiles (after Gene Filter and Log-transformation, Methods) of top 10 marker genes for each of the 10 clusters identified by SC3. The colors of the clusters do not correspond to the ones at the top of (**b**). Note that for cluster 2 no marker genes were found and for cluster 5 only four marker genes were found. (**d**) The outlier cells panel from SC3. Cell outliers in the 10 reference clusters (provided by the authors) - nine adjacent cells in the green cluster were prepared using a different protocol (see text for details).

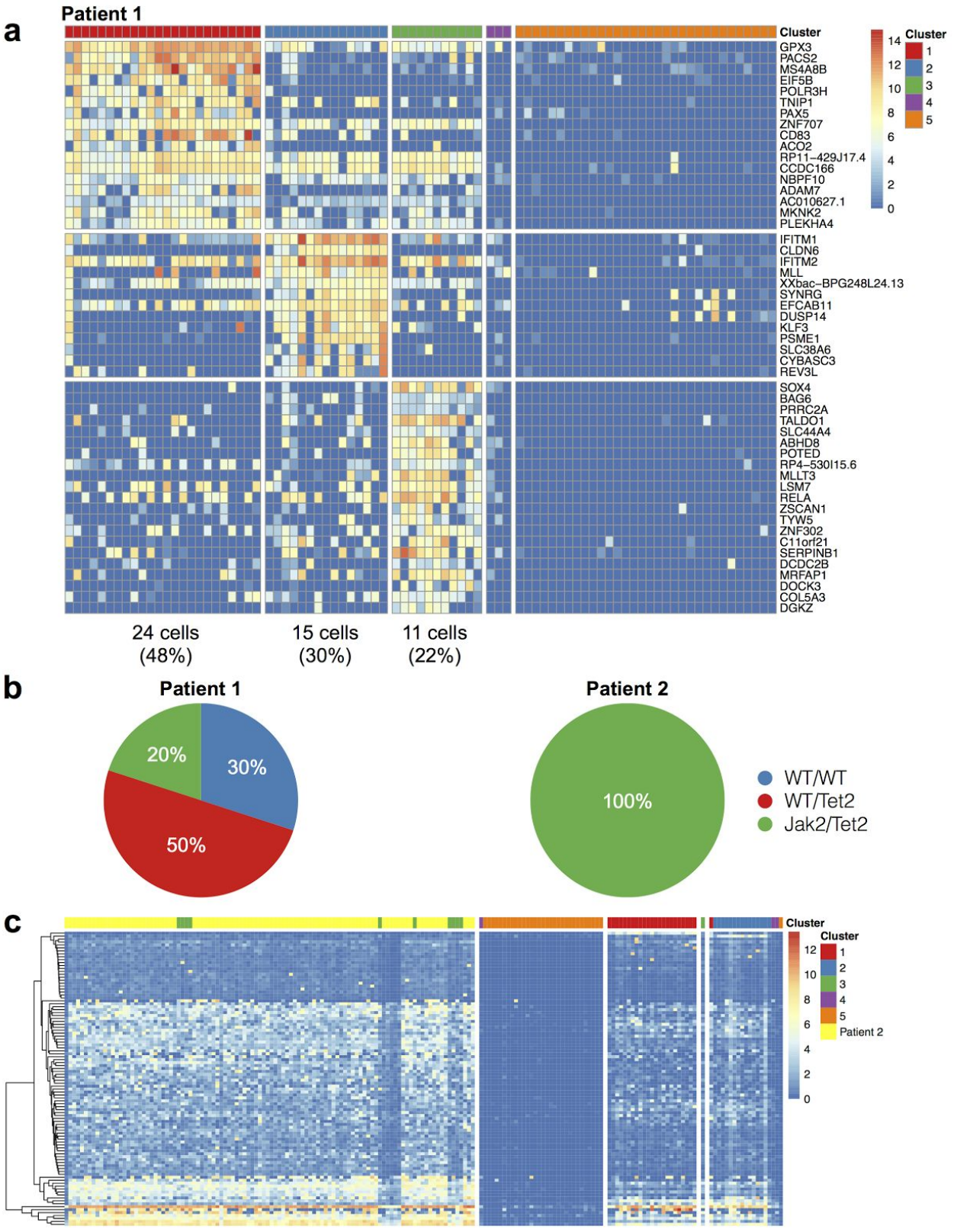


Figure 5. **SC3 clustering of scRNA-seq data from patients.** (**a**) Heatmap of the expression (after Gene Filter and Log-transformation, Methods) of marker genes from clusters 1-3 from patient 1 (clusters 4 and 5 do not have any marker genes). Clusters correspond to k = 5 (Methods). (**b**) Clonal composition of patients 1 and 2 (Methods). (**c**) Heatmap of gene expression (after Gene Filter and Log-transformation, Methods) of a combined dataset (patient 1 + patient 2). Clusters (separated by white vertical lines) correspond to k = 5 (Methods). Cells corresponding to patient 1 are indicated with the same colour as in panel (**a**). Cells from patient 2 are indicated in yellow. On the vertical axis, genes are clustered by k-means with k = 100 and the heatmap represents the expression levels of the cluster centers.