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]

[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

Corresponding author: Martin Hemberg (mh26@sanger.ac.uk)

Abstract

Single-cell RNA-seq (scRNA-seq) enables a quantitative cell-type characterisation 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. We demonstrate that SC3 is capable of identifying subclones based on the transcriptomes from neoplastic cells collected from patients.

Main text

One of the key applications of scRNA-seq is determining cell types based on transcriptome profiles alone through unsupervised clustering[1–3]. A full characterisation of the transcriptional landscape of individual cells holds an enormous potential, both for basic biology and clinical applications. SC3 is an interactive and user-friendly R-package for clustering and its integration with Bioconductor[4] and scater[5] makes it easy to incorporate into existing bioinformatic workflows.

The SC3 pipeline is presented in Fig. 1a, Methods. Each of the steps requires the specification of a number of parameters. Choosing optimal parameter values is difficult and time-consuming. 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 clusterings. SC3 then combines *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.

To constrain the parameter values of the SC3 pipeline, we first considered six publicly available scRNA-Seq datasets[1] (Fig. 1b). The datasets were selected on the basis that one can be highly confident in the cell-labels as they represent cells from different stages, conditions or lines, and thus we consider them as 'gold standard'. To quantify the similarity between the reference labels and the clusters obtained by SC3, 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. For the gold standard datasets, we found that the quality of the outcome as measured by the ARI was sensitive to the number of eigenvectors, *d*, retained after the spectral transformation (Fig. S1, S2). For all six datasets we find that the best clusterings were achieved when *d* is between 4-7% of the number of cells, *N* (Fig. 1c, S3a, Methods). The robustness of the 4-7% region was supported by a simulation experiment where the reads from the six gold standard datasets were downsampled by a factor of ten (Methods and Fig. S3a). We further tested the SC3 pipeline on six other published datasets, where 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 find that SC3 performs well when using *d* in the 4-7% of *N* interval (Fig. S3b). The final step, consensus clustering, improves both the accuracy and the stability of the solution. k-means based methods will typically provide different outcomes

[1] Full references to the datasets can be found in the Supplementary Results

depending on the initial conditions. We find that this variability is significantly reduced with the consensus approach (Fig. 1d).

To benchmark SC3, we considered five other methods: tSNE[6] followed by *k*-means clustering (a method similar to the one used by Grün et al[1]), pcaReduce[7], SNN-Cliq[8], SINCERA[9] and SEURAT[10]. As Fig. 2a shows, SC3 performs better than the five tested methods across all datasets (Wilcoxon signed-rank test p-value < 0.01), with only a few exceptions. In addition to considering accuracy, we also compared the stability of SC3 with other stochastic methods (pcaReduce and tSNE+kmeans, but not SEURAT) by running them 100 times (Fig. 2b, Methods, black dots in Fig. 2a). In contrast to the other methods that rely on different initializations, SC3 is highly stable.

Although SC3's consensus strategy provides a high accuracy, it comes at a moderate computational cost: the run time for N = 2,000 is ~20 mins (Fig. S4a). The main bottleneck is the k-means clustering and by reducing how many different runs are considered it is possible to cluster 5,000 cells in ~20 mins with only a slight reduction in accuracy (Fig. S4b). To apply SC3 to even larger datasets, we have implemented a hybrid approach that combines 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 (SVM, Methods), which is employed to assign labels to the remaining cells. Our result shows that the use of an SVM to predict cell labels works well (Fig. 2c, S4c and Methods). Using the hybrid approach, we were able to analyse a large Drop-Seq dataset with N = 44,808 cells and k = 39 clusters[10] and our results were again in good agreement with the original authors' (Supplementary Results, Methods, Fig. S5, Table S1). The main drawback of the sampling strategy is that one may fail to identify rare cell-types, and when N>>5,000 there is a substantial risk that the sampled distribution will differ significantly from the full distribution (Methods). If the user is trying to identify a rare subpopulation (e.g. cancer stem cells), then methods specifically designed to identify rare cell-types such as RaceID[1] or GiniClust[11] may be more appropriate.

To help the user identify a good choice of *k*, we have implemented a method based on Random Matrix Theory (RMT)[12,13] for determining the number of clusters (Methods). Overall, we find good agreement between these estimates, $\hat{k}$, and the numbers suggested by the original authors (Fig. 2b). Additionally, in the interactive SC3 session the user can explore different choices of *k* in real time, by either assessing the consensus matrix (Fig. 2d), the silhouette index[14] (a measure of how tightly grouped the cells in the clusters are), or the expression matrix.

To help the user interpret the clustering result SC3 can identify differentially expressed genes, marker genes, and outlier cells (Fig. S6, Methods, Table S2). Marker genes are particularly useful since they can be used to uniquely identify a cluster. To illustrate these features, we analysed the Deng[15] dataset tracing embryonic developmental stages. The most stable result for k = 10 is shown in Fig. 2d, and our clusters largely agree with the known sampling timepoints. In total, we identified ~3000 marker genes (Table S3), many of which had been previously reported as specific to the different developmental stages[16,17]. Furthermore, the analysis reveals several genes specific to each developmental stage which had previously not been reported (Table S3). Importantly, when using the reference labels reported by the authors[15], nine cells have high outlier scores (purple cells in Fig. S6c). As it turns out, these were prepared using the Smart-Seq2 protocol instead of the Smart-Seq protocol[8,15].

Finally, we investigated the ability of SC3 to identify subclones based on transcriptomes. 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[18]. From exome sequencing data, we previously identified TET2 and JAK2V61F as the only driver mutations in a large patient cohort[19]. Haematopoietic stem cells (HSCs) are thought to be the cell of origin in myeloproliferative neoplasms. To gain further insight into the transcriptional landscape of patient derived HSCs, we obtained scRNA-seq data from the two patients (Figs. S7a-b, S8, Methods, Table S4). For patient 1 (N = 51), both the silhouette index of SC3 and our RMT method suggested that k = 3, provides the best clustering, revealing three clusters of similar size (Fig. S9). For patient 2 (N = 89) SC3 indicated k=1 (Fig. S10), in agreement with the RMT algorithm, suggesting that one single cluster might best reflect the underlying transcriptional changes.

Since known driver mutations in these patients are the *TET2* and *JAK2V617F* loci[20] we hypothesized that the different clusters correspond to different combinations of mutations within different clones. The genotype composition for each HSC clone was determined by growing individual haematopoietic stem cells into granulocyte/macrophage colonies, followed by Sanger sequencing of the TET2 and JAK2V617F loci (Fig. S7b-c). 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. S7c). 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% (Fig. S7c). 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 patient 2 was 100% mutant for TET2 and JAK2V617F (Fig. S7c), which again was consistent with clustering of *k*=1 suggested by SC3 (Fig. S10). We then analysed the pooled cells from patient 1 and 2. SC3 clustering again suggested *k*=3 (Figs. 3, S11), in agreement with the RMT algorithm. Most importantly, 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, Table S5). Together with additional evidence (Supplementary Results), we conclude that SC3 is able to identify subclones across patients.

Data Availability

All datasets (in Fig. 1b and Macosko dataset) were acquired from the accessions provided in the original publications. According to the 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 accession GSE79102.

Software availability

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

Scripts for figures 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):

1. 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) or directly from Github using commands:

   ```
   install.packages("devtools")
   devtools::install_github("hemberg-lab/SC3", ref = "8a86b60463")
   ```

   In the newer versions the main SC3 pipeline has not been changed.

2. SEURAT (version 1.3) - 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.

Acknowledgements

We would like to thank B. Vangelov, J.-C. Delvenne and R. Lambiotte for fruitful discussions and their help with computational methods. We would also like to 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 helpful comments on the manuscript.

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; V.Y.K. and M.H. led the writing of the manuscript with input from the other authors.

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. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
5. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btw777
6. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
7. Zurauskiene, J. & Yau, C. pcaReduce: Hierarchical Clustering of Single Cell Transcriptional Profiles. *bioRxiv* 026385 (2015). doi:10.1101/026385
8. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv088
9. 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).
10. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of

Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).

11. Jiang, L., Chen, H., Pinello, L. & Yuan, G.-C. GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. *Genome Biol.* **17,** 144 (2016).

12. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

13. Tracy, C. A. & Widom, H. Level-spacing distributions and the Airy kernel. *Commun. Math. Phys.* **159,** 151–174 (1994).

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

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

16. 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).

17. Boroviak, T. *et al.* Lineage-Specific Profiling Delineates the Emergence and Progression of Naive Pluripotency in Mammalian Embryogenesis. *Dev. Cell* **35,** 366–382 (2015).

18. Chen, E., Staudt, L. M. & Green, A. R. Janus kinase deregulation in leukemia and lymphoma. *Immunity* **36,** 529–541 (2012).

19. Ortmann, C. A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372,** 601–612 (2015).

20. Nangalia, J. *et al.* Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* **369,** 2391–2405 (2013).

Figure Legends

Figure 1. **The SC3 framework for consensus clustering.** (**a**) Overview of clustering with SC3 framework (see Methods). The consensus step is exemplified using the Treutlein data. (**b**) Published datasets used to set SC3 parameters. *N* is the number of cells in a dataset; *k* is the number of clusters originally identified by the authors; 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**) Histogram of the *d* values where ARI>.95 is achieved for the gold standard datasets. The black vertical lines indicate the interval *d* = 4-7% of the total number of cells *N,* showing high accuracy in the classification. (**d**) 100 realizations of the SC3 clustering of the datasets shown in (**b**). Dots represent individual clustering runs. Bars correspond to the median of the dots. Red and grey colours correspond to clustering with and without consensus step. The black line corresponds to ARI=0.8. The dashed black line separates gold and silver standard datasets.

Figure 2. **Benchmarking of SC3 against existing methods.** (**a**) SC3, tSNE+kmeans 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, however was optimised over different values of the density parameter *G* (Methods). Each panel shows the ARI (black dots, Methods) between the inferred clusterings and the reference labels. Bars correspond to the median of the dots. For the Pollen and Usoskin datasets all different hierarchies were considered (Data Avaialbility). The black line indicates ARI = 0.8. The dashed black line separates gold and silver standard datasets. (**b**) Number of clusters $\hat{k}$ predicted by SC3, SINCERA and SNN-Cliq for all datasets. Ref is the reference clustering reported by the authors. **(c)** The performance of the hybrid SC3 (Methods). 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. The dashed black line in the legend separates gold and silver standard datasets. (**d**) The consensus matrix as generated by SC3 for the Deng dataset (Methods). The matrix indicates how often each pair of cells was assigned to the same cluster by the different parameter combinations as indicated by the colorbar (1 - always, 0 - never). 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).

Figure 3. **Using SC3 to define subclones from two patients with myeloproliferative neoplasm.** Marker gene expression matrix (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). Only the top 10 marker genes are shown for each cluster.