Cell-Type Annotation

Subjects: Mathematical & Computational Biology | Biology | Biochemistry & Molecular Biology
Contributor: Hongjian Jin , Changde Cheng , Wenan Chen , Xiang Chen

Multicellular organisms consist of cells that can be categorized by their function and morphology. Single-cell transcriptomics makes it possible to individually profile thousands of cells in multiple tissues and organisms within a single experiment. Determining and labeling cell types or states in single cell transcriptomic data is known as cell-type annotation or identification. Several methods are employed for cell-type annotation, including signature scoring, supervised learning, cell-integration-based label transfer, and semi-supervised annotation. Considering the lineage relationships among cell types, hierarchical classification methods are crucial for accurately identifying cell types and subtypes at an optimal clustering resolution. The use of well-curated reference datasets, implementation of quality control measures, and careful consideration of cluster resolutions heavily influence the reliability of cell-type annotation. The aim of cell-type annotation is to gain insights into cell heterogeneity in various biological processes and diseases, with the potential to drive improvements in therapeutic interventions.

scRNA-seq analysis method | cell-type annotation | single-cell data integration

1. Cell Annotation by Signature Scoring

The prevailing method of cell-type annotation consists of unsupervised clustering analysis followed by manual or automatic annotation using a set of known “marker genes”, also known as gene sets, markers, or signatures. An example of this approach is the Seurat function FindMarkers [1], which employs differential expression analysis to identify biomarkers defining clusters. This annotation approach does not necessitate training a model with another “annotated” reference dataset. Still, it heavily relies on existing biological knowledge of known marker genes and involves subjective decision-making, such as choosing the number of clusters (resolution).

Moreover, this process is typically manual, leading to potential time constraints and annotation inconsistency.

1.1. Signature Database

Several databases provide extensive collections of known markers that can aid in cell-type annotation (see **Table 1**). These databases include MSigDB [2], Enrichr ARCHS4 tissues [3], TISSUES 2.0 [4], SaVanT [5], xCell [6], celldex [7], PanglaoDB [8], CellMarker [9][10], SCsig, and CellMatch [11]. Among these, PanglaoDB, CellMarker, SCsig, and CellMatch were specifically developed for scRNA-seq analysis. The scMRMA method utilizes Cell Ontology [12] to reorganize PanglaoDB into a hierarchical structure, enabling consistent representation of cell types across various levels of anatomical granularity.

Table 1. A survey of databases used for cell annotation.

Database	Data Source	Link
PanglaoDB	scRNA-seq	https://panglaodb.se/markers/PanglaoDB_markers_27_Mar_2020.tsv.gz accessed on 25 May 2023
Hierarchical PanglaoDB	scRNA-seq	https://github.com/JiaLiVUMC/scMRMA/tree/main/data accessed on 25 May 2023
Cellmarker	scRNA-seq +bulk RNA-seq	http://bio-bigdata.hrbmu.edu.cn/CellMarker/download/all_cell_markers.txt accessed on 25 May 2023
CellMatch	scRNA-seq +bulk RNA-seq	https://github.com/ZJUFanLab/scCATCH/raw/master/data/cellmatch.rda accessed on 25 May 2023
SCSig	scRNA-seq	https://data.broadinstitute.org/gsea-msigdb/msigdb/supplemental/scsig/1.0/scsig.all.v1.0.symbols.gmt accessed on 25 May 2023
SaVanT	Microarray	http://newpathways.mcdb.ucla.edu/savant-dev/SaVanT_Signatures_Release01.zip accessed on 25 May 2023
MSigDB	Bulk RNA-seq, Microarray	https://data.broadinstitute.org/gsea-msigdb/msigdb/release/7.2/msigdb_v7.2.xml accessed on 25 May 2023
xCell	Bulk RNA-seq	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5688663/bin/13059_2017_1349_MOESM3_ESM.xlsx accessed on 25 May 2023
Enrichr ARCHS4 tissues	Bulk RNA-seq	https://maayanlab.cloud/Enrichr/geneSetLibrary?mode=text&libraryName=ARCHS4_Tissues accessed on 25 May 2023
TISSUES 2.0	Bulk RANseq, Microarray	http://tissues.jensenlab.org/ accessed on 25 May 2023

1.2. Scoring Method

Common scoring methods, like single sample gene set enrichment analysis (ssGSEA, [13]), gene set variation analysis (GSVA, [14]), and Singscore [15], were initially designed for bulk RNA-seq data. The ssGSEA score quantifies the coordinated up- or down-regulation of an input gene signature within a sample. GSVA performs kernel density estimation of the gene expression profile across all samples, and Singscore calculates a normalized mean percentile rank. However, these methods rely on statistical assumptions that do not consider the extensive presence of zero values and missing genes within individual cells across a dataset, making these bulk-sample-based methods prone to dropout effects and therefore suboptimal for scRNA-seq data analysis.

The optimal scenario for scoring genes is when there is a bi-modal distribution, indicating a high expression of signature genes in one cell type but not others. However, at the single-cell level, most genes are either not expressed or exhibit unstable expression patterns. Gene expression analysis is further complicated by dropouts (resulting from low input of RNA amounts), transcriptional stochasticity, and diversity of cell states and identities.

Researchers have made significant efforts to address these challenges in order to improve the evaluation of gene signatures in scRNA-seq data. Several approaches have been developed, including the cell-type activity (CTA) score [8], single cell signature scorer (SCSS, [16]), ModuleScore (implemented in Seurat's AddModuleScore function), AUCell [17], Ucell [18], JASMINE [19], scType [20], scCATCH [11], and scMRMA [21], among others (see **Table 2**). These methods aim to provide improved assessments of gene signatures within scRNA-seq datasets.

Table 2. Scoring methods used for cell annotation.

Method	Description
CTA (Cell-Type Activity)	Sum of the weighted expression
Ucell	Mann–Whitney U statistic
AUCell	Area under the curve
ssGSEA	Rank-based enrichment score
SCSS	Sum of UMI, normalized by library size
GSVA	Kernel density estimation
Singscore	Normalized mean percentile rank
ScType	Cluster summary enrichment score
JASMINE	Approximate mean of gene ranks and the enrichment of the signatures
AddModuleScore (Seurat)	Average expression level
scCATCH	Evidence-based scoring

The cell-type activity (CTA) method calculates an activity score for each cell type by summing the weighted expressions of its marker genes [8]. The SCSS score for a signature in a cell is computed as the sum of all UMI (unique molecular identifier) counts for the genes in the gene set expressed in that cell divided by the sum of total UMI counts in the cell.

Seurat's AddModuleScore function calculates the average expression levels of each signature at the single-cell level, with the aggregated expression of control feature sets subtracted. The analyzed features are grouped into bins based on their average expression, and control features are randomly selected from each bin.

AUCell utilizes the area under the curve (AUC) to determine whether a critical subset of genes in the input gene set is enriched at the top of the ranking for each cell. The AUC reflects the proportion of expressed genes in the signature and their expression values relative to other genes within the cell.

UCell calculates gene signature scores for scRNA-seq data using the Mann–Whitney U statistic, which is correlated with the AUC scores computed by AUCell. JASMINE calculates the approximate mean using gene ranks among expressed genes and measures the enrichment of the signature in expressed genes. These two values are scaled to a range of 0–1 and averaged to obtain the final JASMINE score.

ScType calculates a cell-type-specific marker enrichment score per cluster by computing a cell type specificity score for each marker, and then multiplying these by the z-score of marker gene expression across all cells. The values of each cell signature are summed across cells corresponding to a specific cluster, resulting in the cluster summary enrichment score.

scCATCH employs the evidence-based scoring (ES) process, which utilizes tissue-specific cell taxonomy reference databases (CellMatch) to determine cell types and subtypes in two steps.

Notably, scMRMA utilizes the CTA scoring method with different parameters at different levels (major cell types and subtypes). This approach enables multiresolution cell annotation through iterative clustering and the mapping of clusters to the hierarchical PanglaoDB marker database.

By implementing scoring methods, the annotation process of cells or clusters can be efficiently automated in annotation tools like scType, scCATCH, and scMRMA. Since single-resolution unsupervised clustering cannot capture both global and local biological variances simultaneously, a multi-resolution strategy like scMRMA can achieve more comprehensive and detailed annotation.

The performance of signature-based cell annotation relies on several factors, including gene sets, scoring methods, and the characteristics of the query data. It is important to note that the signature scores obtained may not always be normalized or comparable across different gene sets or datasets. Improving the reproducibility and reliability of cell annotation will require addressing the following general limitations:

- Cell marker databases are compiled from diverse data sources generated using different technologies, each with its own technical biases such as sensitivity, dropouts, and cell population purity. The derived signatures for the same cell type can therefore vary across technologies. Additionally, signatures obtained from bulk RNA-seq or microarray data may not accurately annotate cell types in single-cell data.

- There is a lack of consistent criteria or methods for curating signatures. Gene sets can be derived experimentally, computationally, or manually curated from the literature. Even computational selection methods, such as differential expression analysis, can result in different gene sets due to arbitrary cutoffs (e.g., log2 fold change, false discovery rate, top number of genes).

- The size of gene sets (i.e., the number of genes they contain) varies greatly, making it difficult to compare the scores of different signatures. Smaller gene sets (e.g., size < 20) are more likely to yield cells with unstable scores, while larger gene sets (e.g., size > 100) can provide greater stability for detection and evaluation. It is

often observed that the signature scores of large random gene sets follow an approximately normal distribution, abiding by the central limit theorem.

- Redundancy across gene sets is common in large databases. Since gene sets may share a significant proportion of their constituent genes, scoring results can be dominated by long lists of candidate cell types associated with overlapping signatures, potentially obscuring meaningful cell types that possess only a few marker genes.

- Most databases adopt a flat structure, treating each cell type equally and independently. While this approach can effectively distinguish major cell types, it may struggle to identify cell subtypes due to the lack of relationships between cell types. Hierarchical cell type databases could enhance discrimination of specific cell types or subtypes [21].

- Unstandardized cell nomenclature in certain publications can lead to overlapping or ambiguous anatomy terms or identifiers for cell types. To address this, collaborative efforts such as the Cell Ontology (CL) and The Human Cell Atlas (HCA) have begun to build a high-dimensional compendium of cell information.

For quality control of signature-based annotation, the following measures can be considered:

- Assess the reliability of cell annotation by plotting the score histogram of a specific gene set and examining the distribution of scores within cell types in the dataset.

- Visualize the signature scores or average expression of a gene set in a two-dimensional plot. Calculating the mean expression with library-size normalization provides an intuitive approach.

- Some methods are sensitive to the number of detected genes or dropout rates. Checking marker gene expression through dot plots or stacked violin plots can help to identify potential issues.

- Employ a confusion matrix or mosaic plot to evaluate the final assignment of cell type labels.

By addressing these considerations and implementing quality control measures, the reliability and reproducibility of cell annotation based on signatures can be improved.

2. Cell Annotation by Supervised Learning

In recent years, supervised cell annotation has gained significant attention due to the exponential growth of publicly available single-cell RNA sequencing (scRNA-seq) data, including projects like the Human Cell Atlas (https://www.humancellatlas.org/ accessed on 25 May 2023, [22]), Tabula Muris (https://tabula-muris.ds.czbiohub.org/ accessed on 25 May 2023, [23]), and the Mouse Cell Atlas (https://bis.zju.edu.cn/MCA/ accessed on 25 May 2023 [24]). Supervised learning, a type of machine learning, has been employed to transfer cell type labels from labeled to unlabeled datasets for cell-type annotation. Various common algorithms, such as

Support Vector Machine (SVM, [25]), Random Forest [26], k-nearest neighbors (kNN, [27]), neural networks [28], and deep learning [29], have been utilized in this field.

In general, the process of supervised cell annotation involves several steps. Firstly, a classifier is constructed using a reference dataset of known cell types, which serves as the labeled training set. Secondly, feature selection is performed to identify the most informative features for training the classifier. Thirdly, the classifier is trained using the labeled training set to associate specific features with each cell type. Finally, once the classifier has been trained and evaluated for its accuracy, it can be used to predict the cell type of new cells or clusters in an unannotated dataset.

As these steps require substantial computational expertise, numerous automatic annotation software tools employing different supervised approaches have been actively developed to enable efficient supervised cell annotation.

2.1. Feature Selection

Feature selection is a crucial step in enhancing the performance and interpretability of a model by identifying the most informative variables within a dataset. The primary objective is to reduce the dimensionality of the feature space by eliminating redundant, irrelevant, or noisy features. This reduction not only improves computational efficiency during model training and evaluation but also facilitates more accurate machine learning outcomes.

When it comes to cell-type annotation, known marker genes associated with specific cell types, obtained from external resources, can be directly employed as features. Alternatively, marker genes can be identified through differential expression (DE) analysis, which involves comparing the gene expression levels in a particular cell type against all other cell types using statistical tests like t-tests [30], Wilcoxon signed-rank tests [31], or dedicated packages such as limma [32], DESeq2 [33], or Seurat's FindAllMarkers function.

Certain feature selection methods rely on variance filtering. By establishing a threshold on the variance, features below that threshold are eliminated from the feature set. Bartlett's test [34] is utilized to assess whether the variances across all groups are equal. Additionally, F-statistics are useful if the data follows a normal distribution and the group variances are equal. Several feature ranking methods, such as information gain (Entropy test) [35], chi-square statistics [36], the Kolmogorov–Smirnov (KS) test [37], and the bimodality index [38], can assign scores, ranks, or significance levels to genes based on their relevance to cell-type annotation. Genes with higher scores or significance levels are considered more informative or cell-type specific.

Li et al. [35] pioneered the use of entropy, a measure of dispersion from information theory, to assess the distribution of gene expression levels following a Poisson–Gamma mixture model. The entropy could be estimated directly from the logarithm of the mean gene expression, and genes with larger total entropy differences were found to be more cell-type specific. FEAST [39] applies unsupervised consensus clustering followed by an F-test on the clusters to calculate feature significance and rank features accordingly. Andrews et al. [40] introduced M3Drop, which employs a Bayesian model to estimate the dropout rate for each gene, incorporating its mean expression,

and subsequently performing differential expression analysis to select informative genes. This dropout-based feature selection method demonstrates superior performance compared to variance-based approaches. Lin et al. [41] showed that the differential expression (DE) gene selection method outperformed other tested methods (DE, DD, BD, and DP) in terms of cell-type annotation accuracy (**Table 3**).

Table 3. Methods used for feature selection.

Method	Description	Reference
DE	Differentially expressed genes	[32]
DD	Differentially distributed genes by Kolmogorov–Smirnov test	
DV	Differentially variable genes by Bartlett's test	
BD	Bimodally distributed by bimodality index	[38]
DP	Differentially proportioned genes by chi-squared test	
M3Drop	Dropout-based feature selection	[40]
E-test	Entropy-based feature selection	[35]
FEAST	Unsupervised consensus clustering followed by F-test for ranking features	[39]

2.2. Prediction Model (Classifier)

A variety of methods have been developed to annotate cell types in single-cell transcriptomics data using machine learning models. For instance, scPred [42] employs support vector machine (SVM)-based classifiers on PCA-transformed gene expression matrices. The singleCellNet [43] and scAnnotate [44] methods utilize the Random Forest technique for classification. Garnett [45] trains a multinomial classifier using elastic-net regression [46] to discriminate between different cell types. The L2-regularized logistic regression implemented in cellTypist [47] enables automated annotation of immune cells across human tissues. The scClassify [41] method takes advantage of a k-nearest neighbors (kNN)-based learning algorithm, combining multiple similarity metrics and feature selections. On the other hand, scDeepSort [48] employs a weighted graph neural network, while Cell Blast [49] leverages large-scale reference databases and an autoencoder-based generative model to obtain low-dimensional representations of cells and employs a cell similarity metric for mapping query cells to specific types. SciBET [35] achieves rapid and accurate single-cell-type identification using a multinomial-distribution model and maximum likelihood estimation. Notably, scBERT [50] is an adaptation of the Bidirectional Encoder Representations from Transformers (BERT, [51]) model, originally developed for natural language processing for cell-type annotation. The scBERT method incorporates gene expression data to represent cells and their relationships, demonstrating superior performance in tasks such as novel cell type discovery and robustness, to batch effects, through to pretraining and fine-tuning.

Several supervised cell annotation methods have been specifically developed for single-cell RNA sequencing (scRNA-seq) data (**Table 4**), focusing on the correlation between the target and reference datasets. Notable methods include SingleR [7], CellAssign [52], CHETAH [53], and scmap [54]. SingleR assigns cellular identities to single-cell transcriptomes by comparing them to a built-in reference transcriptome of pure cell types obtained from microarray or bulk RNA-sequencing data. CellAssign employs a probabilistic model that utilizes a marker-based reference for cell type assignment. CHETAH adopts a hierarchical classification approach, allowing cells to be assigned to intermediate or unassigned types through stepwise traversal of the classification tree. Finally, scmap classifies query cells based on their similarity to reference cell types using various correlation measures.

Table 4. Supervised machine learning methods for cell annotation.

Tool	Year	Reference Database	Algorithm	Ref.
SingleR	2019(**)	Built-in celldex (transcriptome of pure cell types)	Spearman	[7]
scmap-cell	2018(**)	Annotated transcriptome	K-nearest neighbor (KNN)	[54]
Garnett	2019(**)	Marker genes	Elastic net regression	[45]
CellAssign	2019(**)	Marker genes	Probabilistic Bayesian model	[52]
scPred	2019(**)	Annotated transcriptome	Support vector machines (SVM)	[42]
singleCellNet	2019(*)	Annotated transcriptome	Random Forest	[43]
CHETAH	2019(*)	Annotated transcriptome	Spearman and confidence	[53]
cellTypist	2022(*)	Annotated transcriptome	L2-regularized logistic regression	[47]
CellBlast	2020	Annotated transcriptome	Neural network-based generative model	[49]
sciBET	2020	Annotated transcriptome	Multinomial-distribution model	[35]
scClassify	2020	Annotated transcriptome	Weighted KNN	[41]
scDeepSort	2021	Annotated transcriptome	Weighted graph neural network	[48]
scBERT	2022	Annotated transcriptome	BERT	[50]
scAnotate	2023	Annotated transcriptome	Random Forest	[44]
TOSICA	2023	Annotated transcriptome	Transformer	[55]

**, citation > 100; *, citation > 50; accessed on 25 May 2023.

Supervised methods are generally not optimized for discovering novel cell types. Without additional configurations to prevent over-classification, any new cell type in the target data may be forced into one of the existing cell types in the reference dataset. However, a common strategy is to set a threshold on the prediction odds, classifying

certain cells as unassigned. This threshold-based approach is implemented in popular tools such as scmap, CellAssign, and CHETAH, allowing the identification of unassigned cells.

The assessment of prediction results can be effectively conducted using multiple established metrics, each providing a unique perspective:

- Accuracy: This metric captures the ratio of correctly classified cell types to the total number of cells, providing a broad view of model performance.
- Adjusted Rand Index (ARI): ARI allows for the comparison of clustering patterns between the predicted and actual (ground truth) classifications. It offers an insight into how closely the model's clustering aligns with the actual data.
- F1 score: The F1 score offers a robust measure of a model's classification accuracy. It amalgamates precision and recall into a single measure by averaging the individual F1 scores for each class. It provides a more nuanced view of model performance, especially in scenarios where class imbalances exist.
- Normalized Mutual Information (NMI): NMI is a metric that quantifies the shared information between the predicted and ground truth distributions. By normalizing against the maximum possible mutual information value, it gives a relative perspective on how much the predicted labels reveal about the actual labels, which is particularly useful in clustering contexts.
- Variation of Information (VI): VI evaluates the degree of difference between predicted and actual labels. It effectively gauges how much the model's classification deviates from the true label distribution.

There are more metrics that have been used to evaluate the performance of cell clustering and annotation; interested readers may consult Hossin et al. [56].

The performance of cell annotation methods is heavily influenced by the quality of annotated reference databases. However, constructing these reference datasets presents several notable challenges. One of these challenges is the unavoidable need for manual cell-type annotation, which can be a time-consuming and subjective process. Additionally, determining the appropriate clustering resolution or the number of cell types in both the reference and query data often relies on subjective choices based on specific study requirements or expert opinions. Another crucial factor affecting classifier accuracy is the quality of the training set. If the reference data is not well curated, the classifier may yield inaccurate results, leading to erroneous cell-type annotations in the query data. These considerations underscore the importance of meticulous curation and careful selection of reference datasets for robust and reliable cell-type annotation.

3. Other Cell Annotation Methods

3.1. Cell-Integration-Based Label Transfer

An alternative method for annotating cells based on transcriptomic data involves integrating a query dataset with a well-established reference dataset using an integration algorithm. This integration enables the annotation of clusters that span both datasets, allowing the transfer of labels from the reference data to the corresponding query cells within the clusters. This approach facilitates the identification of identical, distinct, and novel cell types. However, it is important to note that this method can be computationally demanding. Additionally, integration algorithms may exhibit varying performances, and batch effects or disparities between the reference and query data can introduce challenges.

3.2. Semi-Supervised Annotation

Semi-supervised learning [57][58][59] is a machine learning approach that leverages both labeled and unlabeled data during model training. This technique is particularly valuable when only a limited amount of labeled data is available, as the unlabeled data can enhance the model's understanding of the problem domain. By incorporating unlabeled data, the model can learn more about the underlying patterns and structure of the data, leading to better generalization. This approach is particularly useful when acquiring labeled data is costly or time consuming, as it can make the most of available resources and achieve satisfactory results with a smaller labeled dataset. However, it is important to note that training a semi-supervised model can be computationally intensive [58][60]. Additionally, selecting the appropriate algorithm for a given problem and interpreting the results of such a model can be challenging.

There are two noteworthy recent implementations in this field: SCINA [61] and scNym [62]. SCINA is a semi-supervised model that utilizes an expectation-maximization algorithm [63] to annotate cells at the cluster level. It achieves this by fitting a bimodal distribution to cell type marker genes. On the other hand, scNym is a semi-supervised approach that employs an adversarial neural network [64] to transfer cell identity annotations from one experiment to another. Remarkably, scNym has demonstrated high performance in cell-type annotation across experiments, even when faced with biological and technical differences.

In summary, semi-supervised learning is a valuable technique that can enhance the performance of machine learning models when labeled data is limited. Recent implementations such as SCINA and scNym showcase the potential of semi-supervised approaches in annotating cells at the cluster level and transferring annotations across experiments.

4. Perspective

In many tissues, there are typically a small number of major cell types [65]. These major cell types can further be divided into subtypes in a hierarchical manner, forming what is known as a "cell type hierarchy" [66]. While most supervised methods classify cells directly into a "terminal" cell type, this one-step annotation approach can successfully identify the major cell types but may result in misclassification of similar cell subtypes.

To address this challenge of cell subtyping, and considering the hierarchical relationships between cell types, recent advancements in scientific research have introduced multi-scale or multi-resolution classification frameworks such as scMRMA and scClassify. These frameworks take into account the hierarchical relationships between cell types and aim to improve the accuracy of cell subtyping. Additionally, the divisive hierarchical clustering method uses various marker genes to cluster cells in multiple iterations and at different resolutions, as seen in the co-occurrence clustering algorithm [67] and TooManyCells [68].

Interestingly, a similar approach based on multi-level scale-adaptive clustering has been reported for the unsupervised classification of tumor subtypes using RNA-seq. This approach, known as Resolution-Adaptive Coarse-to-Fine Clusters Optimization (RACCOON, [69]), classified more than 13,000 samples into an eight-level hierarchical tree based on their expression similarities. It successfully generated an atlas consisting of 455 tumor and normal classes. Building upon this extensive hierarchy, the same research group developed a classifier called OTTER for childhood cancer. OTTER is an ensemble of convolutional neural networks that performs robustly across all cancer types.

The choice of cluster resolution in data analysis depends on the specific dataset and research objectives. Low-resolution clustering can impede the accurate identification of distinct cell types, while annotating cells at the single-cell level is susceptible to errors due to stochastic noise. To overcome these challenges, several approaches have been proposed.

A common strategy is to employ validation indices, such as the silhouette score or the gap statistic. These indices evaluate clustering quality by comparing the distances within clusters to those between clusters. A higher score indicates better clustering performance. An example of this approach is scLCA [70], which combines the Tracy–Widom test [71][72][73] based on random matrix theory to determine the number of significant eigenvalues, and the silhouette score to rank the results of spectral clustering. The scLCA approach has demonstrated effectiveness in accurately determining the number of clusters in scRNA-seq data through systematic benchmarking [74].

Another approach involves utilizing visualization tools like t-SNE or UMAP. These techniques aid in identifying clusters that may be excessively small or large, assisting in the refinement of cluster resolution. Optimizing resolution in this manner can yield biologically meaningful and desirable outcomes, especially when considering common dropout events in scRNA-seq data.

Nevertheless, it is important to recognize that, while there are various strategies for optimization and hierarchy, the ultimate decision on cluster resolution remains a subjective judgment that the researcher must make.

Nonetheless, the careful curation, integration, and optimization of hierarchical knowledge databases derived from cell-type ontologies and expression similarities in atlas datasets will have a pivotal impact on the advancement of cell-type annotation methodologies. Moreover, this process will enable us to delve deeper into our comprehension of cell heterogeneity in developmental processes and diseases, ultimately facilitating the development of more effective treatments.

The annotation of new or rare cell types or subtypes presents challenges due to the scarcity of known markers or reference datasets associated with them. In such cases, a combination of approaches can be considered. Initially, a supervised method can be employed to predict the major cell types using a well-established reference dataset. Subsequently, an unsupervised clustering method can be applied to identify subtypes within each major cell type separately. When annotating new or rare cell types, a conservative approach is recommended. It is preferable to omit a cell type lacking solid validation rather than erroneously categorizing a cell as a different type.

References

1. Satija, R.; Farrell, J.A.; Gennert, D.; Schier, A.F.; Regev, A. Spatial Reconstruction of Single-Cell Gene Expression Data. Nat. Biotechnol. 2015, 33, 495–502.
2. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015, 1, 417–425.
3. Lachmann, A.; Torre, D.; Keenan, A.B.; Jagodnik, K.M.; Lee, H.J.; Wang, L.; Silverstein, M.C.; Ma’ayan, A. Massive Mining of Publicly Available RNA-Seq Data from Human and Mouse. Nat. Commun. 2018, 9, 1366.
4. Palasca, O.; Santos, A.; Stolte, C.; Gorodkin, J.; Jensen, L.J. TISSUES 2.0: An Integrative Web Resource on Mammalian Tissue Expression. Database 2018, 2018, bay003.
5. Lopez, D.; Montoya, D.; Ambrose, M.; Lam, L.; Briscoe, L.; Adams, C.; Modlin, R.L.; Pellegrini, M. SaVanT: A Web-Based Tool for the Sample-Level Visualization of Molecular Signatures in Gene Expression Profiles. BMC Genom. 2017, 18, 824.
6. Aran, D.; Hu, Z.; Butte, A.J. XCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. Genome Biol. 2017, 18, 220.
7. Aran, D.; Looney, A.P.; Liu, L.; Wu, E.; Fong, V.; Hsu, A.; Chak, S.; Naikawadi, R.P.; Wolters, P.J.; Abate, A.R.; et al. Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage. Nat. Immunol. 2019, 20, 163–172.
8. 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 2019, 2019, baz046.
9. Zhang, X.; Lan, Y.; Xu, J.; Quan, F.; Zhao, E.; Deng, C.; Luo, T.; Xu, L.; Liao, G.; Yan, M.; et al. CellMarker: A Manually Curated Resource of Cell Markers in Human and Mouse. Nucleic Acids Res. 2019, 47, D721–D728.
10. Hu, C.; Li, T.; Xu, Y.; Zhang, X.; Li, F.; Bai, J.; Chen, J.; Jiang, W.; Yang, K.; Ou, Q.; et al. CellMarker 2.0: An Updated Database of Manually Curated Cell Markers in Human/Mouse and Web Tools Based on ScRNA-Seq Data. Nucleic Acids Res. 2023, 51, D870–D876.

11. Shao, X.; Liao, J.; Lu, X.; Xue, R.; Ai, N.; Fan, X. ScCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. iScience 2020, 23, 100882.
12. Bard, J.; Rhee, S.Y.; Ashburner, M. An Ontology for Cell Types. Genome Biol. 2005, 6, R21.
13. Barbie, D.A.; Tamayo, P.; Boehm, J.S.; Kim, S.Y.; Moody, S.E.; Dunn, I.F.; Schinzel, A.C.; Sandy, P.; Meylan, E.; Scholl, C.; et al. Systematic RNA Interference Reveals That Oncogenic KRAS-Driven Cancers Require TBK1. Nature 2009, 462, 108–112.
14. Hänzelmann, S.; Castelo, R.; Guinney, J. GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. BMC Bioinform. 2013, 14, 7.
15. Foroutan, M.; Bhuva, D.D.; Lyu, R.; Horan, K.; Cursons, J.; Davis, M.J. Single Sample Scoring of Molecular Phenotypes. BMC Bioinform. 2018, 19, 404.
16. Pont, F.; Tosolini, M.; Fournié, J.J. Single-Cell Signature Explorer for Comprehensive Visualization of Single Cell Signatures across ScRNA-Seq Datasets. Nucleic Acids Res. 2019, 47, e133.
17. Aibar, S.; González-Blas, C.B.; Moerman, T.; Huynh-Thu, V.A.; Imrichova, H.; Hulselmans, G.; Rambow, F.; Marine, J.-C.; Geurts, P.; Aerts, J.; et al. SCENIC: Single-Cell Regulatory Network Inference and Clustering. Nat. Methods 2017, 14, 1083–1086.
18. Andreatta, M.; Carmona, S.J. UCell: Robust and Scalable Single-Cell Gene Signature Scoring. Comput. Struct. Biotechnol. J. 2021, 19, 3796–3798.
19. Noureen, N.; Ye, Z.; Chen, Y.; Wang, X.; Zheng, S. Signature-Scoring Methods Developed for Bulk Samples Are Not Adequate for Cancer Single-Cell RNA Sequencing Data. eLife 2022, 11, e71994.
20. Ianevski, A.; Giri, A.K.; Aittokallio, T. Fully-Automated and Ultra-Fast Cell-Type Identification Using Specific Marker Combinations from Single-Cell Transcriptomic Data. Nat. Commun. 2022, 13, 1246.
21. Li, J.; Sheng, Q.; Shyr, Y.; Liu, Q. ScMRMA: Single Cell Multiresolution Marker-Based Annotation. Nucleic Acids Res. 2022, 50, e7.
22. Regev, A.; Teichmann, S.A.; Lander, E.S.; Amit, I.; Benoist, C.; Birney, E.; Bodenmiller, B.; Campbell, P.; Carninci, P.; Clatworthy, M.; et al. The Human Cell Atlas. eLife 2017, 6, e27041.
23. Schaum, N.; Karkanias, J.; Neff, N.F.; May, A.P.; Quake, S.R.; Wyss-Coray, T.; Darmanis, S.; Batson, J.; Botvinnik, O.; Chen, M.B.; et al. Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris. Nature 2018, 562, 367–372.
24. Han, X.; Wang, R.; Zhou, Y.; Fei, L.; Sun, H.; Lai, S.; Saadatpour, A.; Zhou, Z.; Chen, H.; Ye, F.; et al. Mapping the Mouse Cell Atlas by Microwell-Seq. Cell 2018, 172, 1091–1107.e17.
25. Cortes, C.; Vapnik, V. Support-Vector Networks. Mach. Learn. 1995, 20, 273–297.

26. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.
27. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theory 1967, 13, 21–27.
28. McCulloch, W.S.; Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bull. Math. Biophys. 1943, 5, 115–133.
29. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. Nature 2015, 521, 436–444.
30. Student. The Probable Error of a Mean. Biometrika 1908, 6, 1–25.
31. Wilcoxon, F. Individual Comparisons by Ranking Methods. Biom. Bull. 1945, 1, 80–83.
32. Ritchie, M.E.; Phipson, B.; Wu, D.I.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. Nucleic Acids Res. 2015, 43, e47.
33. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. Genome Biol. 2014, 15, 550.
34. Bartlett, M.S. Properties of Sufficiency and Statistical Tests. Proc. R. Soc. Lond. Ser.-Math. Phys. Sci. 1937, 160, 268–282.
35. Li, C.; Liu, B.; Kang, B.; Liu, Z.; Liu, Y.; Chen, C.; Ren, X.; Zhang, Z. SciBet as a Portable and Fast Single Cell Type Identifier. Nat. Commun. 2020, 11, 1818.
36. Pearson, K.X. On the Criterion That a given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling. Lond. Edinb. Dublin Philos. Mag. J. Sci. 1900, 50, 157–175.
37. Massey Jr, F.J. The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc. 1951, 46, 68–78.
38. Wang, J.; Wen, S.; Symmans, W.F.; Pusztai, L.; Coombes, K.R. The Bimodality Index: A Criterion for Discovering and Ranking Bimodal Signatures from Cancer Gene Expression Profiling Data. Cancer Inform. 2009, 7, CIN.S2846.
39. Su, K.; Yu, T.; Wu, H. Accurate Feature Selection Improves Single-Cell RNA-Seq Cell Clustering. Brief. Bioinform. 2021, 22, bbab034.
40. Andrews, T.S.; Hemberg, M. M3Drop: Dropout-Based Feature Selection for ScRNASeq. Bioinformatics 2019, 35, 2865–2867.
41. Lin, Y.; Cao, Y.; Kim, H.J.; Salim, A.; Speed, T.P.; Lin, D.M.; Yang, P.; Yang, J.Y.H. ScClassify: Sample Size Estimation and Multiscale Classification of Cells Using Single and Multiple Reference. Mol. Syst. Biol. 2020, 16, e9389.

42. Alquicira-Hernandez, J.; Sathe, A.; Ji, H.P.; Nguyen, Q.; Powell, J.E. ScPred: Accurate Supervised Method for Cell-Type Classification from Single-Cell RNA-Seq Data. Genome Biol. 2019, 20, 264.
43. Tan, Y.; Cahan, P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. Cell Syst. 2019, 9, 207–213.e2.
44. Ji, X.; Tsao, D.; Bai, K.; Tsao, M.; Xing, L.; Zhang, X. ScAnnotate: An Automated Cell-Type Annotation Tool for Single-Cell RNA-Sequencing Data. Bioinforma. Adv. 2023, 3, vbad030.
45. Pliner, H.A.; Shendure, J.; Trapnell, C. Supervised Classification Enables Rapid Annotation of Cell Atlases. Nat. Methods 2019, 16, 983–986.
46. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. Ser. B Stat. Methodol. 2005, 67, 301–320.
47. Domínguez Conde, C.; Xu, C.; Jarvis, L.B.; Rainbow, D.B.; Wells, S.B.; Gomes, T.; Howlett, S.K.; Suchanek, O.; Polanski, K.; King, H.W.; et al. Cross-Tissue Immune Cell Analysis Reveals Tissue-Specific Features in Humans. Science 2022, 376, eabl5197.
48. Shao, X.; Yang, H.; Zhuang, X.; Liao, J.; Yang, P.; Cheng, J.; Lu, X.; Chen, H.; Fan, X. ScDeepSort: A Pre-Trained Cell-Type Annotation Method for Single-Cell Transcriptomics Using Deep Learning with a Weighted Graph Neural Network. Nucleic Acids Res. 2021, 49, e122.
49. Cao, Z.-J.; Wei, L.; Lu, S.; Yang, D.-C.; Gao, G. Searching Large-Scale ScRNA-Seq Databases via Unbiased Cell Embedding with Cell BLAST. Nat. Commun. 2020, 11, 3458.
50. Yang, F.; Wang, W.; Wang, F.; Fang, Y.; Tang, D.; Huang, J.; Lu, H.; Yao, J. ScBERT as a Large-Scale Pretrained Deep Language Model for Cell Type Annotation of Single-Cell RNA-Seq Data. Nat. Mach. Intell. 2022, 4, 852–866.
51. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
52. Zhang, A.W.; O'Flanagan, C.; Chavez, E.A.; Lim, J.L.P.; Ceglia, N.; McPherson, A.; Wiens, M.; Walters, P.; Chan, T.; Hewitson, B.; et al. Probabilistic Cell-Type Assignment of Single-Cell RNA-Seq for Tumor Microenvironment Profiling. Nat. Methods 2019, 16, 1007–1015.
53. De Kanter, J.K.; Lijnzaad, P.; Candelli, T.; Margaritis, T.; Holstege, F.C.P. CHETAH: A Selective, Hierarchical Cell Type Identification Method for Single-Cell RNA Sequencing. Nucleic Acids Res. 2019, 47, e95.

54. Kiselev, V.Y.; Yiu, A.; Hemberg, M. Scmap: Projection of Single-Cell RNA-Seq Data across Data Sets. Nat. Methods 2018.
55. Chen, J.; Xu, H.; Tao, W.; Chen, Z.; Zhao, Y.; Han, J.-D.J. Transformer for One Stop Interpretable Cell Type Annotation. Nat. Commun. 2023, 14, 223.
56. Hossin, M.; Sulaiman, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. Int. J. Data Min. Knowl. Manag. Process 2015, 5, 1.
57. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A Holistic Approach to Semi-Supervised Learning. Adv. Neural Inf. Process. Syst. 2019, 32, 5049–5059.
58. van Engelen, J.E.; Hoos, H.H. A Survey on Semi-Supervised Learning. Mach. Learn. 2020, 109, 373–440.
59. Zhu, X.J. Semi-Supervised Learning Literature Survey. Available online: https://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf (accessed on 25 May 2023).
60. Killamsetty, K.; Zhao, X.; Chen, F.; Iyer, R. Retrieve: Coreset Selection for Efficient and Robust Semi-Supervised Learning. Adv. Neural Inf. Process. Syst. 2021, 34, 14488–14501.
61. Zhang, Z.; Luo, D.; Zhong, X.; Choi, J.H.; Ma, Y.; Wang, S.; Mahrt, E.; Guo, W.; Stawiski, E.W.; Modrusan, Z.; et al. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. Genes 2019, 10, 531.
62. Kimmel, J.C.; Kelley, D.R. Semisupervised Adversarial Neural Networks for Single-Cell Classification. Genome Res. 2021, 31, 1781–1793.
63. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. Ser. B Methodol. 1977, 39, 1–22.
64. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. J. Mach. Learn. Res. 2016, 17, 1–35.
65. Breschi, A.; Muñoz-Aguirre, M.; Wucher, V.; Davis, C.A.; Garrido-Martín, D.; Djebali, S.; Gillis, J.; Pervouchine, D.D.; Vlasova, A.; Dobin, A.; et al. A Limited Set of Transcriptional Programs Define Major Cell Types. Genome Res. 2020, 30, 1047–1059.
66. Bakken, T.; Cowell, L.; Aevermann, B.D.; Novotny, M.; Hodge, R.; Miller, J.A.; Lee, A.; Chang, I.; McCorrison, J.; Pulendran, B.; et al. Cell Type Discovery and Representation in the Era of High-Content Single Cell Phenotyping. BMC Bioinform. 2017, 18, 559.
67. Qiu, P. Embracing the Dropouts in Single-Cell RNA-Seq Analysis. Nat. Commun. 2020, 11, 1169.
68. Schwartz, G.W.; Zhou, Y.; Petrovic, J.; Fasolino, M.; Xu, L.; Shaffer, S.M.; Pear, W.S.; Vahedi, G.; Faryabi, R.B. TooManyCells Identifies and Visualizes Relationships of Single-Cell Clades. Nat.

Methods 2020, 17, 405–413.

69. Comitani, F.; Nash, J.O.; Cohen-Gogo, S.; Chang, A.I.; Wen, T.T.; Maheshwari, A.; Goyal, B.; Tio, E.S.; Tabatabaei, K.; Mayoh, C.; et al. Diagnostic Classification of Childhood Cancer Using Multiscale Transcriptomics. Nat. Med. 2023, 29, 656–666.
70. Cheng, C.; Easton, J.; Rosencrance, C.; Li, Y.; Ju, B.; Williams, J.; Mulder, H.L.; Pang, Y.; Chen, W.; Chen, X. Latent Cellular Analysis Robustly Reveals Subtle Diversity in Large-Scale Single-Cell RNA-Seq Data. Nucleic Acids Res. 2019, 47, e143.
71. Tracy, C.A.; Widom, H. Level-Spacing Distributions and the Airy Kernel. Commun. Math. Phys. 1994, 159, 151–174.
72. Johnstone, I.M. On the Distribution of the Largest Eigenvalue in Principal Components Analysis. Ann. Stat. 2001, 29, 295–327.
73. Patterson, N.; Price, A.L.; Reich, D. Population Structure and Eigenanalysis. PLoS Genet. 2006, 2, e190.
74. Yu, L.; Cao, Y.; Yang, J.Y.H.; Yang, P. Benchmarking Clustering Algorithms on Estimating the Number of Cell Types from Single-Cell RNA-Sequencing Data. Genome Biol. 2022, 23, 49.