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 Trophoblast 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 cell-type characterisation 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 achieves high accuracy and robustness by consistently integrating different clustering solutions through a consensus approach. Tests on nine published datasets show that SC3 outperforms 4 existing methods, while remaining scalable for large datasets, as shown by the analysis of a dataset containing 44,808 cells. Moreover, an interactive graphical implementation makes SC3 accessible to a wide audience of users, and SC3 also aids biological interpretation by identifying marker genes, differentially expressed genes and outlier cells. We illustrate the capabilities of SC3 by characterising newly obtained transcriptomes from subclones of neoplastic cells 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. One key application of scRNA-seq is the ability to determine cell types based on their 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, but tools from molecular biology have enabled researchers to categorise 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 opens up the possibility 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 its potential to unravel and catalogue the full diversity of cells in the human body.

A full characterisation of the transcriptional landscape of individual cells holds an enormous potential, both for basic biology and clinical applications. An important medical application is cancer, which has long been known to be a heterogeneous disease with multiple subclones coexisting within the same tumor. Until recently, tumor heterogeneity has mainly been assessed at the DNA level by genome sequencing[6–8]. The use of scRNA-seq makes it possible to characterise the transcriptional landscape of the different subclones while utilizing very small number of cells. A better understanding of the transcriptome in the different subclones could thus yield important insights about drug resistance, thereby informing the development of novel therapies. However, due to the large variability in gene expression, identifying subclones from patient transcriptomes remains challenging[9,10].

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 and reliable benchmarks for validation renders this unsupervised clustering a hard problem. For scRNA-seq the challenge is further compounded by technical errors and biases that remain incompletely understood, a high degree of *biological* variability in gene expression[11], and the high dimensionality of the transcriptome.

Although scRNA-seq is a relatively new technology, several groups have developed custom clustering methods for single cell data[2,12–16]. Yet, 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[14,17,18].

We present a novel interactive clustering tool for scRNA-seq data, SC3 (**S**ingle **C**ell **C**onsensus **C**lustering) a user-friendly R-package with a graphical interface. SC3 obtains robust results by combining several well-established techniques using a consensus clustering approach. SC3 has several features to facilitate the evaluation of the clustering quality, and to aid the user in determining the appropriate number of clusters. We demonstrate the performance of SC3 by applying it to nine published datasets and thoroughly benchmarking the method against four other methods. Furthermore, we showcase the scalability of SC3 to very large datasets by analysing a dataset with ~45,000 cells[15]. Crucially, 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. We apply SC3 to the first single cell RNA-Seq data from hematopoietic stem cells isolated directly from patients with myeloproliferative neoplasm. Myeloproliferative neoplasms are a heterogeneous disease, with each patient typically harbouring multiple neoplastic subclones that coexist for long periods of time[19]. Therefore, transcriptome data from patients are typically hard to interpret due to its inherent heterogeneity within these cells. Here, we identify clusters corresponding to different subclones within two patients with different mutational landscapes, and we characterise the differences between their expression profiles.

Results

Consensus clustering as a robust methodology

The output of a scRNA-seq experiment is typically represented as an expression matrix 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, spectral transformations combined with *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 spectral transformation is a nonlinear step whereby the first *d* eigenvectors of the distance matrix are retained 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 particular 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 to obtain a set of optimized clusterings. Instead of trying to pick the optimal solution from this set, we combine *all* the different clustering outcomes into a consensus matrix that summarises how often each pair of cells is located in the same cluster. The final result provided by SC3 is determined by complete-linkage hierarchical clustering of the consensus matrix into *k* groups. 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[9,14,17,18,20–23], ranging from $N = 80$ to $N = 3{,}005$ single cells (Fig. 1b). For each dataset, we used the clustering provided by the authors of the original study as *reference labels* (i.e., the ground truth) against which we compared the clusters obtained by SC3. To quantify the similarity between clusterings, we used the Adjusted Rand Index (ARI, see Methods) which ranges from 1, when the clusterings are identical, to 0 when the similarity is what one would expect by chance. A wide variety of methods and parameters to include within our consensus approach was evaluated by first analysing three of the eight datasets (Fig. 1c). We found that clustering performance was largely unaffected by the cell filter, and for different choices of gene filter, distance metrics and spectral transformations, the ARI values were overall similar (Fig. 1c, S1-S3). In contrast, we found that the quality of the outcome as measured by the 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 for our type of data. It is likely that this range will no longer be optimal for other data, e.g. a small yeast genome or a sample with much more shallow coverage. We thus use this range of parameters within our consensus approach to narrow down the parameter space to be explored. Strikingly, consensus clustering[24] of the solutions obtained in this range further improves the ARI (Fig. S4).

To benchmark SC3, we considered four other methods: tSNE[25] followed by *k*-means clustering (a method similar to those used by Grün et al[1] and Macosko et al[15]), pcaReduce[13], SNN-Cliq[12] and

SINCERA[16]. SC3 performs better than the four tested methods across all datasets, with the exception of the Pollen1 dataset (Fig. 2). In contrast to the other methods that rely on different initializations (pcaReduce and tSNE+kmeans), SC3 is highly stable, and except for the Usoskin and Zeisel data, a single dominant solution is obtained. 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.

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 $O(N^3)$)[26]. This scaling makes it impractical to use the version of SC3 described above for datasets with >1,000 cells. To apply SC3 to larger datasets, we have implemented a hybrid approach that combines unsupervised and supervised methodologies. When the number of cells is large (N>1,000), SC3 selects a small subset of cells uniformly at random, and obtains clusters from this subset as described above. Subsequently, the inferred labels are used to train a support vector machine (SVM, Methods), which is employed to assign labels to the remaining cells. Training the SVM typically takes only a few minutes, thus allowing SC3 to be applied to very large datasets.

To test the SVM in isolation, we used the validation datasets from Fig. 1c with the authors' reference labels 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). For the situation where we first assign labels using the unsupervised clustering described above, we find that it is still possible to achieve an ARI>0.8 with only 20% of the training cells for the larger datasets (Fig. 3b). Overall, the performance of the SVM improves with higher N, and for the larger datasets (Klein and Zeisel), an ARI of 0.8 can be achieved with <10% of the cells used as a training set.

Using this approach, we are able to analyse a large Drop-Seq dataset with N = 44,808 cells and k = 39 clusters[15]. 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. Interestingly, in the case of this dataset SC3 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[15]. This implies that the original cluster 24 may be more heterogeneous than suggested by the authors.

SC3 features an interactive and user-friendly interface

To increase its usability, SC3 features a graphical user interface displayed through a web browser, thus minimizing 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. SC3 will then 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. 4a). 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[27], a measure of how tightly grouped the cells in the cluster are (Fig. S10).

SC3 assists with biological interpretation

A key aspect to evaluate 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 biology of 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[28] 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 all the remaining ones (Fig. 4b). To select marker genes, SC3 uses a binary classifier based on the gene expression to distinguish one cluster from all the others. For each gene, a receiver operator characteristic (ROC) curve is calculated and the area under the ROC curve is used to rank the genes. The area under the ROC curve provides a quantitative measure of how well the gene can distinguish one cluster from the rest. The most significant candidates are then reported as marker genes (Methods). Cell outliers are also identified by SC3 through the calculation of a score for each cell using the Minimum Covariance Determinant[29]. 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. 4c). The outlier score helps to identify cells that could correspond to, e.g., rare cell-types or technical artefacts. In addition, SC3 facilitates obtaining a gene ontology analysis for each cluster by directly exporting the list of marker genes to WebGestalt[30]. All 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, including zygote, 2-cell, 4-cell, 8-cell, 16-cell and blastomere. Based on the silhouette index, SC3 recommends a clustering into either k=2 groups or k=10 groups (Fig. S10). The solution for k=2 identifies one cluster with the blastocyst and one with the remaining cells. The most stable result for k=10 is shown in Fig. 4a, and our clusters largely agree with the known sampling timepoints. However, our results suggest that the difference between the 8 and 16 cell stages is quite small. The latter stages of development are labelled “early”, “mid” and “late” blastomeres, although it is well known that these stages consist primarily of trophoblasts and inner cell mass. Interestingly, SC3 suggests that the mid-blastocyst stage could be split into two groups, which most likely correspond to trophoblasts and the inner cell mass (clusters 3 and 4). This conclusion is supported by the fact that *Sox2* and *Tdgf1* (inner cell mass markers) are listed as marker genes in cluster 4[31] (Table S2). In total, we identified ~3000 marker genes (Table S2), many of which had been previously reported as specific to the different developmental stages[32–36]. Furthermore, the analysis revealed several genes specific to each developmental stage which had previously not been reported (Table S2). Importantly, using the reference labels reported by the authors[22], we identified nine cells with high outlier scores (green cells in Fig. 4c) which were prepared using the Smart-Seq2 protocol instead of the Smart-Seq protocol[12,22], thus demonstrating the ability of SC3 to identify technical outliers.

SC3 characterises myeloproliferative neoplasm subclones

Myeloproliferative neoplasms, a group of diseases characterised by the overproduction of terminally differentiated cells of the myeloid lineage, reflect an early stage of tumorigenesis where multiple subclones are known to coexist in the same patient[19]. From exome sequencing data, we previously identified TET2 and JAK2V61F as the only driver mutations in a large patient cohort [37]. In this paper, we use two patients out of this cohort which harbour TET2 and JAK2V617F mutations in their stem cell compartment. Haematopoietic stem cells (HSCs) are thought to be the cell of origin in MPNs. However, little is known about the transcriptional consequences of driver mutations on the stem cell in HSCs. We obtained scRNA-seq data from HSC from two different patients (Methods). For patient 1 (N = 51), SC3 suggested that k = 3 provides the best clustering, revealing three clusters of similar size (Fig. S14). For patient 2 (N = 89) SC3 indicated k=1, suggesting that one single cluster per patient might best reflect the underlying transcriptional changes.

Since known driver mutations in our patients are the *TET2* and *JAK2V617F* loci[38] we hypothesized that the different clusters correspond to different combinations of mutations within different clones of each patient. Unfortunately, the coverage of the *JAK2V617F* and *TET2* loci was insufficient to reliably determine the genotype of each cell directly from the scRNA-seq data. Instead the genotype composition for each HSC clone was determined by growing individual hematopoietic stem cells into granulocyte/macrophage colonies, followed by Sanger sequencing of the TET2 and JAK2V617F loci (Fig. 5a). In agreement with the clustering defined by SC3, patient 1 (k=3) was found to harbor three different subclones: (i) cells with both TET2 and JAK2V617F mutations, (ii) cells with a TET2 mutation and (iii) wild-type cells (Fig. 5b). Strikingly, the SC3-clusters contain 22%, 29% and 49% of the cells, in excellent agreement with the proportions of each genotype found in the patient, namely 20%, 30% and 50%. Thus, we hypothesize that cluster 1 corresponds to the double mutant, cluster 2 corresponds to cells with only a TET2 mutation, and cluster 3 corresponds to wild-type cells. The HSC compartment of patients 2 was 100% mutant for TET2 and JAK2V617F, which was consistent with clustering of k=1 suggested by SC3 (Fig. S16 and S19).

Three additional lines of evidence support the assumption that SC3 can help to define subclonal composition. Firstly, we analysed data from patient 2, with a dominant double mutant clone, together with all cells from patient 1, harbouring three different clones. SC3 clustering again suggested k = 3 (Fig. 5b, S17). Most importantly, all of the putative double mutant cells from patient 1 were grouped with the cells from patient 2. SC3 also reported 33 marker genes for the putative *TET2* mutant and 202 marker genes for the putative double mutant clone (Fig. 5c, Table S4). Secondly, we used microarray data from BFU-E colonies available for patient 1[37] where the genotype of each clone was linked to a specific transcriptional signature. When comparing differentially expressed genes for the double mutant clone from BFU-E 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 further strengthens our assumption that the double mutant clone identified through clustering by SC3 indeed harbours both mutations. This result demonstrates that SC3 is capable of finding clusters defined by mutations rather than by patient batch. By contrast, none of the other clustering methods was able to clearly identify clusters which could readily be related to the genotype information (Fig. S15, Methods).

Lastly, we performed pathway analysis using marker genes. KEGG Pathway analysis (Methods) of the latter group revealed that one of the most upregulated categories was “Chemokine signalling” (Table S5) with Wikipathway analysis (Methods and Table S6) highlighting interleukin and EPO receptor signalling pathways as being enriched. Therefore, ligands 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 normalized variances revealed that the putative TET2 mutants have more variable transcriptomes than putative wild-type cells (Mann-Whitney test *p*-value <2.2e-16, Fig. S18). In summary, we can, for the first time, confer genotype information from patient data and infer pathways and putative disease genes important for disease maintenance for HSCs.

Discussion

We have presented SC3, an interactive tool for unsupervised clustering of scRNA-seq data. By comparing to several other 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 scale the method to very large experiments, e.g. Drop-seq[15]. 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. The ability of SC3 to identify marker genes and differentially expressed genes is important when analysing complex scRNA-seq datasets. Traditional methods for differential expression analysis[39,40] are impractical since a total of $k(k-1)/2$ comparisons are required, as only two groups at a time are compared. Each comparison provides a list of genes, and extracting the information directly provided by SC3 would require a large amount of post-processing.

A major challenge when developing unsupervised clustering algorithms for scRNA-seq 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 other datasets, however, the labelling was based on a combination of the authors' clustering methods and their biological knowledge. In these latter cases, the labelling is less reliable, and we cannot be certain that the original clustering represents a meaningful ground truth. Reassuringly, the experimental consistency in the original labels correlates well with the observed accuracy of SC3; for the datasets where the labels are almost certainly correct (Ting, Patel and Klein) SC3 obtains almost perfect ARIs close to 1, whereas for the remaining, less certain datasets, SC3 reaches good 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[41] and Davies-Bouldin[42] indexes, can be used to choose *k*, but we did not find any correlation between them and the external indexes (Rand, ARI and Jaccard[43]) (Fig. S5). 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.

Although significant progress has been made on understanding the mutations leading to cancer[19], much less is known about the differences between subclones within the tumour at the transcriptional level. We applied SC3 to scRNA-seq data from two patients diagnosed with myeloproliferative neoplasms. We found strong evidence in support of the hypothesis that the clusters revealed by SC3 directly correspond to subclones identified by independent experiments[37]. Moreover, we used the marker genes identified by SC3 to provide a biological characterisation of the different subclones (Fig. 5b, c). Our results demonstrate that it is possible to identify subclones using scRNA-seq, and that the analysis of the transcriptome can provide significant insights into the functional consequences of different mutations. We aim to investigate the role of the identified marker genes in future studies to further demonstrate the value of the characterisation of the transcriptome of individual subclones.

As sequencing costs decrease, larger scRNA-seq datasets will become increasingly common, 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 analysing 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[17] was acquired from personal communication with Alex A Pollen.

SC3 clustering

SC3 takes as input an expression matrix ***M*** where columns correspond to cells and rows correspond to genes. Each element may correspond either to the number of transcripts present or number of reads depending on the details of the experimental protocol. By default SC3 does not carry out any form of normalization or correction for batch effects. SC3 is based on seven elementary steps. The parameters in each of these steps can be easily adjusted by the user, but are set to sensible default values, determined via cross-validation on the Deng, Pollen and Usoskin datasets.

1. Cell Filter
The cell filter is an optional feature that 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, which contain less than a specified number of expressed genes (genes with measured non-zero expression level). The default in SC3 for the minimum number of expressed genes is set to 2,000.

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 associated graph Laplacian ($\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, where ***I*** is the identity matrix, ***A*** is the distance matrix and ***D*** is the degree matrix of ***A***). The columns of the resulting matrices are then sorted in descending order by their corresponding eigenvalues. The MDS method was not included in SC3 because of poor performance (Fig. S1-3).

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[44]. The maximum number of iterations was set to 10^9 and the number of starts was set to 1,000.

7. Consensus clustering

SC3 computes a consensus matrix using the Cluster-based Similarity Partitioning Algorithm (CSPA)[24]. For each clustering result 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 (Fig. 1a). A consensus matrix is calculated by averaging all similarity matrices. This can be split in 2 steps:

7a. Consensus over the d range

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 time, 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

Consensus over the parameter set takes multiple results of the *consensus over the d range* and includes additional averaging of similarity matrices over the distance measures and transformations. As a final result SC3 uses consensus averaging over all six combinations of distances and transformations by default.

7c. Consensus clustering

The resulting consensus matrix (after *consensus over the parameter set*) 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. 1a).

Fig. S4 shows how the quality and the stability of clustering improves after *consensus clustering*.

Adjusted Rand Index

If cell-labels are available (e.g. from a published dataset) the Adjusted Rand Index (ARI)[45] 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[43]) and internal (Dunn index[41] and Davies-Bouldin index[42]) (Fig. S5) 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 good agreement with each other, there was little or no correlation between the external and the internal indexes.

Benchmarking

Before benchmarking we applied the Gene Filter to all the datasets. For tSNE+k-means, SNN-Cliq and pcaReduce we additionally applied Log-transformation as described above (***M'*** = log2(***M*** + 1)). For SINCERA we used the original z-score normalisation[16] instead of the Log-transformation. For tSNE the Rtsne() function of the Rtsne R package was used with the default parameters.

Support Vector Machines (SVM)

Results shown in Fig. 3 were obtained using the following steps. When a dataset contains more than 1,000 cells, we randomly select and cluster 1,000 cells. Next, a support vector machine (SVM)[46] 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. S7-S9) 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.

Based on the results presented in Fig. 3, we set the default of SC3 so that when N>1,000 only 20% of the cells are used and when N>5,000 only 1,000 cells are used for clustering before training the SVM.

Biological insights

Identification of differential expression

Differential expression is calculated using the non-parametric Kruskal-Wallis test[28], 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 p-values<0.01, corrected for multiple testing (using the default "holm" method of p.adjust() R function) and plots gene expression profiles of the 50 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 advise to use the p-values for ranking the genes only.

Identification of marker genes

For each gene a binary classifier is constructed based on the mean cluster expression values. 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 corresponds to the 99% quantile of the AUROC distributions obtained from 100 random permutations of cluster labels for all datasets (Table S1 and Fig. S11). 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 PCA[47]. Second, robust distances are calculated using the minimum covariance determinant (MCD)[29]. We then used a threshold based on the 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).

Patients

All 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. S12) and grown in a cytokine cocktail designed to promote progenitor expansion as previously described[48]. For scRNA-seq studies, single HSCs were directly sorted into lysis buffer as described in Picelli et al[49].

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 by Ortmann *et al*[37].

Single cell RNA-Sequencing

Single HSCs were sorted into 96-well plates and cDNA generated as described previously[49]. The Nextera XT library making kit was used for library generation as described by Picelli *et al*[49].

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[50] was used to assess the sequence quality. Foreign sequences from the Nextera Transposase agent were discovered and subsequently removed with Trimmomatic[51]. The reads were trimmed to 90 bases to remove low quality positions before being mapped with TopHat[52] to the Ensembl[53] reference genome version GRCh38.77 augmented with the spike-in controls downloaded from the ERCC consortium[54]. Counts of uniquely mapped reads in each protein coding gene and each ERCC spike-in were calculated using SeqMonk (http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk) and were used for further downstream analysis. Quality control of the cells contained two steps: 1.

filtering of cells based on the number of expressed genes; 2. filtering of cells based on the ratio of the total number of ERCC spike-in reads to the total number of reads in protein coding genes. Filtering threshold were manually chosen by visual exploration of the quality control features (Fig. S13). After filtering, 51 and 89 cells were retained from patient 1 and patient 2, correspondingly. The expression values in each dataset were then normalised by first using a size-factor normalisation (from DESeq2 package[55]) to account for sequencing depth variability. Secondly, to account for technical variability, a normalisation based on ERCC spike-ins was performed using the RUVSeq package[56] (RUVg() function with parameter k = 1). For combined patient data, normalisation steps were performed after pooling the cells. The resulting filtered and normalised datasets were clustered by SC3.

Clustering of patient scRNA-seq data by SC3

We clustered scRNA-seq data from patient 1 and patient 2 separately as well as a combined dataset containing data from patient 1 + patient 2. For patient 1 the best clustering was achieved for k = 3 (Fig. S14). Data from patient 2 was homogeneous and we were unable to identify more than one meaningful cluster (Fig. S15). For the combined dataset for patient 1 + patient 2 the best values of k were 2 and 3 (Fig. S17). In both cases all of the cells from cluster 1 in patient 1 were grouped with the cells from patient 2. For k = 3 clusters 1 and 3 of patient 1 were also resolved.

Comparison of clustering of patient 1 scRNA-seq data

Among the tools (SC3, SNN-Cliq and SINCERA) that can predict k, SC3 was the only one predicting 3 clusters for the patient 1 dataset (Fig. S14). SNN-Cliq and SINCERA predicted 2 and 1 clusters, respectively. We compared SC3 clustering of the patient 1 data for k = 2, 3, 4 and 5 to the other methods by evaluating the clustering and its stability (Fig. S15). The stability was defined by running each stochastic method 100 times, then taking one of the solutions and comparing it to the other 99 solutions using the ARI. The stability was defined as “number of times the ARI equals 1” + 1. The stability of the deterministic methods (SNN-Cliq and SINCERA) was set to 100. SC3 identified the same 11 cells of cluster 1 of the patient 1 (red cells on Fig. S15a) at three levels of granularity (k = 3, 4 and 5) with a stability of 100 (Fig. S15b). SINCERA found a similar cluster when k was set to 4 or 5. tSNE+kmeans was able to identify 3 clusters similar to the ones identified by SC3, but the perplexity parameter had to be adjusted manually (perplexity = 16) and the solutions were unstable (stability = 1). pcaReduce was unable to find the 3 clusters and had a stability of less than 20 (exact value depends on the choice of random seed).

Identification of differentially expressed genes from microarray data

The microarray data of patient 1 (EE50) was obtained from Array Express accession number E-MTAB-3086[37]. One replicate (2B) was identified as an outlier and removed. The limma R package[57] was used to identify 932 differentially expressed genes between WT and TET2/JAK2V617F double mutant using an adjusted (by false discovery rate) p-value threshold of 0.1.

Marker genes analysis for patients

For both patients, to increase the number of marker genes, the AUROC threshold was set to 0.7 instead of the default value of 0.85 and the 0.1 *p*-value threshold was chosen.

Pathway enrichment analysis

We utilized WebGestalt[30] web tool to find pathway enrichments in obtained set of marker genes. By default we used the KEGG Analysis and Wikipathways Analysis options with a *p*-value of 0.1.

Contributions

MH conceived the study; VYK, MH, MS, MB and TA contributed to the computational framework; KK and TC performed the experiments for the patient data; KNN helped with the analysis of embryonic mouse data; MB, WR, ARG and MH supervised the research; VYK 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 GSE79102.

Software availability

SC3 is available as a R package at http://bioconductor.org/packages/SC3/.

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, Daniel Ramsköld and Meri Huch 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. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
7. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
8. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (2016).
9. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).
10. Min, J.-W. *et al.* Identification of Distinct Tumor Subpopulations in Lung Adenocarcinoma via Single-Cell RNA-seq. *PLoS One* **10,** e0135817 (2015).
11. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297,** 1183–1186 (2002).
12. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv088
13. Zurauskiene, J. & Yau, C. pcaReduce: Hierarchical Clustering of Single Cell Transcriptional Profiles. *bioRxiv* 026385 (2015). doi:10.1101/026385
14. 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).
15. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
16. 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).

17. 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).
18. Usoskin, D. *et al.* Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* **18,** 145–153 (2015).
19. Chen, E., Staudt, L. M. & Green, A. R. Janus kinase deregulation in leukemia and lymphoma. *Immunity* **36,** 529–541 (2012).
20. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).
21. 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).
22. 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).
23. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell* **161,** 1187–1201 (2015).
24. Strehl, A. & Ghosh, J. Cluster Ensembles --- a Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **3,** 583–617 (2003).
25. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
26. 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).
27. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
28. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **47,** 583–621 (1952).
29. Hubert, M. & Debruyne, M. Minimum covariance determinant. *WIREs Comp Stat* **2,** 36–43

(2010).

30. 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).
31. Marikawa, Y. & Alarcón, V. B. Establishment of trophectoderm and inner cell mass lineages in the mouse embryo. *Mol. Reprod. Dev.* **76,** 1019–1032 (2009).
32. 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).
33. Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2014 update. *Nucleic Acids Res.* **42,** D818–24 (2014).
34. 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).
35. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev. Cell* **35,** 366–382 (2015).
36. Blakeley, P. *et al.* Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142,** 3151–3165 (2015).
37. Ortmann, C. A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372,** 601–612 (2015).
38. Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* **369,** 2391–2405 (2013).
39. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).
40. Delmans, M. & Hemberg, M. Discrete distributional differential expression (D(3)E) - a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* **17,** 110 (2016).
41. 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).

42. Davies, D. L. & Bouldin, D. W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1,** 224–227 (1979).

43. 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).

44. 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).

45. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66,** 846–850 (1971).

46. Ben-Hur, A., Horn, D., Siegelmann, H. T. & Vapnik, V. Support Vector Clustering. *J. Mach. Learn. Res.* **2,** 125–137 (2002).

47. Hubert, M., Rousseeuw, P. J. & Branden, K. V. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics* **47,** 64–79 (2005).

48. 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).

49. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9,** 171–181 (2014).

50. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *Reference Source* (2010).

51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

52. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

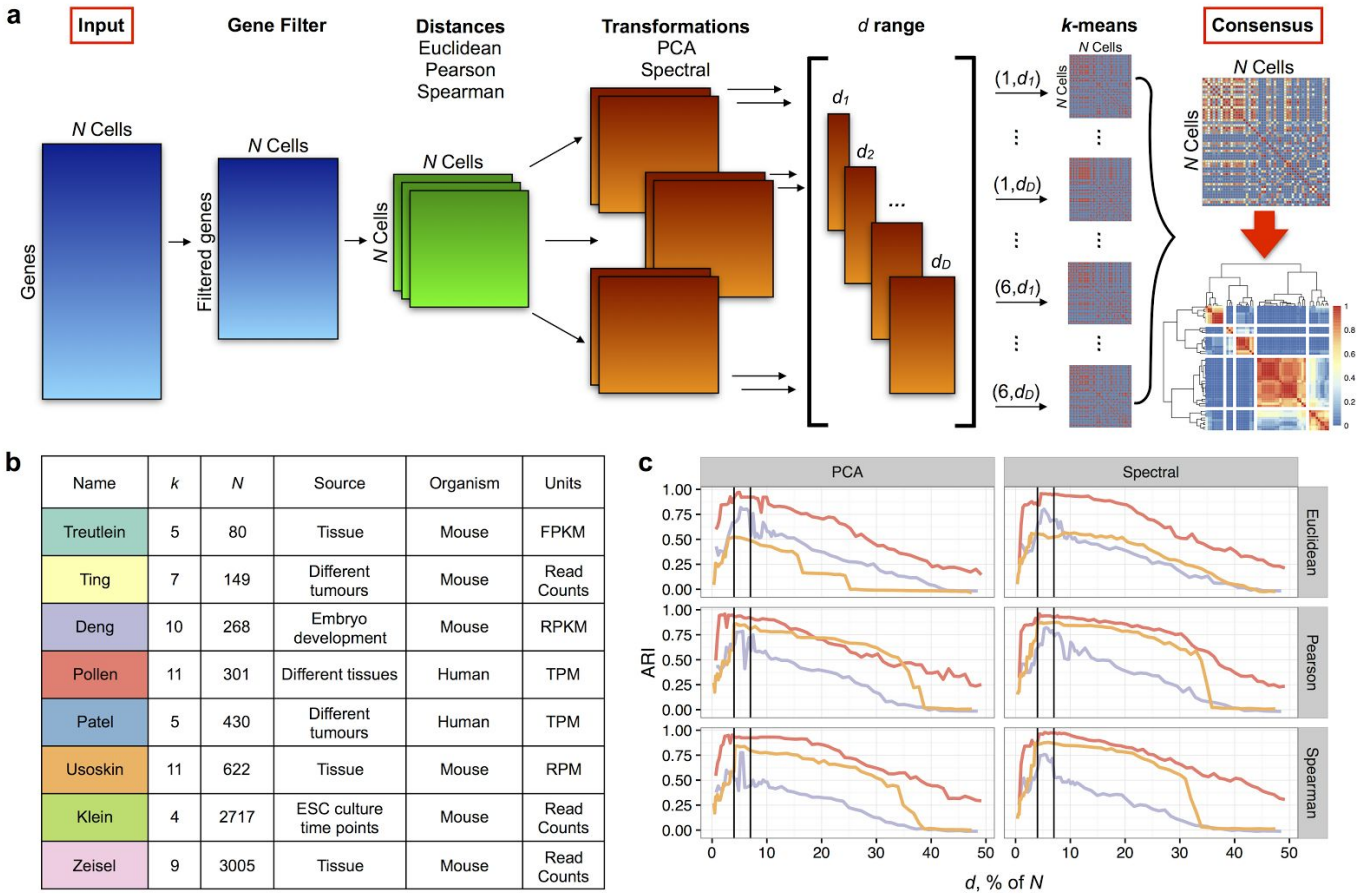
53. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–9 (2015).

54. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat. Methods* **2,** 731–734 (2005).

55. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

56. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32,** 896–902 (2014).

57. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).



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. **The SC3 framework for consensus clustering.** (**a**) Overview of clustering with SC3 framework (see Methods). A total of 6*D* clusterings are obtained, where *D* is the total number of spectral dimensions d_1, …, d_D considered, which are then combined through a consensus step to increase accuracy and robustness. The consensus step is based on the construction of the consensus matrix, exemplified here with the Treutlein data: binary matrices (Methods) corresponding to each clustering solution in are averaged, and the resulting matrix is clustered using hierarchical clustering up to the *k* level of the hierarchy (*k* = 5 in this example). (**b**) Summary of published datasets used to validate SC3. *N* is the number of cells in a dataset; *k* is the number of clusters originally identified by the authors [9,14,17,18,20–23]; Units: 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, TPM is Transcripts Per Million mapped reads. (**c**) Testing the distances, nonlinear transformations and *d* range. Median of ARI over 100 realizations of the SC3 clustering for three validation datasets (Deng, Pollen and Usoskin, colours as in (**b**)). The x-axis shows the number of eigenvectors *d* (see (**a**)) as a percentage of the total number of cells, *N*. The black vertical lines indicate the interval *d* = 4-7% of the total number of cells *N,* showing high accuracy in the classification

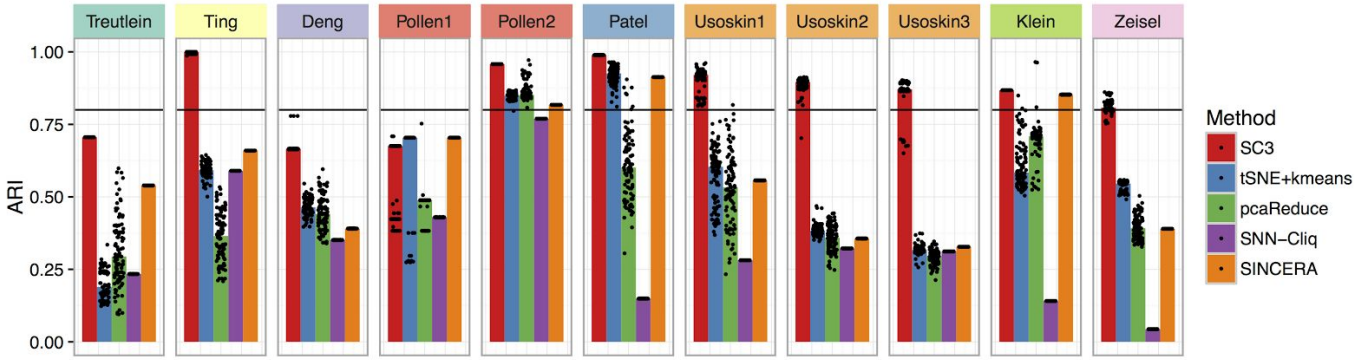


Figure 2. **Benchmarking of SC3 against existing methods.** SC3, tSNE+kmeans and pcaReduce were applied 100 times to each dataset to evaluate accuracy and stability. SNN-Cliq and SINCERA are deterministic and were thus run only once. Each panel shows the similarity between clusterings, as quantified by the Adjusted Rand Index (ARI, see Methods) which ranges from 1, when the clusterings are identical, to 0 when the similarity is what one would expect by chance. For each run, the ARI was calculated comparing the cluster assignments with the reference labels (black dots). The top of each bar corresponds to the median of the distribution of the black dots. For the Pollen and Usoskin datasets we considered all the different hierarchies reported in the original papers (Pollen1 k = 4, Pollen2 k = 11, Usoskin1 k = 4, Usoskin2 k = 8, Usoskin3 k = 11).

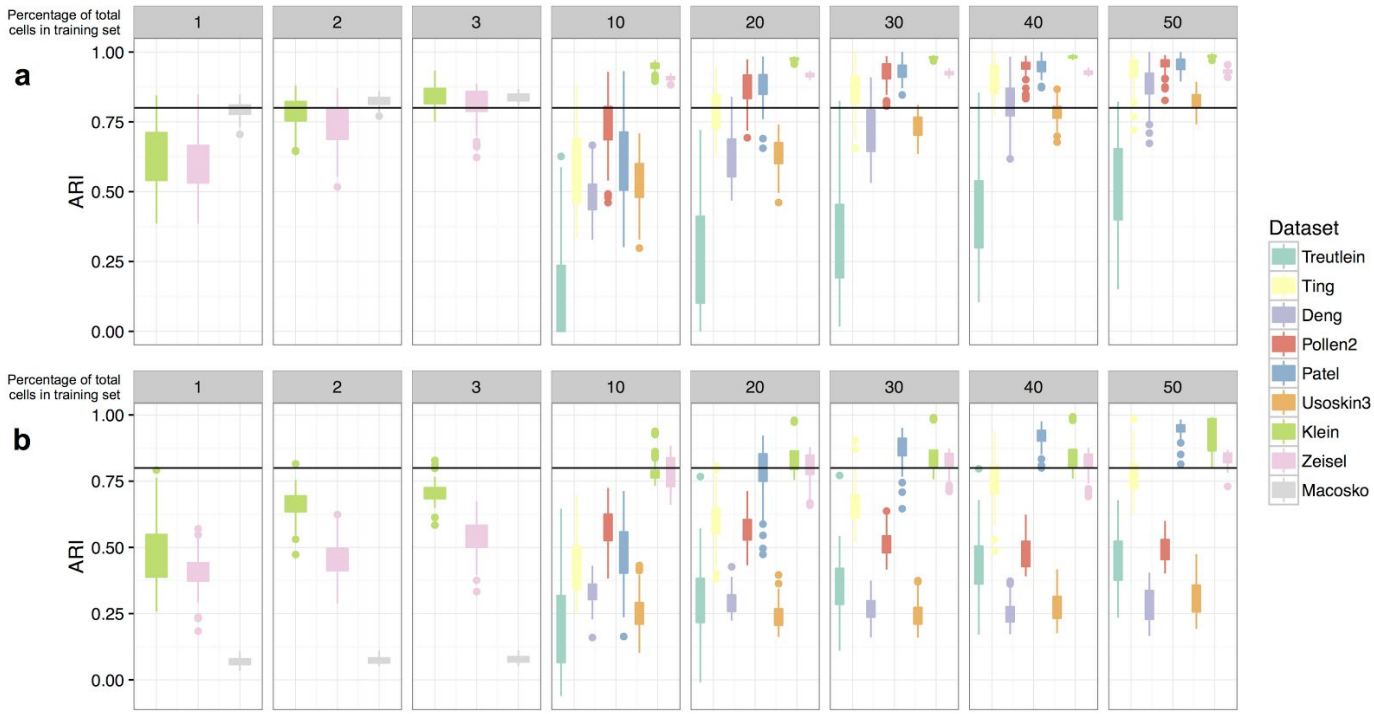


Figure 3. **SC3 can be scaled to large datasets using a hybrid approach.** The numbers inside the grey boxes correspond to the number of cells, as the % of the total number of cells *N,* used to train the linear kernel support vector machine (SVM). As the percentage of training cells increases, the accuracy as defined by the Adjusted Rand Index (ARI) increases. The results indicate that accurate clustering can be achieved with only a small percentage of all cells used for training. (**a**) When the reference labels (ground truth provided by the authors) are used for training, SVM prediction reaches levels close to ARI=1. 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. (**b**) When the labels for training are calculated by SC3 without any other input, the ARI levels still reach high levels of ARI ~0.8.

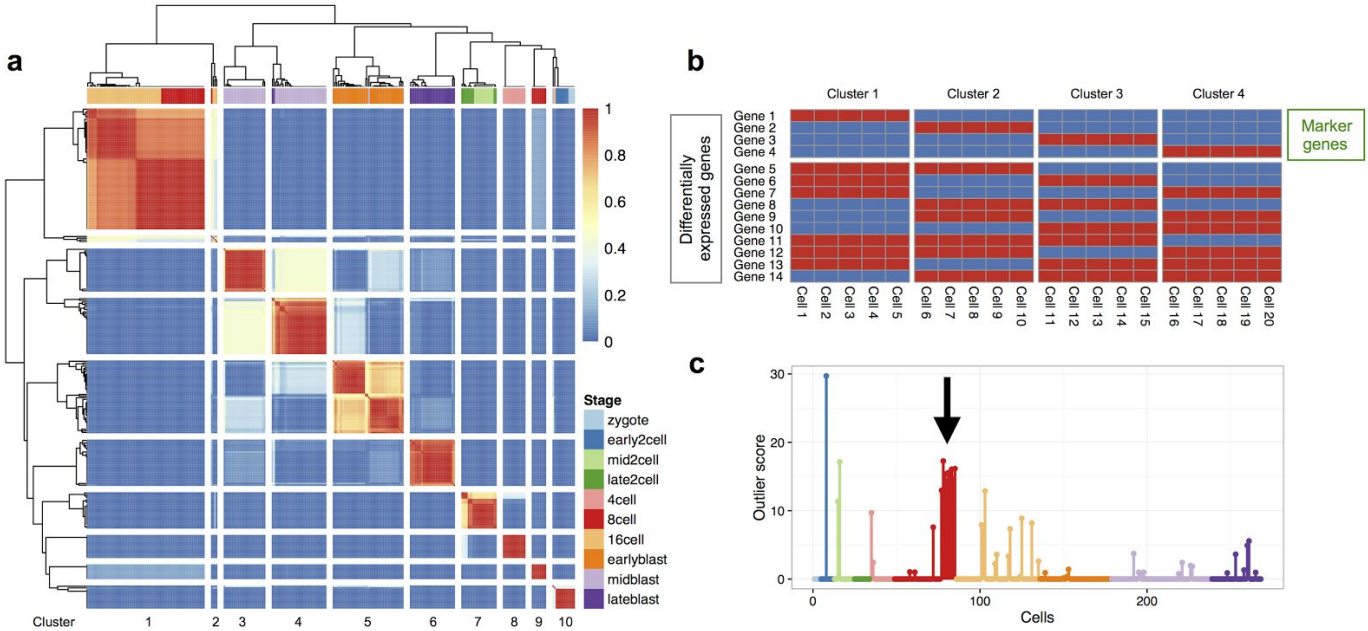


Figure 4. **Applying SC3 to the Deng dataset aids biological interpretation.** (**a**) The consensus matrix panel as generated by SC3. The matrix indicates how often each pair of cells was assigned to the same cluster by the different parameter combinations. Dark red (1) indicates that the cells were always assigned to the same cluster whereas dark blue (0) indicates that they were never assigned to the same cluster. In this case, SC3 finds a clustering with k = 10 clusters, separated by the white lines as visual guides. The colors at the top represent the reference labels, corresponding to different stages of development (see colour guide). (**b**) 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. (**c**) Outlier scores for all N= 268 cells as generated by SC3 (colors correspond to the 10 reference clusters provided by the authors - Stage in (**a**)). 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.

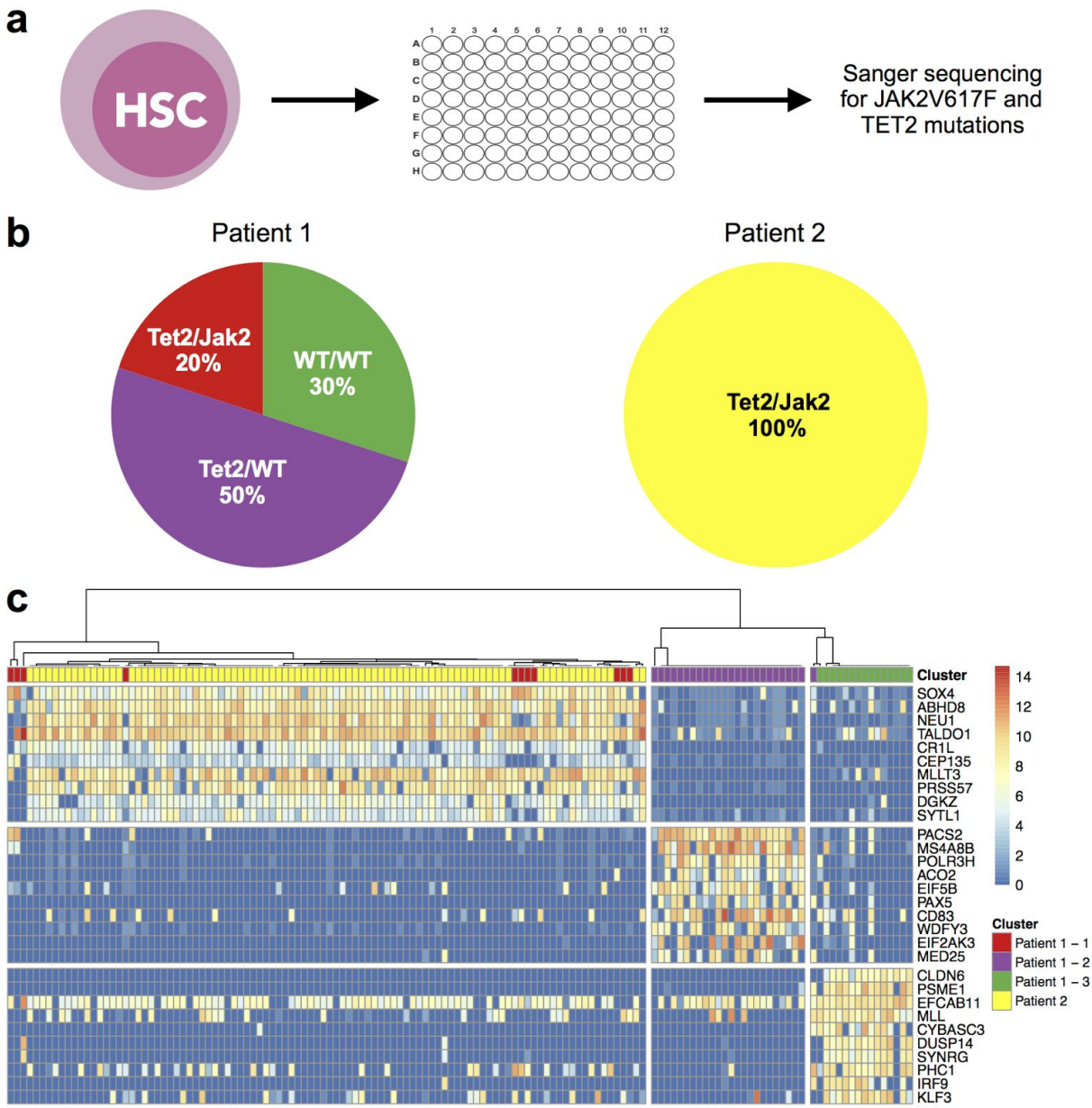


Figure 5. **Using SC3 to define subclones from two patients with myeloproliferative neoplasm.** (**a**) Individual HSCs were placed into wells, grown into granulocyte/macrophage colonies, and the *JAK2V617F* and the *TET2* loci were characterised using Sanger sequencing. (**b**) Clonal composition of patients 1, 2 obtained by independent sequencing experiments of the *JAK2V617F* and the *TET2* loci (Methods). (**c**) Marker gene expression (after Gene Filter and Log-transformation, Methods) of the combined dataset (patient 1 + patient 2). Clusters (separated by white vertical lines) correspond to k = 3 (Methods). Cells corresponding to patient 1 are indicated with the same colour as in panel (**b**). Cells from patient 2 are indicated in yellow. Only the top 10 marker genes are shown for each cluster.