bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CIDR: Ultrafast and accurate clustering through imputation for single cell RNA-Seq data

Peijie Lin[1,2], Michael Troup[1] & Joshua W. K. Ho[1,2]

[1] *Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia.*

[2] *St Vincent's Clinical School, University of New South Wales, Darlinghurst, NSW 2010, Australia.*

Most existing dimensionality reduction and clustering packages for single cell RNA-Seq (scRNA-Seq) data deal with dropouts by heavy modelling and computational machinery. Here we introduce *CIDR* (Clustering through Imputation and Dimensionality Reduction), an ultrafast algorithm which uses a novel yet very simple 'implicit imputation' approach to alleviate the impact of dropouts in scRNA-Seq data in a principled manner. Using a range of simulated and real data, we have shown that *CIDR* outperforms the state-of-the-art methods, namely *t-SNE*, *ZIFA* and *RaceID*, by at least 50% in terms of clustering accuracy, and typically completes within seconds for processing a dataset of hundreds of cells.

scRNA-Seq enables researchers to study heterogeneity between individual cells and define cell types from a transcriptomic perspective. One prominent problem in scRNA-Seq data analysis is the prevalence of dropouts, caused by failures in amplification during the reverse-transcription step in the RNA-Seq experiment. The prevalence of dropouts manifests as an excess of zeros and near zero counts in the dataset, which has been shown to create difficulties in scRNA-Seq data analysis[1,2].

Several packages have recently been developed for the various aspects of scRNA-Seq data analysis[2–4], but few perform pre-processing steps such as dimensionality reduction and clustering, which are critical steps for studying cell type heterogeneity. The state-of-the-art dimensionality reduction package for scRNA-Seq data is *ZIFA*[1]; its use of the expectation-maximization algorithm makes it computationally intensive and hence difficult to cope with the increasingly large scRNA-Seq datasets. Another package *t-SNE*[5] is popular among biologists, but it is not designed specifically for scRNA-Seq data and does not address the issue of dropouts. Regarding clustering

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and cell type classification for scRNA-Seq data, there have only been two packages, *SNN-Cliq*[6] and *RaceID*[7], developed specifically for this purpose. Like *t-SNE*, neither of these algorithms addresses the issue of dropouts.

In contrast to the use of heavy modelling and computational machinery by current state-of-the-art methods, *CIDR* uses a novel yet very simple 'implicit imputation' approach to alleviate the impact of dropouts in a principled manner (Supplementary Fig. 1). *CIDR* first performs a logarithmic transformation on the tag per million (TPM), after which the logTPM for each cell typically displays a bimodal distribution. For each cell C_i, *CIDR* finds a sample-dependant threshold T_i that separates the first and second modes; Supplementary Fig. 2a shows the distribution of tags for a library in a simulated dataset, and the red vertical line indicates the threshold T_i. The entries for cell C_i with an expression of less than T_i are dropout candidates, and the entries with an expression of at least T_i are referred to as 'expressed'. We call this threshold T_i the 'dropout candidate threshold'. Note that dropout candidates include dropouts as well as real low expressions.

Let u be the unobserved real expression of a feature in a cell and let $P(u)$ be the probability of it being a dropout. Empirical evidence suggests that $P(u)$ is a decreasing function[1,2]. *CIDR* uses non-linear least squares to fit a decreasing logistic function to the data (empirical dropout rate versus average of expressed entries) as an estimate for $P(u)$, illustrated by the 'Tornado Plot' Supplementary Fig. 2b for the simulated dataset. Using the whole dataset to estimate $P(u)$, which we denote as $\hat{P}(u)$, makes the reasonable assumption that most dropout candidates in the dataset are actually dropouts, and allows the sharing of information between genes and cells.

$\hat{P}(u)$ is used for imputation in the calculation of the *CIDR* dissimilarity matrix. The dropout candidates are treated as missing values and we will now describe *CIDR*'s pairwise 'implicit' imputation process. Consider a pair of cells C_i and C_j, and their respective observed expressions o_{ki} and o_{kj} for a feature F_k, and let T_i and T_j be dropout candidate thresholds defined as above. Imputation is only applied to dropout candidates, hence the case in which $o_{ki} \geq T_i$ and $o_{kj} \geq T_j$ requires no imputation. Now consider the case in which one of the two expressions is below T_i, say $o_{ki} < T_i$ and $o_{kj} \geq T_j$; in this case o_{ki} needs to be imputed and the imputed value $\hat{o}_{ki}$ is defined

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as the weighted mean

$$\hat{o}_{ki} = \hat{P}(o_{kj})o_{kj} + \big(1 - \hat{P}(o_{kj})\big)o_{ki}. \tag{1}$$

To achieve fast speed in the implementation of the above step, we replace $\hat{P}(u)$ with a much simpler step function $W(u)$, defined as

$$W(u) = \begin{cases} 0, & \hat{P}(u) \leq T, \\ 1, & \hat{P}(u) > T, \end{cases} \tag{2}$$

where T is by default 0.5. We refer to $W(u)$ as the 'imputation weighting function' as it gives us the weights in the weighted mean in the imputation, and we refer to the jump of $W(u)$, i.e. $\hat{P}^{-1}(T)$, as the 'imputation weighting threshold' (Supplementary Fig. 2c). Therefore, the implemented version of Equation (1) is

$$\tilde{o}_{ki} = W(o_{kj})o_{kj} + \big(1 - W(o_{kj})\big)o_{ki}, \tag{3}$$

where $\tilde{o}_{ki}$ is used as the imputed value of o_{ki}. Lastly, if $o_{ki} < T_i$ and $o_{kj} < T_j$, we set both $\tilde{o}_{ki}$ and $\tilde{o}_{kj}$ to be zero.

Then, the dissimilarity between C_i and C_j is calculated as the Euclidean distance using the imputed values. We call this imputation approach 'implicit', as the imputed value of a particular observed expression of a cell changes each time when it is paired up with a different cell. The theoretical justification of this implicit imputation approach can be found in Supplementary Materials.

Dimensionality reduction is achieved by performing principal coordinates analysis on the *CIDR* dissimilarity matrix. It has been known that high dimensionality has adverse effects on clustering results, and clustering performed on the reduced dimensions improves the results[8]. *CIDR* performs hierarchical clustering on the first few (by default 4) principal coordinates, and decides the number of clusters based on the Calinski-Harabasz Index[9].

For evaluation, we have created a realistic simulated scRNA-Seq dataset. We set the number of markers for each cell type low to make it a difficult dataset to analyse. Supplementary Fig. 2a

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

shows the distribution of tags for one randomly chosen library in this simulated dataset. The spike on the left is typical for scRNA-Seq datasets and the tags in this spike are dropout candidates. We have compared *CIDR* with the standard principal component analysis implemented by the R function *prcomp*, two state-of-the-art dimensionality reduction algorithms – *t-SNE* and *ZIFA*, and the recently published scRNA-Seq clustering package *RaceID*. Since *prcomp*, *ZIFA* and *t-SNE* don't perform clustering, for the purpose of comparison, we apply the same hierarchical clustering procedure used by *CIDR* to the first four principal components output by each of the algorithms. We use the Adjusted Rand Index[10] to measure the accuracy of clustering.

As shown in Fig. 1, the only algorithm that displays three clearly recognisable clusters in the first two dimensions is *CIDR*; it is also the only algorithm that correctly identifies the number of clusters. *CIDR*'s accuracy in cluster membership assignment is reflected by an Adjusted Rand Index much higher than the other four compared algorithms (Fig. 1f). *CIDR* outputs all the principal coordinates as well as a plot showing the proportion of variation explained by each of the principal coordinates (Supplementary Fig. 2d). Supplementary Fig. 2f shows the result when the number of principal coordinates used in clustering is altered from the default value of 4 to 2, based on an inspection of the proportion of variation plot.

We have applied *CIDR* and the four compared algorithms on two biological datasets where the cell types are known. In these studies, cell types were determined through a multi-stage process involving additional information such as cell type molecular signatures. For the purpose of evaluation and comparison, we have applied each of the compared algorithms only once in an unsupervised manner to test how well each algorithm can recover the cell type assignments in the two studies.

Fig. 2 shows the comparison results for the human brain scRNA-Seq dataset[11]. In this dataset there are 420 cells in 8 cell types after we exclude hybrid cells. Determining the number of clusters is known to be a difficult issue in clustering; *CIDR* has managed to identify 7 clusters in the brain dataset, which is very close to 8, the number of annotated cell types in this dataset. *CIDR* has also identified the members of each cell type largely correctly, as reflected by an Adjusted Rand Index

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

close to 0.9, which is a greater than 50% improvement over the second best algorithm (Fig. 2f). In the two dimensional visualization by *CIDR* (Fig. 2e), the first principal coordinate separates neurons from other cells, while the second principal coordinate separates adult and fetal neurons. Note that *t-SNE* is nondeterministic and it outputs dramatically different plots after repeated runs with the same input and the same parameters (Supplementary Fig. 4).

The human pancreatic islet scRNA-Seq dataset[12] has a smaller number of cells – 60 cells in 6 cell types after we exclude undefined cells and bulk RNA-Seq samples. *CIDR* is the only algorithm that displays clear and correct clusters in the first two dimensions (Supplementary Fig. 6). Regarding clustering accuracy, *CIDR* outperforms the second best algorithm by more than 80% in terms of Adjusted Rand Index (Supplementary Fig. 6f).

CIDR has ultrafast runtime, which is vital given the rapid growth in the size of scRNA-Seq datasets. The runtime comparison between *CIDR* and the other four algorithms is shown in Table 1. Across three datasets, *CIDR* takes only seconds to run on a standard laptop. It is faster than *prcomp* for two of the three datasets, and it is faster than all other compared algorithms for all three datasets; in particular, it's more than 400-fold faster than *ZIFA*.

Data pre-processing steps such as dimensionality reduction and clustering are important in scRNA-Seq data analysis because detecting clusters can greatly benefit subsequent analyses. For example, clusters can be used as covariates in differential expression analysis[3], or co-expression analysis can be conducted within each of the clusters separately[13]. Certain normalization procedures should be performed within each of the clusters[14]. Therefore, the vast improvement CIDR has over existing tools will be of interest to both users and developers of scRNA-Seq technology.

Methods

Dropout Candidates To determine the dropout candidate threshold that separates the first two modes in the distribution of tags (logTPM) of a library, *CIDR* finds the minimum point between the two modes in the density curve of the distribution. The Epanechnikov kernel is used in the

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

kernel density estimation. For robustness, after calculating all the dropout candidate thresholds, the top and bottom 10 percentiles of the thresholds are assigned the 90th percentile and the 10th percentile threshold values respectively. *CIDR* also gives the user the option of calculating the dropout candidate thresholds for only some of the libraries and in this option the median of the calculated thresholds is taken as the dropout candidate threshold for all the libraries.

Dimensionality Reduction A modified version of the *pcoa* function in the R package *ape* is used to perform principal coordinates analysis on the *CIDR* dissimilarity matrix. Because the *CIDR* dissimilarity matrix does not in general satisfy the triangle inequality, the eigenvalues can possibly be negative. This doesn't matter as only the first few principal coordinates are used in both visualization and clustering, and their corresponding eigenvalues are positive. Negative eigenvalues are discarded in the calculation of the proportion of variation explained by each of the principal coordinates. Some clustering methods require the input dissimilarity matrix to satisfy the triangle inequality. To allow integration with these methods, *CIDR* gives the user the option of Cailliez correction[15], implemented by the R package *ade4*. The corrected *CIDR* dissimilarity matrix doesn't have any negative eigenvalues.

Clustering By default, the first four principal coordinates are used to generate a distance matrix for clustering. *CIDR* outputs a plot that shows the proportion of variation explained by each of the principal coordinates, and the user is encouraged to inspect this plot and possibly alter the number of principal coordinates used in clustering. Supplementary Fig. 2d is the proportion of variation plot for the simulated dataset, and in this case an obvious good choice for the number of principal coordinates to be used in clustering is 2; Supplementary Fig. 2f shows the result when the number of principal coordinates used in clustering is altered from the default value of 4 to 2. Hierarchical clustering is performed using the R package *NbClust*. *CIDR*'s default clustering method for hierarchical clustering is 'ward.D2'[16], and the number of clusters is decided according to the Calinski-Harabasz Index[9]. Upon user request, *CIDR* can output a Calinski-Harabasz Index versus the number of clusters plot (Supplementary Fig. 2e); if needed, the user can overwrite the number of clusters.

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Simulation Study Simulated log tags are generated from a log-normal distribution. For each cell type, an expected library, i.e. the true distribution of log tags, is first generated, and then dropouts and noise are simulated. For each cell type, the expected library includes a small number of differentially expressed features (e.g. genes, transcripts) and markers; by markers we mean features that are expressed in one cell type and zeros in all the other cell types.

A probability function $\pi(x)$, where x is an entry in the expected library, is used to simulate dropouts. $\pi(x)$ specifies how likely an entry becomes a dropout, so intuitively it should be a decreasing function. In our simulation, we use a decreasing logistic function. The parameters of the logistic function can be altered to adjust the level of dropouts. After the simulation of dropouts, Poisson noise is added to generate the final distribution for each library.

Biological Datasets Tag tables from two recent scRNA-Seq studies (human brain[11] and human pancreatic islet[12]) were downloaded from the data repository NCBI Gene Expression Omnibus (GSE67835, GSE73727). The raw tag tables were used as the inputs for *CIDR*. For other dimensionality reduction and clustering algorithms, rows with tag sums less than or equal to 10 were deleted. Log tags, with base 2 and prior count 1, were used as the inputs for *ZIFA*, as suggested by the *ZIFA* documentation. Datasets transformed by logTPM were used as inputs for *prcomp* and *t-SNE.*

Implementation *CIDR* can be downloaded at http://github.org/VCCRI/CIDR.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Dr Joshua Ho (email: j.ho@victorchang.edu.au).

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1: **Performance evaluation with simulated data.** Simulated scRNA-Seq dataset parameters: 3 cell types, 50 cells in each cell type, 20,000 non-differentially expressed features, 150 differentially expressed features and 10 markers for each cell type. The three colors denote the three true cell types; while the different plotting symbols denote the clusters output by each algorithm. (a) - (e) Clustering output by each of the five compared algorithms; (f) Adjusted Rand Index is used to compare the accuracy of the clustering output by each of the compared algorithms.

Figure 2: **Performance evaluation with the human brain scRNA-Seq dataset.** In this dataset there are 420 cells in 8 cell types after the exclusion of hybrid cells. The different colors denote the cell types annotated by the study[11]; while the different plotting symbols denote the clusters output by each algorithm. (a) - (e) Clustering output by each of the five compared algorithms; (f) Adjusted Rand Index is used to measure the accuracy of the clustering output by each of the compared algorithms.

Table 1: Runtime comparison between *CIDR* and four other algorithms (standard laptop: 2.8 GHz Intel Core i5 (I5-4308U), 8GB DDR3 RAM).

Dataset	Number of Cells	CIDR	prcomp	t-SNE	RaceID	ZIFA
Pancreatic Islet	60	5.5**s**	2.8**s**	8.5**s**	48.6**s**	40.1**mins**
Simulation	150	2.5**s**	2.6**s**	16.0**s**	20.7**s**	31.1**mins**
Brain	420	8.1**s**	13.2**s**	77.2**s**	89.0**s**	68.7**mins**

1. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology* **16**, 1–10 (2015).

2. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11**, 740–742 (2014).

3. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**, 1–13 (2015).

4. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* **33**, 155–160 (2015).

5. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/068775; this version posted August 10, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–80 (2015).

7. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).

8. Ronan, T., Qi, Z. & Naegle, K. M. Avoiding common pitfalls when clustering biological data. *Science Signaling* **9**, re6 (2016).

9. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* **3**, 1–27 (1974).

10. Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193–218 (1985).

11. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* **112**, 7285–7290 (2015).

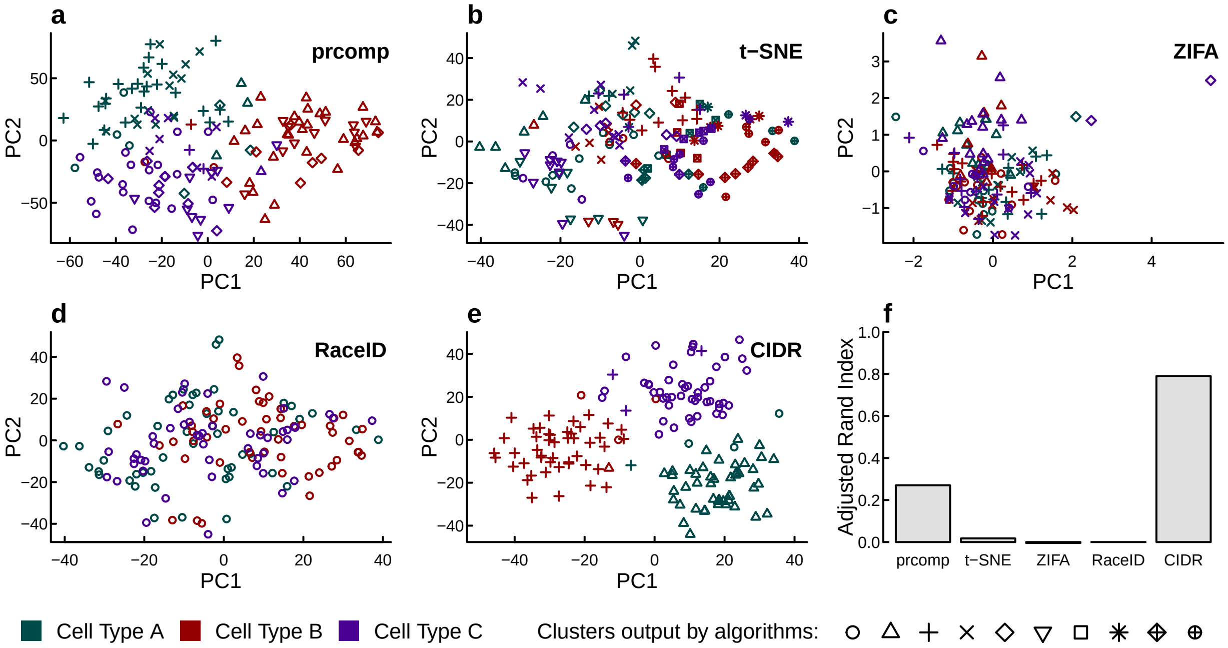
12. Li, J. *et al.* Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Reports* **17**, 178–187 (2016).

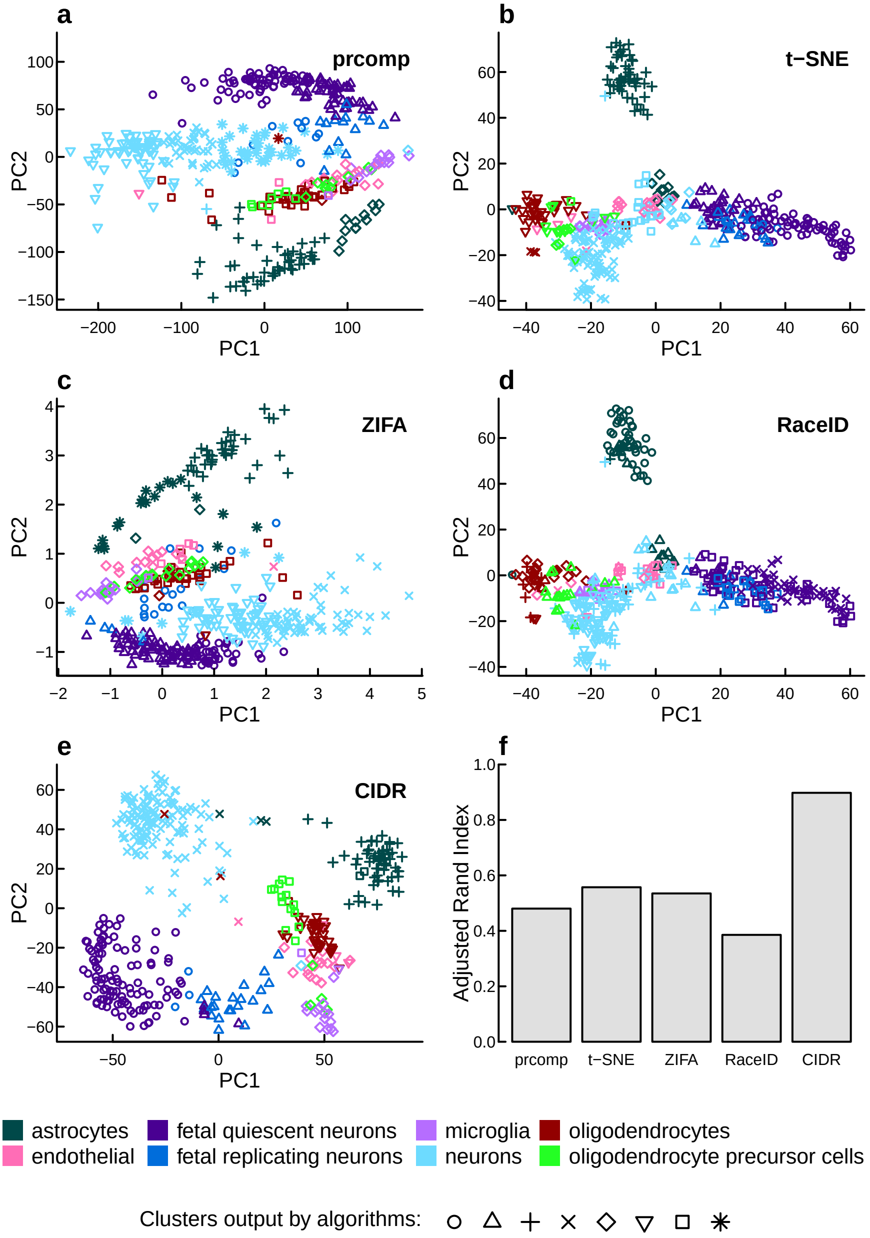
13. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491–1498 (2015).

14. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* **17**, 1 (2016).

15. Cailliez, F. The analytical solution of the additive constant problem. *Psychometrika* **48**, 305–308 (1983).

16. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification* **31**, 274–295 (2014).


a
prcomp
PC2
50
0
−50
−60
−40
−20
0
20
40
60
PC1
b
t−SNE
PC2
40
20
0
−20
−40
−40
−20
0
20
40
PC1
c
ZIFA
PC2
3
2
1
0
−1
−2
0
2
4
PC1
d
RaceID
PC2
40
20
0
−20
−40
−40
−20
0
20
40
PC1
e
CIDR
PC2
40
20
0
−20
−40
−40
−20
0
20
40
PC1
f
Adjusted Rand Index
1.0
0.8
0.6
0.4
0.2
0.0
prcomp
t−SNE
ZIFA
RaceID
CIDR
Cell Type A
Cell Type B
Cell Type C
Clusters output by algorithms:
○ △ + × ◇ ▽ □ ✳ ⊕


a
prcomp
PC2
PC1
100
50
0
−50
−100
−150
−200
−100
0
100
b
t−SNE
60
40
20
0
−20
−40
−40
−20
0
20
40
60
c
ZIFA
4
3
2
1
0
−1
−2
−1
0
1
2
3
4
5
d
RaceID
e
CIDR
60
40
20
0
−20
−40
−60
−50
0
50
f
Adjusted Rand Index
1.0
0.8
0.6
0.4
0.2
0.0
prcomp
t−SNE
ZIFA
RaceID
CIDR
astrocytes
endothelial
fetal quiescent neurons
fetal replicating neurons
microglia
neurons
oligodendrocytes
oligodendrocyte precursor cells
Clusters output by algorithms:
○ △ + × ◇ ▽ □ ✳