SINGLE-CELL OMICS

Challenges in unsupervised clustering of single-cell RNA-seq data

Vladimir Yu Kiselev, Tallulah S. Andrews and Martin Hemberg *

Abstract | Single-cell RNA sequencing (scRNA-seq) allows researchers to collect large catalogues detailing the transcriptomes of individual cells. Unsupervised clustering is of central importance for the analysis of these data, as it is used to identify putative cell types. However, there are many challenges involved. We discuss why clustering is a challenging problem from a computational point of view and what aspects of the data make it challenging. We also consider the difficulties related to the biological interpretation and annotation of the identified clusters.

Unsupervised clustering
The process of grouping objects based on similarity but without any ground truth or labelled training data.

The cell can be considered the fundamental unit in biology. For centuries, biologists have known that multicellular organisms are characterized by a plethora of distinct cell types. Although the notion of a cell type is intuitively clear, a consistent and rigorous definition has remained elusive. Cells can be distinguished by their size and shape using a microscope, and attributes based on their physical appearance have traditionally been the primary determinant of cell type. Later, discoveries in molecular biology made it possible to characterize cell types on the basis of the presence or absence of surface proteins. However, surface proteins represent only a small fraction of the proteome, and it is likely that important differences are not manifested at the cell membrane.

Advances in microfluidics have made it possible to isolate a large number of cells, and along with improvements in RNA isolation and amplification methods, it is now possible to profile the transcriptome of individual cells using next-generation sequencing technologies. Technological developments have advanced at a breathtaking speed. The first single-cell RNA sequencing (scRNA-seq) experiment was published in 2009, and the authors profiled only eight cells[1]. Only 7 years later, 10X Genomics released a data set of more than 1.3 million cells[2]. Thus, we are now in an era where large volumes of scRNA-seq data make it possible to provide detailed catalogues of the cells found in a sample.

For researchers to be able to take full advantage of these rich data sets, efficient computational methods are required. There are several steps involved in the computational analysis of scRNA-seq data, including quality control, mapping, quantification, normalization, clustering, finding trajectories and identifying differentially expressed genes (FIG. 1). The steps upstream of clustering may have a substantial impact on the outcome, and for each step numerous tools are available. Moreover, there are also software packages that implement the entire clustering workflow, for example, Seurat[3], scanpy[4] and SINCERA[5]. We encourage the reader to consult recently published overviews of this workflow[6–10], as this Review focuses on clustering alone. As clustering is the key step in defining cell types based on the transcriptome, one must carefully consider both the computational and biological aspects.

The ability to define cell types through unsupervised clustering on the basis of transcriptome similarity has emerged as one of the most powerful applications of scRNA-seq. Broadly speaking, the goal of clustering is to discover the natural groupings of a set of objects[11]. Defining cell types on the basis of the transcriptome is attractive because it provides a data-driven, coherent and unbiased approach that can be applied to any sample. This opportunity has spurred the creation of several atlas projects[12–17], most notably the Human Cell Atlas[18]. These atlas projects aim to build comprehensive references for all cell types present in an organism or tissue at various stages of development. In addition to providing a deeper understanding of the basic biology, atlases will also be useful as references for disease studies. For a cell atlas to be of practical use, reliable methods for unsupervised clustering of the cells will be one of the key computational challenges.

Although considerable progress has been made in terms of clustering algorithms over the past few years, a number of questions remain unanswered. In particular, there is no strong consensus about what is the best approach or how cell types can be defined based on scRNA-seq data. In this Review, we discuss several computational and biological aspects related to clustering. We first discuss the types of available clustering methods and when it is appropriate to use them, because one of the underlying assumptions is that discrete clusters are present in the data. Next, we outline why unsupervised clustering is a difficult problem and what considerations need to be taken from both experimental and computational points of view. We then discuss the challenges

Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK.

**e-mail: mh26@sanger.ac.uk*

https://doi.org/10.1038/s41576-018-0088-9

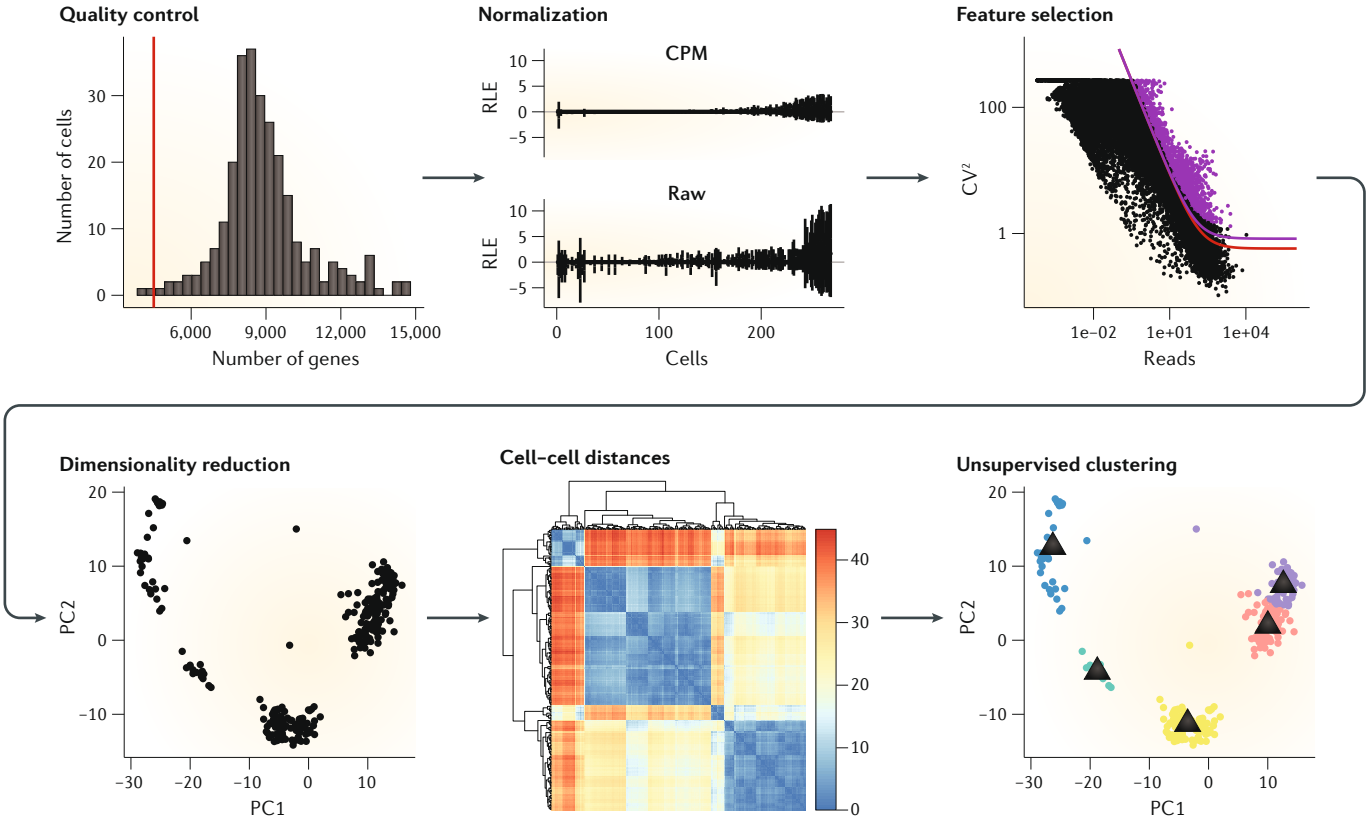

Fig. 1 | **Example data analysis workflow for scRNA-seq.** Overview of the workflow for the computational analysis of single-cell RNA sequencing (scRNA-seq) data leading up to unsupervised clustering. First, unreliable cells (and possible doublets) are removed through quality control. The cleaned data set is then normalized to correct for differences in read coverage and other technical confounders. Feature selection and dimensionality reduction isolate the most informative genes and strongest signals from background noise, respectively. Cell–cell distances are then calculated in the lower dimensional space and used to either construct a cell–cell distance graph or used directly by clustering algorithms to assign cells to clusters. Some methods will compute the distances before the dimensionality reduction. CPM, counts per million; CV, coefficient of variation; PC, principal component; RLE, relative log expression.

Feature selection
A collection of statistical approaches that identify and retain only variables that are most relevant to the underlying structure of the data set.

Dimensionality reduction
A collection of statistical approaches that reduces the number of variables in a data set. It often refers specifically to methods that recombine the original variables into a new set of non-redundant variables. Dimensionality reduction can help in identifying important patterns and reducing the amount of computations needed.

Greedy
An algorithm that, at each step, chooses the option that leads to the greatest reduction of the cost function. Greedy algorithms are often fast, but they may fail to find the optimal solution.

involved in the biological interpretation and annotation of the results. Finally, we discuss how clustering approaches are likely to evolve over the coming years.

What clustering strategies are available?

Many clustering algorithms are generic in the sense that they can be applied to any type of data that are equipped with a measure of distance between data points. Owing to the large number of genes assayed in scRNA-seq, that is, the high dimensionality, distances between data points (that is, cells) become similar, which is known as the 'curse of dimensionality'[19]. Consequently, differences in distances tend to be small and thus not reliable for identifying cell groups (FIG. 2). The application of feature selection and/or dimensionality reduction (FIG. 1) may reduce the noise and speed up calculations. Feature selection involves identifying the most informative genes, for example, the ones with the highest variance[20], whereas dimensionality reduction, for example, principal component analysis (PCA), projects data into a lower dimensional space. Many tools use variants of the standard methods: SC3 uses a small subset of principal components and pcaReduce applies PCA iteratively. Subsequently, distances are calculated in the lower dimensional space or by using only the selected genes. There are several different choices available, including Euclidean distance, cosine similarity, Pearson's correlation and Spearman's correlation. The main advantage of the three latter measures is their scale invariance, that is, they consider relative differences in values, making them more robust to library or cell size differences.

Diverse types of clustering methods are available (FIG. 3). The most popular clustering algorithm is *k*-means (FIG. 3b), which iteratively identifies *k* cluster centres (centroids), and each cell is assigned to the closest centroid. The standard method for *k*-means, known as Lloyd's algorithm[21], has the advantage of scaling linearly with the number of points, which means that it can be applied to large data sets. However, Lloyd's algorithm is greedy, and the method is not guaranteed to find the global minimum. These drawbacks can be overcome by repeated application of *k*-means using different initial conditions or upstream processing and finding the consensus, as performed by SC3 (REF.[22]). Another disadvantage of *k*-means is its bias towards identifying equal-sized clusters, which may result in rare cell types being hidden among a larger group. To overcome these issues, RaceID[23] augments *k*-means with outlier

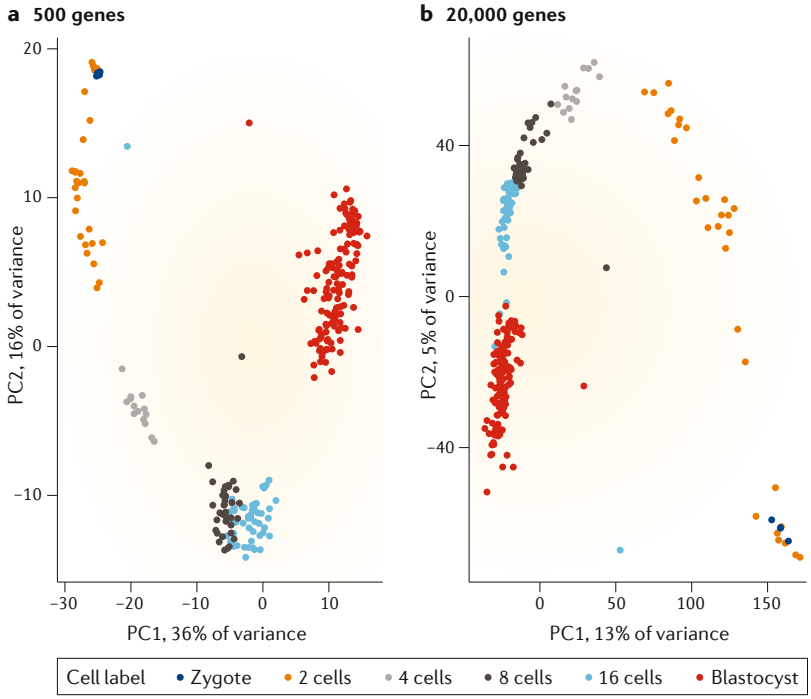

Fig. 2 | **Illustration of the curse of dimensionality.** Six separate populations of cells should ideally be distinguishable. Principal component (PC) analysis (PCA) plots of the Deng data set[42] using 500 (part **a**) and 20,000 (part **b**) of the most variable genes. When using a large number of features, clusters are less distinct, as indicated by the shorter distances between clusters (for example, the 4-cell stage is not as isolated). Consequently, unsupervised clustering becomes more challenging.

detection to identify rare cell types, whereas SIMLR[24] adapts *k*-means by simultaneously training a custom distance measure.

Another widely used generic clustering algorithm frequently adapted for scRNA-seq is hierarchical clustering (FIG. 3c), which sequentially combines individual cells into larger clusters (agglomerative) or divides clusters into smaller groups (divisive). An important shortcoming is that both time and memory requirements scale at least quadratically with the number of data points, which means that it is prohibitively expensive to use hierarchical clustering for large data sets. CIDR[25] adapts hierarchical clustering for scRNA-seq by adding an implicit imputation of zeros into the distance calculation, which gives more stable estimates of cell–cell distances in low-depth samples. Many scRNA-seq tools expand upon the idea of hierarchical clustering by carrying out dimensionality reduction after each merge or split. This iterative strategy improves the ability to identify small clusters, and it is used by BackSPIN[26] and pcaReduce[27] and in a study by Tasic et al.[28] (TABLE 1).

Owing to the limitations of *k*-means and hierarchical clustering, particularly for large data sets, it has become increasingly popular to apply community-detection-based algorithms to scRNA-seq data. Community detection is a variant on the idea of clustering that is specifically applied to graphs. Instead of identifying groups of points that are close together, community detection identifies groups of nodes that are densely connected. To apply such methods to scRNA-seq data, it is necessary to construct a *k*-nearest-neighbours graph. The choice of how many nearest neighbours (denoted by *k)* to include when constructing the single-cell graph affects the number and size of the final clusters. To improve robustness to outliers, the graph is often reweighted based on the shared nearest neighbours of each pair of cells (FIG. 3d,e).

As some of the graph data sets available are extremely large, for example, those representing social networks or hyperlinks on the World Wide Web, several of the algorithms for community detection have been developed with an emphasis on speed and scalability[29]. In contrast to the methods based on hierarchical clustering that return the partitions at all levels, most graph-based methods return only a single solution, which allows for faster run times. An advantage is that most graph-based methods do not require the user to specify the number of clusters to identify, instead employing indirect resolution parameters. Only the Louvain algorithm has been widely applied to scRNA-seq data, despite many others being available[30], some of which have demonstrated better performance in benchmarking studies[31]. The combination of shared-nearest-neighbour graphs and Louvain community detection was first applied to scRNA-seq data in the PhenoGraph[32] method, and this approach has since been incorporated into Seurat[3] and scanpy[4].

There are several different user-friendly clustering methods available today (TABLE 1), and to help researchers determine which one is most suitable, recent studies have provided quantitative benchmarks[33–35]. Owing to their speed and scalability, the clustering methods that are part of the scanpy and Seurat packages are popular choices for large data sets. However, it has been shown that clustering based on the Louvain method does not perform as well for smaller data sets[36]. More generally, finding a clustering method that is best for all situations may be futile because it has been shown that it is impossible for a single algorithm to achieve the full range of desired properties[37]. In fact, formal analysis cautions against comparing algorithms on the basis of a narrow set of criteria because no method can perform well for all problems[38].

Discrete versus continuous cell grouping

One shortcoming of most clustering methods is that they will partition the data, regardless of whether or not there are any biologically meaningful groups present. Although some methods (for example, SC3, SINCERA and pcaReduce) can determine that only a single group is present, clustering methods often mistake random noise for true structure because of heuristic optimization. Thus, if there are no discrete groups of cells present in the data, then clustering is not an appropriate approach. An example of a situation when clusters may not be present is when considering differentiation trajectories[39]. Instead, cells can be placed on a continuum connecting two or more end states. When analysing such continuous processes, the commonly used approach is to forego clustering and instead order the cells along a one-dimensional manifold ('pseudotime')[39–41].

It is not always clear a priori whether clustering or pseudotime analysis is the most appropriate approach. For example, consider the study by Deng et al. of early

Graphs
Each graph consists of a set of nodes connected to each other with a set of edges. In single-cell RNA sequencing, nodes are cells, and edges are determined according to cell–cell pairwise distances.

Heuristic optimization
A method for solving a problem that is designed to sacrifice accuracy in favour of speed. These methods are often based on approximations and cannot be guaranteed to find the best solution.

mouse development, which included cells from the 1-cell, 2-cell, 4-cell, 8-cell and 16-cell states and from the blastoderm[42]. On the one hand, it is reasonable to apply clustering, expecting to find groups representing the different discrete stages of development. On the other hand, one would expect that a pseudotime ordering should be able to align the cells in accordance with their developmental stage. We used two popular methods for pseudotime ordering to analyse these data, and the results suggest that the inferred order is in good agreement with the cluster labels provided by the authors (FIG. 4). Similarly, it has been shown that unsupervised clustering algorithms provide good results for this data set[22].

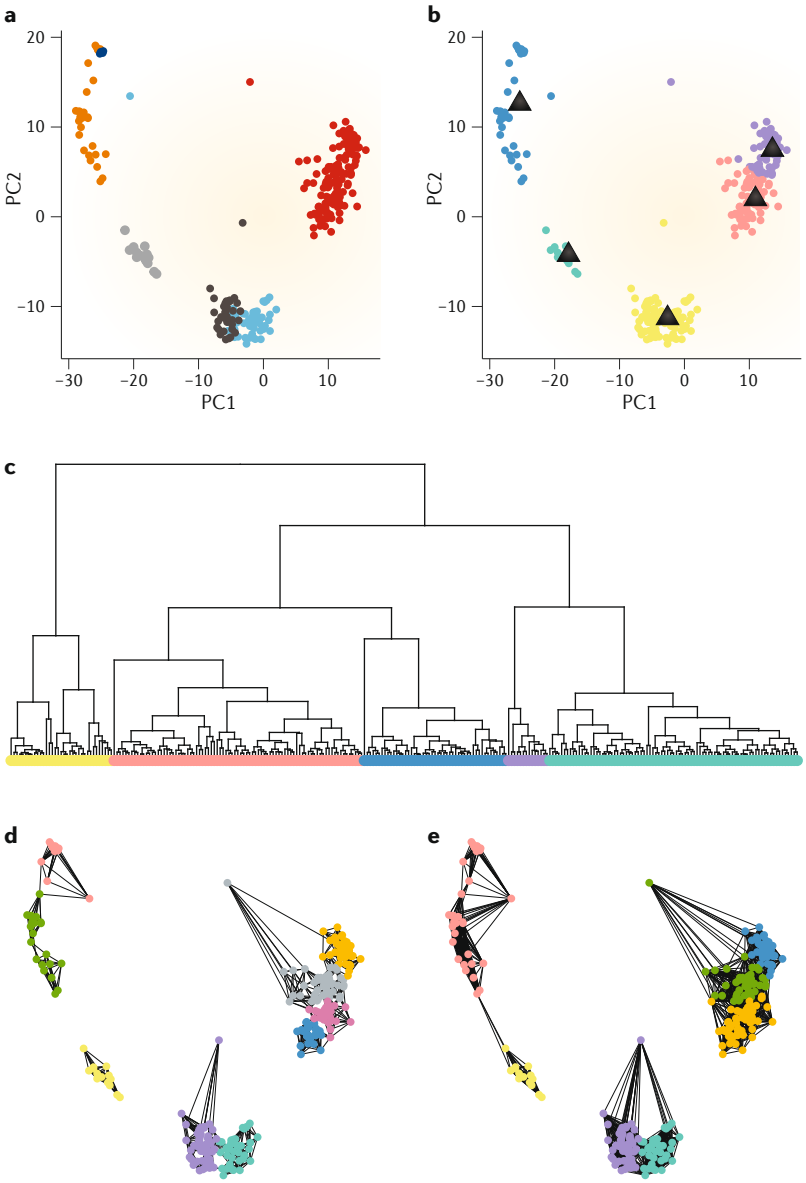

Fig. 3 | **Clustering methods for scRNA-seq.** Representation of different clustering approaches for single-cell RNA sequencing (scRNA-seq) using the Deng data set[42] of early mouse embryo development. **a** | True clusters, as defined by the authors, are based on the developmental stage (colours are the same as in FIG. 2). **b** | *k*-means separates cells into $k=5$ groups. Because *k*-means assumes equal-sized clusters, the larger group of blastocysts is split from the other cell groups before the 8-cell and 16-cell stages are separated from each other. **c** | Complete-linkage hierarchical clustering creates a hierarchy of cells that can be cut at different levels (the result for $k=5$ is indicated by the coloured bars at the bottom). Cutting farther down the tree would reveal finer substructures within the clusters. **d**,**e** | Louvain community detection[29] is applied to a shared-nearest-neighbour graph connecting the cells and finds tightly connected communities in the graph (number of nearest neighbours used to construct the graph is five for part **d** and ten for part **e**). Increasing the number of neighbours when constructing the cell–cell graph indirectly decreases the resolution of graph-based clustering. Each clustering algorithm was implemented in R (igraph for parts **d** and **e**) and applied to the first two principal components (PCs) of the data.

Some authors have developed strategies that bridge the pseudotime and the clustering approaches. Tasic et al. left out a subset of cells as the data were repeatedly clustered[28]. This bootstrapping strategy allowed them to categorize cells as stably assigned to the same cluster versus the cells that were assigned to different clusters. On the basis of this characterization, these authors labelled cells as either stable or transient, with the latter assumed to represent cells transitioning between two cell types. This strategy is a version of soft or fuzzy clustering whereby cells are assigned to groups of different probabilities[43]. A novel method by Wolf et al.[44] provides a coarse-grained graph representation in which cells are assigned either to nodes (which represent stable groups) or to connecting edges. This method is an advance over methods such as Mpath[45] and TSCAN[41], which first find discrete clusters and then subsequently infer a graph structure connecting the clusters. Taken together, both the discrete view and the continuous view of the underlying structure can be informative, and it is advisable to explore both when faced with a situation where the choice is not obvious.

Technical challenges

Owing to the low initial amounts of RNA obtained from a single cell, scRNA-seq data generally exhibit higher levels of noise and more zero values (known as dropouts) than RNA-seq data from bulk cell populations. It is not uncommon to have >50% of the entries in a count matrix equal to zero[46]. There are three explanations for why dropouts are observed: first, the transcript was not present and the zero is thus an accurate representation of the state of the cell; second, the sequencing depth was low, and, although it was present, the transcript is not reported; and third, as part of the library preparation, the transcript was not captured or failed to amplify. Moreover, dropouts introduce computational challenges, as some methods are poorly equipped to deal with data that deviate greatly from a multivariate normal distribution. There are several statistical methods available for imputing zeros[25,47,48], but they all rely on pre-existing cell–cell or gene–gene correlations in the data to infer the appropriate imputed value.

Estimating technical noise in scRNA-seq data is challenging because each individual cell is a biological, not a technical, replicate. However, through the use of endogenous spike-in RNA[49], several noise models have been developed[50–54]. These can be used to estimate the robustness of clusters by adding simulated noise to data sets and reapplying the clustering workflow, as implemented in BEARscc[54].

One type of technical noise that may arise because of the experimental design is often referred to as a batch effect[55] (FIG. 5). Batch effects refer to changes in gene expression that are due to experimental factors, for example, the time of the experiment, the laboratory where it

Table 1 | **Clustering methods for scRNA-seq**

Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3 (REF.[22])	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
SIMLR[24]	2017	Data-driven dimensionality reduction + *k*-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR[25]	2017	PCA + hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust[75]	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce[27]	2016	PCA + *k*-means + hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al.[28]	2016	PCA + hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN[41]	2016	PCA + Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath[45]	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN[26]	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID[23], RaceID2 (REF.[115]), RaceID3	2015	*k*-Means	Detects rare cell types, provides estimation of *k*	Performs poorly when there are no rare cell types
SINCERA[5]	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Cliq[80]	2015	Graph-based	Provides estimation of *k*	High complexity, not scalable

DBSCAN, density-based spatial clustering of applications with noise; PCA, principal component analysis; scRNA-seq, single-cell RNA sequencing.

was carried out, the person carrying out the experiment or the lane used in the sequencing machine[56]. Several studies have suggested that batch effects can have a large impact on clustering[55,57–59]. The best strategy for avoiding batch effects is to have a balanced experimental design so that samples are split across experimental batches[60]. In such cases, it is fairly simple to regress out batch effects. However, in some cases, for example, when working with perishable samples, this strategy may not be feasible.

It is also important to pay close attention to how the samples are handled, as this can have a major impact. When acquiring postmortem samples, RNA may degrade non-uniformly, and it is known that sensitivity can vary between tissues[61,62]. Moreover, dissociation of sensitive tissues, such as neuronal cells, may activate expression of immediate early genes or other stress-response genes[63]. Adding inhibitors or preserving cells through freezing or chemical fixation may reduce the effects of handling; however, efforts to optimize such protocols for scRNA-seq are still ongoing.

Considering the high level of noise in scRNA-seq experiments, one must ask whether each cluster corresponds to a true biological effect or whether the cluster arose because of technical artefacts, for example, droplets containing two cells (doublets)[64]. Doublets arising from cells of two distinct cell types can be easily mistaken for rare transitional cells, as they will exhibit a phenotype that is intermediate between the two originating cell types. Some plate-based or microfluidic-chip-based protocols allow imaging of captured cells before lysis, which can facilitate the identification of doublets. Owing to the wide range of cell sizes and sequencing depths in scRNA-seq studies, it is computationally challenging to identify doublets[65]. Several tools have been developed whereby synthetic doublets are generated computationally for a given data set, and an algorithm is trained to identify them and is then applied to the original data[66–68]. As there are many other technical confounders, it is important to evaluate factors such as mitochondrial RNA, experimental batch, sequencing depth and the number of genes detected across clusters to ensure that none of these aspects drives the clustering. It has also been suggested that highly expressed genes, for example, ribosomal genes, may have an exaggerated effect on clustering[69].

Biological challenges

In addition to technical noise, transient biological states can mask the underlying cell identity. A well-documented example is the cell-cycle phase, which can confound cell-type identity in differentiating T cells[70]. Tools such as scLVM[70] or cyclone[71] can regress out cell-cycle effects and provide a corrected transcriptome. However, it is not always clear whether a specific signature should be considered a confounder. For example,

Bootstrapping
A statistical approach in which data sets are randomly sampled and reanalysed to assess the robustness of a result.

Gaussian mixture model
A statistical model of one or more normal distributions. When fitted to data, each normal distribution can be interpreted as a distinct cluster of points.

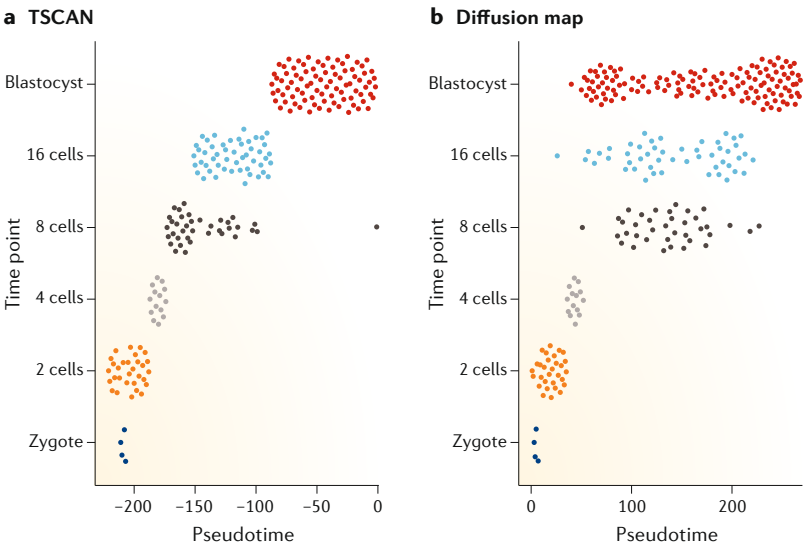

Fig. 4 | **Comparison of clustering and pseudotime methods.** Pseudotime cell trajectories were identified in the Deng et al. data set[42] by two trajectory methods: TSCAN[41] (part **a**) and diffusion maps (part **b**)[114]. Each dot represents a cell, and the colour corresponds to the annotation provided by Deng et al. The cells have been grouped along the *y*-axis according to the cell type and placed in temporal order. The *x*-axis corresponds to the inferred pseudotime, and ideally, the groups of cells should fall on the diagonal. Neither of the methods performs well across all times, but TSCAN performs better for the later time points (4-cell stage and beyond), whereas diffusion maps does well up to the 8-cell stage.

in cancer proliferation, signatures are biologically relevant and can correlate with cell-type identity rather than mask it[72]. Similarly, overall RNA content or cell size can confound clustering analyses but in many cases may reflect true differences in cell type[73]. Determining which biological signals should be considered confounders or valid cell-type differences is contingent on the particular question or system under consideration; thus, analysis pipelines must be customizable for different situations.

The heterogeneity of most tissues presents an additional challenge. One of the most well-studied systems is human blood, and it has been shown that cell-type frequencies span at least two orders of magnitude[74]. The recently published mouse cell atlas with ~300,000 cells profiled[14] shows a similar range. However, it is very likely that larger studies with deeper sequencing will reveal additional rare cell types, pushing the range of frequencies three or four orders of magnitude. As many methods work best when clusters are approximately equal in size, tools such as GiniClust[75] and RaceID[23] have been specifically tailored to identify rare cell types. Unfortunately, a better ability to distinguish rare cell types comes at the cost of poorer performance when clustering more frequent cell types. To deal with these situations, many authors have adopted a divide-and-conquer strategy, whereby large clusters identified after an initial clustering are subsequently reclustered[76,77]. This tactic is useful because biological samples frequently have multiple levels of functional specialization; for instance, neurons share specific functional characteristics that are distinct from those of various glial cell types but also contain distinct subtypes with more specialized functions, such as excitatory or inhibitory properties. However, determining when a large cluster should or should not be reclustered is difficult.

Computational challenges

Many scRNA-seq data sets are very large, with hundreds of thousands of cells, presenting both challenges and opportunities. A large data set ensures that analyses will have high power and improves the ability to detect rare cell types. Although it is possible to cluster such large data sets in a time span of hours[3,4], visualizing and interpreting the clustering results is difficult. Linear transformations, such as PCA, are unable to accurately capture relationships between cells because of the high levels of dropout and noise. Nonlinear techniques are more flexible, as they can provide outcomes that are often more aesthetically pleasing and easier to interpret by visual inspection. The most commonly used nonlinear dimensionality reductions are tSNE[78] and UMAP[79]. The main limitation of these methods is that they contain parameters that are required to be manually defined by the user and can strongly affect the visualization. As the guidelines for choosing the parameters are vague, the possibility of achieving a wide range of outcomes remains open.

Most clustering methods include one or more parameters that can be chosen by the user to determine the resolution of the clustering. The choice of parameter often has a large effect on the outcome. Selecting the resolution of clustering is often referred to as choosing *k*. For some methods, for example, *k*-means clustering, this choice is made explicitly by the user, but for other methods, the decision can be indirect, for example, choosing the number of nearest neighbours when constructing a graph. There are computational methods available to help guide the choice of *k*[22,23,80]. Many of these methods are based on the idea of calculating a cluster quality score and identifying an 'elbow,' that is, the point where the score plateaus. These scores tend to favour a fairly coarse resolution, with clearly separated clusters rather than closely related or overlapping cell types. As there is no consensus on the correct method for choosing *k*, judgement from the researcher is required. If there are reasons to believe that a sample is heterogeneous or if one is interested in uncovering new subtypes, then it is advisable to use a high *k* or a method that is tailored towards the discovery of rare cell types. Moreover, if the cells are sequenced to only a shallow depth, then it is less likely that a fine-grained clustering strategy will work.

Perhaps the most challenging aspect of scRNA-seq analysis (and this is not restricted to clustering) is how to validate a computational analysis method. The best strategy currently available is to have a set-up where the cell types are known through other means, for example, by selecting cells from distinct cell lines[81,82], using tissues that are very well studied and understood (for example, peripheral blood mononuclear cells[74]) or considering cells taken from the earliest stages of embryonic development[42,83]. These data sets serve as reliable ground truth, but one of the drawbacks is that they are unlikely to be as complex or challenging as some tissue samples.

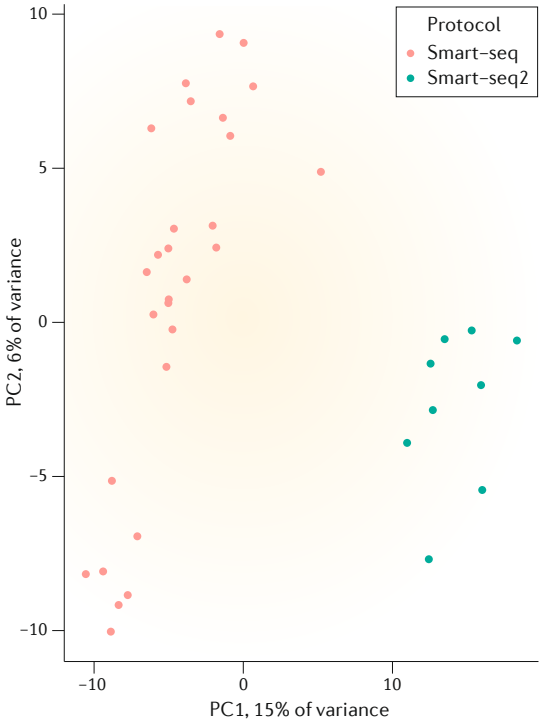

Fig. 5 | **Illustration of batch effects.** For the Deng data set[42], the authors report that, in the 8-cell stage, nine cells were sequenced using a different protocol, resulting in a batch effect between the two sets of cells, as seen by their distinct groupings in the principal components (PCs) plot.

Another drawback is that many of the suitable data sets are quite small, making it difficult to test methods at the kinds of scale that are relevant for current experiments. Another very useful strategy is to use spatial methods, for example, seqFISH[84], RNAscope[85] and merFISH[86]. As these methods do not rely on sequencing, they are orthogonal and a positive result should be considered strong validation. However, the limited number of mRNAs that can be profiled with these technologies and the costs and challenges involved in setting up the assays mean that their use is currently limited.

Biological interpretation and annotation

Although clustering methods partition cells according to transcriptional similarity, they leave it to the user to provide the biological interpretation. Analysing and understanding each cluster comprise an often time-consuming process that involves manually searching the literature and various databases. It is frequently assumed that each cluster will correspond to one or more cell types. However, there are no fixed criteria or rules for designating a cluster as a specific cell type, and there is no centralized database of known cell types and their characteristics. Instead, for most fields, there is an implicit understanding among researchers regarding the nature of the most important cell types and what genes they express. For many biological systems, relying on this type of 'folklore' for cell-type annotation appears to work well in practice. As an example, in October 2016, scRNA-seq studies of the adult human pancreas were published by five separate groups[87–91]. Although the work was carried out independently, the choices of cell-type labels were very consistent[92]. Similarly, a recent study[93] has suggested that neuronal clusters identified in different studies correspond well. However, the good correspondence may reflect a bias due to the existing literature.

The most relevant information that can be extracted from a cluster is the set of RNAs that are present or absent. The genes that are highly expressed and make it possible to distinguish one cluster from the others are often referred to as marker genes. A popular approach is to use gene ontology analysis to identify the terms most enriched for the marker genes of each cluster[94]. The gene ontology terms may give an indication of what biological process is most relevant for the cells. Alternatively, these genes can be compared with those referenced in existing literature or used for validation experiments.

As more and more data become available, one strategy will be to compare newly identified clusters against previously annotated data sets (FIG. 6). However, several studies have demonstrated that batch effects can be substantial, resulting in cells from the same tissue clustering by experimental origin rather than by biological similarity[58,59] (FIG. 5). Thus, comparing samples collected by different research groups or using different protocols remains challenging. Currently, there are two main strategies available for joint analysis of data sets: merging and projection.

If two samples are taken from a similar biological origin (for example, the same healthy tissue), then it is useful to merge the data from both samples before clustering. However, to make this possible, experimental batch effects must first be taken into account. The methods that have been developed for this task[58,59] try to distinguish components of the variability that is common between two data sets from the variability that is unique to one data set. It is assumed that the variability found across data sets originates from biological processes, for example, differences between cell types. By contrast, the unique components are assumed to be due to experimental artefacts and should be removed before merging two data sets.

Instead of computationally merging data sets directly, one can instead project the cells from one data set to another[92,95,96]. The projection strategy is favourable when one of the data sets is very large and reanalysis would be costly. Projection corresponds to a nearest-neighbour-finding problem where the goal is to find the best match to a query cell among a set of cells that have been previously clustered and annotated. The main limitation is that cells derived from a novel cell type that is not present in the reference may be incorrectly projected or simply fail to project. Thus, the generation of comprehensive atlases of cell types will greatly facilitate this strategy (FIG. 6).

Another helpful type of resource that aims to reuse already collected data is cell ontology databases[97]. The main challenge in the creation of these databases is to describe cell states or transitions on the basis of gene expression in single cells in a meaningful and comprehensive manner. Integration of cell ontologies and scRNA-seq atlases will enable the systematic

Cell ontology
A hierarchical organization of controlled vocabulary to describe properties of (and relationships between) different cell types.

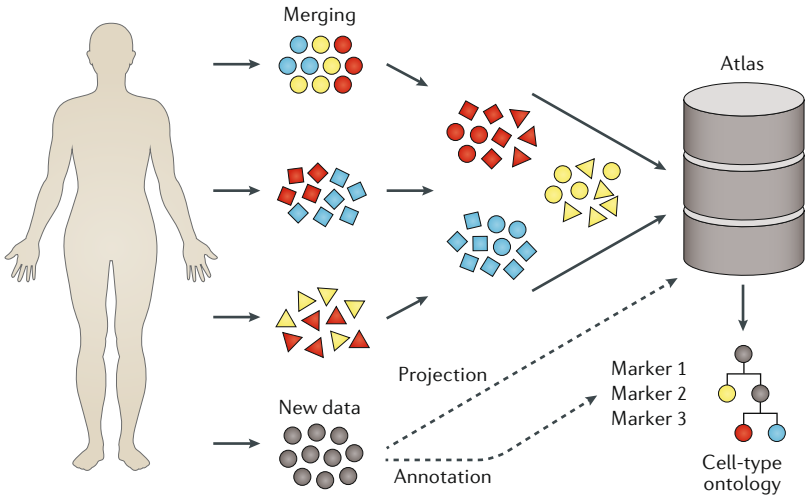

Fig. 6 | **Schematic overview of clustering and annotation in the context of a cell atlas project.** Many data sets (each data set represented by a different shape) are acquired from multiple tissues, and these are clustered and merged to identify consistent cell types across multiple data sets (shared cell types shown in the same colour) and stored in a database. Manual and computational curation of the atlas is used to inform the construction of a cell-type ontology and to identify robust markers for each cell type. New samples (grey circles) can be annotated with the standardized nomenclature defined in the ontology either using the cell-type markers or by projecting directly onto the atlas to identify the cells that are most similar.

identification of marker genes for annotating other data sets[98]. Similar to gene ontologies, cell ontologies are hierarchical, and thus they are able to describe relationships between cell types and at multiple resolutions. Comparison to the ontology will make it easier to put novel cell types into context and relate them to existing knowledge[99] (FIG. 6). Cell ontologies will be helpful to ensure that annotations are consistent. Considering the scale and complexity of some of the data sets being collected today, this attainment of consistency is a formidable challenge. For example, several collections of cells, each containing hundreds of thousands of cells from the mouse brain, have recently been released[12–14,100]. Each of these data sets has been clustered into hundreds of cell types, but it remains unclear how well they match. There are also computational methods available that can consider other types of information to aid interpretation. SCENIC[101] uses putative regulatory binding sites found in promoter regions to identify shared regulatory networks. In addition to providing additional information for the clustering, knowing which transcription factors are most important for a particular cell-type identity can facilitate the biological interpretation. PAGODA[51] identifies overdispersed gene sets, which can be related to functional modules.

When does a cluster represent a new cell type?

A central aim of scRNA-seq analysis is often to define cell type using unsupervised clustering based on the whole transcriptome[102]. However, for a new cell type to be accepted, it is necessary to go beyond characterization of the transcriptome. Researchers must demonstrate that the newly identified cluster is also functionally distinct. There are no universally applicable rules that can be applied here, and which assay is appropriate depends on the biological context.

To date, there are already several studies available demonstrating that this approach can be successful[103–105]. One of the most striking findings was made by Villani et al.[76], who discovered several new cell subpopulations in human blood. Although the study considered only ~2,400 cells, a relatively modest number by today's standards, sorting based on several markers and deep sequencing using full-length transcripts provided a high-quality data set. The novel clusters were shown to be distinct according to several properties, including morphology, stimulation by pathogens and ability to activate T cells.

Although the principles of defining new cell types are clear, there are many practical challenges, both experimental and computational, as we have shown in this Review. Thus, there are many issues that must be resolved to reach a consensus on how to best define cell types on the basis of the transcriptome profile. Some debates will be technical, focusing on questions about how to best choose the number of clusters or what quality of antibody is required for a validation experiment. By contrast, some of the questions will be more philosophical, for example, what assays are relevant for the given context, what magnitude of qualitative and quantitative differences is required, and whether a transient difference merits designation as a new cell type or should be considered a change in cell state.

We anticipate that the large number of data sets collected through cell atlas projects will enable a more stringent definition of cell types than what we have today. For example, with samples from across the whole body of an organism, it should be possible to determine how many marker genes are required to uniquely identify a specific cell type. Furthermore, it will be possible to ensure that definitions are consistent across tissues and species.

Outlook

Unsupervised clustering is likely to remain a central component of scRNA-seq analysis. As much of the downstream analysis is carried out based on the clusters, the final conclusions may be strongly affected by the clustering. It is likely that several different algorithms will be in use for clustering in the foreseeable future. To some extent, this diversity will reflect the fact that some methods will perform better for certain types of data, for example, sparse sequencing data from droplet microfluidics approaches versus deeper sequencing data from Smart-seq2 protocols. However, owing to the complex nature of the clustering problem, it is unlikely that one method will be deemed superior to all others.

The specifics of the clustering challenge will evolve as new technologies are introduced. In addition to having to cope with larger and larger data sets, there will be new modalities to consider. One interesting line of research is into so-called multi-omics methods, that is, assays that measure more than one aspect of the cell, such as the DNA methylome, open chromatin or proteome[106–108]. The additional layers of omics data will provide information about the phenotype that is not manifested by

the transcriptome, for example, by identifying active enhancers that help to influence the ability of the cell to respond to external stimuli[109]. Another important technological development is spatial methods[110–113]. Although current approaches are limited in terms of spatial resolution or the number of transcripts that can be profiled, they provide important information that is inaccessible by spatially naive scRNA-seq methods. Incorporating spatial information will be important for clustering. For example, some groups of cells that are difficult to distinguish on the basis of their transcriptomes may occupy different positions in the tissue or be surrounded by distinct neighbours.

In addition to developing methods for carrying out the clustering, there is also a need for methods that will facilitate biological interpretation and annotation. Hopefully, collaborations through cell ontology and cell atlas projects will ensure greater consistency with regard to both clustering and analyses. As discussed here, there is a need for the community to agree on suitable criteria about what constitutes a cell type based on the transcriptome, what assays are required for functional validation, how to select marker genes and what nomenclature to use when assigning names.

Published online: 07 January 2019

1. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
2. 10x Genomics. 10X Genomics single cell gene expression datasets. *10xgenomics* https://support.10xgenomics.com/single-cell-gene-expression/datasets (2017).
3. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
5. 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).
6. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
7. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. [version 2; referees: 3 approved, 2 approved with reservations]. *F1000Res* **5**, 2122 (2016).
8. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
9. Satija, R. SEURAT - R toolkit for single cell genomics: single cell integration in Seurat v3.0. *satijalab.org* https://satijalab.org/seurat/ (2015).
References 4 and 10 are unsupervised clustering methods based on the Louvain method that have been shown to perform very well for large scRNA-seq data sets.
10. Kiselev, V. et al. Analysis of single cell RNA-seq data course. *hemberg-lab.github* https://hemberg-lab.github.io/scRNA.seq.course/ (2018).
11. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31**, 651–666 (2010).
12. Quake, S. R., Wyss-Coray, T., Darmanis, S. & The Tabula Muris Consortium. Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a Tabula Muris. Preprint at *bioRxiv* https://doi.org/10.1101/237446 (2017).
13. Zeisel, A. et al. Molecular architecture of the mouse nervous system. Preprint at *bioRxiv* https://doi.org/10.1101/294918 (2018).
14. Han, X. et al. Mapping the mouse cell atlas by Microwell-Seq. *Cell* **172**, 1091–1107 (2018).
References 13–15 are large collections of scRNA-seq data from mouse, and they give an indication of what a full atlas could look like.
15. Reid, A. J. et al. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *eLife* **7**, e33105 (2018).
16. Davie, K. et al. A single-cell transcriptome atlas of the aging *Drosophila* brain. *Cell* **174**, 982–998 (2018).
17. Cusanovich, D. A. et al. The *cis*-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
18. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
19. Bellman, R. *Dynamic Programming* (Courier Corporation, 2013).
20. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
21. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28**, 129–137 (1982).
22. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
SC3 is a user-friendly clustering method that works very well for smaller data sets.
23. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
24. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14**, 414–416 (2017).
25. Lin, P., Troup, M. & Ho, J. W. K. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18**, 59 (2017).
26. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
27. Žurauskienė, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* **17**, 140 (2016).
28. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
29. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, 10008 (2008).
30. Xie, J., Kelley, S. & Szymanski, B. K. Overlapping community detection in networks. *ACM Comput. Surv.* **45**, 1–35 (2013).
31. Lancichinetti, A. & Fortunato, S. Community detection algorithms: a comparative analysis. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **80**, 056117 (2009).
32. Levine, J. H. et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
33. Mereu, E. et al. matchSCore: matching single-cell phenotypes across tools and experiments. Preprint at *bioRxiv* https://doi.org/10.1101/314831 (2018).
34. Freytag, S., Lonnstedt, I., Ng, M. & Bahlo, M. Cluster headache: comparing clustering tools for 10X single cell sequencing data. Preprint at *bioRxiv* https://doi.org/10.1101/203752 (2017).
35. Menon, V. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief. Funct. Genom.* **17**, 240–245 (2018).
36. Fortunato, S. & Barthélemy, M. Resolution limit in community detection. *Proc. Natl Acad. Sci. USA* **104**, 36–41 (2007).
37. Kleinberg & Jon. *An impossibility theorem for clustering* (2002).
38. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Computat.* **1**, 67–82 (1997).
39. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. Preprint at *bioRxiv* https://doi.org/10.1101/276907 (2018).
40. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
41. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).
42. 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).
43. Peters, G., Crespo, F., Lingras, P. & Weber, R. Soft clustering – fuzzy and rough approaches and their extensions and derivatives. *Int. J. Approx. Reason.* **54**, 307–322 (2013).
44. Wolf, F. A. et al. Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Preprint at *bioRxiv* https://doi.org/10.1101/208819 (2017).
45. Chen, J., Schlitzer, A., Chakarov, S., Ginhoux, F. & Poidinger, M. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat. Commun.* **7**, 11988 (2016).
46. Andrews, T. S. & Hemberg, M. Dropout-based feature selection for scRNASeq. Preprint at *bioRxiv* https://doi.org/10.1101/065094 (2018).
47. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
48. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
49. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
50. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
51. Fan, J. et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241–244 (2016).
52. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
53. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* **14**, 565–571 (2017).
54. Severson, D. T., Owen, R. P., White, M. J., Lu, X. & Schuster-Böckler, B. BEARscc determines robustness of single-cell clusters using simulated technical replicates. *Nat. Commun.* **9**, 1187 (2018).
55. Buttner, M., Miao, Z., Wolf, A., Teichmann, S. A. & Theis, F. J. Assessment of batch-correction methods for scRNA-seq data with a new test metric. Preprint at *bioRxiv* https://doi.org/10.1101/200345 (2017).
56. Gilad, Y. & Mizrahi-Man, O. A reanalysis of mouse ENCODE comparative gene expression data. [version 1; referees: 3 approved, 1 approved with reservations]. *F1000Res* **4**, 121 (2015).
57. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
58. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
59. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
References 59 and 60 present the first two methods for correcting batch effects to merge samples.
60. Baran-Gale, J., Chandra, T. & Kirschner, K. Experimental design for single-cell RNA sequencing. *Brief. Funct. Genom.* **17**, 233–239 (2018).

61. Gallego Romero, I., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 42 (2014).
62. Ferreira, P. G. et al. The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Commun.* **9**, 490 (2018).
63. Wu, Y. E., Pan, L., Zuo, Y., Li, X. & Hong, W. Detecting activated cell populations using single-cell RNA-seq. *Neuron* **96**, 313–329 (2017).
64. Petukhov, V. et al. dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.* **19**, 78 (2018).
65. Ilicic, T. et al. Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
66. DePasquale, E. A. K. et al. DoubletDecon: cell-state aware removal of single-cell RNA-seq doublets. Preprint at *bioRxiv* https://doi.org/10.1101/364810 (2018).
67. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. Preprint at *bioRxiv* https://doi.org/10.1101/357368 (2018).
68. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Preprint at *bioRxiv* https://doi.org/10.1101/352484 (2018).
69. Freytag, S., Tian, L., Lönnstedt, I., Ng, M. & Bahlo, M. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Res* **7**, 1297 (2018).
70. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
71. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
72. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
73. Cole, M. B. et al. Performance assessment and selection of normalization procedures for single-cell RNA-seq. Preprint at *bioRxiv* https://doi.org/10.1101/235382 (2017).
74. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
75. 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).
76. Villani, A.-C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017).
This study is a good example of how scRNA-seq was used to identify new cell types, which were subsequently confirmed by functional assays.
77. Campbell, J. N. et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci.* **20**, 484–496 (2017).
78. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Machine Learn. Res.* **9**, 2579–2605 (2008).
79. McInnes, L. & Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://arxiv.org/abs/1802.03426 (2018).
80. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).
81. Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).
This study shows that shallow sequencing can be sufficient to distinguish cell types.
82. Kolodziejczyk, A. A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* **17**, 471–485 (2015).
83. Fan, X. et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16**, 148 (2015).
84. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
85. Wang, F. et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagn.* **14**, 22–29 (2012).
86. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
87. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
88. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394 (2016).
89. Wang, Y. J. et al. Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* **65**, 3028–3038 (2016).
90. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).
91. Xin, Y. et al. RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.* **24**, 608–615 (2016).
92. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
93. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
References 93 and 94 present methods for comparing clusters across data sets without merging.
94. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
95. Sato, K., Tsuyuzaki, K., Shimizu, K. & Nikaido, I. CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA-sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/374462 (2018).
96. Srivastava, D., Iyer, A., Kumar, V. & Sengupta, D. CellAtlasSearch: a scalable search engine for single cells. *Nucleic Acids Res.* **46**, W141–W147 (2018).
97. Meehan, T. F. et al. Logical development of the cell ontology. *BMC Bioinformatics* **12**, 6 (2011).
98. Aevermann, B. D. et al. Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Hum. Mol. Genet.* **27**, R40–R47 (2018).
99. Bakken, T. et al. Cell type discovery and representation in the era of high-content single cell phenotyping. *BMC Bioinformatics* **18**, 559 (2017).
100. Saunders, A. et al. A single-cell atlas of cell types, states, and other transcriptional patterns from nine regions of the adult mouse brain. Preprint at *bioRxiv* https://doi.org/10.1101/299081 (2018).
101. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
102. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
103. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
104. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
105. Pal, B. et al. Construction of developmental lineage relationships in the mouse mammary gland by single-cell RNA profiling. *Nat. Commun.* **8**, 1627 (2017).
106. Hu, Y. et al. Single cell multi-omics technology: methodology and application. *Front. Cell Dev. Biol.* **6**, 28 (2018).
107. Bock, C., Farlik, M. & Sheffield, N. C. Multi-omics of single cells: strategies and applications. *Trends Biotechnol.* **34**, 605–608 (2016).
108. Macaulay, I. C., Ponting, C. P. & Voet, T. Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* **33**, 155–168 (2017).
109. Ostuni, R. et al. Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157–171 (2013).
110. Gao, S. et al. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018).
111. Edsgärd, D., Johnsson, P. & Sandberg, R. Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods* **15**, 339–342 (2018).
112. Moncada, R. et al. Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. Preprint at *bioRxiv* https://doi.org/10.1101/254375 (2018).
113. Pandey, S., Shekhar, K., Regev, A. & Schier, A. F. Comprehensive identification and spatial mapping of habenular neuronal types using single-cell RNA-seq. *Curr. Biol.* **28**, 1052–1065 (2018).
114. Angerer, P. et al. *destiny*: diffusion maps for large-scale single-cell data in R. *Bioinformatics* **32**, 1241–1243 (2016).
115. Grün, D. et al. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19**, 266–277 (2016).

Acknowledgements
The authors thank J. Elias for help with the figures. They also thank D. McCarthy for helpful discussions and J. Westoby for feedback on the manuscript.

Author contributions
All authors contributed to all aspects of the manuscript.

Competing interests
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information
Nature Reviews Genetics thanks A. Ziesel and the other, anonymous reviewer(s) for their contribution to the peer review of this work.

RELATED LINKS
BackSPIN: https://github.com/linnarsson-lab/BackSPIN
CIDR: https://github.com/VCCRI/CIDR
GiniClust: https://github.com/lanjiangboston/GiniClust
mpath: https://github.com/JinmiaoChenLab/Mpath
PhenoGraph: https://github.com/jacoblevine/PhenoGraph
RaceID: https://github.com/dgrun/RaceID
RaceID2: https://github.com/dgrun/StemID
RaceID3: https://github.com/dgrun/RaceID3_StemID2
SC3: http://bioconductor.org/packages/release/bioc/html/SC3.html
scanpy: https://github.com/theislab/scanpy
Seurat (latest): https://satijalab.org/seurat/
SIMLR: https://bioconductor.org/packages/release/bioc/html/SIMLR.html
SINCERA: https://github.com/xu-lab/SINCERA
SNN-Cliq: http://bioinfo.uncc.edu/SNNCliq/
TSCAN: https://bioconductor.org/packages/release/bioc/html/TSCAN.html