SC3: consensus clustering of single-cell RNA-seq data

Vladimir Yu Kiselev[1], Kristina Kirschner[2], Michael T Schaub[3,4], Tallulah Andrews[1], Andrew Yiu[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]

Single-cell RNA-seq enables the quantitative characterization of cell types based on global transcriptome profiles. We present single-cell consensus clustering (SC3), a user-friendly tool for unsupervised clustering, which achieves high accuracy and robustness by combining multiple clustering solutions through a consensus approach (http://bioconductor.org/packages/SC3). We demonstrate that SC3 is capable of identifying subclones from the transcriptomes of neoplastic cells collected from patients.

A key advantage of single-cell RNA sequencing (scRNA-seq) is that it can be used to determine cell types in an unbiased way by submitting transcriptomes to unsupervised clustering[1–3]. A full characterization of the transcriptional landscape of individual cells holds enormous potential for both basic biology and clinical applications. However, *de novo* identification and characterization of cell types requires robust and accurate computational methods. We have developed SC3, an interactive and user-friendly R package for clustering (**Supplementary Software 1** and see http://bioconductor.org/packages/SC3 for the latest version). Its integration with Bioconductor[4] and scater[5] makes it easy to incorporate into existing workflows.

Each step of the SC3 pipeline (**Fig. 1a** and Online Methods) requires the user to specify a number of parameters, which can be difficult and time-consuming to optimize. To avoid this problem, SC3 utilizes a parallelization approach whereby a significant subset of the parameter space is evaluated simultaneously to obtain a set of clusterings. SC3 then combines all the different clustering outcomes into a consensus matrix that summarizes how often each pair of cells is located in the same cluster. The final result is determined by complete-linkage hierarchical clustering of the consensus matrix into k groups.

To constrain parameter values in the SC3 pipeline, we first considered six publicly available scRNA-seq datasets[6–11] featuring high-confidence cell labels (since they include cells from different stages, conditions or lines) that can be considered gold standards (**Fig. 1b** and **Supplementary Results 1**). To quantify the similarity between the reference labels and the clusters obtained by SC3, we used the adjusted Rand index (Online Methods), which ranges from 0 for a level of similarity expected by chance to 1 for identical clusterings. For the gold-standard datasets, we found that the quality of the outcome as measured by the adjusted Rand index was sensitive to the number of eigenvectors, d, retained after spectral transformation (**Supplementary Figs. 1** and **2**). For all six datasets, we found that the best clusterings were achieved when d was between 4% and 7% of the number of cells, N (**Fig. 1c**, **Supplementary Fig. 3a** and Online Methods). The robustness of the 4–7% range was supported by a simulation experiment in which the reads from the six gold-standard datasets were downsampled by a factor of ten (**Supplementary Fig. 3a**). We further tested the SC3 pipeline on six other published datasets[12–17], in which the cell labels can only be considered 'silver standard' since they were assigned using computational methods and the authors' knowledge of the underlying biology. Again, we found that SC3 performed well when using d in the 4–7% of N interval (**Supplementary Fig. 3b**). The final step, consensus clustering, improved both the accuracy and the stability of the solution. k-means-based methods typically provide different outcomes depending on the initial conditions. We found that this variability was significantly reduced with the consensus approach (**Fig. 1d**).

To benchmark SC3, we considered five other methods: tSNE[18] followed by k-means clustering (t-SNE + k-means; similar to the method used by Grün *et al.*[1]), pcaReduce[19], SNN-Cliq[20], SINCERA[21] and SEURAT[22]. SC3 performed better than the five tested methods across all benchmark datasets (Wilcoxon signed-rank test, $P < 0.01$), with only a few exceptions (**Fig. 2a**). In addition to considering accuracy, we also compared the stability of SC3 with other stochastic methods (pcaReduce and tSNE + k-means but not SEURAT) by running them 100 times (**Fig. 2a**,**b** and Online Methods). In contrast to the other methods that rely on different initializations, SC3 was highly stable.

Although SC3's consensus strategy provided high accuracy, it came at a moderate computational cost: the run time for 2,000 cells was ~20 min (**Supplementary Fig. 4a**). The main bottleneck was the k-means clustering. By reducing how many runs were considered, it was possible to cluster 5,000 cells in ~20 min with only a slight reduction in accuracy (**Supplementary Fig. 4b**). To apply SC3 to even larger datasets, we implemented a hybrid approach that combines

[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, Namur, Belgium. [4]ICTEAM, Université Catholique de Louvain, Louvain-la-Neuve, 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. Correspondence should be addressed to M.H. (mh26@sanger.ac.uk).

RECEIVED 28 NOVEMBER 2016; ACCEPTED 1 MARCH 2017; PUBLISHED ONLINE 27 MARCH 2017; DOI:10.1038/NMETH.4236

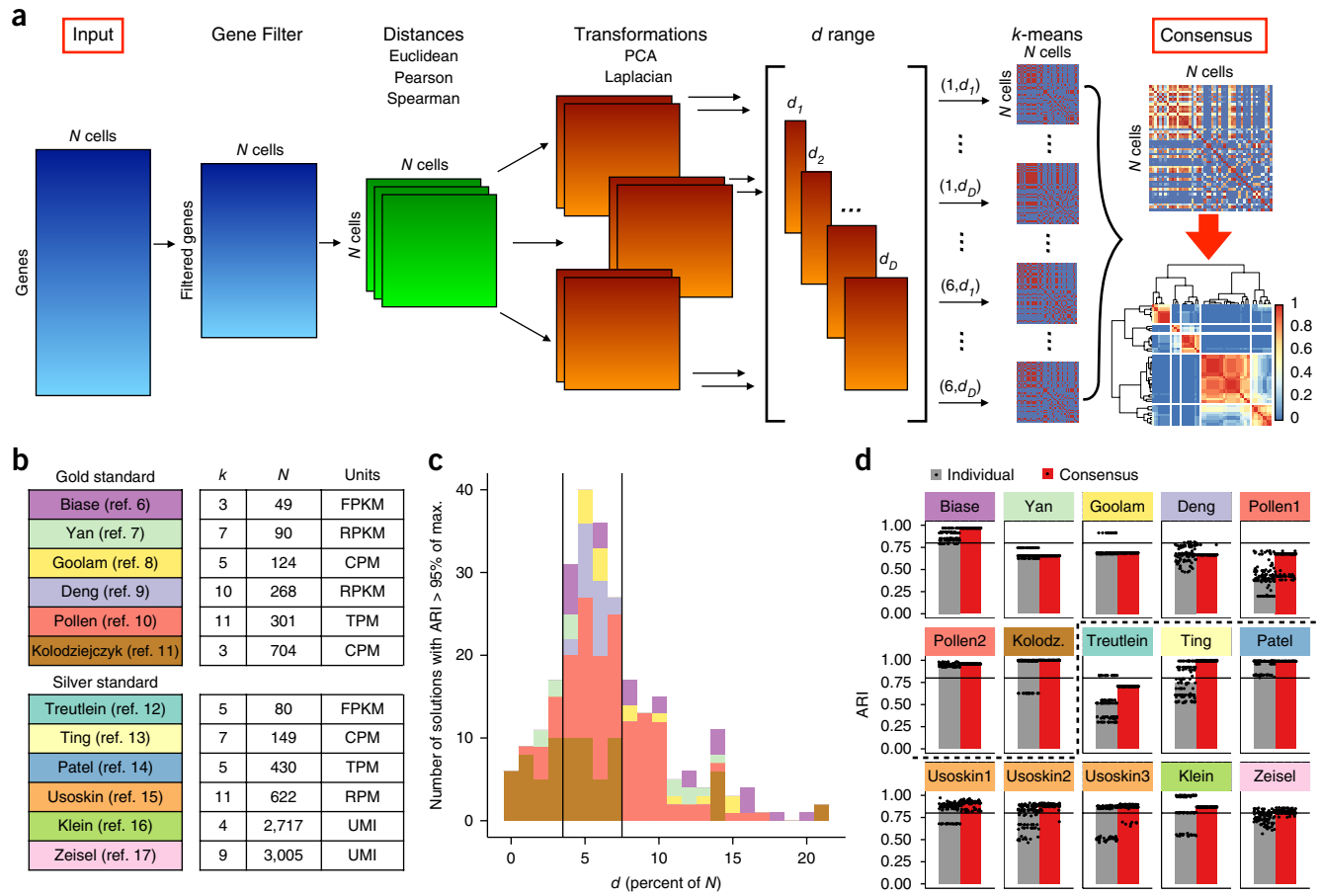
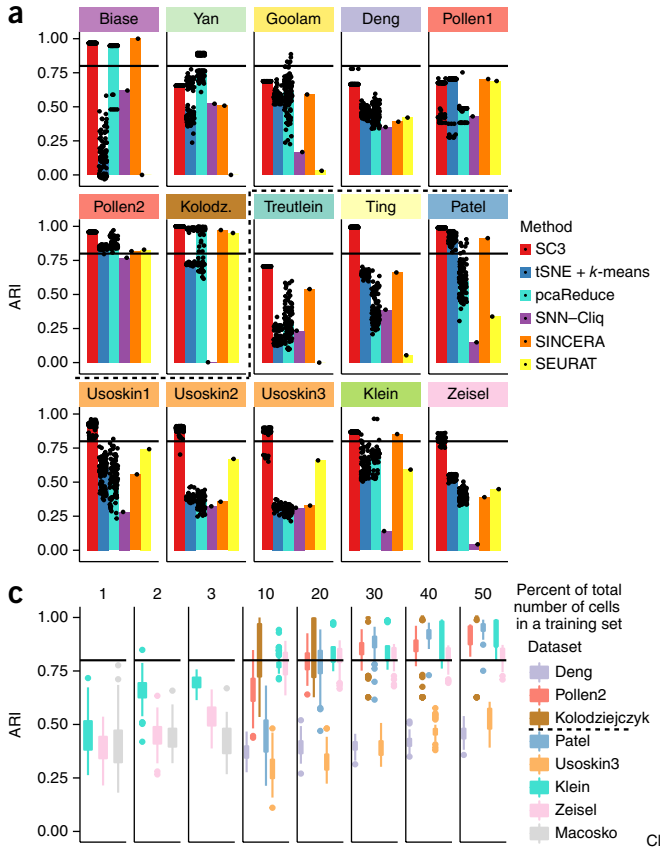


Figure 1 | The SC3 framework for consensus clustering of scRNA-seq data. (**a**) Overview of clustering with SC3. Results of the consensus step are shown for the Treutlein[12] data. (**b**) Published datasets used to set SC3 parameters. *N*, number of cells; *k*, number of clusters originally identified by the authors; RPKM, reads per kilobase of transcript per million mapped reads; RPM, reads per million mapped reads; FPKM, fragments per kilobase of transcript per million mapped reads; TPM, transcripts per million mapped reads; UMI, unique molecular identifiers; CPM, counts per million mapped reads. (**c**) Eigenvector (*d*) values that achieve adjusted Rand index (ARI) > 0.95 on gold-standard datasets. Black vertical lines indicate the interval *d* = 4–7% of *N*, showing high accuracy in the classification. (**d**) 100 realizations of the SC3 clustering of the datasets in **b**. Dots represent individual clustering runs and bars represent the median. Red and gray correspond to clustering with and without consensus step, respectively. The solid black line corresponds to ARI = 0.8. The dashed black line separates gold- and silver-standard datasets.

unsupervised and supervised methodologies. SC3 selects a subset of 5,000 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, which then assigns labels to the remaining cells (Online Methods). The hybrid approach worked well to predict cell labels (**Fig. 2c** and **Supplementary Fig. 4c**). We were able to analyze a Drop-seq dataset with 44,808 cells and 39 clusters[22], generating results in good agreement with the original study (**Supplementary Results**, **Supplementary Fig. 5** and **Supplementary Table 1**). The main drawback of the sampling strategy is that rare cell-types may not be identified, and when the number of cells greatly exceeds 5,000, there is a substantial risk that the sampled distribution will differ significantly from the full distribution (Online Methods). For identifying rare subpopulations (for example, cancer stem cells), methods specifically designed for this purpose, such as RaceID[1] or GiniClust[23], may be more appropriate.

To help users choose an optimal number of clusters, we have implemented a method based on random matrix theory (RMT)[24,25] (Online Methods). Overall, we found good agreement between RMT estimates and cluster numbers suggested by the original authors (**Fig. 2b**). SC3 is also interactive, allowing users to explore different choices of *k* in real time by assessing the consensus matrix (**Fig. 2d**), the silhouette index[26] (a measure of how tightly grouped the cells in the clusters are) or the expression matrix.

SC3 can help to interpret the results of clustering by identifying differentially expressed genes, marker genes and outlier cells (**Supplementary Fig. 6**, **Supplementary Table 2** and Online Methods). Marker genes are particularly useful since they can be used to uniquely identify a cluster. To illustrate these features, we analyzed the Deng[9] dataset tracing embryonic developmental stages. The most stable result for $k = 10$ generated clusters that largely agreed with known sampling timepoints (**Fig. 2d**). We identified ~3,000 marker genes (**Supplementary Table 3**), many of which had been previously reported as developmental stage-specific[27,28] and several of which were stage-specific but had not been previously reported (**Supplementary Table 3**). Notably, when using published reference labels[9], we identified nine cells with high outlier scores (**Supplementary Fig. 6c**), which turned



b

	Estimation of k				Solution stability		
Gold standard	Ref	SC3	SINCERA	SNN-Cliq	SC3	tSNE + k-means	pcReduce
Biase	3	3	5	6	1	0.12	0.82
Yan	7	6	6	11	1	0.53	0.89
Goolam	5	6	4	21	1	0.8	0.57
Deng	10	9	3	20	0.99	0.81	0.54
Pollen	11	11	9	14	0.79 1	0.76 0.95	0.89 0.88
Kolodziejczyk	3	10	18	2	1	0.87	0.85
Silver standard							
Treutlein	5	3	19	3	1	0.68	0.35
Ting	7	10	10	13	1	0.89	0.62
Patel	5	17	10	25	1	0.93	0.44
Usoskin	11	11	11	20	0.93 0.95 0.95	0.84 0.77 0.74	0.37 0.39 0.42
Klein	4	18	7	305	1	0.59	0.8
Zeisel	9	30	8	330	0.95	0.89	0.43

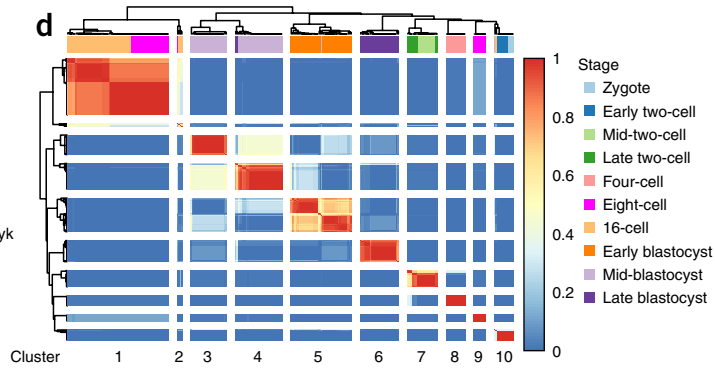


Figure 2 | Benchmarking of SC3 against existing methods. (**a**) SC3, tSNE + *k*-means and pcaReduce were applied 100 times to each dataset. SNN-Cliq and SINCERA are deterministic and were run only once. SEURAT was also run once but was optimized over values of the density parameter *G* (Online Methods). Dots represent ARI between inferred clusterings and reference labels; bars correspond to median ARI. The solid black line indicates ARI = 0.8. The dashed black line separates gold- and silver-standard datasets. (**b**) The number of clusters $\hat{k}$ predicted by SC3, SINCERA and SNN-Cliq for all datasets. Ref, reference clustering reported by the authors. Stability is defined as $N_c/100$, where N_c is the number of times the most frequent solution was found from 100 runs. (**c**) Performance of the SC3 hybrid approach. Dots represent outliers higher (or lower) than the highest (or lowest) value within 1.5× the interquartile range (IQR). The solid black line indicates ARI = 0.8. The dashed black line in the legend separates gold- and silver-standard datasets. (**d**) Consensus matrix as generated by SC3 for the Deng[9] dataset, indicating how often each pair of cells was assigned to the same cluster by the different parameter combinations (1, always; 0, never). Colors at the top represent reference labels corresponding to stages of development.

out to have been prepared using the Smart-seq2 protocol instead of the Smart-seq protocol[9,20].

Finally, we investigated the ability of SC3 to identify subclones based on transcriptomes. Myeloproliferative neoplasms, a group of diseases characterized by the overproduction of terminally differentiated myeloid cells, reflect an early stage of tumorigenesis in which multiple subclones are known to coexist in the same patient[29]. Myeloproliferative neoplasms are thought to originate from hematopoietic stem cells (HSCs). To gain further insight into the transcriptional landscape of patient-derived HSCs, we obtained scRNA-seq data from two patients (**Supplementary Figs. 7** and **8**, **Supplementary Table 4** and Online Methods). For patient 1 (N = 51), the silhouette index and the RMT method suggested that three clusters were optimal, and SC3 produced three clusters of similar size (**Supplementary Fig. 9**). For patient 2 (N = 89), SC3 generated a single cluster (**Supplementary Fig. 10**), in agreement with the RMT algorithm.

Since *TET2* and *JAK2V617F*[30,31] are the only loci with known driver mutations in these two patients, we hypothesized that clusters corresponded to clones with different combinations of mutations. The genotype composition of each HSC clone was determined by growing individual HSCs into granulocyte and macrophage colonies, followed by Sanger sequencing of the *TET2* and *JAK2V617F* loci (**Supplementary Fig. 7b**,**c**). In agreement with SC3 clustering, patient 1 was found to harbor three different subclones: (i) cells with mutations in both loci, (ii) cells with a *TET2* mutation and (iii) wild-type cells. Strikingly, the SC3 clusters contained 22%, 29% and 49% of the cells, respectively, in excellent agreement with the 20%, 30% and 50% found in the patient (**Supplementary Fig. 7c**). The HSC compartment of patient 2 was 100% mutant for *TET2* and *JAK2V617F* (**Supplementary Fig. 7c**), again consistent with SC3 clustering (**Supplementary Fig. 10**). We then analyzed the pooled cells from patients 1 and 2. SC3 clustering again suggested k = 3 (**Fig. 3** and **Supplementary Fig. 11**), in agreement with the RMT algorithm. Notably, all of the putative double-mutant cells from patient 1 were grouped with the double-mutant cells from patient 2. SC3 reported 33 marker genes for the putative *TET2* mutant and 202 marker genes for the putative double mutant clone (**Fig. 3** and **Supplementary Table 5**). Together with additional evidence (**Supplementary Results** and **Supplementary Fig. 12**), we conclude that SC3 is able to identify subclones across patients.

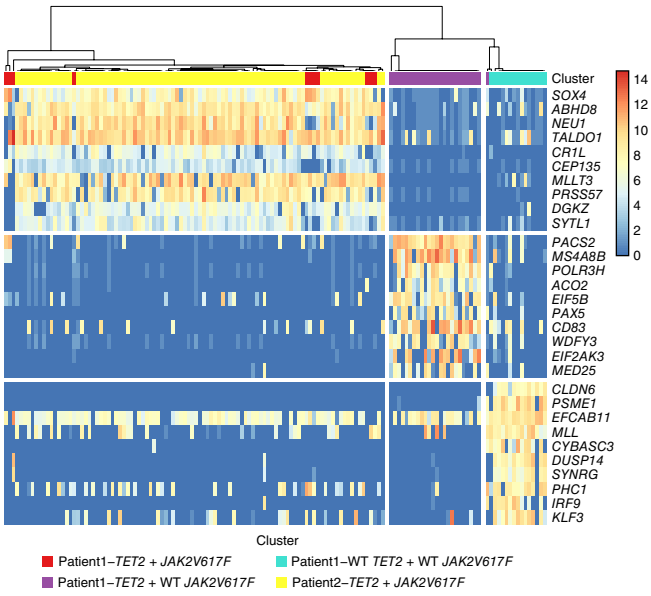


Figure 3 | SC3 defines subclones from two patients with myeloproliferative neoplasm. Marker-gene expression matrix (after gene filter and log-transformation; see Online Methods) of a combined dataset of patient 1 and patient 2. Clusters (separated by white vertical lines) correspond to $k = 3$. Only the top 10 marker genes are shown for each cluster. WT, wild type (i.e., no mutation).

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank B. Vangelov, J.-C. Delvenne and R. Lambiotte for fruitful discussions and for their help with computational methods. We also thank D. Flores Santa Cruz, D. Dimitropolou and J. Grinfeld for technical assistance with experiments. We thank I. Vasquez-Garcia, D. Harmin, M. Kosicki, D. Ramsköld and M. Huch for comments on the manuscript. V.Y.K., T.A., A.Y. and M.H. are supported by Wellcome Trust Grants. K.N.N. is supported by the Wellcome Trust Strategic Award 'Single cell genomics of mouse gastrulation'. M.T.S. acknowledges support from FRS-FNRS; the Belgian Network DYSCO (Dynamical Systems, Control and Optimisation), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian State Science Policy Office; and the ARC (Action de Recherche Concerte) on Mining and Optimization of Big Data Models, funded by the Wallonia-Brussels Federation. M.B. acknowledges support from EPSRC (grant EP/N014529/1). T.C. was funded through a core funded fellowship by the Sanger Institute and a Chancellor's fellowship from the University of Edinburgh. K.K. and A.R.G. are 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. W.R. was supported by BBSRC (grant ref. BB/K010867/1), the Wellcome Trust (grant ref. 095645/Z/11/Z), EU BLUEPRINT and EpiGeneSys.

AUTHOR CONTRIBUTIONS

M.H. conceived the study; V.Y.K., M.H., M.T.S., M.B., T.A. and A.Y. contributed to the computational framework; K.K. and T.C. performed the experiments for the patient data; K.N.N. helped with the analysis of embryonic mouse data; M.B., W.R., A.R.G. and M.H. supervised the research; and V.Y.K. and M.H. led the writing of the manuscript with input from the other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Grün, D. *et al. Nature* **525**, 251–255 (2015).
2. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
3. Mahata, B. *et al. Cell Rep.* **7**, 1130–1142 (2014).
4. Gentleman, R.C. *et al. Genome Biol.* **5**, R80 (2004).
5. McCarthy, D.J., Campbell, K.R., Lun, A.T.L. & Wills, Q.F. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btw777 (2017).
6. Biase, F.H., Cao, X. & Zhong, S. *Genome Res.* **24**, 1787–1796 (2014).
7. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
8. Goolam, M. *et al. Cell* **165**, 61–74 (2016).
9. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. *Science* **343**, 193–196 (2014).
10. Pollen, A.A. *et al. Nat. Biotechnol.* **32**, 1053–1058 (2014).
11. Kolodziejczyk, A.A. *et al. Cell Stem Cell* **17**, 471–485 (2015).
12. Treutlein, B. *et al. Nature* **509**, 371–375 (2014).
13. Ting, D.T. *et al. Cell Rep.* **8**, 1905–1918 (2014).
14. Patel, A.P. *et al. Science* **344**, 1396–1401 (2014).
15. Usoskin, D. *et al. Nat. Neurosci.* **18**, 145–153 (2015).
16. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
17. Zeisel, A. *et al. Science* **347**, 1138–1142 (2015).
18. van der Maaten, L. & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
19. Zurauskiene, J. & Yau, C. *BMC Bioinformatics* http://doi.org/10.1186/s12859-016-0984-y (2016).
20. Xu, C. & Su, Z. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btv088 (2015).
21. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. & Xu, Y. *PLoS Comput. Biol.* **11**, e1004575 (2015).
22. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
23. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. *Genome Biol.* **17**, 144 (2016).
24. Patterson, N., Price, A.L. & Reich, D. *PLoS Genet.* **2**, e190 (2006).
25. Tracy, C.A. & Widom, H. *Commun. Math. Phys.* **159**, 151–174 (1994).
26. Rousseeuw, P.J. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
27. Guo, G. *et al. Dev. Cell* **18**, 675–685 (2010).
28. Boroviak, T. *et al. Dev. Cell* **35**, 366–382 (2015).
29. Chen, E., Staudt, L.M. & Green, A.R. *Immunity* **36**, 529–541 (2012).
30. Ortmann, C.A. *et al. N. Engl. J. Med.* **372**, 601–612 (2015).
31. Nangalia, J. *et al. N. Engl. J. Med.* **369**, 2391–2405 (2013).

ONLINE METHODS

SC3 clustering. SC3 takes as input an expression matrix, $\boldsymbol{M}$, in which columns correspond to cells and rows correspond to genes/transcripts. Each element of $\boldsymbol{M}$ corresponds to the expression of a gene/transcript in a given cell. By default, SC3 does not carry out any form of normalization or correction for batch effects. SC3 is based on five 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 the gold-standard datasets (see main text).

1. *Gene filter.* The gene filter removes genes/transcripts that are either expressed (expression value > 2) in less than X% of cells (rare genes/transcripts) or expressed (expression value > 0) in at least (100 – X)% of cells (ubiquitous genes/transcripts). By default, X is set at 6. The motivation for the gene filter is that ubiquitous and rare genes are most often not informative for clustering. We also explored all three parameters defined in the gene filter (expression thresholds of rare and ubiquitous genes/transcripts and the percentage X) and found that in general the gene filter did not affect the accuracy of clustering (**Supplementary Fig. 3c**). However, the gene filter significantly reduced the dimensionality of the data, thereby speeding up the method.

For further analysis, the filtered expression matrix $\boldsymbol{M}$ is log-transformed after adding a pseudocount of 1: $\boldsymbol{M}' = \log_2(\boldsymbol{M} + 1)$.

2. *Distance calculations.* Distances between the cells (i.e., columns) in $\boldsymbol{M}'$ are calculated using the Euclidean, Pearson and Spearman metrics to construct distance matrices.

We investigated the impact of dropouts on distance calculations by considering a modified distance metric that ignores dropouts. This was done by excluding genes that were not expressed in at least one cell from the distance calculation. We found that this did not improve the performance (**Supplementary Fig. 3d**).

3. *Transformations.* All distance matrices are then transformed using either principal component analysis (PCA) 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 $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{A}$ is a similarity matrix ($\boldsymbol{A} = e^{-\boldsymbol{A}'/\max(\boldsymbol{A}')}$), where $\boldsymbol{A}'$ is a distance matrix) and $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{A}$, a diagonal matrix that contains the row-sums of $\boldsymbol{A}$ on the diagonal ($\boldsymbol{D}_{ii} = \Sigma_j \boldsymbol{A}_{ij}$). The columns of the resulting matrices are then sorted in ascending order by their corresponding eigenvalues.

4. *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[32]. By default, the maximum number of iterations is set to 10^9 and the number of starts is set to 1,000.

5. *Consensus clustering.* SC3 computes a consensus matrix using the cluster-based similarity partitioning algorithm (CSPA)[33]. For each individual 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 of individual clusterings. To reduce computational time, if the length of the d range (D in **Fig. 1a**) is more than 15, a random subset of 15 values selected uniformly from the d range is used.

The resulting consensus matrix is clustered using hierarchical clustering with complete agglomeration, and the clusters are inferred at the k level of hierarchy, where k is defined by the user (**Fig. 1a**). In principle, the k used for the hierarchical clustering need not be the same as the k used in step 5. However, for simplicity in SC3 the two parameters are constrained to have the same value. **Figure 1d** shows how the quality and the stability of clustering improves after consensus clustering.

Adjusted Rand index. If cell-labels are available (for example, from a published dataset) the adjusted Rand index (ARI)[34] can be used to calculate similarity between the SC3 clustering and the published clustering. ARI is defined as follows. Given a set of n elements and two clusterings of these elements, the overlap between the two clusterings can be summarized in a contingency table, in which each entry denotes the number of objects in common between the two clusterings. The ARI can then be calculated as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where n_{ij} are values from the contingency table, a_i is the sum of the ith row of the contingency table, b_j is the sum of the jth column of the contingency table and $\binom{\ }{\ }$ denotes a binomial coefficient.

Since the reference labels are known for all published datasets, ARI is used for all comparisons throughout the paper.

Downsampling of the gold-standard datasets. For each gene i and each cell j, the downsampled expression value was generated by drawing from a binomial distribution with parameters $P = 0.1$ and $n = \text{round}(\boldsymbol{M}_{ij})$.

Additional validation of SC3 pipeline. Additionally, we investigated the impact of dropouts by considering a modified distance metric that ignores dropouts, but we found that this did not improve the performance (**Supplementary Fig. 3d** and Online Methods).

Identification of a suitable number of groups $\hat{k}$. Matrix $\boldsymbol{Z}$ is obtained from $\boldsymbol{M}'$ by subtracting the mean and dividing by the s.d. for each column (z-score). Next, the eigenvalues of $\boldsymbol{X} = \boldsymbol{Z}^T\boldsymbol{Z}$ are calculated. The number of clusters, $\hat{k}$, is determined by the number of eigenvalues that are significantly different with $P < 0.001$ from the Tracy–Widom distribution[24,25] with mean $\left(\sqrt{n-1} + \sqrt{p}\right)^2$ and s.d.

$$\left(\sqrt{n-1} + \sqrt{p}\right) \cdot \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{p}}\right)^{\frac{1}{3}},$$

where n is the number of genes/transcripts and p is the number of cells.

Benchmarking. For each dataset we used the expression units provided by the authors of that set (**Fig. 1b**). The gene filter was applied to all the datasets. For tSNE + k-means, SNN-Cliq and pcaReduce, the same log-transformation as in SC3 ($\boldsymbol{M}' = \log_2(\boldsymbol{M} + 1)$)

was applied. For SINCERA, we used the original *z*-score normalization[21] instead of the log-transformation. For tSNE, the Rtsne R package was used with the default parameters. For SEURAT, we used the original Seurat R package (version 1.3): we performed tSNE embedding with the default parameters once (following the authors' tutorial at http://satijalab.org/seurat/seurat_clustering_tutorial_part1.html) and then clustered the data using the DBSCAN algorithm multiple times, during which we varied the density parameter G in the range 10^{-3}–10^{3} to find a maximal ARI (this ARI is presented in **Fig. 2a**). SEURAT was not able to find more than one cluster for the smallest datasets (Biase, Yan, Goolam, Treutlein and Ting) leading to very small ARI scores. For all methods we supplied the k used by the original authors.

Cluster stability. We calculated stability of clustering solutions by running each method 100 times and finding the most frequent solution and the number of times (N_c) it appeared. The stability measure shown in **Figure 2b** is then calculated as $N_c/100$.

Support vector machines (SVM). When using SVM, a specific fraction of the cells is selected at random with uniform probability. Next, an SVM[35] model with a linear kernel 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.

Identification of rare cell-types. To specifically evaluate the sensitivity of SC3 for identifying rare cell-types, we carried out a synthetic experiment in which cells from one cell-type were removed iteratively from the Kolodziejczyk and Pollen datasets. For the Pollen dataset, all but 1–7 of the cells in one of the 11 clusters were removed. The limit of 7 cells corresponds to the size of the smallest cluster in the original data. Subsequently, SC3 was run using $k = 11$, and we asked whether or not the cells of the rare cell-type were located in a separate cluster. This was repeated 100 times for each cell-type; **Supplementary Figure 4d** reports the percentage of runs in which the rare cells were found together in a cluster with no other cells. Note that the ARI is a poor indicator of the ability to identify rare cells, since this measure is relatively insensitive to the behavior of a small fraction of the cells. For the Kolodziejczyk dataset, we used a similar strategy, but we allowed for 1–101 cells in the rare group. For the Pollen dataset, SC3 can detect clusters containing ~1% of the cells, whereas for the Kolodziejczyk dataset ~10% of the cells are required (**Supplementary Fig. 4d**). We hypothesize that the ability to identify rare cells reflects the origins of the two datasets; the Pollen data is more diverse, as it represents 11 different cell lines, while the Kolodziejczyk data comes from one cell-type grown in three different conditions.

For the hybrid SC3 approach, with 30% of cells used to train the SVM, we were able to calculate the probability of including the rare cell-types in the training set analytically by multiplying the data from **Supplementary Figure 4d** by the probability of all rare cells to be included in the drawn sample (30% of all cells). This probability was calculated using the hypergeometric distribution R function: phyper($n.rare.cells - 1$, $n.rare.cells$, $n.other.cells$, $0.3 \times (n.other.cells + n.rare.cells)$, lower.tail = F), where $n.rare.cells$ is the number of rare cells and $n.other.cells$ is the number of other cells in the dataset (**Supplementary Fig. 4e**).

Biological insights. SC3 can identify differentially expressed genes as genes that vary between two or more clusters. Accordingly, marker genes are identified as genes that are highly expressed in only one of the clusters and are able to distinguish one cluster from all the remaining ones (**Supplementary Fig. 6a**). Cell outliers are identified through the calculation of a score for each cell using the minimum covariance determinant[36]. 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.

Identification of differential expression. Differential expression is calculated using the nonparametric Kruskal–Wallis test, an extension of the Mann–Whitney test for tests of more than two groups. The Kruskal–Wallis test has the advantage of being nonparametric, but as a consequence, it is not well suited for situations in which many genes have the same expression value. 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 $P < 0.01$, corrected for multiple testing (using the default 'holm' method of the p.adjust() R function), and plots gene expression profiles of the 50 most significant differentially expressed genes. Note that calculating differential expression after clustering can introduce a bias in the distribution of P-values, and thus we advise using 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 using the Wilcoxon signed-rank test, comparing gene ranks in the cluster with the highest mean expression with all others (P-values are adjusted using the default 'holm' method of the p.adjust() R function). Genes with areas under the ROC curve (AUROC) > 0.85 and with $P < 0.01$ are defined as marker genes. The AUROC threshold corresponds to the 99th percentile of the AUROC distributions obtained from 100 random permutations of cluster labels for all datasets (**Supplementary Table 2** and **Supplementary Fig. 6b**). SC3 provides a visualization of the gene expression profiles for the top 10 marker genes of each obtained cluster.

Cell outlier detection. Outlier cells are detected by first taking an expression matrix of each individual cluster (all cells with the same labels) and reducing its dimensionality using the robust method for PCA (ROBPCA)[37]. This method outputs a matrix with N rows (number of cells in the cluster) and P columns (retained number of principal components after running ROBPCA). SC3 then uses $P = \min(P, 3)$ first principal components for further analysis. If ROBPCA fails to perform or $P = 0$, SC3 shows a warning message. We found (results not shown) that this usually happened when the distribution of gene expression in cells was too skewed toward 0. Second, robust distances (Mahalanobis) between the cells in each cluster are calculated from the reduced expression matrix using the minimum covariance determinant (MCD)[36]. We then used a threshold based on the Q% quantile of the chi-squared distribution (with p degrees of freedom) to define outliers. By default $Q = 99.99$, but it can be manually adjusted by a user. Finally, we define an outlier score as the difference between the square root of the robust distance and the square root of the Q% quantile of the chi-squared distribution (with p degrees of freedom). The outlier score is plotted as a bar plot (**Supplementary Fig. 6c**).

doi:10.1038/nmeth.4236

Gene and pathway enrichment analysis. We used the g:Profiler web tool[38] to perform gene and pathway enrichment analysis in all obtained sets of genes.

Analysis of the Macosko dataset. To analyze the Drop-seq dataset we followed the procedure used by Macosko *et al.*[22] and selected the 11,040 cells in which more than 900 genes were expressed. Moreover, due to the low read depth, the gene filter was removed. We then sampled 5,000 cells and clustered using SC3, including the SVM step, 100 times. All 100 solutions were consistent with each other, resulting in an average ARI of 0.58, and they were sufficiently accurate compared to the reference authors' clustering, yielding an average ARI of 0.54 (**Supplementary Fig. 5a**). Since each of the 100 solutions were different, we added an additional consensus clustering step using the 'best of *k*' consensus algorithm[39]. This approach provided a single solution based on the 100 different solutions and was as accurate as the individual solutions, with an ARI of 0.52 (the actual labels are presented in **Supplementary Table 1**). The SC3 consensus solution splits the large original cluster (cluster 24 with 29,400 cells) hierarchically into two clusters of smaller sizes (18,105 + 10,558 = 28,663 cells; clusters 4 and 8 in **Supplementary Fig. 5b**). Additional gene and pathway enrichment analysis for the differentially expressed genes between the two clusters is presented in **Supplementary Table 1**. If more than 75% of the cells from the reference cluster were shared with the SC3 cluster, we defined these two clusters as matched. In total, 31 reference clusters were matched to the SC3 clusters.

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 hematopoietic stem and progenitor cells. Cell populations were derived from peripheral blood enriched for hematopoietic 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 (**Supplementary Fig. 7a**,**b**) 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 chosen and DNA-isolated for Sanger sequencing for *JAK2V617F* and *TET2* mutations as previously described by Ortmann *et al.*[30].

Single cell RNA-sequencing. Single HSCs were sorted into 96-well plates and cDNA generated as described previously[41]. The Nextera XT library-making kit was used for library generation as described by Picelli *et al.*[41].

Processing of scRNA-seq data from HSCs. We sequenced 96 single cell samples per patient with tow sequencing lanes per sample, yielding a variable number of reads (mean = 2,180,357; s.d. = 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], using the parameters HEADCROP:19 ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 TRAILING:28 CROP:90 MINLEN:60 to trim the reads to 90 bases, before mapping with TopHat[44] to the Ensembl reference genome version GRCh38.77, augmented with the spike-in controls downloaded from the ERCC consortium. 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 comprised two steps: (i) filtering cells based on the number of expressed genes and (ii). filtering cells based on the ratio of the total number of ERCC spike-in reads to the total number of reads in protein-encoding genes. Filtering thresholds were manually chosen by visual exploration of the quality control features (**Supplementary Fig. 8**). After filtering, 51 and 89 cells were retained from patient 1 and patient 2, respectively. The expression values in each dataset were then normalized by first using a size-factor normalization (from DESeq2 package[45]), to account for sequencing depth variability, and then by using a normalization based on ERCC spike-ins, performed using the RUVSeq package[46] (RUVg() function with parameter $k = 1$), to account for technical variability. For combined patient data, normalization steps were performed after pooling the cells. The resulting filtered and normalized datasets were clustered by SC3. Potential biases in cell filtering on the proportions of cells in the clusters of patient 1 are considered in **Supplementary Results 2**. The cluster of lower cell quality was separated from the other biologically meaningful clusters of patient 1 and did not change the total proportion of the biologically meaningful clusters. **Supplementary Results 3** shows that SC3 clustering results of patient 1 did not depend on the normalization procedure.

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, in agreement with the RMT algorithm, the best clustering was achieved for $k = 3$ (**Supplementary Fig. 9**). Data from patient 2 was homogeneous, and SC3 was unable to identify more than one meaningful cluster (**Supplementary Fig. 10**), again in agreement with the RMT algorithm. For the combined dataset for patient 1 + patient 2, the best values of the silhouette index were obtained when k was 2 or 3 (**Supplementary Fig. 11**). 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 2 and 3 of patient 1 were also resolved (**Fig. 3**). The RMT algorithm also provided $k = 3$ for the merged patient 1 + patient 2 dataset.

Comparison of clustering of patient 1 scRNA-seq data. Results of the clustering of patient 1 data by other methods and their comparisons are SC3 is presented in **Supplementary Results 4** and **5**.

Identification of differentially expressed genes from microarray data. The microarray data of patient 1 was obtained from Array Express, under accession number E-MTAB-3086 (ref. 30). One replicate (2B) was identified as an outlier and removed. The 'limma' R package[47] was used to identify 932 differentially expressed genes between WT and *TET2*/*JAK2V617F* double-mutants 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 *P*-value threshold was set at 0.1.

Data availability. All datasets (in **Fig. 1b** and the Macosko dataset) were acquired from the accession numbers provided in the original publications. According to their respective authors, the

Pollen dataset contains two distinct hierarchies and the cells can be grouped either into 4 or 11 clusters, and the Usoskin dataset contains three hierarchies and the cells can be grouped either into 4, 8 or 11 clusters. scRNA-seq data for patient 1 and 2 is available from GEO under accession code GSE79102. Source data files for **Figures 1–3**, and **Supplementary Figure 1–7** and **12** are available online.

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

Scripts for figure generation are available at http://github.com/hemberg-lab/SC3-paper-figures. At the time of writing the manuscript, the following old versions of some of the tools were used (these tools have been updated/upgraded since then):

- SC3 (1.1.2 ≤ Version < 1.1.5). These versions of SC3 can be installed:
  - from source/binary files from Bioconductor http://bioconductor.org/packages/3.3/bioc/html/SC3.html
  - from GitHub using commands:
    - install.packages("devtools")
    - devtools::install_github("hemberg-lab/SC3", ref = "8a86b60463")
  - SC3 v.1.1.2 source and DESCRIPTION files can be found in **Supplementary Software 1**.
  - In the newer versions, the main SC3 pipeline has not been changed.
- SEURAT (version 1.3), which can be installed from GitHub:
  - install.packages("devtools")
  - devtools::install_github("satijalab/Seurat", ref = "da6cd08")
  - In the newer versions of SEURAT, a different algorithm is used for clustering.
- Source files used for generating **Supplementary Results 2–5** can be found in **Supplementary Software 2**.

32. Hartigan, J.A. & Wong, M.A. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108 (1979).
33. Strehl, A. & Ghosh, J. *J. Mach. Learn. Res.* **3**, 583–617 (2003).
34. Hubert, L. & Arabie, P. *J. Classif.* **2**, 193–218 (1985).
35. Ben-Hur, A., Horn, D., Siegelmann, H.T. & Vapnik, V. *J. Mach. Learn. Res.* **2**, 125–137 (2001).
36. Hubert, M. & Debruyne, M. *WIREs Comp Stat* **2**, 36–43 (2010).
37. Hubert, M., Rousseeuw, P.J. & Branden, K.V. *Technometrics* **47**, 64–79 (2005).
38. Reimand, J. *et al. Nucleic Acids Res.* **44**, W83–W89 (2016).
39. Goder, A. & Filkov, V. Consensus clustering algorithms: comparison and refinement. in *Proceedings of the Meeting on Algorithm Engineering & Experiments* 109–117 (Society for Industrial and Applied Mathematics, 2008).
40. Petzer, A.L., Zandstra, P.W., Piret, J.M. & Eaves, C.J. *J. Exp. Med.* **183**, 2551–2558 (1996).
41. Picelli, S. *et al. 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. *Bioinformatics* **30**, 2114–2120 (2014).
44. Trapnell, C., Pachter, L. & Salzberg, S.L. *Bioinformatics* **25**, 1105–1111 (2009).
45. Love, M.I., Huber, W. & Anders, S. *Genome Biol.* **15**, 550 (2014).
46. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. *Nat. Biotechnol.* **32**, 896–902 (2014).
47. Ritchie, M.E. *et al. Nucleic Acids Res.* **43**, e47 (2015).

doi:10.1038/nmeth.4236