scmap: projection of single-cell RNA-seq data across data sets

Vladimir Yu Kiselev, Andrew Yiu & Martin Hemberg

Single-cell RNA-seq (scRNA-seq) allows researchers to define cell types on the basis of unsupervised clustering of the transcriptome. However, differences in experimental methods and computational analyses make it challenging to compare data across experiments. Here we present scmap (http://bioconductor.org/packages/scmap; web version at http://www.sanger.ac.uk/science/tools/scmap), a method for projecting cells from an scRNA-seq data set onto cell types or individual cells from other experiments.

As more and more scRNA-seq data sets become available, there is a growing need to compare data sets of similar biological origin collected by different labs, and to ensure that annotations and analyses are consistent. Moreover, as large references such as the Human Cell Atlas (HCA)[1] become available, it will be important to project cells from a new sample—a disease tissue, for example—onto existing references in order to characterize differences in composition or detect new cell types (**Fig. 1a**). Conceptually, such projections are similar to the popular BLAST[2] method, which quickly finds the closest match in a database of nucleotide or amino acid sequences.

Projecting a new cell, *c*, onto a reference data set allows one to identify the cluster or cell most similar to *c*, that is, its nearest neighbor (Online Methods). We carry out a search by cluster, referred to here as scmap-cluster, in which we represent each cluster by its centroid (a vector of the median value of the expression of each gene) and measure the similarity between *c* and each cluster centroid or cell. The nearest cluster can be searched for exhaustively because the number of clusters is typically much smaller than the number of cells in the reference. To speed up the search for the nearest cell, we carry out an approximate nearest neighbor search using a product quantizer[3]. Moreover, instead of using all genes to calculate the similarity, we use unsupervised feature selection to include only the genes that are most relevant for the underlying biological differences, which allows us to overcome batch effects[4].

Here we investigated three different strategies for feature selection: random selection, highly variable genes (HVGs)[5] and genes with a higher number of dropouts (zero expression) than expected (determined using M3Drop)[4]. To increase speed, we modified M3Drop to fit a linear model instead of the Michaelis–Menten model (Online Methods and **Supplementary Fig. 1a**). For the number of features, we used the top 100, 200, 500, 1,000, 2,000, 5,000, or all genes. We calculated similarities by using the cosine similarity and Pearson and Spearman correlations, which are restricted to the interval $[-1, 1]$ and are thus insensitive to differences in scale between data sets. We required that at least two of the similarities be in agreement, and that at least one be >0.7. If these criteria were not met, then *c* was labeled as "unassigned" to indicate that it did not correspond to any cell type present in the reference. For the approximate nearest neighbor search, which we refer to as scmap-cell, we carried out a form of *k*-nearest -neighbor classification with only cosine similarity. For a cell type to be assigned, we required that the three nearest neighbors have the same cell type and that the highest similarity among them be >0.5.

To validate the projections, we considered 17 previously published data sets (**Supplementary Table 1**). We evaluated feature-selection methods in a self-projection experiment. We randomly selected 70% of the cells from the original sample for the reference, and projected the remaining 30%, with clusters as defined by the original authors. To quantify mapping accuracy, we used Cohen's κ (ref. 6), which is a normalized index of agreement between sets of labels that accounts for the frequency of each label. A value of 1 indicates that the projection assignments were in complete agreement with the original labels, whereas 0 indicates that the projection assignment was no better than a random guess. The dropout-based method for feature selection had the best performance, and, somewhat surprisingly, we also found that random selection was better than the HVG approach (**Supplementary Fig. 1b**). Furthermore, the dropout-based method performed consistently well when 100–1,000 features were selected. The dropout-based method performed better than the HVG method because it selects genes that are either absent or present in each cluster, and these genes provide a more reliable signal for the separation of groups of cells[4]. We also considered two commonly used supervised methods for assigning labels to new samples: random forest classifier (RF) and support vector machine (SVM). We trained these classifiers on the reference and then applied them to the held-out cells as before. For the self-projection experiment, RF and SVM performed slightly better than scmap-cluster and scmap-cell for all three feature-selection methods (**Supplementary Fig. 1b**).

As positive controls, we considered seven pairs of data sets (**Supplementary Table 2**) that we expected to correspond well on the basis of similar sample origins. The positive controls were more realistic than for the self-projection because they included systemic differences such as batch effects. For example, for three

Wellcome Sanger Institute, Hinxton, UK. Correspondence should be addressed to M.H. (mh26@sanger.ac.uk).

RECEIVED 26 JULY 2017; ACCEPTED 9 MARCH 2018; PUBLISHED ONLINE 2 APRIL 2018; DOI:10.1038/NMETH.4644

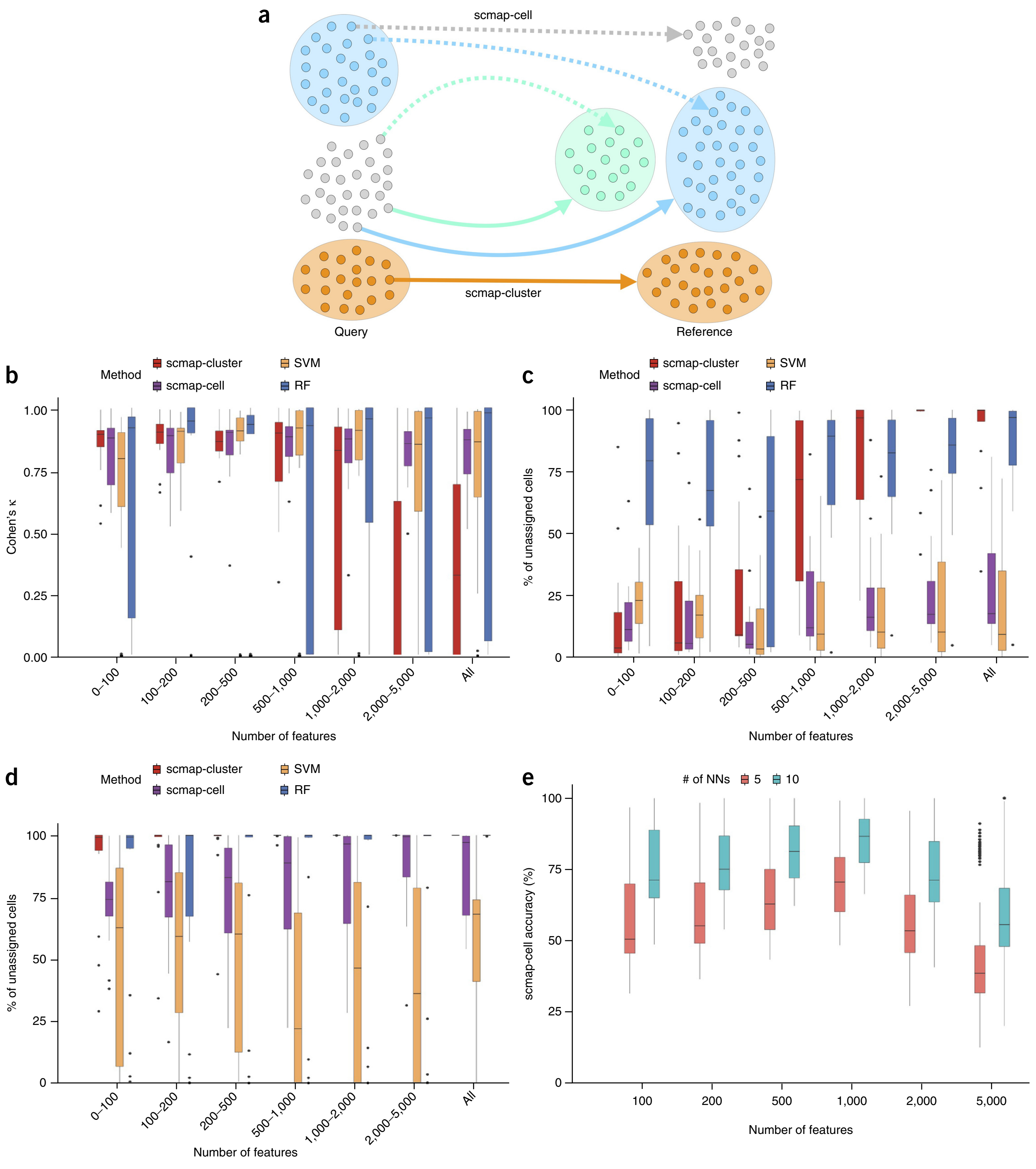
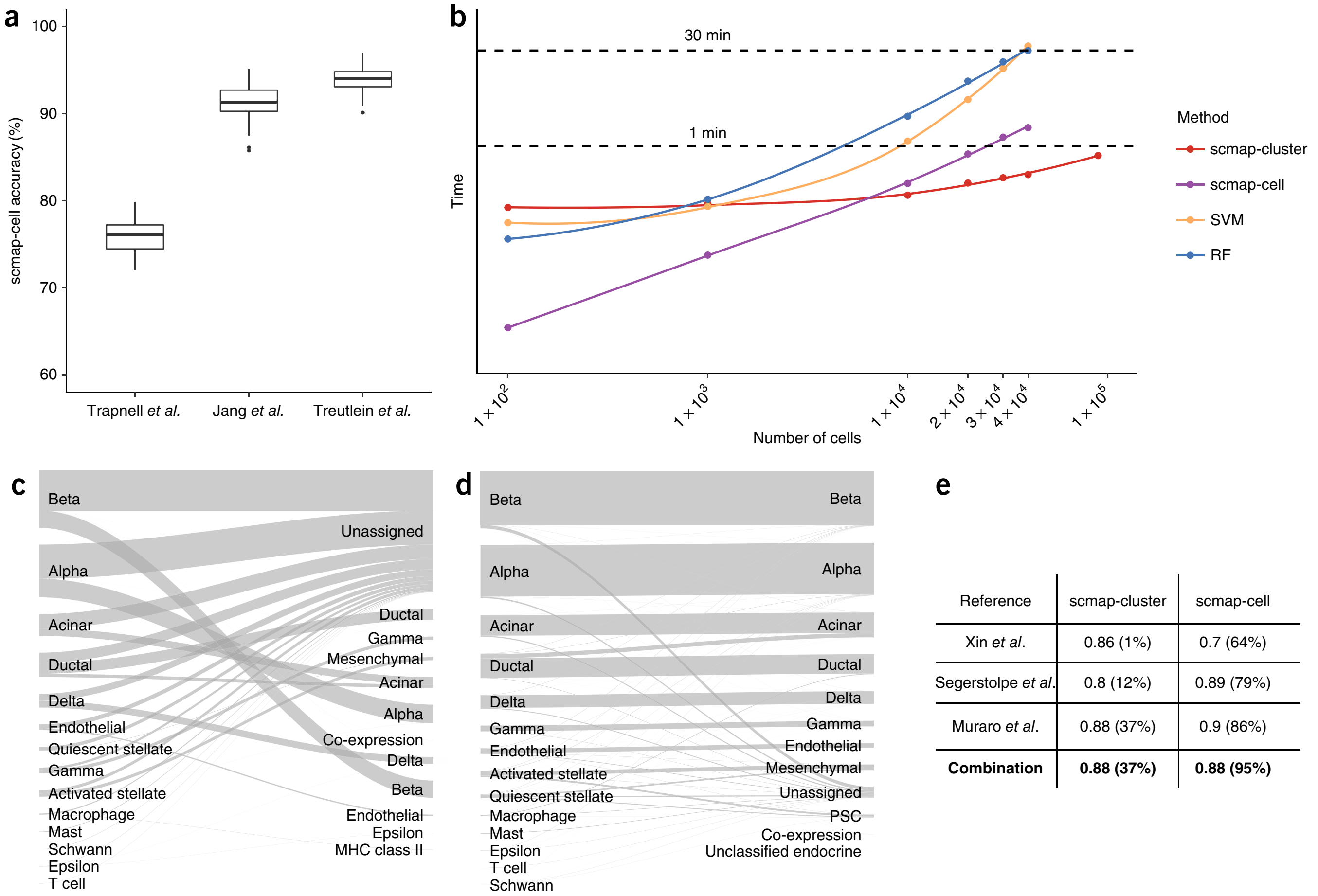


Figure 1 | scmap use and performance. (**a**) scmap can map individual cells from a query sample to cell types in the reference (scmap-cluster; solid lines) or to individual cells in a reference (scmap-cell; dashed lines). Colors correspond to previously identified cell types; gray denotes unknown cell types. (**b**,**c**) Cohen's κ (**b**) and the percentage of unassigned cells (**c**) for *n* = 14 positive controls listed in **Supplementary Table 2**. (**d**) The percentage of unassigned cells in *n* = 18 negative controls listed in **Supplementary Table 3**. (**e**) Accuracy of scmap-cell search for nearest neighbors (NNs), calculated to show how often the true nearest neighbor was found among the five or ten nearest cells for the data sets listed in **Supplementary Table 1**, except those of Shekhar *et al.*[24] and Macosko *et al.*[25]. For **b**–**e**, projections were performed in both directions, dropout-based feature selection was used for all methods (Online Methods) and scmap-cell was run once for each data set cpair, except for **e**, where scmap-cell was run 100 times for each data set. Box plots show the median (center line), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points.



Reference	scmap-cluster	scmap-cell
Xin *et al.*	0.86 (1%)	0.7 (64%)
Segerstolpe *et al.*	0.8 (12%)	0.89 (79%)
Muraro *et al.*	0.88 (37%)	0.9 (86%)
Combination	**0.88 (37%)**	**0.88 (95%)**

Figure 2 | scmap for combined references. (**a**) scmap-cell accuracy for three data sets[11–13] with differentiation trajectories, calculated to show how often the true nearest neighbor is found among the ten nearest cells (1,000 dropout-selected features were used for projections, and scmap-cell was run $n = 100$ times for each data set). Box plots indicate the median (center lines), interquartile range (hinges) and 1.5 times the interquartile range (whiskers); outlier data beyond this range are plotted as individual points. (**b**) Single CPU laptop run times for creating the reference (for scmap) and training classifiers (for SVM and RF), for different numbers of cells in the reference data set. For all methods, 1,000 features were used. For 10^5 cells, scmap-cell failed owing to a lack of memory. Solid lines are local polynomial regression fitting with span = 1, plotted by the geom_smooth() function of the R ggplot2 package (https://cran.r-project.org/package=ggplot2). (**c**,**d**) Results of scmap-cluster projection of the Baron *et al.*[18] data set to the Xin *et al.*[19], Segerstolpe *et al.*[20] and Muraro *et al.*[26] pancreas data sets using a combination strategy (Sankey diagram) (**c**) and scmap-cell projection (Sankey diagram) (**d**). (**e**) Results of scmap-cluster and scmap-cell projections of the Baron *et al.* data set to the three human pancreas data sets (Reference), and results of the "Combination" projection. Under "scmap-cluster" and "scmap-cell", the first number is the κ value, and the second is the percentage of assigned cells.

of the pairs, one data set was collected via a full-length protocol, and the other was collected via a protocol that used unique molecular identifiers. Despite the substantial differences between the protocols[7–9], both scmap-cluster and scmap-cell on average achieved $\kappa > 0.75$ and assignment rates above 75% when 100–500 features were used (**Fig. 1b**,**c**, **Supplementary Figs. 2** and **3**). Although RF achieved the highest κ, it also had the lowest assignment rate (<50%), which indicates that it achieves high specificity at the cost of low sensitivity. An important feature of scmap is its robustness to gene dropouts, as both the centroids and the nearest neighbor relations are unaffected by the increased frequency of zeros (**Supplementary Fig. 4**).

As negative controls, we projected data sets with origins not related to the reference (e.g., retina onto pancreas; **Supplementary Table 3**). We found that both scmap versions categorized >80% of the cells as unassigned when the number of features used was >100 (**Fig. 1d**). Notably, SVM had a much smaller fraction of unassigned cells than RF and scmap, which indicates that it is too lenient in assigning matches. After comparing self-projection experiments and positive and negative controls, we concluded that scmap with 500 features provides the best performance by balancing high sensitivity and specificity with a low false positive rate.

We evaluated scmap-cell by asking how often it is able to identify the true nearest neighbor, which we defined by calculating the nearest neighbor exactly among one of the five or ten nearest cells. For 15 of the 17 data sets used earlier, scmap-cell had an average accuracy of 64% or 80% when 5 or 10 neighboring cells were compared, respectively (**Fig. 1e**). Although scmap-cell identified the correct cluster for the remaining two data sets, which were generated by Drop-seq, it achieved an accuracy of only ~20% for identification of the nearest neighbor. We hypothesize that Drop-seq data performed worse because they represented approximately tenfold more cells but a similar sequencing depth compared with the inDrop data sets. In this situation, the linearized dropout-based feature selection performs worse because there are fewer outliers in the expression–dropout plot (**Supplementary Fig. 1**). We thus suggest that deeper sequencing is required for scmap-cell to reliably identify nearest neighbors.

When discrete cluster labels are not available, as for continuous differentiation trajectories[10], scmap-cluster cannot be used, and one must instead use scmap-cell. For trajectories from mouse myoblast differentiation[11], mouse embryonic stem cell differentiation[12] and mouse fibroblast-to-neuron reprogramming[13] data sets, scmap-cell correctly identified the nearest neighbor in 76%, 91% and 94% of the cases, respectively (**Fig. 2a**).

An important feature of scmap is its speed. Feature selection and centroid calculation take around 20 s for 40,000 cells with scmap-cluster, and scmap-cell takes less than 1 min to create the index, whereas it takes almost 30 min to train an RF or SVM (**Fig. 2b**). For all four methods, the time to project the new cells was negligible, which means that they are very fast with a precomputed reference. The scmap-cluster index is ~5,000 times smaller than the original expression matrix; scmap-cluster will thus be applicable to very large data sets, as complexity scales with the number of reference clusters. The scmap-cell index is ~500-fold smaller than the original expression matrix (**Supplementary Table 1**).

Large references, including the HCA, will be conglomerations of data sets collected by different groups. The merging of different scRNA-seq data sets remains an open problem[14–16], but the results from our study suggest that samples with similar origins are largely consistent[17] (**Fig. 1b**). Instead of correcting for batches and merging, one can create a composite reference and compare the new cells to each data set separately. When there are multiple data sets in the reference, scmap reports the best match for each. Thus, if a cell shows a high degree of similarity to clusters with similar annotations from different data sets, the confidence of the mapping increases. To illustrate mapping to multiple data sets, we considered the pancreas data set from Baron *et al.*[18], because it had the most unassigned cells when projected to the other pancreas data sets included in our study. When we combined all projections (Online Methods), the fraction of unassigned cells decreased from 99% (ref. 19) and 88% (ref. 20) to 63% without making κ worse. Interestingly, for this example scmap-cell performed better than scmap-cluster (**Fig. 2c–e**). Because the reference used by scmap is modular and can be extended without recalculation of the features or centroids for the previously processed data sets, construction of a composite reference works well when the reference is expected to grow over a long period of time.

We have implemented scmap as an R package (**Supplementary Software**), and it has been included in Bioconductor to facilitate incorporation into bioinformatic workflows (http://bioconductor.org/packages/scmap). Because scmap is integrated with scater[21], it can easily be combined with many other popular computational scRNA-seq methods. Moreover, scmap is available via the web (http://www.sanger.ac.uk/science/tools/scmap), where users can either upload their own reference or use a collection of data sets from this paper for which the features and centroids have been precalculated (Online Methods).

Because of differences in experimental conditions, comparison among scRNA-seq data sets remains challenging. However, for researchers to take advantage of large references such as the HCA, fast, robust and accurate methods for merging[22,23] and projecting cells across data sets are required. scmap is unique as a widely applicable projection method that can identify the best-matching cell type or individual cell in the reference. We have demonstrated that scmap can be used to compare samples of similar origin collected by different groups, as well as to compare cells to a large reference composed of multiple data sets.

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 T. Andrews, K.N. Natarajan, G. Parada, M. Schaub, M. Stubbington, V. Svensson, J. Westoby and F. Wünnermann for helpful discussions, feedback on the manuscript and testing of the cloud implementation of scmap. Amazon Web Services (AWS) Cloud provided credits for running the scmap server for 1 year. V.Y.K., A.Y. and M.H. were supported by core funding to the Wellcome Sanger Institute provided by the Wellcome Trust.

AUTHOR CONTRIBUTIONS

M.H. conceived the study and supervised the research; V.Y.K., A.Y. and M.H. contributed to the computational framework; V.Y.K. and M.H. wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

1. Regev, A. *et al. eLife* **6**, e27041 (2017).
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403–410 (1990).
3. Jégou, H., Douze, M. & Schmid, C. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 117–128 (2011).
4. Andrews, T.S. & Hemberg, M. *BioRxiv.* Preprint at https://www.biorxiv.org/content/early/2017/05/25/065094 (2016).
5. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
6. Cohen, J. *Psychol. Bull.* **70**, 213–220 (1968).
7. Picelli, S. *et al. Nat. Protoc.* **9**, 171–181 (2014).
8. Hashimshony, T. *et al. Genome Biol.* **17**, 77 (2016).
9. Klein, A.M. *et al. Cell* **161**, 1187–1201 (2015).
10. Wagner, A., Regev, A. & Yosef, N. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
11. Trapnell, C. *et al. Nat. Biotechnol.* **32**, 381–386 (2014).
12. Jang, S. *et al. elife* **6**, e20487 (2017).
13. Treutlein, B. *et al. Nature* **534**, 391–395 (2016).
14. La Manno, G. *et al. Cell* **167**, 566–580 (2016).
15. Tung, P.-Y. *Sci. Rep.* **7**, 39921 (2017).
16. Camp, J.G. *et al. Nature* **546**, 533–538 (2017).
17. Crow, M., Paul, A., Ballouz, S., Huang, Z.J. & Gillis, J. *Nat. Commun.* **9**, 884 (2018).
18. Baron, M. *et al. Cell Syst.* **3**, 346–360 (2016).
19. Xin, Y. *et al. Cell Metab.* **24**, 608–615 (2016).
20. Segerstolpe, Å. *et al. Cell Metab.* **24**, 593–607 (2016).
21. McCarthy, D.J., Campbell, K.R., Lun, A.T.L. & Wills, Q.F. *Bioinformatics* **33**, 1179–1186 (2017).
22. Butler, A. & Satija, R. *BioRxiv.* Preprint at https://www.biorxiv.org/content/early/2017/07/18/164889 (2017).
23. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2017/07/18/165118 (2017).
24. Shekhar, K. *et al. Cell* **166**, 1308–1323 (2016).
25. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
26. Muraro, M.J. *et al. Cell Syst.* **3**, 385–394 (2016).

ONLINE METHODS

Data sets. All data sets and cell-type annotations were downloaded from their public accessions. The data sets were converted into Bioconductor SingleCellExperiment (http://bioconductor.org/packages/SingleCellExperiment) class objects (details are available on our data set website, https://hemberg-lab.github.io/scRNA.seq.datasets). In the data set from Segerstolpe *et al.*[20], cells labeled as "not applicable" were removed because it was unclear how to interpret this label and what it should be matched to in the other data sets. In the data set from Xin *et al.*[19], cells labeled as "alpha.contaminated," "beta.contaminated," "gamma.contaminated" and "delta.contaminated" were removed because they were likely to correspond to cells of lower quality. In the following data sets, similar cell types were merged:

- In the data set from Deng *et al.*[27], "zygote" and "early2cell" were merged into a "zygote" cell type, "mid2cell" and "late2cell" were merged into a "2cell" cell type, and "earlyblast," "midblast" and "lateblast" were merged into the "blast" cell type.
- All bipolar cell types in the data set from Shekhar *et al.*[24] were merged into the "bipolar" cell type.
- In the data set from Yan *et al.*[28], "oocyte" and "zygote" cell types were merged into the "zygote" cell type.

Feature selection. To select informative features, we used a method conceptually similar to M3Drop[4] to relate the mean expression (E) and the dropout rate (D). We used a linear model to capture the relationship of $\log(E)$ and $\log(D)$, and after fitting a linear model using the lm() command in R, we selected important features as the top N residuals of the linear model (**Supplementary Fig. 1a**). The features are selected only from the reference data set, and those that are absent or zero in the projection data set are further excluded before scmap is run. All three feature-selection methods are described in **Supplementary Note 1**.

Reference centroid. In scmap-cluster, each cell type in the reference data set is represented by its centroid, that is, the median value of gene expression for each feature gene across all cells in that cell type.

Approximate nearest neighbor search using product quantizer. scmap-cell performs a fast approximate k-nearest-neighbor search using product quantization[3]. The original algorithm, built around the Euclidean distance, was adapted to incorporate the cosine distance, which helps to protect against batch effects and scaling inconsistencies between data sets. The product quantizer creates a compressed index where every cell in the reference is identified with a set of sub-centroids found via k-means clustering based on a subset of the features. Through concatenation of the sub-centroids, a close approximation of the original expression vector is obtained. When the reference is searched for the nearest neighbors to a query cell, the approximations provided by the sub-centroids are used instead of the individual cells in the reference. Because the number of centroids can be made much smaller than the original number of cells in the data set, the method provides a substantial reduction in both computation time and storage requirements compared to exact search.

Projection data set. For projection of a data set to a reference data set, similarities between each cell and all centroids of the reference data set are calculated, using only the common selected features. Three similarity measures are used: Pearson, Spearman and cosine. The cell is then assigned to the cell type that corresponds to the highest similarity value. However, scmap-cluster requires that at least two similarity measures agree with each other; otherwise the cell is marked as "unassigned." Additionally, if the maximum similarity value across all three similarities is below a similarity threshold (the default is 0.7), then the cell is also marked as "unassigned." Because only the cosine similarity measure was calculated for scmap-cell, the default threshold of 0.5 was used, and the nearest three neighbors were required to be in agreement with respect to cell type in order for the cell to be assigned. Positive and negative control plots corresponding to **Figure 1c–e** for different values of the similarity/probability threshold (0.5, 0.6, 0.8 and 0.9; see subsection "SVM and RF" below) are shown in **Supplementary Figure 2**.

SVM and RF. The scmap projection algorithm was benchmarked against SVM[29] (with a linear kernel) and RF[30] (with 50 trees) from the R packages e1071 and randomForest, respectively. The classifiers were trained on all cells of the reference data set, and a cell type for each cell in the projection data set was predicted by the classifiers. Additionally, a threshold (default value of 0.7) was applied to the probabilities of assignment: if the probability was less than the threshold, the cell was marked as "unassigned."

Sensitivity to sequencing depth and dropouts. We artificially increased the dropout rate in the positive control data sets (**Supplementary Table 2**) by randomly setting 10%, 30% and 50% of the nonzero expression values to zero (**Supplementary Fig. 4a**,**b**). scmap was run 100 times for each box.

Projection based on multiple data sets. When the reference contains multiple data sets collected from similar samples by different groups in addition to all similarities, for each cell scmap also reports a top cell-type match based on the highest value of similarities across all reference cell types. A similarity threshold of 0.7 is also applied in this case.

scmap on the cloud. An example of a cloud version of scmap is available at http://www.sanger.ac.uk/science/tools/scmap. Instructions for setup on a user's personal web cloud environment are available on our github page (https://github.com/hemberg-lab/scmap-shiny). An extended tutorial on how to use scmap can be found in **Supplementary Note 2**.

Figures. All data and scripts used to generate figures in this paper are available at https://github.com/hemberg-lab/scmap-paper-figures. All methods were run on a MacBook Pro laptop (Mid 2014) with 2.8 GHz Intel Core i7 processor, 16 GB 1600 MHz DDR3 of RAM.

Life Sciences Reporting Summary. Further information on experimental design is available in the **Life Sciences Reporting Summary**.

Data availability. All data used here are from published studies, and information about their original publication can be found in **Supplementary Table 1**. Source data for **Figures 1** and **2** and **Supplementary Figures 1–4** are available online.

27. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. *Science* **343**, 193–196 (2014).

28. Yan, L. *et al. Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).

29. Ben-Hur, A., Horn, D., Siegelmann, H.T. & Vapnik, V. *J. Mach. Learn. Res.* **2**, 125–137 (2001).

30. Breiman, L. *Mach. Learn.* **45**, 5–32 (2001).

doi:10.1038/nmeth.4644

nature research

Corresponding author(s): DBPR, NMETH-BC32156B

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

No experiments in study

2. Data exclusions

Describe any data exclusions.

No experiments in study

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

No experiments in study

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No experiments in study

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No experiments in study

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
□	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
□	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
□	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Nature Methods: doi:10.1038/nmeth.4644

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

http://bioconductor.org/packages/scmap (v. 1.1.5)
https://github.com/hemberg-lab/scmap (v. 1.1.5)
R packages e1071 (v. 1.6-8) and randomForest (v. 4.6-12)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used in the study

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No eukaryotic cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

No animals were used in this study

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This study did not involve human research participants