nature protocols

REVIEW ARTICLE

https://doi.org/10.1038/s41596-021-00534-0

Check for updates

# Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods

Zoe A. Clarke[1,2,8], Tallulah S. Andrews[2,3,4,8], Jawairia Atif[3,4,8], Delaram Pouyabahar[1,2,8], Brendan T. Innes[1,2], Sonya A. MacParland[3,4,5]✉ and Gary D. Bader[1,2,6,7]✉

**Single-cell transcriptomics can profile thousands of cells in a single experiment and identify novel cell types, states and dynamics in a wide variety of tissues and organisms. Standard experimental protocols and analysis workflows have been developed to create single-cell transcriptomic maps from tissues. This tutorial focuses on how to interpret these data to identify cell types, states and other biologically relevant patterns with the objective of creating an annotated map of cells. We recommend a three-step workflow including automatic cell annotation (wherever possible), manual cell annotation and verification. Frequently encountered challenges are discussed, as well as strategies to address them. Guiding principles and specific recommendations for software tools and resources that can be used for each step are covered, and an R notebook is included to help run the recommended workflow. Basic familiarity with computer software is assumed, and basic knowledge of programming (e.g., in the R language) is recommended.**

Single-cell genomics enables the molecular profiling of thousands of cells in a single experiment[1–3] to create comprehensive maps of cellular heterogeneity in multicellular systems[4,5]. In particular, single-cell RNA sequencing (scRNA-seq) and single-nuclei RNA sequencing can be used to measure single-cell transcriptomes and map novel cell types[6], states[7] and dynamics[8] in a wide variety of tissues and organisms.

Single-cell transcriptomics data are often presented as a two-dimensional 'map' organizing cells on the basis of the similarity of their gene expression profiles. Data visualized in this way naturally identify groups (or 'clusters') of highly similar cells, as well as gradients and other transcript-based patterns. Such artifacts must be interpreted and annotated to define cell types and states to support biological discovery (Fig. 1). Standard experimental protocols and analysis workflows detail how to create single-cell transcriptomic maps from tissues[9–12]. Briefly, tissues are dissociated into single cells and profiled using a single-cell transcriptomic technology. Computational analysis is then used to perform quality control filtering on the results (e.g., removing low-quality cells), quantify the expression of each mapped gene in each cell[13], identify groups of similar cells using a clustering algorithm[14–18] and visualize all cells in two dimensions using techniques such as t-distributed stochastic neighbor embedding (t-SNE)[19] or uniform manifold approximation and projection (UMAP)[20] to produce an unannotated 'single-cell map' image (Box 1)[21]. To interpret this map biologically, it is necessary to determine which cell types or cell states are represented by clusters or other patterns (e.g., gradients) observed in the data. These interpretations can then be labeled on the map, which helps place them in a conceptual framework useful for better understanding tissue biology. This tutorial offers a guide to the map interpretation and labeling process, starting from clustered data and resulting in a completely annotated single-cell map (Fig. 1). The general workflow for annotating cells in scRNA-seq data has three major steps: automatic annotation, manual annotation and verification (Fig. 2). First, automatic annotation uses a predefined set of 'marker genes' (i.e., genes that are specifically expressed in a known cell type) or reference single-cell data (i.e., an existing expertly annotated single-cell map) to identify and label individual cells or cell clusters by matching their gene expression patterns (signatures) to those of known cell types. A second major step is manual annotation, which involves studying genes and gene functions specific to each cell cluster or pattern to verify automatic cell annotations and identify novel cell types

[1]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [2]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [3]Ajmera Transplant Centre, Toronto General Hospital Research Institute, Toronto, Ontario, Canada. [4]Department of Immunology, University of Toronto, Toronto, Ontario, Canada. [5]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [7]Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada. [8]These authors contributed equally: Zoe A. Clarke, Tallulah S. Andrews, Jawairia Atif, Delaram Pouyabahar.
✉e-mail: s.macparland@utoronto.ca; gary.bader@utoronto.ca

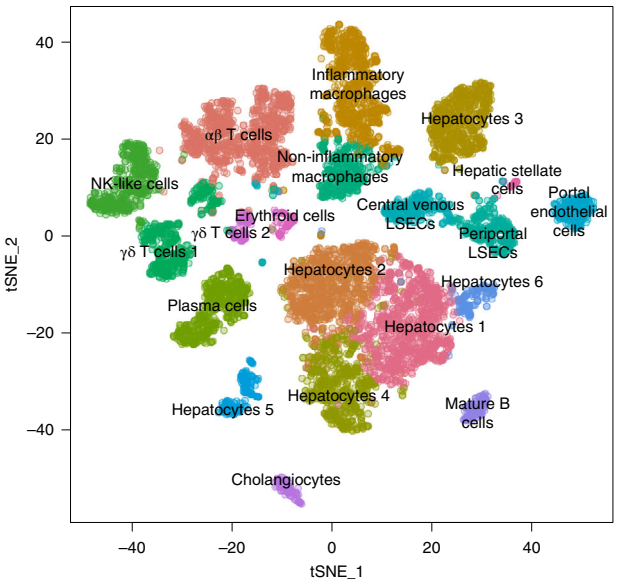



**Fig. 1 | An annotated single-cell transcriptomic map.** A completely annotated single-cell transcriptomic map of the human liver from data generated using scRNA-seq applied to five human liver samples (8,444 cells) reported in MacParland et al.[100] and visualized using a t-SNE plot. LSECS, liver sinusoidal endothelial cells; NK, natural killer.

and states. Finally, verification can confirm the identity and function of select cell types using independent methods, such as new validation experiments.

## Step 1: automatic cell annotation

Automatic cell annotation is an efficient way to label cells or cell clusters using a computer algorithm and an appropriate set of prior biological knowledge. The general principle is to identify a gene expression signal (pattern or signature) in a single cell or cell cluster that matches a characteristic gene expression signature of a known cell type or state; the cell or cluster is then assigned the respective label. Labels often have an associated confidence score.

There are two major automatic cell annotation approaches. One is to use known marker genes for each of the cell types that are likely to be found in the sample to be annotated (referred to as 'marker-based automatic annotation'). In this case, known relationships between marker genes and cell types are obtained from databases, such as SCSig[22], PanglaoDB[23] and CellMarker[24], or manually from the literature. Then, cells or clusters are labeled according to the marker genes they characteristically express. The second approach is to compare single-cell RNA-seq data to be annotated (the 'query' data set) to an existing, similar, expertly annotated scRNA-seq data set (the 'reference' data set) and transfer the label from a reference cell or cluster to a sufficiently similar one in the query (referred to as 'reference-based automatic annotation'). Reference single-cell data are obtained from sources such as Gene Expression Omnibus (GEO)[25], the Single Cell Expression Atlas[26] or cell atlas projects[27,28].

Automatic cell annotation methods can be applied to individual cells (either before or after clustering) or to clusters of cells, which occurs only after clustering the cells. In the case of annotating clusters, the gene expression profile for each cluster is determined by averaging the expression profiles of all cells within the cluster. Annotating individual cells is ideal, because this reduces the chance of missing important differences between cells. However, some scRNA-seq experimental data are based on low numbers of transcript reads per cell; thus, there may be insufficient data for cell-based annotation to function correctly, making clustered data sets easier to work with. Annotating clusters is faster, because there are fewer clusters than cells to process; it can also be more accurate than the single-cell approach, considering that it is based on more reliable expression level estimates averaged across all cells in a cluster. However, not all cells can be easily grouped into clusters, especially for dynamic systems like developing tissues[29] or tissues that contain gene expression gradients[30,31].

A major challenge with automatic cell annotation is that many cell types do not have well-characterized gene expression signatures, resulting in incomplete or inaccurate labeling for

### Box 1 | Visualizing single-cell data in a 2D projection

An scRNA-seq data set is typically visualized as a 2D scatter plot where cells (points) with similar transcriptomes are placed near each other. This 2D representation is projected from a higher dimensional space where each cell is described by the expression of thousands of genes, each of which is considered a separate dimension. The three most popular projection methods used for scRNA-seq data are t-SNE[19], UMAP[20] and PCA[110]. t-SNE (Fig. 6c) is a nonlinear projection that preserves local groups of similar cells, while equalizing the density of cells within each group[111]. The scale of a 'local group' is controlled by the 'perplexity' parameter, with higher values creating larger local groups[112]. t-SNE effectively visualizes distinct robust clusters, making it easy to observe discrete cell types; however, global relationships between cell types are not maintained, and thus cluster-to-cluster relationships cannot be inferred and may be misleading. Cell subtypes can be combined into one large cluster or split into distinct plot regions depending on the perplexity[111].

UMAP (Extended Data Fig. 1) is a nonlinear projection method that differentiates discrete cell clusters[20]. UMAP is typically regarded as better for visualizing global relationships and gradients than t-SNE, although these differences are probably due to default parameters[113]. UMAP is often less computationally intensive to run than t-SNE[114].

PCA (Fig. 6b) performs a linear transformation of normalized and scaled scRNA-seq data, to identify independent principal components (PCs) that capture major axes of variation in the data, which could represent biological factors, like cell types and states, or technical factors[115]. PCs are ranked in decreasing order of variance, and typically the first two PCs are used to visualize the data, but more can be considered to detect more subtle expression patterns between cells[110]. PCA can be useful for visualizing cell gradients and states.

Although these methods visually group similar cells and help visualize clusters, they do not define clusters and, therefore, are not clustering algorithms. Cell-clustering algorithm output is typically visualized as colors on the plot, and these colors may or may not correspond to patterns observed in the 2D plot.

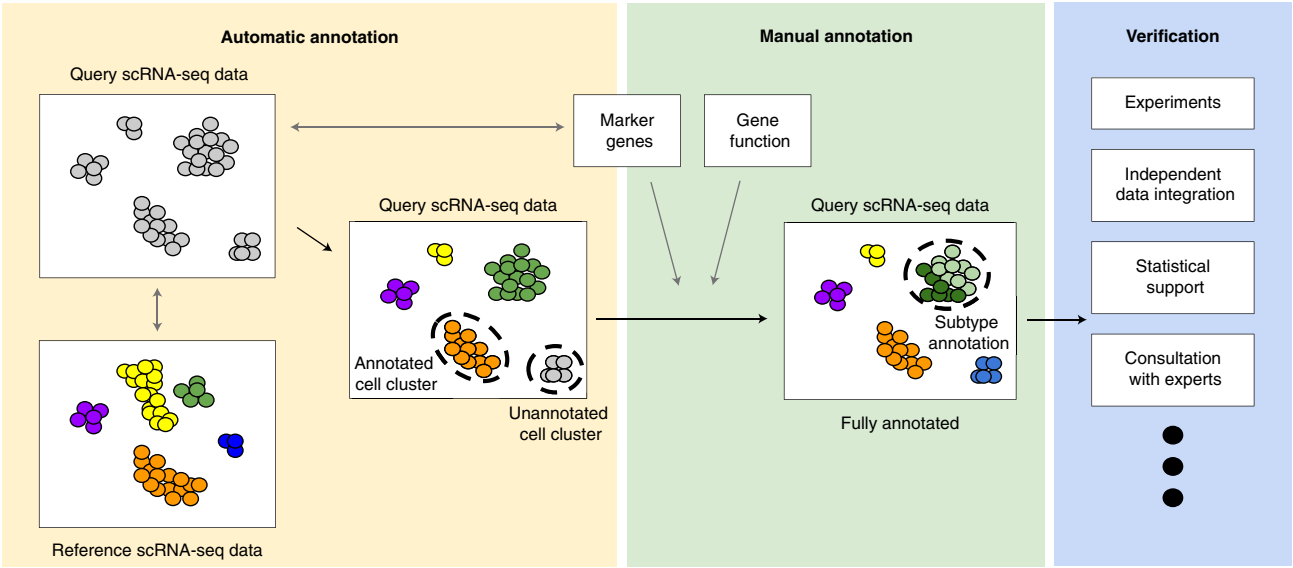



**Fig. 2 | Cell annotation workflow.** The recommended cell annotation process is composed of three major steps: automatic cell annotation, manual cell annotation and verification. The scRNA-seq data typically enter the workflow as a clustered gene-by-cell matrix, which is visualized using a dimensionality reduction method. An automatic cell annotation method is used to annotate cells either by comparison of the data with annotated reference data (e.g., a single-cell atlas) or using known marker genes indicative of a specific cell type. Manual annotation confirms or provides further detail for annotated cells or clusters or identifies the cell type of unlabeled clusters. Cell type can be manually inferred using a combination of marker genes, pathway analysis and differentially expressed genes with known functional information. Cell annotations are often verified using independent sources, such as new validation experiments, or comparison to complementary data, such as spatial transcriptomics data.

some cells. Automated methods typically work better for major cell types and may not be able to effectively distinguish subtypes. Thus, automatic cell annotation is useful to quickly identify known cell types and highlight unknown cell types for further exploration. The main caveats and recommendations for automatic cell annotation are summarized in Table 1.

### Marker-based automatic annotation

Marker-based automatic annotation labels cells or cell clusters on the basis of the characteristic expression of known marker genes. To be successful, the marker gene or gene set (a collection of marker genes) should be specifically and consistently expressed in a given cell, cluster or class of cells (e.g., immune cells). Markers are readily available for well-characterized organisms and cell types (e.g., human peripheral blood mononuclear cell (PBMC) samples[32]). Marker-based automatic annotation works well once a relevant and sufficiently large set of marker genes is collected[33].

To label individual cells, one of the most reliable marker-based annotation tools is semi-supervised category identification and assignment (SCINA)[34]. SCINA assumes that each marker follows a bimodal gene expression distribution, where one peak corresponds to cells from the associated cell type, and the other peak contains the rest of the cells in the experiment. A cell of a particular type is assumed to have expression in the upper part of this distribution for all the markers of that cell type, consequently requiring markers provided as input to SCINA to be specific to only one cell type. AUCell[35] is another good marker-based labeling method that classifies individual cells or clusters. AUCell ranks the genes in each cell by decreasing expression value, and cells are labeled according to their most active (highly expressed) marker gene sets. AUCell works best with cell types that have a sufficiently large set of marker genes such that multiple markers are detected in each cell. It has the advantage of scoring a whole set of marker genes at once, which may increase sensitivity over methods that examine each marker gene independently.

To label whole clusters, gene set variation analysis[36] (GSVA) has been benchmarked to be fast and reliable[37]. GSVA works similarly to AUCell: given a database of marker gene sets, it identifies sets that are enriched in the gene expression profile of a cluster. The GSVA software has the practical advantage that it can annotate all clusters in one operation.

Marker-based automatic cell annotation methods often have the advantage that they will assign labels only to cells associated with known markers, and other cells will remain unlabeled[33]. However, this depends on the specific tool and the parameters used; see Table 2 and Supplementary Table 1 for details on which tools have the option to leave cells unlabeled. A disadvantage of these tools is that markers are not easily accessible for all cell types.

### Reference-based automatic cell annotation

Reference-based cell annotation is based on the concept of 'guilt by association', whereby a cell or cluster label in the reference data is transferred to an unlabeled cell or cluster in the query data with a similar gene expression profile. Consequently, this approach is possible only if high-quality and relevant annotated reference single-cell data are available. Studying the original clustering and annotation steps performed on the reference data can help determine its quality and ensure that errors in the

**Table 1 | Comparison of the caveats and recommendations for different approaches to cell annotation**

| Stage of analysis | Aspect of analysis | Potential caveats | Recommendation |
|---|---|---|---|
| Automatic cell annotation | All automatic methods | Fast, but not effective for poorly characterized cells | Use manual annotation for poorly characterized cells |
| | Annotating clusters | May miss important differences between cells | Use automatic annotation of clusters to get a general idea of cell type and then refine labels manually. In addition, use multiple cluster-based methods and compare results |
| | Annotating individual cells | Ideal, but requires high reads per cell | Experiments with low reads per cell require cluster-based annotation |
| | Marker-based annotation methods | Marker genes not easily accessible for all cell types; may result in conflicting or absent cell labels | Requires expert knowledge to curate more extensive marker lists |
| | Reference-based annotation methods | Perform poorly with incomplete or poorly matched reference data, which may result in conflicting or absent cell labels | Use well-matched reference data or marker-based methods if such data are unavailable |
| | | Often requires batch correction, which may reduce the accuracy of results | Analyze the reference data for strong biological signals. Use a good experimental protocol that will prevail over batch effects |
| | | Mistakes in reference data get carried over to results | Analyze reference data for potential errors before using |
| | Comparing results from different automatic annotation methods | Results may not agree with each other | Compare confidence scores of respective labels and consider label agreement (majority rule); resolve conflicts using manual annotation |
| | | | Consider the possibility of cell subtypes, new cell types or gradients and cell states |
| Expert manual cell annotation | All manual methods | Slow, labor-intensive | Whenever possible, begin with automatic annotation to determine general cell labels |
| | | Subjective | Work with an expert; consider multiple cell-type conclusions |
| | Marker-based annotation | Cell types not distinguishable by a single marker | Use multiple markers for each cell type |
| | | Known markers not distinguishing cell types | Curate larger lists of markers from the literature, additional experiments or experts |
| | | Conflicting marker gene sets between sources | Select a marker gene set that best represents the biological signal being looked for in the data (e.g., if looking for cell subtypes, use more extensive gene sets than what is used for general cell-type annotation) |

reference will not be propagated to new data. Tissue-specific reference data can be obtained from public databases (e.g., the GEO[25] or the Expression Atlas[26]) or large cell atlas projects (e.g., the Human Cell Atlas[27], the Tabula Muris or Mouse Cell Atlas[5] or others[4,28,38–40]), although the required associated cell annotations are not always easily available. These atlases typically contain hundreds of thousands of cells and dozens of different annotated cell types.

scmap[41] is one of the best performing tools for reference-based automatic cell or cluster annotation in terms of both accuracy of assigned labels and avoiding incorrect labeling of novel cell types[33]. Other tools for reference-based automatic annotation include SingleCellNet[42] and SingleR[43]. SingleCellNet has high accuracy when all cell types are well represented in the reference data but has low accuracy if the reference data are incomplete or represent a poor match[33]. The main advantage of SingleR is that a reasonable, general reference data set is included with the tool, but this may not perform as well as a reference specifically matched to the query data set. An alternative to using specific software packages for reference-based cell annotation is to train a machine learning tool, such as a support vector machine[44] or random forest classifier[45], on selected reference data. This model can then be applied to classify cells or clusters as specific cell types in novel data. These methods can outperform any of the prepackaged automatic-annotation software tools[33] but require substantial computational expertise to use.

Another approach to reference-based cell annotation is to integrate a query data set with a reference data set using an integration algorithm, enabling clusters to be identified that span both data sets (Box 2). The reference labels can then be transferred to within-cluster query data cells. This approach supports the identification of novel cell types, distinct cell types and gradients in cell state but can be computationally expensive to run, and additional problems, such as overintegration, may be encountered.

**Table 2 | Summary of referenced annotation tools**

| Tool | Type | Language | Resolution | Approach | Allows 'None' | Notes |
|---|---|---|---|---|---|---|
| singleCell Net[42] | Reference based | R | Single cells | Relative-expression gene pairs + random forest | Yes, but rarely does so even when it should[33] | 10–100× slower than other methods; high accuracy |
| scmap-cluster[41] | Reference based | R | Single cells | Consistent correlations | Yes | Fastest method available; balances false-positives and false-negatives; includes web interface for use with a large pre-built reference or custom reference set |
| scmap-cell[41] | Reference based | R | Single cells | Approximate nearest neighbors | Yes | Assigns individual cells to nearest neighbor cells in reference; allows mapping of cell trajectories; fast and scalable |
| singleR[43] | Reference based | R | Single cells | Hierarchical clustering and Spearman correlations | No | Includes a large marker reference; does not scale to data sets of ≥10,000 cells; includes web interface with marker database |
| Scikit-learn[102] | Reference based | Python | Multiple possible | k-nearest neighbors, support vector machine, random forest, nearest mean classifier and linear discriminant analysis | (Optional) | Expertise required for correct design and appropriate training of classifier while avoiding overtraining |
| AUCell[103] | Marker based | R | Single cells | Area under the curve to estimate marker gene set enrichment | Yes | Because of low detection rates at the level of single cells, it requires many markers for every cell type |
| SCINA[34] | Marker based | R | Single cells | Expectation maximization, Gaussian mixture model | (Optional) | Simultaneously clusters and annotates cells; robust to the inclusion of incorrect marker genes |
| GSEA/GSVA[36,104] | Marker based | R/Java | Clusters of cells | Enrichment test | Yes | Marker gene lists must be reformatted in GMT format. Markers must all be differentially expressed in the same direction in the cluster |
| Harmony[105] | Integration (Box 2) | R | Single cells | Iterative clustering and adjustment | Yes | Integrates only lower-dimensional projection of the data; seamlessly integrated into Seurat pipeline; may overcorrect data |
| Seurat-canonical correlation analysis[106] | Integration (Box 2) | R | Single cells | MNN anchors + canonical correlation analysis | Yes | Accuracy depends on the accuracy of MNN anchors, which are automatically-identified corresponding cells across data sets |
| mnnCorrect[107] | Integration (Box 2) | R | Single cells | MNN pairs + singular value decomposition | Yes | Accuracy depends on the accuracy of MNN pairs (cells matched between data sets). Referred to in Box 2 |
| Linked inference of genomic experimental relationships (LIGER)[108] | Integration (Box 2) | R | Single cells | Non-negative matrix factorization | Yes | Allows interpretation of data set–specific and shared factors of variation. Referred to in Box 2 |

MNN, mutual nearest neighbor.

## Box 2 | Correcting confounding factors

ScRNA-seq data contain a mix of biological (e.g., cell types, states, age, sex and disease condition) and technical (e.g., batch effects) factors[116]. It is important to correct for undesired (i.e., 'confounding') factors while maintaining biological signals of interest. Confounding factors can either be regressed out of the data or adjusted for when integrating data from multiple samples (Fig. 7). Batch effects can be identified when cells from different batches form distinct stripes within groups or completely separate groups in a 2D visualization (Box 1). Harmony, mnnCorrect, Seurat v3 and LIGER are among the top-performing scRNA-seq integration or batch-correction tools[117].

Harmony[105] iteratively merges data sets represented by top PCs, which are then used to cluster cells. Each cell is iteratively adjusted on the basis of an estimated correction vector to shift it closer to the center of its cluster until convergence. MNN approaches, such as mnnCorrect/FastMNN[107] or Seurat v3[21], identify the most similar cells (MNNs), called 'anchors', across data sets that are used to estimate and correct the cell type–specific batch effects. LIGER[108] identifies shared (common biology) and unique (biological or technical) factors between data sets using non-negative matrix factorization. LIGER is recommended when specific cell types appear to be present in some of the data sets and missing in others[117].

Integration methods can suffer from overcorrection, where different cell types are merged, or undercorrection, when resulting clusters contain cells from only one input data set. Multiple integration methods may need to be evaluated to find a balance that best represents the data.

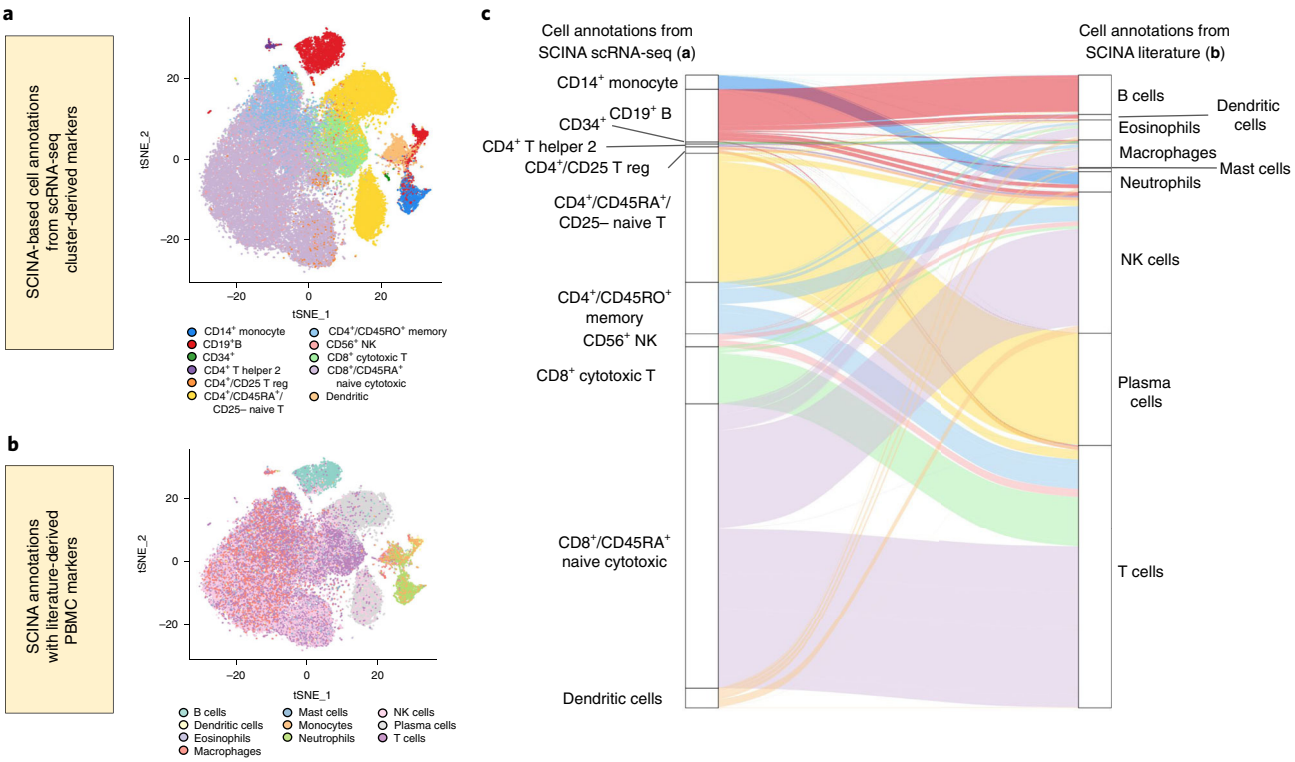



**Fig. 3 | Automatic annotation results depend on the marker genes used.** PBMCs from a 10x Genomics 3′ sequencing protocol[109] (68,579 cells) were automatically annotated with SCINA. The user provides lists of marker genes associated with all expected cell types, and SCINA assigns cell-type labels to individual cells on the basis of the expression levels of the marker genes. SCINA was provided with the top 20 marker genes from each previously annotated scRNA-seq cell cluster along with their associated cell type[109] (**a**) and literature-derived PBMC cell-type markers gathered by Diaz-Mejia et al.[37] (**b**). **c**, Sankey plot colored by the cell-type labels found in **a** compares the assigned cell-type labels between those from SCINA annotations **a** and **b**. Changes in label assignment demonstrate the variability of automatic annotation on the basis of the marker genes chosen by the user to represent each cell type or subtype. The PBMC data set is available from https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/fresh_68k_pbmc_donor_a. NK, natural killer.

### Refining automatic annotation

Benchmarking studies show variable performance of automatic annotation tools, depending on the data set and distinctiveness of the gene expression profiles of the cell types to be annotated[33,37]. For instance, distinguishing T cells from B cells is relatively straightforward, but automatic tools sometimes cannot accurately distinguish $CD8^+$ cytotoxic T cells from natural killer cells (Fig. 3). Thus, we recommend applying multiple, complementary annotation tools with multiple available marker gene databases to a single data set.

When applying multiple cell annotation methods to a data set, cells or clusters will acquire multiple, sometimes conflicting, cell-type labels. A set of annotations on a cell or cluster can easily be resolved to a single label if all labels are in agreement. If conflicts exist, most tools provide label confidence scores that can be used to identify a single high-scoring label. However, confidence scores are not standardized between tools, so they are often not comparable. Conflicts can also be resolved via a majority-rule approach, which selects the most frequent label (Fig. 4), or

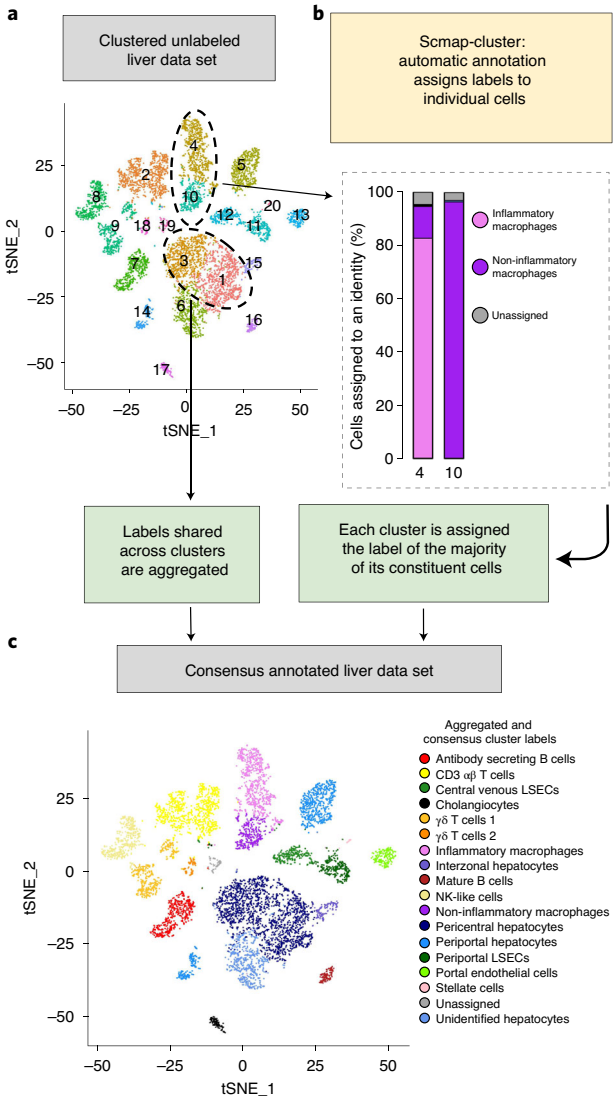



**Fig. 4 | Refining cluster labels from automatic annotation.** Raw scmap-cluster[41] annotations provide every cluster with a unique label on the basis of the identity assigned to most cells in each cluster. These labels can be refined by aggregating identical labels, and label confidence can be assessed by viewing the proportion of cell identities in each cluster ('majority rule'). **a** and **c**, t-SNE maps of the liver transcriptomic map constructed from the data set reported in MacParland et al. 2018[100] (8,444 cells). **a**, The cells are colored by the clusters identified by Seurat with the use of default parameters. Each cluster was assigned a cell-type label by scmap-cluster. Two clusters (labeled '4' and '10') are identified as different immune cells with labels of various levels of support, further described in **b**. **b**, scmap-cluster identifies cluster labels as well as individual cell labels. Cluster 4 has ~83% of cells identified as inflammatory macrophages. Cluster 10 consists almost entirely of non-inflammatory macrophages with some cells unassigned. **c**, Consensus cluster-level annotations are determined after aggregating identical cluster labels (e.g., clusters 1 and 3 from **a**) and using a majority rule (selecting a cluster label from the majority of cells assigned to a cell type)[100]. LSECs, liver sinusoidal endothelial cells; NK, natural killer.

percent agreement between methods. If no label can be confidently decided, the cells or cluster must be manually annotated.

Conflicting annotations within a cluster may reflect important information about that cluster, such as whether it contains cell subtypes. However, if subtypes cannot be clearly defined, a more general cell-type annotation may be more appropriate. For example, if a cluster is annotated as regulatory T cells, naive T cells and helper T cells by different methods, it may be most appropriate to assign the general label of 'T cells'. In this case, the original clustering parameters should be altered to better capture cell subtypes (see 'The impact of experimental and analysis parameters on annotation').

If the conflicting annotations are not subtypes of the same cell type, then the cluster may represent an intermediate cell state or gene-expression gradient. Because many automatic annotation tools assume discrete cell types, they often assign clusters or cells within a larger gradient to a well-defined endpoint. However, gradients often contain cells of various phenotypes; thus, multiple methods may assign the same cell to different ends of the gradient. Recommendations for handling gradients are discussed in Annotating cell states and gradients. Alternatively, a conflicting label on a cell could indicate that the cell is actually a doublet; a scenario in which two or more cells of different types are captured by the same cell-barcode. This case can be detected using doublet-finding methods[46–48].

Most automatic annotation tools are designed to annotate individual cells (Table 2 and Supplementary Table 1). Advantages of this approach are the ability to identify insufficiently resolved cell types and cellular gradients and the independence to choose clustering resolution, feature selection and dimensionality reduction parameters. Interestingly, the resulting annotations can be used to inform these analysis choices. For example, cell annotation could help optimize the clustering process to result in one cluster per cell type.

Finally, a cluster may have a novel cell identity that is absent from the reference data. This often results in widely varying results from automatic annotation methods or insufficient confidence for any tool to assign any label. In such situations, manual annotation must be performed.

## Step 2: expert manual cell annotation

Although automated cell annotation methods are convenient and systematic, they require an appropriate reference database and do not always result in high-confidence annotations. When these methods result in lower confidence, conflicting or absent cell labels, expert manual annotation is required. In manual cell annotation, cells are manually examined for clues to their function using various resources, following the same principles as marker-based automatic annotation. Manual annotation usually operates at the cluster level for convenience, but rare cells can be individually examined. Expert manual annotation is typically regarded as the gold-standard method for annotating cells; however, it is slow and labor intensive and can be subjective.

If automatic annotation has not been performed, marker-based annotation should first be manually applied. Usually, each known marker gene is individually visualized on the 2D projected data map (Box 1) to create a 'gene expression overlay'

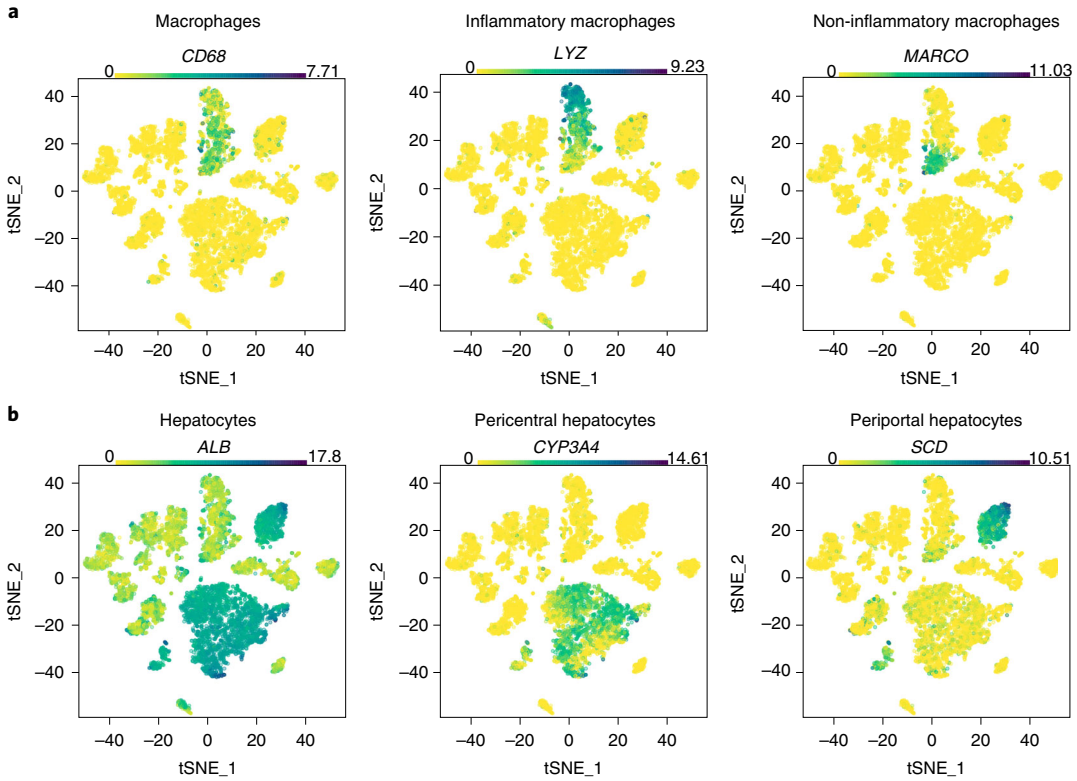



**Fig. 5 | Visualizing well-known markers.** Human liver marker genes are visualized as gene expression data overlaid on a t-SNE plot. The various markers represent: immune cells and immune cell subtypes that can be easily identified by well-known markers (**a**) and hepatocytes and inferred hepatocyte subtypes that are difficult to identify because the markers are not as well characterized and may vary across individual samples (**b**). Gene expression data are reported in MacParland et al.[100].

plot (Fig. 5). The entire list of markers may also be simultaneously visualized across clusters as a heat map (Extended Data Fig. 1) or dot plot (Fig. 6). A dot plot is more informative than a heat map, because it can communicate mean detected gene expression levels and the proportion of cells in a cluster in which each gene is detected, whereas a heat map typically describes only average gene expression levels per cluster. If many marker genes for a known cell type are highly expressed across cells in a cluster, this is often sufficient support for it to be labeled as that cell type. Easy-to-use software such as the free Loupe Cell Browser for 10x Genomics scRNA-seq data supports this visualization and analysis process. Challenges in this approach are that well-known markers are often too few in number to completely annotate an scRNA-seq data set, and some well-known markers may not be as specific within an scRNA-seq data set as expected. Additional markers often must be manually found via searching the literature and mining existing single-cell transcriptomic data for gene expression signatures related to the query data set. Master transcription factors that drive cell fate[49,50] often make better gene expression markers than cell-surface proteins that are commonly used to classify cell populations with methods such as like flow cytometry[5,39], presumably because mRNA and protein levels may not be strongly correlated[51]. Furthermore, there may not be any single distinguishing gene expression marker, in which case, multiple genes must be used together to distinguish a cell type from others in the data.

The ideal primary source for cell-defining genes is a single-cell atlas from a relevant organism, organ and disease context. In the absence of this, gene expression markers can be collected from bulk RNA-seq data from purified cell populations isolated from the same tissue source[52]. Given that protein expression may correlate with mRNA expression, protein expression markers can be gathered and used as potential gene expression markers[53] from published evidence of staining patterns within the tissue (i.e., using immunohistochemistry or immunofluorescence), flow cytometry and western blots. Integrating markers from independent sources can be challenging because of conflicts between lists. For instance, PanglaoDB[23] contains 220 markers for B cells, and CellMarker[24] contains 1,426 markers, yet only 66 are shared. If species-specific data are scarce, then data can be transferred by orthology from model organisms (Box 3) or other models (e.g., in vitro cell culture or organoids).

Ideally, each cluster will uniquely express the markers of one cell type. However, in some instances a cluster may not express markers of any known cell type; conversely, it may express markers of more than one cell type. Clusters that express markers of more than one cell type may represent doublets. Typically, such clusters will be very small compared to the clusters of true single cells, and they may express more genes than single cells. There are various doublet detection tools that can help determine if a cluster is composed of doublets[46–48]. If a cluster does not express markers

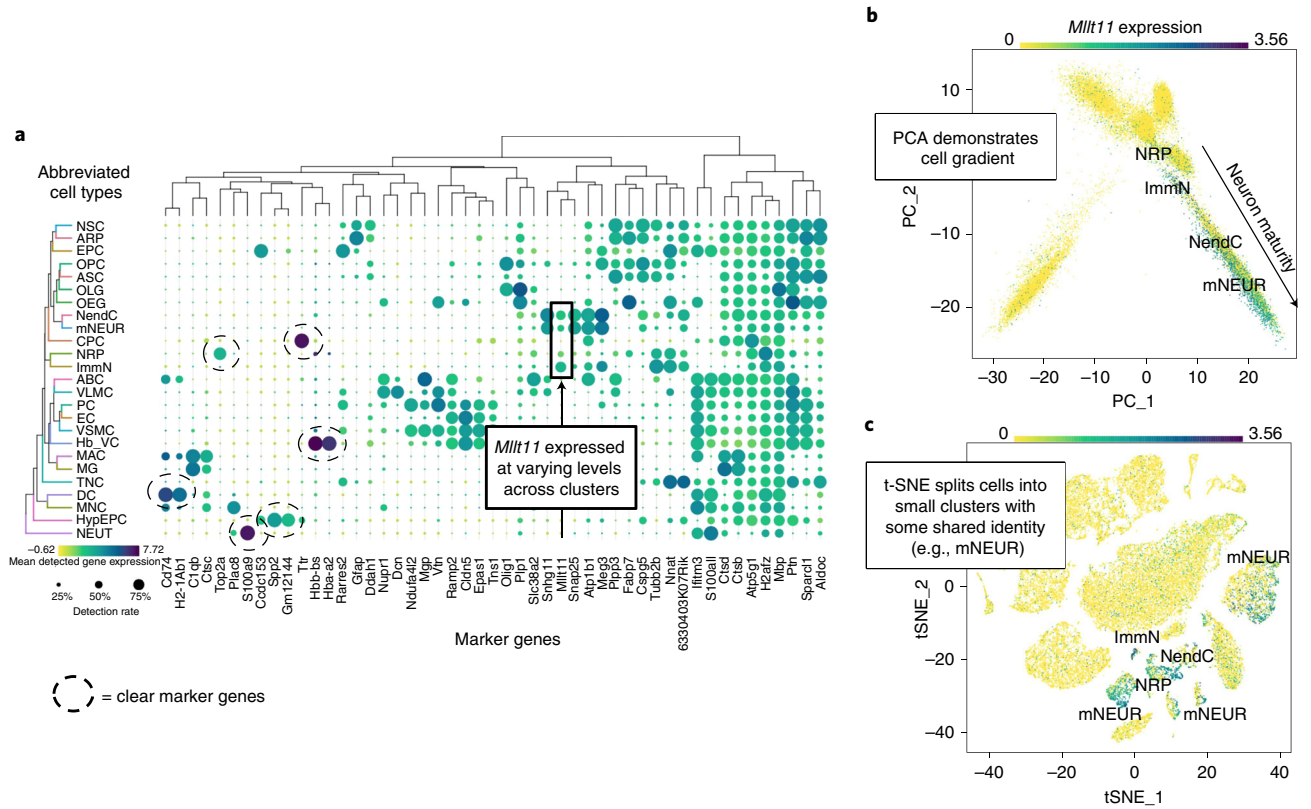



**Fig. 6 | How to identify and visualize a cell-type gradient.** Gradients across cell types can be identified by: finding marker genes that are expressed at varying levels across clusters (**a**), observing clear gradients of marker gene expression between cell types in principal component analysis (PCA) gene expression overlay plots (**b**) and identifying closely related cell types that are split across clusters in a t-SNE plot (**c**). **a**, Dot plots of marker genes distinguish clusters of scRNA-seq data from young and old mouse brains reported in Ximerakis et al.[101] (37,069 cells). **b** and **c**, These data can be visualized in two dimensions with gene expression of *Mllt* overlaid across cells; this occurs in a gradient across neuronal-restricted precursors (NRP), immature neurons (ImmN), neuroendocrine cells (NendC) and mature neurons (mNEUR). ABC, arachnoid barrier cells; ARP, astrocyte-restricted precursors; ASC, astrocytes; CPC, choroid plexus epithelial cells; DC, dendritic cells; EC, endothelial cells; EPC, ependymocytes; Hb_VC, hemoglobin-expressing vascular cells; HypEPC, hypendymal cells; MAC, macrophages; MG, microglia; MNC, monocytes; NEUT, neutrophils; NSC, neural stem cells; OEG, olfactory ensheathing glia; OLG, oligodendrocytes; PC, pericytes; TNC, tanycytes; VLMC, vascular and leptomeningeal cells; VSMC, vascular smooth muscle cells.

## Box 3 | Cell annotation across species

Sometimes the best reference single-cell map to use for cell annotation is from a different organism. To use such a reference for cell annotation, genes from the query organism must be mapped by orthology to genes from the reference, using databases such as Ensembl[118] or EggNOG[119] or tools such as OrthoFinder[120], before being input to data integration or marker-based annotation methods. Typically, one-to-one orthologs are used[100,121], which better ensures the conservation of function, although it is possible to use one-to-many and many-to-many relationships to increase gene coverage. The latter can be accomplished by grouping paralogs to create artificial gene groupings called 'meta-genes'[122]. If homologous genes are unavailable, genes from each species can be aggregated into pathways or 'biological process activities', which are compared across species to improve sensitivity of cross-species mapping[123]. If the query species genome has not been sequenced, RNA transcripts can be assembled de novo from the entire pool of RNA-seq reads from all cells, which are then used to quantify gene expression and identify orthologs. Evolutionarily close reference species should be chosen; otherwise, integration may not be able to map similar cell types for annotation transfer.

of any known cell type, it may contain poor-quality cells or represent a novel cell type.

Once cell-type information from known markers is exhausted, cells that have not been confidently annotated must be manually examined, cluster by cluster. Potential novel markers are identified by computing differential expression between a cluster and all other cells[54–56] (Fig. 6 and Extended Data Fig. 1). All marker genes are then manually researched to find functional information that may help identify the cell type of the cluster with which they are associated. Pathway enrichment analysis should also be applied to each cluster to identify cluster-specific pathways using standard workflows[57] and tools such as gene set variation analysis (GSVA)[36] or single-sample gene set enrichment analysis (ssGSEA)[58]. Pathway enrichment analysis simultaneously scores multiple functionally related genes for gene expression activity within a cluster at once and can be more sensitive than individual gene-based analysis.

**Fig. 7 | Batch correction.** Three PBMC samples were assayed with the 10x platform using different library construction protocols: 5′ (green, 7,726 cells), 3′ V1 (orange, 5,419 cells) and 3′ V2 (blue, 7,726 cells). **a** and **b**, UMAP diagrams showing clusters annotated by the SCINA using PBMC markers collected by Diaz-Mejia et al.[37] **c**–**f**, Bar graphs indicating the proportions of cells per cluster. The left column (**a**,**c** and **e**) shows the data merged without batch correction, and the right column (**b**,**d** and **f**) shows the data integrated using the Harmony batch correction method. Before Harmony, cells group by experimental protocol, clusters rarely contain cells from multiple experiments and multiple clusters of the same cell type exist (e.g., there is one B-cell cluster for each experimental protocol, circled in **e**). After Harmony, cells group by cell type, clusters contain cells from various protocols and fewer clusters share a cell identity. ▶

Some cells may be challenging to annotate, including novel cell types, which can be described on the basis of the function of genes they express. Furthermore, it can be particularly difficult to differentiate between tissue-resident cells (e.g., tissue-resident macrophages) and non-tissue-resident cells (e.g., monocytes circulating in the blood) of the same overall type. One approach to identify tissue-resident cells is to modify the experimental design to remove passenger cells from the tissue in question with a perfusion step. However, the number and types of cells removed by flushing will depend on the specific tissue and protocol. In situations where this flushing is not possible, one may profile PBMCs from the same individual and then subtract those cell signatures from the tissue map using the cell annotation methods mentioned above.

Ultimately, when annotating cell types, it is prudent to use standard nomenclature, such as from the Cell Ontology, which is a hierarchically organized controlled vocabulary of cell types and subtypes[59–61]. This enables maps to be more easily integrated across studies.

### Annotating cell states and gradients

When analyzing and characterizing novel cell types, it is important to determine whether they represent a stable cell type or contain multiple cell states. The definitions of cell type and state are not yet standardized, but a stable cell type may be expected to have homogeneous gene expression across a cluster and be compact in a 2D projection plot (Box 1), whereas cell gradients appear as a spread-out string of cells and cell states (e.g., cell cycle state) (Fig. 6). Expression gradients indicate continuous differences that are present in the cell population, which could represent states like the cell cycle[62], immune activation[63], spatial patterning[64] or transient developmental stages[65,66]. Care must be taken to distinguish biologically meaningful cell states and experimental batch effects, which can manifest in a similar way (Box 2, Fig. 7).

Annotating the intermediate stages of a gradient is often difficult, because these regions rarely express unique marker genes. It is often easier to label the ends of a gradient and then characterize intermediate stages using the order of specific genes that mark these ends as increasing or decreasing across the gradient. Extracting the cells in the gradient and performing principal component analysis (PCA) on them is often a useful visualization for gradients, because it preserves the large-scale distances between cells (Fig. 6). There are currently no automated gradient annotation methods; thus, gradients must be manually annotated, making use of known structure and cell-type transitions relevant to the particular experiment[7,59].

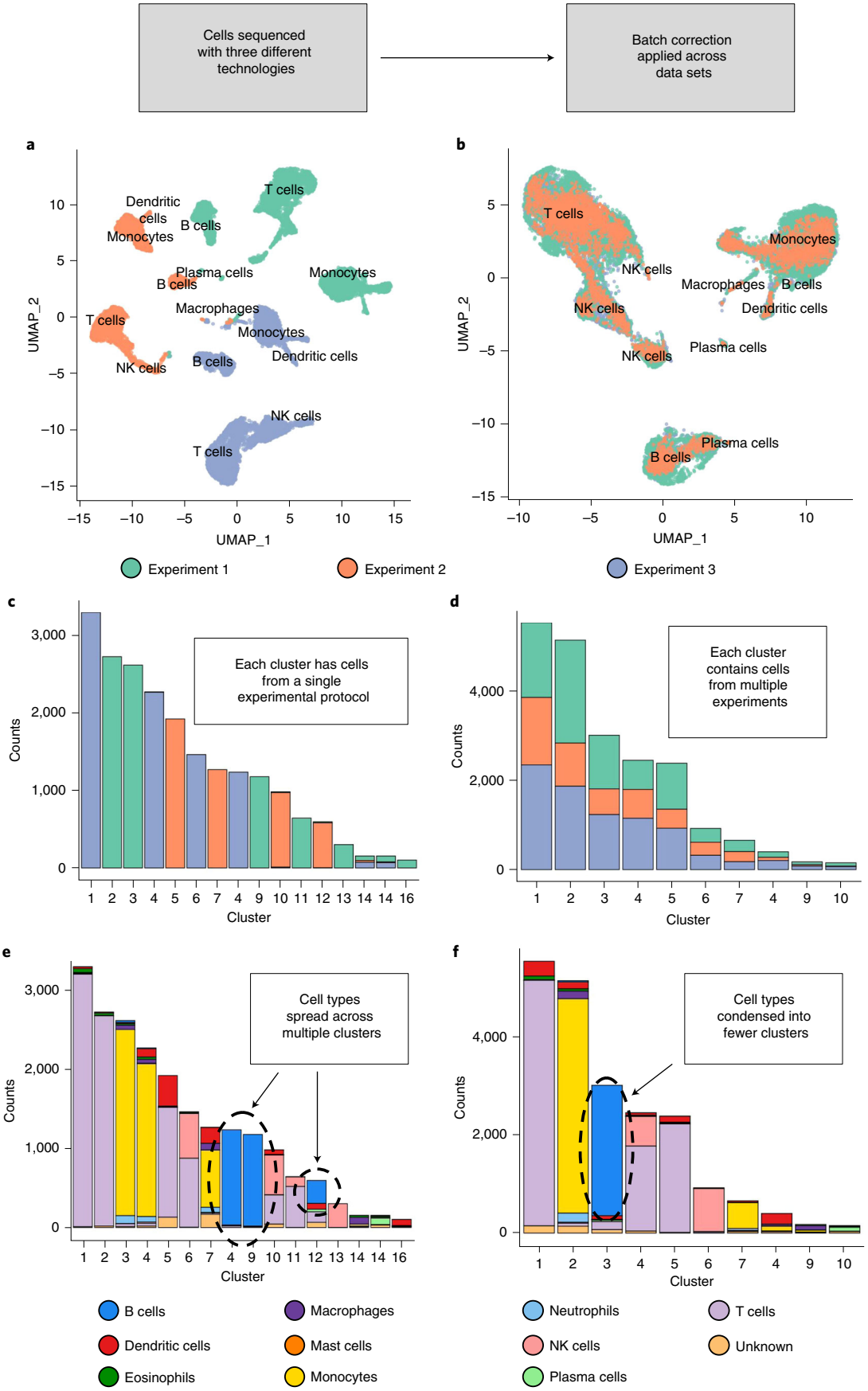
Similarly, homogeneous or similar cell states or cell types are often difficult to annotate because they share many of the same marker genes (Fig. 3). For instance, when annotating T cells within a tissue sample, it is common for all the T-cell subtypes to exhibit common T-cell markers; the subtype-specific markers are hidden within or below the general T-cell signal. In this case, it is often useful to subcluster the population, or to test each subpopulation specifically against the other related clusters to identify the subtype-specific markers. Very fine distinctions between highly similar cell types may not be visible transcriptionally and may be visible only in other genomic layers, such as chromatin state (assay for transposase-accessible chromatin using sequencing (ATAC-seq) and DNA methylation).

## Step 3: annotation verification

The above tools and approaches can provide confident cell-type labels for scRNA-seq data. However, because of the various challenges discussed above, it is important to confirm cell annotation labels using independent methods such as statistical methods[67] or by consulting an expert. Furthermore, because mRNA measurements only partially define cell type and function, important conclusions about novel cell types must be experimentally validated.

As an example, cell-type labels of tissue-resident immune cells can be refined using T-cell receptor (TCR)[68] and B-cell receptor[69] clonotyping, to examine the transcriptional signature of T and B cells as stratified by the TCRs and B-cell receptors, respectively, that they express. For instance, mucosal-associated invariant T cells express the marker genes *SLCA4A10* and *KLRB1*[70], which can be identified in an scRNA-seq experiment, as well as the known semi-invariant TCRs that are found in MAIT cells (*TRAV1-2/TRAJ12/20/3*), which can be revealed by TCR clonotyping. In addition, identifying the B-cell receptor repertoire within single-cell data sets enables annotation of naive versus mature B cells. Naive B cells express both IgM and IgD heavy chains, whereas mature B cells, which have undergone antibody class switching by V(D)J recombination, express IgG, IgA or IgE heavy chains. Other traditional methods to increase cell annotation confidence include in vitro functional assays, such as cytokine secretion, proliferative capacity and cytotoxic potential measures; imaging experiments, such as fluorescence in situ hybridization[71] of source tissue samples[72]; and single-cell qPCR to verify the co-expression of a novel combination of marker genes in a larger number of samples[73,74]. Complementary single-cell genomic methods are also useful, such as cellular indexing of transcriptomes and epitopes by sequencing[51], which simultaneously immunophenotypes cell surface proteins and measures scRNA-seq; single-cell ATAC-seq, which maps chromatin state; and spatial transcriptomics, which combines cell imaging and scRNA-seq to measure spatial transcript patterns and cell morphology in one experiment[75].

In the context of tumor biology, mutations are important to distinguish cancer and normal cell types. Genetic alterations such


Cells sequenced with three different technologies
Batch correction applied across data sets
a
b
UMAP_1
UMAP_2
Dendritic cells
Monocytes
B cells
T cells
Plasma cells
Macrophages
NK cells
Experiment 1
Experiment 2
Experiment 3
c
d
Counts
Cluster
Each cluster has cells from a single experimental protocol
Each cluster contains cells from multiple experiments
e
f
Cell types spread across multiple clusters
Cell types condensed into fewer clusters
B cells
Dendritic cells
Eosinophils
Macrophages
Mast cells
Monocytes
Neutrophils
NK cells
Plasma cells
T cells
Unknown

as single-nucleotide variants and copy-number variants (CNVs) can be detected in single-cell data using tools developed for bulk RNA-seq data[76], despite challenges with sequencing coverage. CNV inference methods identify consistently up or down expression values relative to a reference across large numbers of genomically contiguous genes to call amplification or deletion events, respectively. HoneyBADGER[77] and CaSpER[78] methods predict CNVs using other cells in the data as a reference, whereas InferCNV[79,80] uses a given set of normal cells.

## Experimental design considerations

### The impact of experimental and analysis parameters on annotation

Cell-type annotation quality is affected by many data analysis pipeline parameters, such as data filtering and data quality settings, and the selected clustering resolution. Quality control filtering often involves removing cells from the data set where the marker genes are highly enriched in mitochondrial, heat shock, or other stress-response genes[81–83], but this must be balanced to retain important biological signals that should be kept and annotated[84].

Choosing an appropriate clustering resolution is critical for accurate cell annotation. If the clustering resolution is too low, rare cell types may be merged with larger clusters, or related cell subtypes may be merged with each other. If the clustering resolution is too high, a single cell type may be split across multiple clusters with few unique markers that are often a result of experimental noise rather than distinguishing biological function. To identify rare cell types, it may be necessary to use a feature selection tool that specifically identifies markers of rare cell types (e.g., GiniClust[85]) before clustering the cells. However, this can lead to overclustering of data sets that do not contain rare cell types. If cell-based annotation identifies multiple cell types within a cluster, then increasing the clustering resolution or subsetting the cluster and rerunning the clustering on the resulting smaller group of cells to create a zoomed-in map can help isolate these unique cell groups. Tools such as scClustViz[86], Seurat[21] and clustree[87] help select an appropriate clustering resolution.

Cell-based automatic annotation tools are often useful for choosing an appropriate clustering resolution, because the results are independent of the clustering pipeline. Thus, the clustering parameters can be tuned to identify clusters that correctly segregate cells annotated to different types. Alternatively, the presence of cluster-specific differentially expressed genes can be used to tune the clustering parameters either by gradually increasing or decreasing the resolution until the maximum number of clusters that still exhibit unique differentially expressed genes is identified.

In some cases, the original gene expression matrix generated by droplet-based technologies can be contaminated by cell-free, or 'ambient', mRNA within the cell suspension. Frequently, the ambient RNA is derived from one or more cell types that are sensitive to the tissue dissociation or cell-handling steps of the scRNA-seq experiment and break apart. As a result, markers of the contaminating cell types may be spread to all other cell clusters, which will interfere with marker determination. It is possible to estimate and correct the background contamination using methods such as SoupX[88], which looks for nonspecific expression of cell-type markers, or CellBender[89], which uses machine learning to learn and correct cell expression profiles. However, care is needed to avoid over- or undercorrecting the data.

### Workflow recommendations

The preferred starting approach for transcriptomic cell map annotation depends on the level of computational skills of the annotator. We recommend starting with automatic annotation because it is fast and reproducible, thus efficient for large data sets with many samples. Automated methods require programming, database and data science skills to operate (mainly using R or Python programming languages). A little programming knowledge goes a long way, because many recommended software packages are well documented and easy to use with basic programming knowledge. We strongly recommend that anyone working regularly with single-cell genomics data learn programming. R programming is a good language to start with, because of its prevalent use in single-cell genomics and ease of use. This recommendation may change, because an increasing number of point-and-click tools are being developed that package automated methods into easy-to-use workflows[90–92]. A second recommendation is to use a powerful computer with lots of memory (e.g., 128 GB of RAM), because current analysis and visualization methods load all data in memory for processing.

If needed, the map can be completely annotated manually by investigating gene expression patterns of cells and associated gene functions using point-and-click software (e.g., Loupe Browser, GSEA and Cerebro[93]) without programming skills. This process is easier for those knowledgeable about the biology and markers of the cells in the sample, and it is sufficient for many projects but is time consuming, especially if it must be repeated for multiple analysis parameters, such as different cluster resolutions. Even if automatic cell annotation is used, some manual annotation is usually needed because of the incompleteness of known marker databases and single-cell atlases on which automated methods depend.

Manual annotation should begin by identifying major well-known cell types, indicated by clearly defined, discrete cell clusters, because these are easiest to work with. Each cluster presents its own challenge; in the same experiment, one cluster could be annotated easily using a single well-known marker, and another may require iterations of data preprocessing pipeline optimization to accurately annotate. More challenging aspects of the map, such as cell subtypes, gradients, highly homogeneous data or poorly defined clusters, can then be progressively annotated. Sometimes, it is useful to split the data into broad subsets (e.g., immune, endothelial and tumor) and apply the recommended workflow on each one. Consider each label on the map a scientific claim that a functionally distinct biological entity (e.g., cell type) exists and that this must be supported by evidence.

In addition, not all tools are applicable to all data sets; it is important to consider the availability of reliable known markers and high-quality reference data sets or if there is sufficient diversity in a sample to detect differentially expressed genes

before applying methods that rely on that information. Using an approach based on known marker genes when there are not any repeatedly reproduced markers for the cell type of interest may lead to false conclusions. Likewise, annotating cells or cell clusters with a poor quality or inaccurately labeled reference data set will likely lead to the propagation of incorrect cell type identifications. As a final example, calculating differentially expressed genes on a largely homogeneous sample will typically result in a list of markers that are false positives or genes subject primarily to experimental or technical noise that are unrelated to the actual biology.

## Concluding remarks

Although the field of single-cell genomics is rapidly advancing and new technologies are being developed that will improve our ability to interpret, annotate and validate single-cell maps, we expect the overall workflow described here to remain valid over time. We expect major improvements in automatic annotation because of rapidly growing reference atlases, improvements to resources like the Cell Ontology[59] and improved data set integration algorithms. These methods will also need to scale up to much larger data set sizes with millions of cells[94,95]. New experimental technologies are being developed to measure more molecular details about each cell, including multi-omics technologies (e.g., mRNA, ATAC-seq[96], methylation[97] and surface proteins[98]) that can measure multiple types of information about individual cells, and these are expected to greatly improve our ability to understand multicellular systems. For instance, epigenetic information will help define stem cell subtypes that are not detectable using transcriptomics[99]. Data sets acquired from millions of cells across hundreds of patients will create computational challenges related to efficient analysis and annotation, requiring analysis to be performed on high-performance computing or cloud computing systems. In addition, meta-analyses across many single-cell maps will more clearly identify cell-type markers (e.g., macrophage or endothelial) across tissues and states (e.g., inflammation). We also expect the focus of map interpretation to gradually shift to comparisons across disease, age, or other conditions, as the number of samples per study increases.

## Data availability

The data used to generate this tutorial are openly available at the following sources. The sequence data used to generate Figs. 1 and 4 are available from MacParland et al.[100] through the NCBI GEO accession GSE115469. The analyzed data from which the map was directly created can also be accessed interactively as the R package HumanLiver from https://github.com/BaderLab/HumanLiver. The sequence data used to generate Figs. 3 and 7 are available from 10x Genomics and can be downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets. The sequence data used to generate Fig. 6 are available through the NCBI GEO accession GSE129788, as reported by Ximerakis et al.[101]. The analyzed data can be accessed interactively at http://shiny.baderlab.org/AgingMouseBrain/. The human bulk RNA-seq data used to generate the reference data set in the accompanying R code (https://github.com/BaderLab/CellAnnotationTutorial and https://codeocean.com/capsule/d67541eb-43f8-4cae-a258-5ef0069e5301/) are available from the Database of Immune Cell Expression and can be downloaded in R through the package 'celldex'[43] by the command DatabaseImmuneCellExpressionData(). The query data set used in the accompanying R code is available from 10x Genomics and can be downloaded from https://cf.10xgenomics.com/samples/cell-exp/1.1.0/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz. The collection of PBMC marker genes used in the accompanying R code is available from Diaz-Mejia et al.[37] with read data from NCBI Sequence Read Archive accession number SRX1723926. The supplementary data from the Diaz-Mejia et al. paper can be accessed at https://zenodo.org/record/3369934/#.X3CGN5NKjGI. The Gene Matrix Transposed (GMT) file used for pathway analysis in the accompanying R code can be downloaded from http://download.baderlab.org/EM_Genesets/current_release/Human/symbol/Pathways/Human_MSigdb_March_01_2021_symbol.gmt.

## Code availability

An R script that implements the main workflow described in this proposal is available at https://github.com/BaderLab/CellAnnotationTutorial, and a version-controlled copy is available through Code Ocean at https://codeocean.com/capsule/d67541eb-43f8-4cae-a258-5ef0069e5301/.

## References


1. Sasagawa, Y., Hayashi, T. & Nikaido, I. Strategies for converting RNA to amplifiable cDNA for single-cell RNA sequencing methods. *Adv. Exp. Med. Biol.* **1129**, 1–17 (2019).
2. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
3. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
5. Tabula Muris Consortium. et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* **562**, 367–372 (2018).
6. Grün, D. et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525**, 251–255 (2015).
7. Xia, B. & Yanai, I. A periodic table of cell types. *Development* **146**, dev169854 (2019).
8. Schiebinger, G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928–943.e22 (2019).
9. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643.e4 (2017).
10. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
11. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, 1–14 (2018).
12. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
13. Henry, G. H., Mathews, J. A. & Malladi, V. S. BICF Cellranger count analysis workflow (version publish_1.2.0). Zenodo. https://zenodo.org/record/3373749#.YGzmGhRucdU (2019).
14. 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).
15. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
16. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).
17. Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **20**, 273–282 (2019).
18. Menon, V. Clustering single cells: a review of approaches on high- and low-depth single-cell RNA-seq data. *Brief. Funct. Genomics* **17**, 240–245 (2018).
19. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* (86), 2579–2605 (2008).
20. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
21. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
22. Gene Set Enrichment Analysis. Archived: SCSig collection: Signatures of Single Cell Identities; https://www.gseamsigdb.org/gsea/msigdb/supplementary_genesets.jsp#SCSig
23. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)* **2019**, baz046 (2019).
24. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).
25. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
26. Papatheodorou, I. et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* **48**, D77–D83 (2020).
27. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
28. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
29. Yuzwa, S. A. et al. Developmental emergence of adult neural stem cells as revealed by single-cell transcriptional profiling. *Cell Rep.* **21**, 3970–3986 (2017).
30. Kurial, S. N. T. & Willenbring, H. Transcriptomic traces of adult human liver progenitor cells. *Hepatology* **71**, 1504–1507 (2020).
31. Stanley, G., Gokce, O., Malenka, R. C., Südhof, T. C. & Quake, S. R. Continuous and discrete neuron types of the adult murine striatum. *Neuron* **105**, 688–699.e8 (2020).
32. Satpathy, A. Curated, multi-omic, ML-driven single-cell atlas for characterizing the human immune system across disease states. *J. Immunol.* **204**, 11–159.11 (2020).
33. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
34. Zhang, Z. et al. SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes (Basel)* **10**, 531 (2019).
35. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
36. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
37. Diaz-Mejia, J. J. et al. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. [version 3; peer review: 2 approved, 1 approved with reservations]. *F1000Res.* **8**, ISCB Comm J-296 (2019).
38. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
39. Han, X. et al. Mapping the mouse cell atlas by Microwell-seq. *Cell* **172**, 1091–1107.e17 (2018).
40. Regev, A. et al. *The Human Cell Atlas White Paper*. https://doi.org/10.17863/CAM.40032 (2017).
41. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
42. Tan, Y. & Cahan, P. SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* **9**, 207–213.e2 (2019).
43. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
44. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
45. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
46. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019).
47. Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150–1158 (2020).
48. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337.e4 (2019).
49. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
50. Niwa, H. The principles that govern transcription factor network functions in stem cells. *Development* **145**, dev157420 (2018).
51. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
52. Clark, J. Z. et al. Representation and relative abundance of cell-type selective markers in whole-kidney RNA-Seq data. *Kidney Int.* **95**, 787–796 (2019).
53. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
54. Dal Molin, A., Baruzzo, G. & Di Camillo, B. Single-cell RNA-sequencing: assessment of differential expression analysis methods. *Front. Genet.* **8**, 62 (2017).
55. Adossa, N. A., Schauser, L., Gregersen, V. G. & Elo, L. L. Feature extraction approach in single-cell gene expression profiling for cell-type marker identification. Preprint at *bioRxiv* https://doi.org/10.1101/686659 (2019).
56. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
57. Reimand, J. et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **14**, 482–517 (2019).
58. Barbie, D. A. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
59. Diehl, A. D. et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
60. Meehan, T. F. et al. Logical development of the cell ontology. *BMC Bioinformatics* **12**, 6 (2011).
61. Aevermann, B. D. et al. Cell type discovery using single-cell transcriptomics: implications for ontological representation. *Hum. Mol. Genet.* **27**, R40–R47 (2018).
62. Hsiao, C. J. et al. Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Res.* **30**, 611–621 (2020).
63. Azizi, E. et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell* **174**, 1293–1308.e36 (2018).
64. Adler, M., Korem Kohanim, Y., Tendler, A., Mayo, A. & Alon, U. Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. *Cell Syst.* **8**, 43–52.e5 (2019).

65. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res.* **5**, F1000 Faculty Rev-182 (2016).
66. Schumacher, L. J. Neural crest migration with continuous cell states. *J. Theor. Biol.* **481**, 84–90 (2019).
67. Chung, N. C. Statistical significance of cluster membership for unsupervised evaluation of cell identities. *Bioinformatics* **36**, 3107–3114 (2020).
68. Rosati, E. et al. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).
69. Setliff, I. et al. High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell* **179**, 1636–1646.e15 (2019).
70. Park, D. et al. Differences in the molecular signatures of mucosal-associated invariant T cells and conventional T cells. *Sci. Rep.* **9**, 7094 (2019).
71. Moter, A. & Göbel, U. B. Fluorescence in situ hybridization (FISH) for direct visualization of microorganisms. *J. Microbiol. Methods* **41**, 85–112 (2000).
72. Ren, X. et al. Reconstruction of cell spatial organization from single-cell RNA sequencing data based on ligand-receptor mediated self-assembly. *Cell Res.* **30**, 763–778 (2020).
73. Porter, J. R., Telford, W. G. & Batchelor, E. Single-cell gene expression profiling using FACS and qPCR with internal standards. *J. Vis. Exp*. (120), 55219 (2017).
74. Wu, A. R. et al. Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
75. 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).
76. Liu, F. et al. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol.* **20**, 242 (2019).
77. Fan, J. et al. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* **28**, 1217–1227 (2018).
78. Serin Harmanci, A., Harmanci, A. O. & Zhou, X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat. Commun.* **11**, 89 (2020).
79. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
80. Tickle, T., Gc Ti, Brown, M. & Haas, B. InferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV (Klarman Cell Observatory, Broad Institute of MIT and Harvard, 2019).
81. AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An introduction to the analysis of single-cell RNA-sequencing data. *Mol. Ther. Methods Clin. Dev.* **10**, 189–196 (2018).
82. van den Brink, S. C. et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
83. Zhao, Q. et al. A mitochondrial specific stress response in mammalian cells. *EMBO J.* **21**, 4411–4419 (2002).
84. Guantes, R. et al. Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res.* **25**, 633–644 (2015).
85. 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).
86. Innes, B. T. & Bader, G. D. scClustViz – Single-cell RNAseq cluster assessment and visualization. *F1000Res.* **7**, ISCB Comm J-1522 (2018).
87. Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *Gigascience* **7**, giy083 (2018).
88. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151 (2020).
89. Fleming, S. J., Marioni, J. C. & Babadi, M. CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets. Preprint at *bioRxiv* https://doi.org/10.1101/791699 (2019).
90. Mohanraj, S. et al. Crescent: cancer single cell expression toolkit. *Nucleic Acids Res.* **48**, W372–W379 (2020).
91. David, F. P. A., Litovchenko, M., Deplancke, B. & Gardeux, V. ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. *Nucleic Acids Res.* **48**, W403–W414 (2020).
92. Franzén, O. & Björkegren, J. L. M. alona: a web server for single-cell RNA-seq analysis. *Bioinformatics* **36**, 3910–3912 (2020).
93. Hillje, R., Pelicci, P. G. & Luzi, L. Cerebro: interactive visualization of scRNA-seq data. *Bioinformatics* **36**, 2311–2313 (2020).
94. Zhang, A. W. et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat. Methods* **16**, 1007–1015 (2019).
95. Miao, Z. et al. Putative cell type discovery from single-cell gene expression data. *Nat. Methods* **17**, 621–628 (2020).
96. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
97. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
98. Baron, M. & Yanai, I. New skin for the old RNA-Seq ceremony: the age of single-cell multi-omics. *Genome Biol.* **18**, 159 (2017).
99. Guilhamon, P. et al. Chromatin blueprint of glioblastoma stem cells reveals common drug candidates for distinct subtypes. Preprint at *bioRxiv* https://doi.org/10.1101/370726 (2018).
100. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).
101. Ximerakis, M. et al. Single-cell transcriptomic profiling of the aging mouse brain. *Nat. Neurosci.* **22**, 1696–1708 (2019).
102. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. (85), 2825–2830 (2011).
103. Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).
104. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007).
105. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
106. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
107. 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).
108. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887.e17 (2019).
109. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
110. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometr. Intell. Lab. Syst.* **2**, 37–52 (1987).
111. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
112. Halladin-Dąbrowska, A., Kania, A. & Kopeć, D. The t-SNE algorithm as a tool to improve the quality of reference data used

in accurate mapping of heterogeneous non-forest vegetation. *Remote Sens. (Basel)* **12**, 39 (2019).
113. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both *t*-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2019).
114. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
115. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **26**, 303–304 (2008).
116. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19**, 562–578 (2018).
117. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
118. Clamp, M. et al. Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* **31**, 38–42 (2003).
119. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
120. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
121. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
122. Geirsdottir, L. et al. Cross-species single-cell analysis reveals divergence of the primate microglia program. *Cell* **179**, 1609–1622.e16 (2019); erratum: **181**, 746 (2020).
123. Ding, H., Blair, A., Yang, Y. & Stuart, J. M. Biological process activity transformation of single cell gene expression for cross-species alignment. *Nat. Commun.* **10**, 4899 (2019).

## Acknowledgements

This project has been made possible in part by grant CZF2019-002429 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. We acknowledge Shirley Hui, Daniel Stueckmann and Ronald Xie for their help in reviewing the tutorial.

## Author contributions

Z.A.C., T.S.A., J.A., D.P. and S.A.M. initially wrote different sections of the document and collaboratively joined them together. Z.A.C., T.S.A., J.A., D.P. and B.T.I. designed the figures. Z.A.C. and G.D.B. organized the process and undertook major edits. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41596-021-00534-0.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41596-021-00534-0.

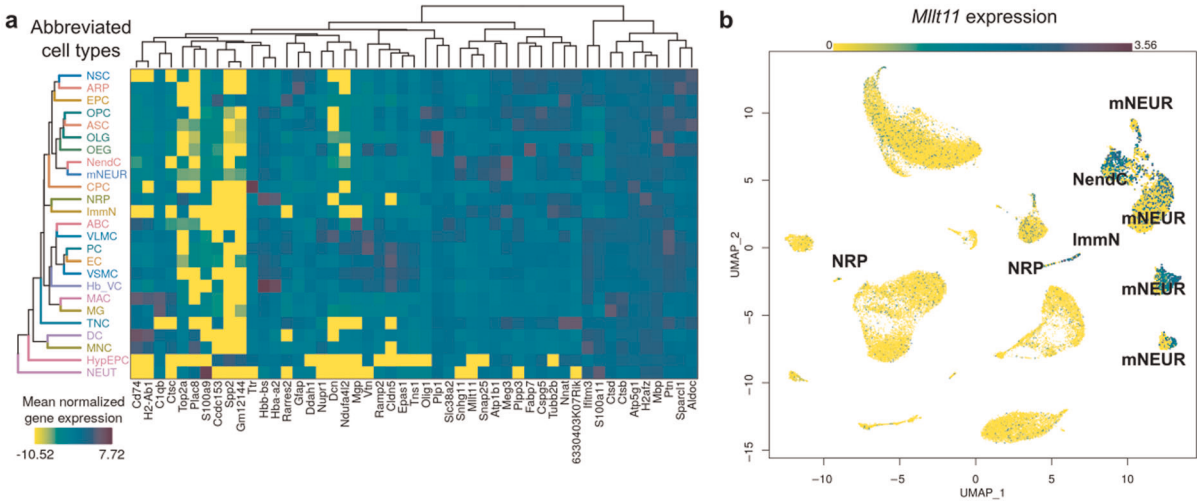
**Correspondence and requests for materials** should be addressed to S.A.MP. or G.D.B.

**Peer review information** *Nature Protocols* thanks Åsa Björklund and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

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

Received: 26 July 2020; Accepted: 12 March 2021;
Published online: 24 May 2021




**Extended Data Fig. 1 | Heat map and UMAP visualizations.** An extension of Fig. 6 incorporating the visualization of marker genes for identified cell types as a heat map (**a**) and a Mllt11 expression overlaid on a UMAP plot (**b**). Mllt11 is expressed at various levels across clusters, suggesting a cell-type gradient across NRP, ImmN, NendC and mNEUR. Both plots are generated from scRNA-seq data from young and old mouse brains[101].