OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf207
https://doi.org/10.1093/bib/bbaf207
Review

An overview of computational methods in single-cell transcriptomic cell type annotation

Tianhao Li[1], Zixuan Wang[2], Yuhang Liu[3], Sihan He[1], Quan Zou[4], Yongqing Zhang[1,*]

[1]School of Computer Science, Chengdu University of Information Technology, No. 24 Block 1, Xuefu Road, 610225 Chengdu, China
[2]College of Electronics and Information Engineering, Sichuan University, No. 24 South Section 1, 1st Ring Road, 610065 Chengdu, China
[3]Faculty of Applied Sciences, Macao Polytechnic University, 999078 Macao, China
[4]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Shahe Campus: No. 4, Section 2, North Jianshe Road, 611731 Chengdu, China

*Corresponding author. School of Computer Science, Chengdu University of Information Technology, 610225 Chengdu, China.
E-mail: zhangyq@cuit.edu.cn

Abstract

The rapid accumulation of single-cell RNA sequencing data has provided unprecedented computational resources for cell type annotation, significantly advancing our understanding of cellular heterogeneity. Leveraging gene expression profiles derived from transcriptomic data, researchers can accurately infer cell types, sparking the development of numerous innovative annotation methods. These methods utilize a range of strategies, including marker genes, correlation-based matching, and supervised learning, to classify cell types. In this review, we systematically examine these annotation approaches based on transcriptomics-specific gene expression profiles and provide a comprehensive comparison and categorization of these methods. Furthermore, we focus on the main challenges in the annotation process, especially the long-tail distribution problem arising from data imbalance in rare cell types. We discuss the potential of deep learning techniques to address these issues and enhance model capability in recognizing novel cell types within an open-world framework.

Keywords: scRNA-seq; cell type annotation; long-tail distribution; dynamic clustering; continual learning; open-world cell recognition

Introduction

Single-cell type annotation plays a critically prospective role across various research areas within the biomedical field [1, 2]. Although traditional wet-lab approaches, such as immunohistochemistry and fluorescence-activated cell sorting, are reliable, their lengthy development cycles and high costs pose significant challenges for single-cell annotation research [3, 4]. In contrast, single-cell RNA sequencing (scRNA-seq) technology [5] can precisely capture the high variability in gene expression across single cells in the transcriptome by analyzing mRNA levels in individual cells [6, 7] (as illustrated in Fig. 1A). Based on these gene expression data, computational methods can effectively identify and differentiate between various cell types and states [8], revealing their specific functions within complex tissues [9]. This computational approach offers unprecedented potential for exploring cell population heterogeneity and achieving precise annotation.

In recent years, computational annotation methods have demonstrated high accuracy across extensive gene expression profile datasets [10–13], significantly enhancing the efficiency and reliability of annotation processes (the process as shown in Fig. 1B). Depending on the specific applications of transcriptomic gene expression data, current computational methods can generally be classified into four categories. (i) Specific gene expression-based methods employ known marker gene information to manually label cells by identifying the characteristic gene expression patterns of specific cell types [14]. (ii) Reference-based correlation methods categorize unknown cells into corresponding known cell types based on the similarity of gene expression patterns to those in a preconstructed reference library [15]. (iii) Data-driven reference methods predict cell types by training classification models on pre-labeled cell type datasets [16]. (iv) Large-scale pretraining-based methods use large-scale unsupervised learning to capture the deep relationships between cell types by studying generic cell features and gene expression patterns [17].

Several important reviews have systematically examined the development and application of computational methods for single-cell type annotation modeling. Pasquini *et al.* [18] conducted an in-depth analysis of early annotation methods for scRNA-seq data, with a focus on the evolution of automated annotation strategies, including marker gene databases, correlation analysis, and supervised classification, and their applications in cell type identification. The methods reviewed by Pasquini *et al.* [18] established the foundational framework for single-cell data analysis, providing theoretical and technical support for further methodological improvements. Similarly, Cheng *et al.* [19] comprehensively summarized annotation methods based on gene signatures, the application of feature databases, and the crucial role of supervised learning in automated cell type annotation.

Received: December 11, 2024. **Revised:** March 14, 2025. **Accepted:** April 01, 2025

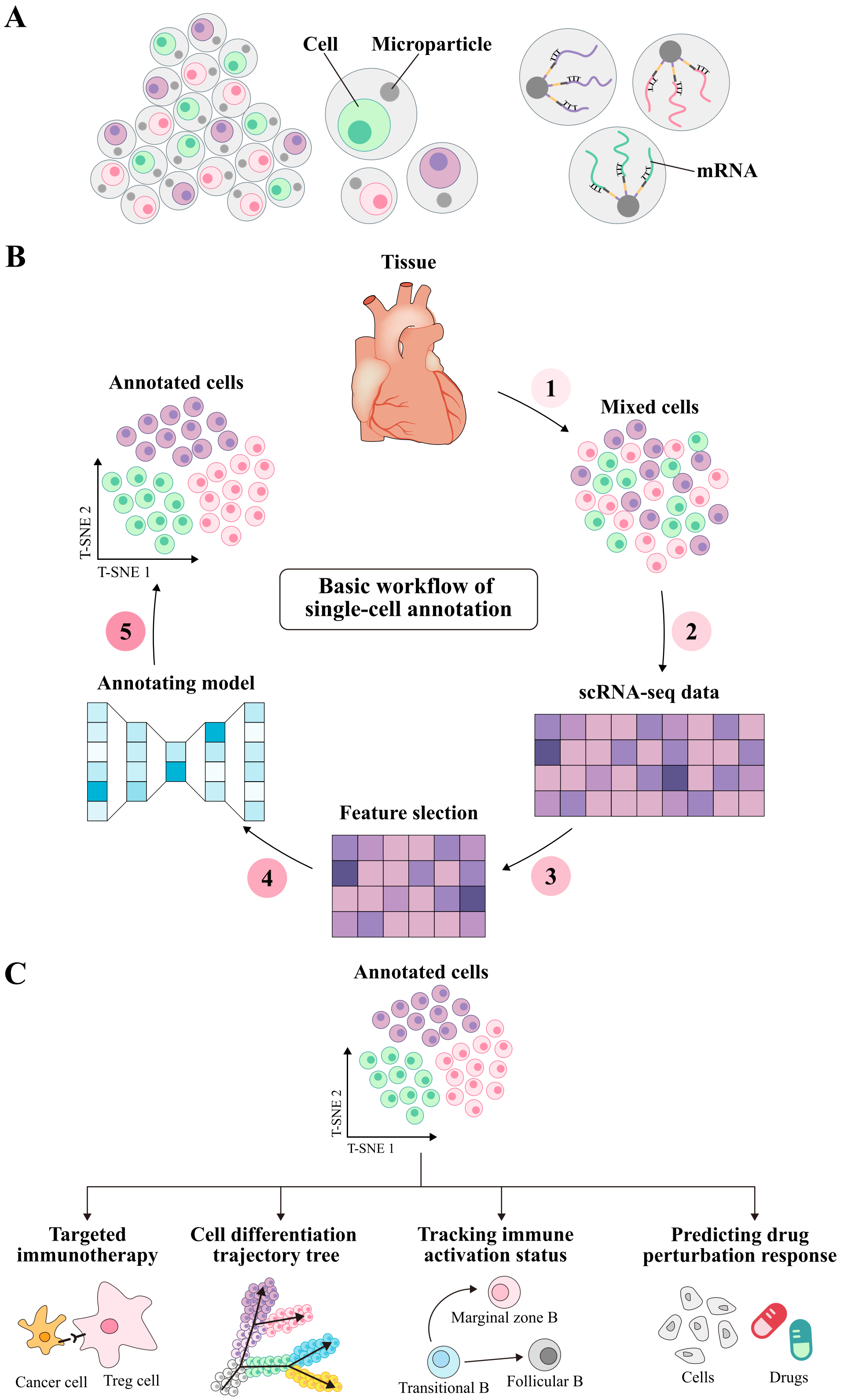

Figure 1. Principles of single-cell annotation. (A) Shows the mRNA extracted from scRNA-seq. As the transcriptional product of cells, mRNA reflects the heterogeneity of gene expression and provides important information for cell type annotation and gene function research. (B) Illustrates the basic workflow of single-cell type annotation. Cells are first extracted from tissue, and single-cell sequencing is performed to obtain a gene expression matrix. High-variance genes are then selected for feature selection. Next, annotation models are used to predict cell types, and the annotation results are finally visualized using dimensionality reduction algorithms such as T-SNE. (C) Demonstrates the applications of single-cell type annotation in various fields, including targeted therapy strategies in the immune microenvironment, cell developmental trajectory reconstruction in developmental biology, immune cell activation state tracking in immunology, and drug perturbation response prediction in precision medicine.

Table 1. Comprehensive databases for cellular and transcriptomic research.

Database	Data type	Species	Info	Tissues/cell types	Ref
HCA	Single cell RNAseq	Human	Multi-organ datasets	33 organs	[28]
MCA	Single cell RNAseq	Mouse	Multi-organ dataset	98 major cell types	[29]
Tabula Muris	Single cell RNAseq	Mouse	Multi-organ datasets	20 organs and tissues	[30]
Allen Brain Atlas	Single nuclei RNAseq	Human and mouse	Brain datasets	69 neuronal cell types	[31]
CellMaker 2.0	Marker genes	Human and mouse	Marker database	467 (human), 389 (mouse)	[32]
PanglaoDB	Marker genes	Human	Marker database	155 cell types	[24]
CancerSEA	Marker genes	Human cancer	Marker database	14 cancer functional states	[33]
Immune Cell Atlas	Single cell RNAseq	Human	Immune cell datasets	Immune system cells	[34]
Human Cell Landscape	Single cell RNAseq	Human	Human atlas of immune cells	Immune cells across tissues	[35]
GEO	RNAseq, microarray	Human, mouse, various	Gene expression profiles	Multiple organs and tissues	[36]
GTEx	RNAseq, genomics	Human	Tissue-specific gene expression	54 tissues	[37]

Their discussion included techniques for improving annotation accuracy through marker gene databases and scoring methods, as well as an analysis of the application of supervised learning in feature selection to optimize model performance and enhance interpretability. These reviews have provided an overview of the current landscape of single-cell type annotation from the perspective of automated annotation strategies and model applications. However, they primarily focus on the frameworks and applications of early methods, with limited discussion of emerging deep learning models, especially in addressing generalization over long-tail distributions [20], open-world data [21], and multi-omics data integration [22]. Consequently, there is a pressing need to integrate the latest computational approaches in single-cell type annotation, delve into the key challenges currently facing the field, and propose potential solutions.

In this work, we provide a comprehensive summary to better understand how to predict single-cell types based on transcriptomic gene features, thereby supporting subsequent single-cell analyses (as shown in Fig. 1C). First, we introduce the existing computational methods for single-cell type annotation, outline the contexts in which each method is applicable, and summarize their primary limitations. Following this, we provide an overview of biological databases used for single-cell type annotation and the processing workflows for scRNA-seq data. Building on this foundation, we explore the key challenges faced by current research and propose potential opportunities to advance single-cell type annotation studies.

Characteristics and challenges of single-cell transcriptomic data

The accumulation of large-scale single-cell transcriptome data has laid the foundation for the rapid development of cell type annotation methods [23]. Marker gene databases, such as PanglaoDB [24] and CellMarker [25], played a crucial role in the early stages by assisting in the identification of known cell types. However, as research progressed, single-cell gene expression profiles, with their comprehensive depiction of cellular heterogeneity, gradually became the core element of annotation models. The combination of marker genes and gene expression profiles has continuously driven the advancement of annotation technologies [26, 27]. Table 1 summarizes the commonly used public databases, which provide vital support for innovation and future exploration in the single-cell field.

Impact of sequencing platforms on cell type annotation

The rapid advancement of scRNA-seq has provided a powerful tool for dissecting cellular heterogeneity, state transitions, and their roles in complex biological processes. At its core, scRNA-seq involves extracting mRNA from individual cells, reverse-transcribing it into cDNA, and obtaining gene expression profiles of single cells through high-throughput sequencing. Compared to traditional bulk RNA-seq, scRNA-seq can resolve subtle differences in gene expression at the single-cell level, enabling precise characterization of cell types, developmental states, and dynamic changes during specific biological processes. This high-resolution sequencing technology has played a crucial role in fields such as tumor microenvironments, immune cell populations, and developmental biology.

Despite the significant advancements in scRNA-seq technology that have enhanced cell type annotation capabilities, differences among sequencing platforms have profoundly impacted annotation outcomes. Various platforms, such as 10x Genomics and Smart-seq, exhibit distinct data characteristics due to differences in their sequencing principles. For instance, 10x Genomics [38] relies on droplet-based encapsulation for high-throughput sequencing, enabling rapid profiling of large cell populations but often resulting in higher data sparsity. In contrast, Smart-seq [39] employs a full-transcriptome amplification strategy, detecting more genes with higher sensitivity, which aids in identifying rare transcripts. However, these technical differences worsen key challenges in scRNA-seq: sparsity, heterogeneity, and batch effects. In cross-platform applications, these factors frequently result in inconsistent annotation performance.

Specifically, the lower gene detection rate of the 10x Genomics platform may hinder the model's ability to capture key marker genes of rare cell types, while the Smart-seq platform, capable of detecting more genes, may reveal finer-grained cell subpopulations that exceed the classification capacity of pre-trained models. Additionally, differences in sequencing depth, primer bias, and other factors often result in significant batch effects across platforms, compromising the comparability of gene expression profiles. Without effective preprocessing strategies, such as batch correction or cross-platform normalization, these systemic biases can directly undermine the model's generalization ability. Collectively, these issues contribute to the reduced stability of existing annotation models in diverse data environments, representing one of the core challenges in scRNA-seq data analysis.

Dynamic updates and sustainability of marker gene

Marker genes play a central role in single-cell research, with their specific expression significantly enhancing the accuracy of cell type annotation and functional analysis. For example, CD133, as a stem cell marker [40], is widely used in stem cell identification and behavioral studies [41, 42], while CD3 [43] and CD19 [44] are classical markers for T cells and B cells, respectively, forming the foundation for the classification and functional analysis of immune cells. These marker genes, through stable and specific expression, provide researchers with a quick and reliable means of analyzing complex cell populations. However, existing marker gene databases, such as CancerSEA [33], CellMarker 2.0 [32], and PanglaoDB [24], have notable limitations, including the absence of certain marker genes, outdated data, and a lack of consistency across samples, which restrict their performance in handling novel cell types or rare cell populations.

In recent years, the introduction of deep learning technologies, such as the self-attention mechanisms of Transformer [45] models, has shown significant advantages in gene selection and feature discovery. For instance, methods like SCTrans [46] leverage attention mechanisms to capture gene combinations that are frequently focused on in gene expression profiles, identifying specific genes highly consistent with marker gene databases and expanding the understanding of previously unseen cell types. This approach not only compensates for the shortcomings of marker gene databases but also provides a powerful tool for discovering new marker genes in an open-world context. In the future, combining the automatic feature selection capabilities of deep learning models with biological validation from experts will enable the dynamic updating of marker gene databases, thereby continuously improving their utility and accuracy in single-cell annotation. This direction will provide essential support for identifying unknown cell types and analyzing complex cellular heterogeneity.

Data preprocessing before annotation

The preprocessing pipeline in single-cell data analysis forms the foundation for ensuring the accuracy of cell type annotation. First, quality control (QC) is performed by evaluating metrics such as the number of detected genes [47], total molecule count, and the proportion of mitochondrial gene expression, thereby eliminating low-quality cells and technical artifacts. Data filtering further refines the dataset by removing noise samples, such as doublets or high-noise cells, thereby improving data quality [48]. Next, normalization removes technical biases, ensuring that gene expression levels are comparable across different cells, thus enabling cross-sample analysis for annotation models [49]. Finally, feature selection identifies highly variable genes (HVGs), highlighting gene expression signals relevant to cell-type specificity and providing essential inputs for capturing biological heterogeneity in models [50]. Figure 2 illustrates this systematic preprocessing workflow, emphasizing the critical role of each step in enhancing single-cell annotation accuracy.

Batch effect correction methods

The sparsity of scRNA-seq data primarily arises from both technical noise, such as low mRNA capture efficiency, and biological factors, including the absence of low-abundance transcripts. This results in a high proportion of zero values in the gene expression matrix, which interferes with the identification of rare cell types and weakens the accuracy of gene co-expression network construction. To address this issue, researchers have proposed multi-level solutions. SCTransform [51] corrects technical biases by modeling the mean-variance relationship of gene expression, effectively reducing the influence of sequencing depth on data quality. Discriminative component analysis (DCA) [52] mitigates data sparsity by leveraging intercellular expression similarity to impute missing values, thereby improving the detection of rare cell types. Additionally, dimensionality reduction methods like PHATE [53] enhance the topological structure of the data, optimizing cell trajectory inference.

Beyond sparsity, the high heterogeneity and batch effects in scRNA-seq data present fundamental analytical challenges. Differences in sequencing platforms, such as the droplet-based 10x Genomics and the full-transcriptome Smart-seq, introduce significant platform-specific variations that exacerbate data heterogeneity. Further discrepancies arise from variations in experimental batches, sample sources, and sequencing depth, leading to batch effects that complicate the direct integration of scRNA-seq datasets from different experiments.

To address these challenges, researchers have developed various cross-batch integration strategies. Mutual nearest neighbors (MNN) [54] constructs a linear mapping model by pairing cells across datasets to eliminate nonlinear shifts, making it particularly effective for small-scale batch differences. Harmony [55] applies iterative soft clustering and latent space alignment to remove systematic technical biases while preserving biologically meaningful variation. The Seurat [56] integration tool utilizes canonical correlation analysis (CCA) to identify dataset-wide anchors and employs a shared nearest neighbor (SNN) graph to achieve robust integration of high-dimensional sparse data [57].

Experimental results indicate that combining these methods, such as applying SCTransform for normalization before integrating data with Harmony, significantly improves data retention, enhances the resolution of downstream clustering, and effectively mitigates batch shifts across platforms. However, excessive imputation may introduce spurious associations, underscoring the need for cross-validation strategies, such as holding out a gene validation set, to strike a balance between data completeness and biological authenticity.

Methods of single-cell type annotation

Single-cell type annotation plays a crucial role in unraveling cellular heterogeneity and advancing single-cell analysis [58, 59]. With the continuous advancement of computational methods, annotation approaches have diversified, resulting in several primary strategies [18, 60]. Currently, these methods can be categorized into four major types. In the following sections, we will discuss the representative models within each category in detail, examining the specific problems they address, their applicable contexts, and their respective strengths and limitations. Furthermore, we have consolidated these four methods into the two annotation workflows illustrated in Fig. 3: one relying on specific gene databases as advisory resources, and the other leveraging previously annotated cell type samples as references.

Methods based on specific gene expression

In single-cell transcriptomics, specific gene markers are categorized into marker genes and gene signatures [61, 62]. Accordingly, cell annotation methods based on specific gene expression can be classified into two distinct approaches. The marker gene approach

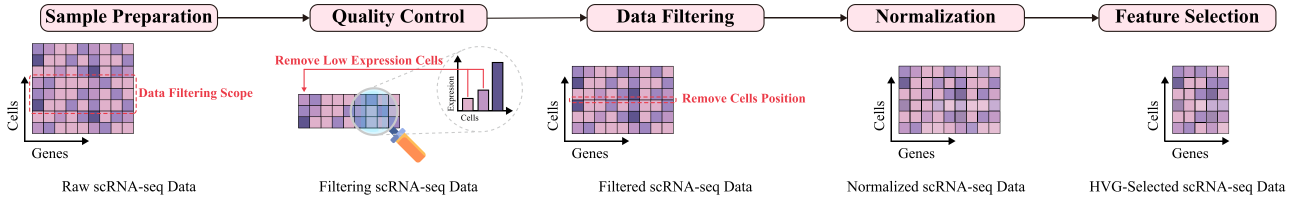

Figure 2. Data preprocessing workflow for single-cell type annotation. Sample data undergo quality control to identify and remove cells with low expression or those requiring exclusion for other reasons. Subsequently, the remaining cell data is subjected to log normalization, and a specific number of highly variable genes are selected based on task requirements, completing the core steps of data preprocessing.

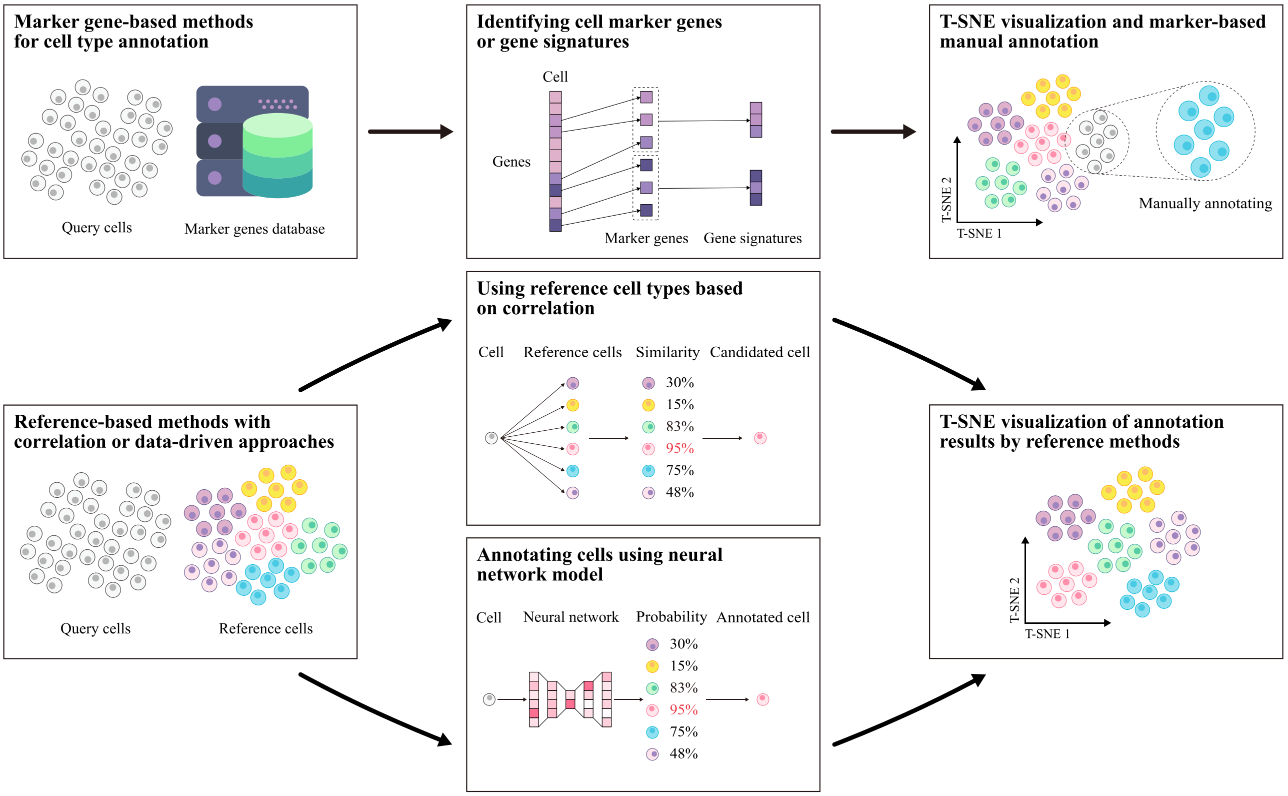
Marker gene-based methods for cell type annotation
Query cells
Marker genes database
Identifying cell marker genes or gene signatures
Cell
Genes
Marker genes
Gene signatures
T-SNE visualization and marker-based manual annotation
T-SNE 2
T-SNE 1
Manually annotating
Using reference cell types based on correlation
Cell
Reference cells
Similarity
Candidated cell
30%
15%
83%
95%
75%
48%
Reference-based methods with correlation or data-driven approaches
Query cells
Reference cells
T-SNE visualization of annotation results by reference methods
T-SNE 2
T-SNE 1
Annotating cells using neural network model
Cell
Neural network
Probability
Annotated cell
30%
15%
83%
95%
75%
48%

Figure 3. Flowchart of single-cell type annotation methods. This chart depicts two main workflows: one using specific gene databases and the other referencing annotated cell-type samples. The specific gene-based method clusters cells and uses marker genes for annotation, while the reference-based method matches cell data to reference databases via correlation or data-driven models. Results are visualized with dimensionality reduction techniques like t-SNE.

relies on the specific expression of a single gene within a particular cell type, typically used for rapid differentiation of well-defined cell types [63, 64]. In contrast, the gene signature approach identifies a set of genes co-expressed within a given cell type, offering a more comprehensive characterization of cell features [65, 66]. This method is particularly advantageous for the identification of cell subtypes and low-abundance cell populations. The standard schematic of these methods is presented in Fig. 4. Table 2 and the subsequent sections provide a detailed overview of these techniques.

Marker gene-based methods

Marker gene-based cell annotation methods typically combine unlabeled data with partially annotated information to accommodate dataset complexity. These methods leverage gene expression patterns for precise cell type identification but face challenges with complex cell populations where cellular subtypes show minimal differences or high data noise [2, 67]. In open world settings, where new cell types lacking marker genes emerge, the identification accuracy of traditional methods declines [68]. Rare cell types with long-tail distributions are also prone to being overlooked in the annotation process.

To address these challenges, a series of improved methods have emerged in recent years, which can be categorized into clustering-based methods, such as Seurat [69], and probabilistic model-based methods [70, 71], including CellAssign [14] and scSorter [68]. Among these, Seurat, which annotates cell types based on clustering and marker genes, remains the most reliable approach. Seurat first normalizes and performs dimensionality reduction (e.g. PCA, UMAP) on single-cell data, followed by clustering analysis to group cells. It then identifies marker genes for each cluster through differential expression analysis and compares them with known marker genes. By integrating prior biological knowledge, Seurat ultimately assigns cell types to each cluster. CellAssign combines a Bayesian probabilistic model with the expectation-maximization (EM) algorithm [72] to compute the posterior probability of each cell belonging to different cell types based on a predefined set of marker genes, thereby enabling cell type

Table 2. Techniques for single-cell type annotation methods based on specific gene expression, including their key algorithm, programming language, and feature and input characteristics.

Model	Year	Key algorithm	Language	Feature and input characteristics
scSorter [68]	2021	Marker gene scoring	R	Focuses on marker gene expression, tailored for scRNA-seq data
SCINA [76]	2019	Expectation-maximization (EM), bimodal distribution	R	Marker gene scoring with bimodal distribution, designed for scRNA-seq
Seurat v3 [56]	2019	kNN graph, transfer learning	R	Integrates multi-omics data (scRNA-seq, scATAC-seq), supports cross-platform fusion
ScType [64]	2022	Louvain clustering	R	Focused on cancer cell annotation, optimized for scRNA-seq datasets
CellID [77]	2021	SVD, multiple correspondence analysis (MCA)	R, C++	Gene signature identification, applies to scRNA-seq
CellAssign [14]	2019	Bayesian inference for marker-based classification	R	Marker-based probabilistic classification, suitable for scRNA-seq
scCATCH [79]	2020	Evidence-based marker scoring	R	Focuses on evidence-weighted marker gene prioritization, suitable for scRNA-seq

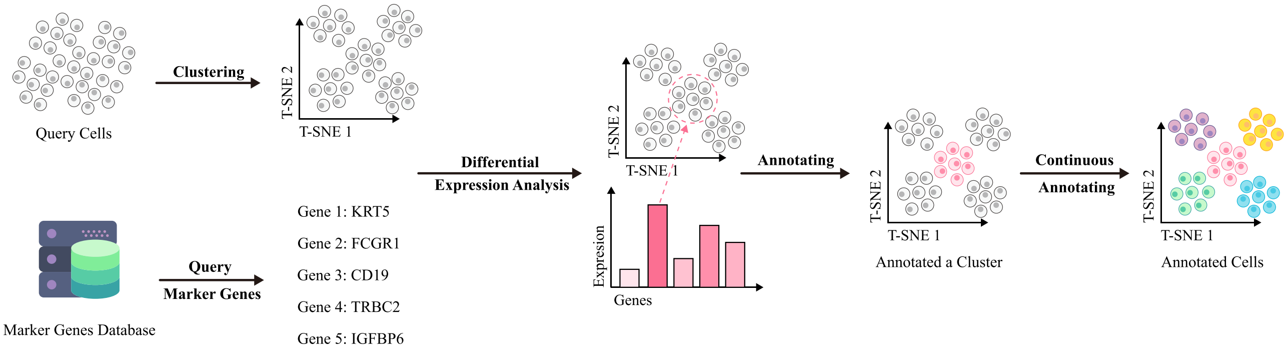

Figure 4. Basic workflow of annotation methods based on specific gene expression. Cell samples are first clustered using a clustering algorithm, then specific cell types within each cluster are identified by querying differential expression genes from biomarker databases.

assignment. Additionally, it supports an “unassigned” state [73], allowing the model to recognize novel cell types that may not be included in the predefined marker gene list, making it suitable for large-scale and complex datasets. However, its performance may be limited when marker genes are missing, expression noise is high, or when dealing with rare and previously unseen cell types. scSorter constructs a semi-supervised classification framework, leveraging both marker and non-marker gene expression information to enhance classification robustness. While maintaining marker gene guidance, scSorter also incorporates auxiliary information from non-marker genes, improving its ability to classify cells. In particular, when marker gene expression is low or data sparsity is high, scSorter remains effective in capturing cell type characteristics and enhances the identification of lowly expressed marker genes.

While these methods advance complex cell type annotation, further improvement is needed to effectively address challenges with new cell types in open-world contexts, rare cell types, and incomplete marker genes. Future research may focus on expanding marker gene databases and developing more robust algorithms to tackle these challenges.

Gene signature-based methods

Gene signature-based cell annotation methods are an evolution of traditional marker gene approaches, aiming to overcome limitations associated with relying on a single specific gene [74, 75]. By integrating a group of co-expressed genes, gene signature methods provide a more comprehensive cellular profile, enabling more accurate annotation of complex cell types and their subtypes. SCINA [76] and CellID [77] exemplify leading strategies in this area. SCINA employs a semi-supervised algorithm that combines gene signatures with an EM strategy [72], effectively enhancing the detection of distinct cellular characteristics and excelling in annotating low-abundance cell types. CellID, on the other hand, uses multiple correspondence analysis (MCA) [78] for dimensionality reduction, preserving the diversity of gene expression patterns and achieving greater stability and consistency across various experimental conditions and parameter settings, which is especially important for cross-dataset analysis. Despite the improved annotation accuracy brought about by enhanced gene feature detection, gene signature methods still encounter critical challenges. On one hand, in identifying rare cell types with long-tail distributions, gene expression heterogeneity can limit their performance. In open-world settings lacking known gene combinations, gene signature methods, being similar to traditional marker gene approaches, demonstrate limited adaptability for recognizing unknown cell types. Overall, future research should focus on optimizing these methods by addressing data heterogeneity to develop more precise and broadly applicable annotation approaches.

Methods based on reference and correlation analysis

The correlation-based reference methods for cell type annotation infer cell types by evaluating gene expression similarities between target cells and known reference datasets (refer to

Table 3. Techniques for single-cell type annotation models based on correlation methods, including their approach, programming language, and key descriptions.

Model	Year	Technology	Language	Description
CHETAH [10]	2019	Classification Tree	R	Uses a classification tree for annotation, identifying novel cell types.
SingleR [15]	2019	Spearman correlation	R	Calculates Spearman correlation for matching.
scamp [11]	2018	Correlation, KNN	R	Combines cosine, Spearman, and Pearson correlation with KNN.
Cell BLAST [84]	2020	GAN	Python	Employs a generative adversarial network for low-dimensional embeddings and unseen cell identification.
scMatch [86]	2022	Correlation	Python	Leverages Spearman and Pearson correlation for large-scale datasets.
scLearn [90]	2020	DCA	R	Uses discriminative component analysis with automatic threshold selection.
ClustifyR [91]	2020	Correlation	R	Integrates multiple data sources with Spearman, Pearson, Kendall, and cosine correlation.

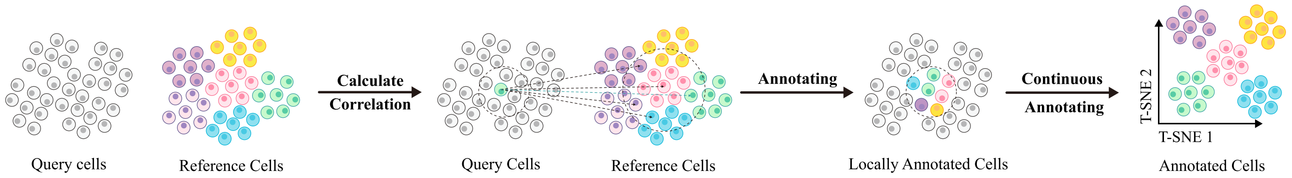

Figure 5. Basic workflow of reference-based annotation methods using correlation. The process begins by establishing correlation relationships between the query cells to be annotated and the reference cell samples. The most similar reference cells are then selected as the basis for determining the cell types of the query cells. This workflow is subsequently extended to annotate all query cell samples.

Table 3). These methods are generally categorized into two strategies: single-cell similarity analysis and centroid-based similarity analysis. The former is ideal for high-resolution single-cell annotation, while the latter is better suited to large-scale cell population analysis. Common similarity metrics, including Pearson correlation coefficient [80], Spearman rank correlation coefficient [81], and cosine similarity [82], offer precise quantification of expression profile similarities across cells. Figure 5 provides an intuitive and illustrative representation of the basic workflow of reference-based annotation methods using correlation.

Early correlation-based reference tools like scmap [11] used K-nearest neighbor (KNN) [83] algorithms to match cell types, annotating based on similarity measures. However, when dealing with complex and highly heterogeneous tumor samples, these methods faced considerable uncertainty. To address this limitation, improved tools have been developed. For instance, CHETAH [10] employs a hierarchical classification tree to progressively match cell types, enhancing its capacity to analyze high-heterogeneity samples, particularly in tumor classification. Cell BLAST [84], on the other hand, introduces generative adversarial networks (GANs) [85] to dynamically adjust the model to new data, demonstrating strong adaptability in multi-source data integration scenarios. scMatch [86] tackles the challenge of annotating low-coverage scRNA-seq data by computing gene expression similarity with large reference datasets like FANTOM5 [87], thus improving robustness to high-dimensional sparse data.

Although these methods represent a significant improvement over traditional marker gene-dependent models, avoiding the limitations of over-relying on databases, and making progress in multi-source data integration and high heterogeneity data classification, they still face several challenges. Specifically, the generalization ability of current methods remains insufficient, particularly in handling batch effects across sequencing standards and species. Therefore, future research could focus on incorporating the concept of continual learning [88, 89], expanding the available scRNA-seq datasets for reference, and enhancing the generalization and continual learning capabilities of correlation-based reference methods.

Methods based on data-driven references

Data-driven methods leverage extensive datasets to enable machine learning models to automatically extract features for cell type annotation. In contrast to specific gene expression and correlation-based reference methods, data-driven approaches offer superior flexibility, capable of autonomously uncovering complex patterns within data. This adaptability effectively addresses limitations of traditional methods in capturing cellular diversity and complexity [17]. Conventional approaches rely heavily on manually selected marker genes or predefined reference sets, making it challenging to comprehensively represent high-dimensional data, often leading to the omission of rare cell types [92]. By contrast, data-driven methods achieve substantial gains in annotation accuracy and generalizability through deep feature extraction [93]. Figure 6 illustrates the basic implementation workflow of such methods, while Table 4 present the advantages and applicability of these methods across various implementation strategies.

In the early stages of single-cell annotation research, traditional machine learning methods, such as support vector machines (SVM) [94] and random forests (RF) [95], were widely applied. For instance, representative methods like scPred [96] and SingleCellNet [97] utilized SVM and RF classifiers to analyze gene expression data. Compared to approaches based on marker genes and correlation, these machine learning strategies exhibited greater flexibility and efficiency. By leveraging supervised learning to extract features from annotated data, these methods effectively mitigated noise and sparsity in gene expression data to some extent, demonstrating strong performance on early single-cell datasets. However, their ability to handle data sparsity heavily relied on feature engineering, particularly the selection of highly variable genes (HVGs). In 2017, McCarthy *et al.* [98] proposed a standard HVG selection procedure that

Table 4. Techniques for single-cell type annotation models based on data-driven reference methods, including their approach, programming language, characteristics, and learning types.

Model	Year	Technology	Language	Characteristics	Learning Type
mtANN [16]	2023	AE, Ensemble model	Python	Ensemble learning, multi-model approach	Supervised learning
scMMT [117]	2024	Convolutional neural network (CNN)	Python	CITE-seq, scRNA-seq, protein prediction	Supervised learning
scMGCN [118]	2024	Graph convolutional network (GCN)	Python	Multi-view learning, single-cell data integration	Semi-supervised learning
scSemiCluster [103]	2020	Deep clustering algorithm	Python	Structural regularization, clustering	Semi-supervised learning
TOSICA [13]	2023	Transformer	Python	Transformer architecture, interpretable annotation	Supervised learning
CAMLU [119]	2022	AE, support vector machine (SVM)	R	Iterative feature selection, novel cell identification	Supervised learning
scAnno [120]	2023	Deconvolution	R	Supervised classification, cell type identification	Supervised learning
scDeepSort [102]	2021	Graph neural network (GNN)	Python	Pre-trained model, weighted GNN	Supervised learning
scTransSort [105]	2023	Transformer, CNN	Python	Gene expression embeddings, data sparsity reduction	Supervised learning
TripletCell [121]	2023	k-nearest neighbors (KNN)	Python	Deep metric learning, triplet loss	Supervised learning
CALLR [122]	2021	Laplacian, logistic regression	R	Graph Laplacian, sparse logistic regression	Supervised learning
scGAD [107]	2023	K-means	Python	Generalized annotation, clustering labels	Unsupervised learning
SciBet [12]	2020	Multinomial distribution model	R, C++	Multinomial distribution, maximum likelihood estimation	Supervised learning
scDeepInsight [123]	2023	CNN	Python	Image transformation, supervised annotation, data integration	Supervised learning
CIForm [106]	2023	Transformer	Python	Transformer, patch concept, computational complexity reduction	Supervised learning
scPred [96]	2019	SVM	R	Unbiased feature selection, probabilistic machine learning	Supervised learning
ItClust [124]	2021	Confidence score	Python	Iterative transfer learning, fine-tuning	Transfer learning
scGCN [125]	2021	GCN, mutual nearest neighbors (MNN)	Python	Semi-supervised GCN, mixed graph	Semi-supervised learning
scNym [126]	2021	Generative adversarial network (GAN)	Python	Adversarial training, pseudo-labels	Unsupervised learning
ACTINN [127]	2020	Artificial neural network (ANN)	Python	Minimal prior knowledge, flexible learning	Supervised learning
SingleCellNet [97]	2019	Random forest (RF)	R	Top-pair transformation, discriminative gene pairs	Supervised learning
scArches [109]	2022	Variational autoencoder (VAE)	Python	Transfer learning, efficient construction	Transfer learning
scNAME [128]	2022	K-means	Python	Contrastive learning, neighborhood-based methods	Unsupervised learning
scLearn [90]	2020	Discriminative component analysis (DCA)	R	Threshold selection, novel cell identification	Supervised learning
SC3 [129]	2017	K-means	R	Gene filtering, consensus clustering	Unsupervised learning
scziDesk [130]	2020	AE, soft K-means	Python	Denoising autoencoders, soft K-means, clustering	Unsupervised learning
SCTrans [46]	2024	Transformer	Python	Multi-scale Transformer, gene sub-vectors	Supervised learning
scEvolve [113]	2024	Prototypical contrastive replay	Python	Forgetting mitigation, memory buffer	Continual learning
scTab [131]	2024	Transformer	Python	Feature attention, data augmentation	Supervised learning
scPOT [111]	2023	Optimal transport (OT)	Not found	Novel type discovery, automatic cell type count estimation	Supervised learning
scDET [132]	2024	AE, K-means	Not found	Distribution-independent framework, contrastive learning, long-tail identification	Unsupervised learning

retained genes with the highest coefficient of variation across cells, typically comprising 10%–20% of all genes. This approach efficiently filtered out low-information loci, reducing the feature dimensionality of RF classifiers by 80%–90% while maintaining over 90% classification accuracy on normalized datasets [99]. This strategy was later adopted by deep learning methods, such as single-cell variational inference (scVI) [100], where the encoder preferentially processed the HVG subset. Although

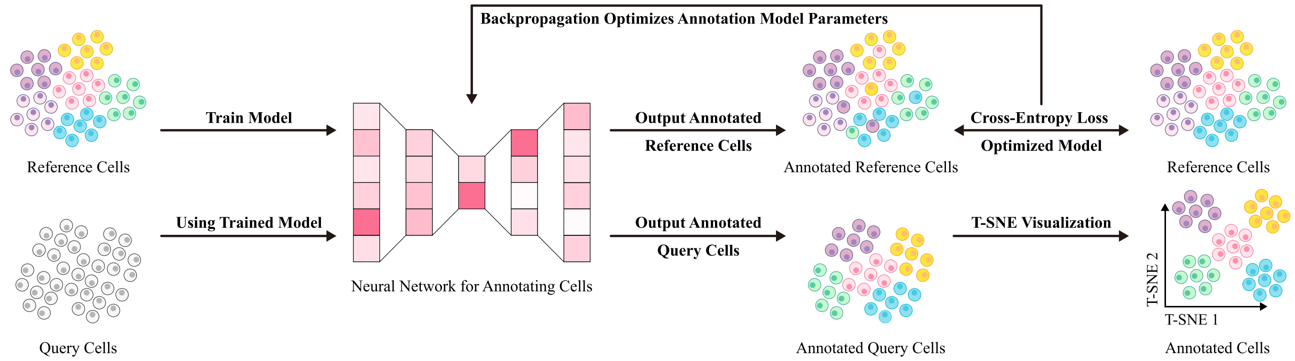

Figure 6. Basic workflow of data-driven reference methods. First, reference data with well-annotated labels are input into a neural network model for training, enabling the model to learn to identify cell types based on gene expression differences under a supervised learning paradigm. Next, query cell data are fed into the trained neural network model to achieve precise cell type annotation.

HVG selection alleviated certain sparsity issues, the increasing throughput of single-cell sequencing has introduced new challenges, particularly the issue of zero inflation in cross-platform data integration. For example, in T-cell subtype classification, training on mixed datasets from 10x Genomics and Smart-seq2 platforms resulted in a 15%–22% decline in the recall rate of SVM [101]. Furthermore, these methods have gradually revealed limitations in batch effect correction, adaptation to cross-dataset distribution shifts, and the identification of rare cell types.

To overcome these challenges, the advent of deep learning has driven substantial advancements in single-cell type annotation. Deep learning enables automatic feature extraction, addressing traditional machine learning methods' deficiencies in batch effect control and generalization. For example, scDeepSort [102] employs a weighted graph neural network to handle complex inter-data relationships, significantly enhancing annotation accuracy without the need for additional reference data. Similarly, scSemiCluster [103] utilizes semi-supervised learning and structural similarity regularization to further mitigate batch effect issues and improve adaptability to diverse datasets. However, while deep learning has advanced generalization performance, it still faces challenges in capturing rare cell types within long-tail distributions [104]. These models often exhibit a tendency to focus on mainstream features in the data, with limited attention to the feature expressions of rare types.

Addressing this, Transformer [45] models have gradually entered the field of single-cell annotation, offering new strategies to address the challenges of rare cell types within long-tail distributions. The self-attention mechanism of Transformers allows them to flexibly focus on critical features in the data, making them particularly suited for capturing the feature expressions of rare cell types. For example, mtANN [16] and TOSICA [13] integrate self-attention mechanisms with multi-gene selection strategies, significantly enhancing the recognition of rare cell types. scTransSort [105] further optimizes sparse data handling, enabling the model to extract more comprehensive feature representations, thereby improving annotation efficiency and robustness. Additionally, CIForm [106] introduces a "patch" concept, effectively reducing computational complexity, thus providing new methods for large-scale single-cell data analysis. Overall, the Transformer architecture not only strengthens long-tail distribution recognition but also enhances accuracy in cell annotation tasks.

Beyond the issue of long-tail distributions, single-cell annotation must also contend with the challenge of identifying unknown cell types in an open-world context. In response, semi-supervised and unsupervised learning strategies are being explored. scGAD [107] uses K-means [108] clustering to generalize potential unknown cell types, allowing the model to distinguish new cell types rather than merely labeling them as "unassigned." Moreover, scArches [109] combines variational autoencoders (VAE) [110] with transfer learning to generate cross-platform reference maps, further enhancing model generalizability across different data platforms. In addition, scPOT [111] employs an optimal transport (OT) [112] framework to accurately annotate and identify unknown cell types, providing an innovative solution for rare type recognition within open sets.

Meanwhile, data-driven methods often exhibit limited flexibility when applied to unseen or external datasets. These methods are susceptible to overfitting the training data, making it difficult to maintain stable and high performance on novel datasets. In contrast, unsupervised approaches based on marker genes or gene signatures typically demonstrate greater robustness and adaptability when processing new data. With the integration of continual learning into the single-cell field, scEvolve [113] represents the first model to achieve single-cell incremental learning [114] and improve predictive generalization through data replay. Extensive evaluations on a series of rigorously curated benchmark datasets consistently demonstrate that scEvolve can continuously assimilate scRNA-seq data from different batches and sequencing platforms over prolonged periods, effectively identifying diverse cell types across various tissues. Furthermore, it alleviates the overfitting risks and generalization limitations inherent to data-driven methods while mitigating catastrophic forgetting when incorporating new datasets. Thus, continual learning provides a promising avenue for advancing data-driven methodologies, fostering enhanced flexibility and superior generalization capabilities.

Despite significant improvements in annotation accuracy and generalizability, data-driven methods' reliance on data quality still poses a risk of information loss. Future directions include integrating multi-omics data to address information gaps, leveraging self-supervised learning [115] to maximize the utility of unlabeled data, applying knowledge distillation [116] to facilitate cross-model knowledge transfer, and adopting continual learning to enhance model adaptability to new data. These advancements aim to provide richer contextual information for single-cell annotation, further improving model adaptability and accuracy, and delivering more comprehensive and flexible solutions for cell type identification.

Table 5. Techniques for single-cell type annotation models based on large-scale pretraining methods, including their approach, programming language, parameter size, input modality, multi-task capabilities, and explainability.

Model	Year	Technology	Language	Parameters (Estimation)	Input Modality	Multi-task (Tasks)	Explainability (From papers)
scBERT [17]	2022	BERT	Python	5M	scRNA-seq	No	Yes, attention weights for gene relevance, identifying key genes for cell types
scGPT [133]	2024	GPT	Python	38M	Multi-omics (scRNA-seq, scATAC-seq, protein)	Yes, annotation, perturbation analysis, multi-batch integration	Yes, gene pathway interpretation via latent features, identifying gene interactions
scFoundation [134]	2024	Transformer	Python	100M	scRNA-seq	Yes, annotation, clustering, drug response prediction	No, but provides cell and gene embeddings for downstream analysis
scRobust [135]	2024	Transformer	Python	18M	scRNA-seq or scATAC-seq	Yes, annotation, drug tolerance, scATAC-seq analysis	Yes, maximizing the highly unique genes of each cell

Methods based on large-scale pretraining

To address the common issue of information loss in traditional machine learning methods, large-scale pretraining approaches have emerged as an effective solution [115]. These methods leverage self-supervised learning to extract underlying gene expression patterns and cellular features from vast amounts of unlabeled data, effectively reducing the information loss typically encountered in high-dimensional data processing. By capturing complex relationships and latent structures within the data without requiring manual labeling, self-supervised learning not only compensates for missing information but also significantly improves model generalization, enabling the identification of a broader range of complex cellular characteristics (as detailed in Table 5). The basic workflow of this approach is illustrated in Fig. 7.

In recent years, several large-scale pretrained models for single-cell annotation, such as scBERT [17], scGPT [133], and scFoundation [134], have made remarkable advances. Through self-supervised learning, these models extract gene expression patterns and cellular features from large-scale unlabeled data, effectively overcoming the limitations of traditional methods with respect to information loss. A primary advantage of these approaches lies in their reliance on extensive unlabeled data for pretraining [136], which allows models to automatically capture deep structures within data and learn more intricate cellular features, thus enhancing cell type recognition capabilities and circumventing information loss caused by high data dimensionality or limited labeling. Moreover, research shows that larger model parameters often yield better performance, as increased model capacity enables richer feature extraction. Additionally, large-scale pretrained models exhibit strong transferability, demonstrating robustness and adaptability across various tasks and datasets, thereby advancing the field of single-cell annotation.

Although large-scale pretraining methods have made significant advances in improving annotation accuracy and generalization, they still face several challenges. First, these methods require high-quality data [137] and substantial computational resources, particularly when handling large-scale datasets. Second, their generalization remains limited [138] when applied to highly heterogeneous or noisy data, especially across different biological conditions and experimental platforms. Additionally, as model parameters scale up, computational and storage costs increase significantly, restricting their practical feasibility [139]. To address these issues, scRobust introduces strategies such as random gene subset pretraining, multi-task collaborative optimization, a highly unique gene-driven dynamic input mechanism, and a lightweight model architecture. These innovations effectively mitigate the sensitivity of traditional self-supervised methods to data quality, excessive computational demands, and limited cross-platform generalization, providing an efficient and robust solution for single-cell analysis. While large-scale pretraining, which uses self-supervised learning to extract deep gene expression patterns from unlabeled data, reduces information loss in data-driven methods, its limitations in data quality, computational efficiency, and cross-dataset generalization remain unresolved.

Experimental evaluation of single-cell annotation

Evaluation metrics

The performance of single-cell annotation models is typically evaluated based on their performance on test data, to assess the model's applicability to new data. Cross-validation (CV) is commonly used for model evaluation [140], where the data are split into training and test sets. The training data are used for model learning, while the test data are used to assess the model's performance. K-fold cross-validation is a popular method, where the dataset is divided into K equal parts. Each time, one part is selected as the test set, and the remaining $K-1$ parts are used as the training set. This process is repeated K times, with each subset being used as the test set in turn, and the average result from the K tests is taken as the model's evaluation score. To balance computational efficiency and evaluation quality, K is usually chosen as 5 or 10 [141].

In single-cell type classification tasks, classification performance can be measured using various evaluation metrics, most of which are based on a "confusion matrix" that includes four key elements: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Based on these values, key performance indicators such as $Accuracy$, $Precision$, $Recall$, and $F1-score$ can be calculated. The formulas for these calculations are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

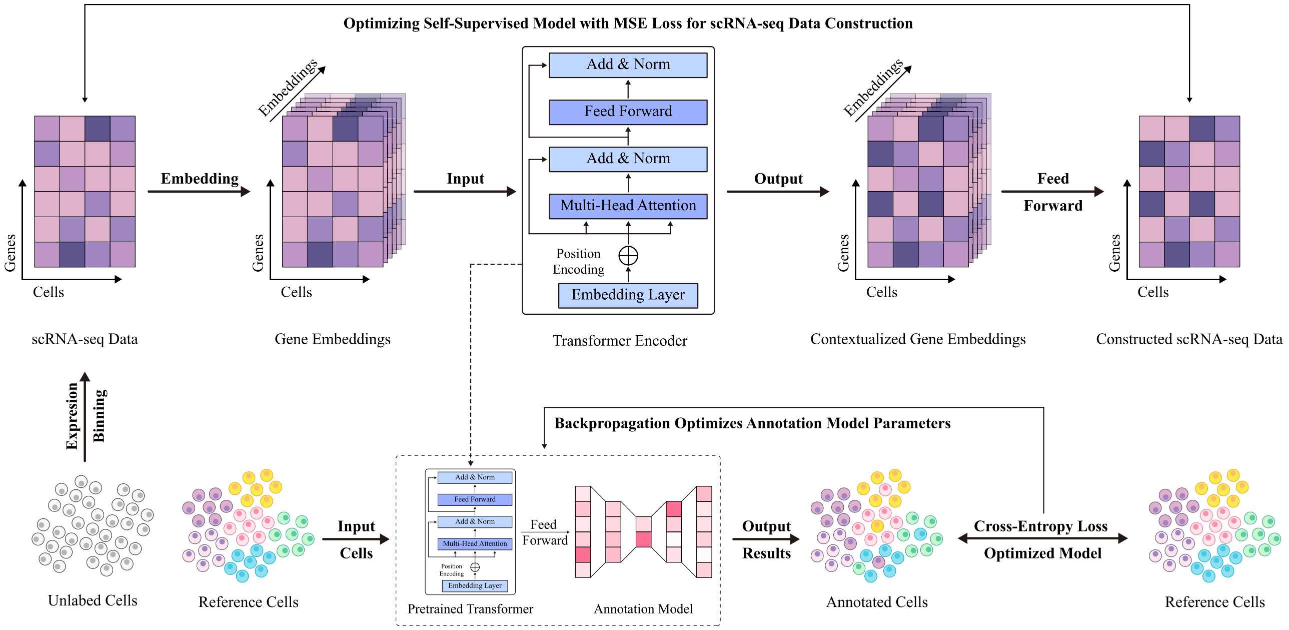

Figure 7. Basic workflow of large-scale pretraining methods. This approach begins by extracting scRNA-seq data from large-scale unlabeled single-cell samples as a comprehensive feature foundation. Using gene embeddings, an encoding-decoding strategy is employed to reconstruct scRNA-seq data in a self-supervised learning framework, while simultaneously pretraining a Transformer encoder as a deep feature extraction model. The pretrained model is then applied to cell type annotation tasks under a data-driven supervised learning paradigm.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

Performance evaluation

In evaluating the performance of various single-cell annotation methods, we adopted the benchmark results reported by Lin *et al.* [46], conducting a comprehensive analysis of these methods across multiple datasets. Figure 8 presents a comparison of their performance in terms of accuracy and F1 scores. The results indicate that deep learning-based approaches, such as SCTrans [46] and scBERT [17], demonstrate a clear advantage, consistently achieving superior performance across diverse datasets and exhibiting exceptional generalization capabilities. By contrast, traditional methods, including Seurat [69] and the gene signature-based CellID [77], show greater variability in performance, particularly with limited adaptability to cross-dataset scenarios.

The boxplot in Fig. 9 further clarifies this trend, revealing that deep learning models show higher stability across datasets, while traditional methods exhibit greater fluctuation. Overall, deep learning methods outperform traditional computational methods in terms of robustness and generalization across multiple datasets, with the latter showing some advantages on certain datasets but lacking overall stability.

Challenges and opportunities

Despite significant progress in single-cell type annotation, several pressing challenges remain, primarily including the issue of long-tail distribution in datasets, the ability to generalize to unseen cell types, and the effective annotation of new sequencing datasets using existing models.

Enhancing single-cell annotation with multi-source data perception

In single-cell type annotation, traditional single-omics methods, due to their reliance on data from a single source, often struggle to fully capture the complex features of cells. For example, scRNA-seq data can reveal transcriptional features of cells but lacks information on other important aspects, such as chromatin accessibility [142] and protein expression [143]. This limitation results in less accurate annotations, particularly for rare cell types or subtypes, especially in tissues with high heterogeneity.

To overcome these limitations, the concept of multi-source perception advocates for the integration of multiple omics data, expanding the model's understanding of cellular features across multiple layers. By leveraging the complementary advantages of various omics sources, models can capture the relationships between them, thus providing a more comprehensive perspective for cell annotation [144]. Currently, methods like scJoint [145] and TotalVI [146] have made progress in this area. For instance, scJoint integrates scRNA-seq and single-cell assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) data into a shared latent space, facilitating the combination of different omics data. TotalVI, based on a variational autoencoder [110] model, combines transcriptomics and proteomics, reducing biases from differences in omics technologies. However, constructing robust latent spaces in high-dimensional, sparse multi-omics data and ensuring effective retention of features from all omics remain major challenges for multi-source perception.

Notably, existing studies have systematically demonstrated that multi-omics integration can significantly enhance the annotation performance of rare cells. For instance, CITE-seq technology, which combines transcriptomic and proteomic data, successfully identified <0.5% circulating NK cells that were

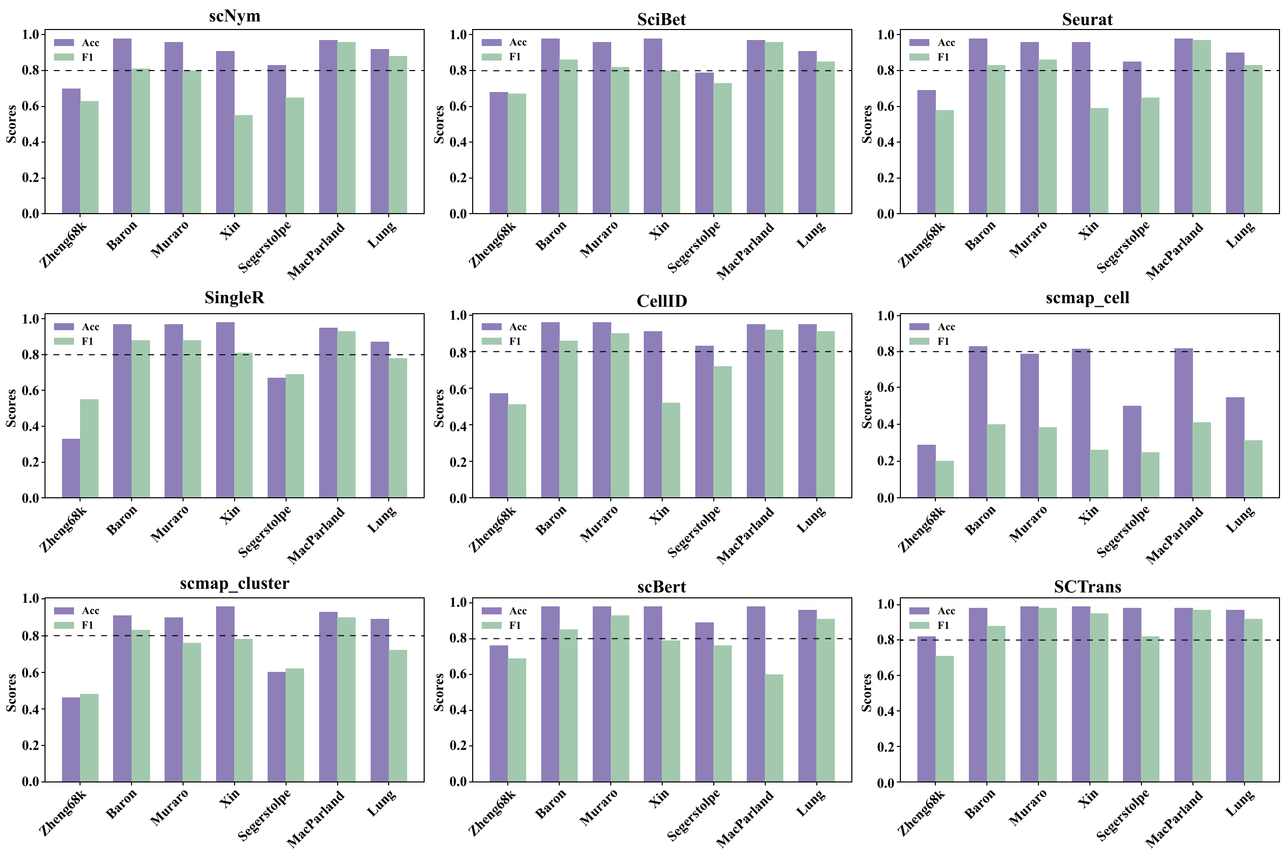

Figure 8. Comparison of annotation performance of different single-cell annotation methods across multiple datasets. The figure presents bar charts evaluating the performance of nine methods on seven benchmark datasets, where a higher bar indicates better performance of the method.

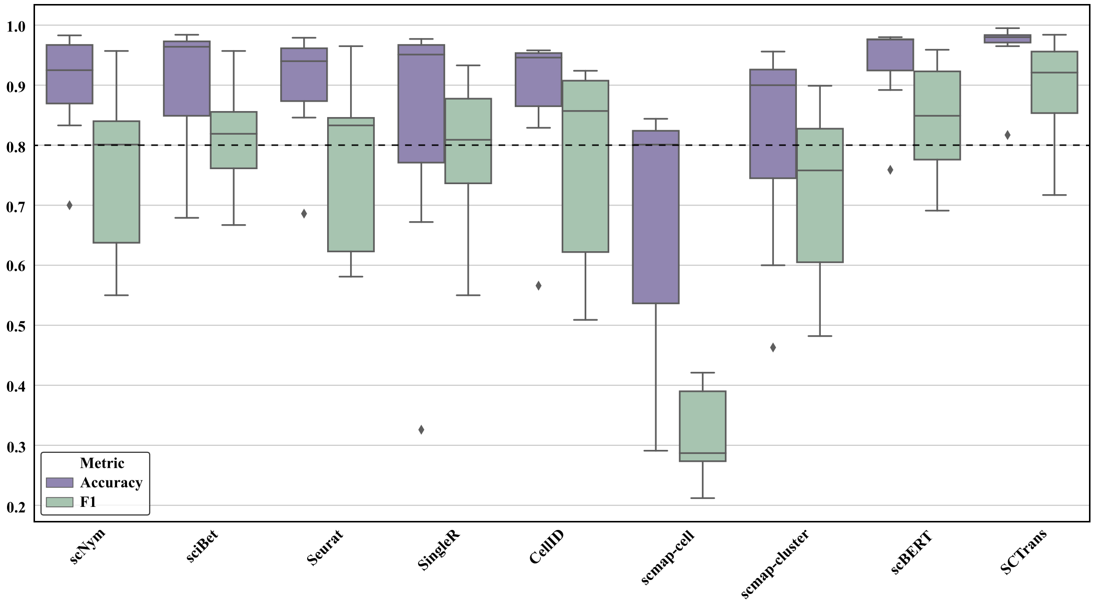

Figure 9. Comparison of stability of different single-cell annotation methods across multiple datasets. The figure shows the stability of each method across different benchmark datasets using boxplots. A higher position and smaller range between the upper and lower quartiles of the boxplot indicate better stability of the method.

previously missed by single-omics approaches [147]. Similarly, after integrating transcriptomic and epigenomic data, the MOFA+ framework improved the clustering purity of 2% endocrine precursor cells from 68% to 92% [148]. In synthetic data testing, multi-omics approaches increased the recall rate for 1% rare cells by 27% compared to single-omics methods [149], and multimodal integration demonstrated statistically significant advantages in identifying low-abundance cell populations (<5%) [57]. These findings suggest that multi-source perception not only expands the feature space but also enhances rare cell resolution through cross-validation of signals across different omics layers.

To further improve the application of multi-source perception in single-cell annotation, strategies such as self-supervised learning [115] and knowledge distillation [116] can be introduced to enhance the model's deeper understanding of multi-omics features. For example, the multi-source perception process could be simulated in the latent space, allowing the model to adaptively learn core features of cell types from each omic layer, thereby preserving the unique information of each omics dataset during integration. Additionally, evaluating cross-dataset generalization ability could improve the adaptability of multi-omics methods across different experimental conditions and technical platforms, ultimately enhancing annotation accuracy and robustness. Such improvements would help multi-source perception methods perform better in identifying heterogeneous samples and rare cell types.

Long-tail distribution and optimization strategies for rare cell type recognition

In single-cell type annotation, the long-tail problem is a significant challenge, referring to the relatively limited sample size of rare cell types within the dataset, which reduces model accuracy in identifying these types. This data imbalance not only affects model generalizability but can also result in the loss of important biological information. To address this challenge, scNAME [128] uses a weighted soft K-means clustering algorithm [108] that groups cells toward the most similar centers, while neighborhood contrastive learning [150] minimizes the distance between homologous cells and maximizes the distance between unrelated ones, enabling distinctive representation of rare cells. Meanwhile, scBERT [17], a large-scale pretrained model, employs a bidirectional performer encoder architecture [151, 152] to capture contextual information in cell expression data, deeply learning cell representations so that the model's attention mechanism focuses on rare cell type distributions, improving the recognition of these cell types.

Although deep learning has made remarkable progress in single-cell annotation, as exemplified by scBERT's enhancement in recognizing rare cell types, current methods remain constrained by high resource demands and limited efficiency. Firstly, supervised classification models like scBERT depend heavily on large amounts of labeled data, which is particularly challenging for rare cell types where samples are already sparse, limiting model performance on long-tail distributions. Secondly, these models often require substantial computational resources and long pretraining and fine-tuning times, significantly increasing training costs and limiting their applicability in resource-constrained settings. Current deep learning methods show insufficient flexibility and adaptability to effectively meet the demands of rare cell type identification in settings with limited data and dynamic environments.

To improve the recognition accuracy of rare cell types, we propose targeted solutions from three perspectives: data volume, feature representation, and learning difficulty. First, meta-learning [153] and few-shot learning [154] effectively address data scarcity. Meta-learning enables the model to quickly adapt to new tasks, allowing it to identify rare cell types with minimal labeled data, while few-shot learning optimizes model structure to maintain efficient learning even with limited data. Second, for feature representation, attention-based feature selection automatically filters marker genes specific to each cell type, constructing an optimized gene set that more accurately captures key characteristics of rare cell types, alleviating the long-tail distribution problem. Lastly, curriculum learning introduces complex tasks in stages, helping the model progressively grasp features of rare cell types, enhancing learning stability and accuracy. Combined, these strategies significantly enhance model performance in scarce data scenarios, advancing single-cell annotation techniques.

Exploring the synergy between dynamic clustering and annotation

In the single-cell type annotation task, balancing clustering and annotation has become a critical issue. In the broader context of cell type annotation, once the model identifies known cell types, unseen cell types are marked as "unassigned," requiring further clustering to identify potential clusters [107]. However, as samples are progressively excluded during the annotation process, the distribution of the remaining data dynamically changes, which affects the stability of clustering, especially when determining the optimal number of clusters. Most existing methods rely on static clustering setups and lack mechanisms for adjusting parameters to account for dynamic changes, making it difficult to optimize clustering number while excluding annotated samples, which ultimately leads to instability in both clustering structure and annotation outcomes. Therefore, the synergy between clustering and annotation is particularly important.

Recent research advancements indicate that clustering methods based on contrastive learning can effectively address this challenge. For instance, scRobust [135], leveraging a self-supervised contrastive learning framework, demonstrates exceptional robustness and adaptability to novel cell types under dynamic data distributions. Experimental results show that in the Zheng 68K dataset, scRobust achieves an identification accuracy of 0.28 for the rare CD4+ T Helper 2 cells, significantly outperforming methods such as Concerto [155], CIForm [106], and TOSICA [13], all of which have accuracy rates below 0.10. Moreover, in the Muraro dataset, scRobust attains a perfect accuracy of 1.0 in identifying epsilon cells, whereas other methods fail to detect this cell type (accuracy = 0). These findings validate the effectiveness of contrastive learning in capturing latent relationships among similar cells, thereby enhancing clustering algorithms' adaptability to data sparsity and dynamic changes, ultimately providing robust technical support for the efficient identification of unannotated cells.

To address this issue, we propose several strategies to optimize the balance between clustering and annotation. First, adaptive clustering algorithms can dynamically adjust the number and structure of clusters, allowing real-time responses to changes in sample distribution and improving the resolution of previously unseen cell types. Second, automatic clustering optimization based on latent features can be employed, utilizing deep learning to extract cellular features and perform clustering in latent space, ensuring the stability of clustering performance even when data is progressively removed. Finally, contrastive learning-based clustering methods serve as another effective strategy, leveraging advanced models such as scRobust to align global gene information with local features, enabling the capture of multidimensional biological characteristics (e.g. cell subtype-specific pathways and sample-specific markers) in sparse data environments. These strategies not only provide novel technical pathways for dynamic optimization but also establish a foundation for improving annotation accuracy and clustering stability, ultimately achieving a synergistic integration of clustering and annotation.

Balancing knowledge retention and adaptation in continual learning with the surge in single-cell data

In the context of the rapid accumulation of single-cell sequencing data, continual learning has become a key strategy to enhance the generalization ability and adaptability of single-cell annotation models [88, 89]. With the continuous increase in new sequencing data and cell types, existing models need to be frequently updated. However, direct re-training is time-consuming and may result in forgetting of prior knowledge. The core of continual learning is enabling models to leverage experience gained from previous tasks to help them learn new tasks, thus allowing knowledge to accumulate over time. By progressively absorbing new data, continual learning helps models retain existing knowledge while adapting to new information, thus expanding their recognition capabilities. This approach is particularly suited for handling the rapidly growing single-cell multi-omics data.

In this context, the scEvolve [113] method was proposed to address the continual learning challenges in single-cell annotation. Based on incremental learning principles, it employs prototype comparison and rehearsal learning strategies to mitigate knowledge forgetting. When new data is introduced, scEvolve ensures that the model integrates information on new cell types while maintaining its performance on old cell types by replaying known cell type data. This strategy enhances the model's adaptability and generalization capabilities, improving both the efficiency and accuracy of single-cell annotation.

However, current research on continual learning in single-cell annotation is still in its infancy, and related methods remain underexplored. There is substantial room for improvement to achieve more robust knowledge expansion and cross-dataset transferability. Therefore, it is necessary to combine some incremental learning strategies [114] to establish a more effective balance between old and new knowledge. Incremental learning emphasizes preserving and optimizing old knowledge while absorbing new knowledge to address the "catastrophic forgetting" problem [156]. For example, knowledge distillation [116] strategies can effectively transfer old knowledge to student models, ensuring they retain the ability to recognize old cell types when absorbing new information, thereby reducing the risk of forgetting. Additionally, dynamic network expansion methods allow models to adjust their network structure when recognizing new cell types, minimizing interference with existing parameters, while regularization methods provide stability constraints to ensure that key weights remain unchanged during updates, helping prevent conflicts between old and new knowledge. Through the combination of these strategies, continual learning in single-cell annotation will have stronger knowledge retention and adaptation capabilities, providing higher accuracy and stability for processing the ever-growing single-cell sequencing data.

Heterogeneity of unseen cells and their potential decoding from an open-world perspective

From an open-world perspective, one of the core and cutting-edge challenges in single-cell annotation is the effective identification and annotation of unseen cell types. Unseen cell types typically refer to novel cell populations that are absent in the labeled reference dataset but present in the query dataset to be annotated. Notably, while traditional marker gene-based methods are constrained by their dependence on prior knowledge when handling novel cell types, they exhibit distinct advantages in wet-lab validation and cross-platform dataset stability. In particular, under scenarios with significant batch effects, these methods often demonstrate superior interpretability and reliability by leveraging explicit biomarker-based matching. In many biological research contexts, especially within the tumor microenvironment, such novel cell types may contain critical information impacting disease progression or therapeutic responses. Failure to accurately identify these cell types may lead to an incomplete understanding of cellular heterogeneity, potentially overlooking cell populations essential to disease progression and their defining characteristics. Current data-driven methods, such as mtANN [16], scLearn [90], and scBERT [17], employ classification thresholds to label samples below the threshold as "unassigned." While this dynamic discrimination mechanism expands the scope of recognition, its biological interpretability still requires improvement compared to marker gene-based approaches, which often necessitate manual validation. Particularly when faced with platform-specific variations or technical noise, these two methodological paradigms tend to exhibit complementary strengths: data-driven methods excel at capturing complex expression patterns, whereas marker gene-based approaches, which provide verifiable biological anchors, enhance annotation reliability.

Against this backdrop, hybrid strategies that integrate different technical approaches have emerged as a key area of exploration in open-world single-cell annotation. For instance, scGAD [107] introduces an anchor-pairing strategy, which seamlessly incorporates prior knowledge from reference datasets while preserving the advantages of data-driven learning. This hybrid approach inherits the stability of marker gene-based methods while retaining the sensitivity of machine learning models in detecting novel patterns. Experimental results indicate that this method effectively links reference and target datasets, utilizing known labels to aggregate potential novel cell types. However, purely data-driven methods still face inherent challenges in biological interpretability, particularly in extracting specific gene expression signatures. Existing models often fail to achieve the level of precision required for wet-lab validation, underscoring the necessity of incorporating marker gene validation at critical points in the annotation process.

Future research should focus on developing an integrated framework that combines the strengths of both approaches. One promising direction is to collaboratively validate key genes identified via attention mechanisms against authoritative marker gene databases (e.g. PanglaoDB) [24], establishing a bidirectional closed-loop mechanism of "data-driven discovery: marker gene validation." This strategy would enhance the interpretability of novel cell cluster features while improving model robustness against batch effects [157]. Furthermore, in exploring adaptive learning algorithms with limited labeled data, a hierarchical validation system inspired by marker gene-based methods could be employed: at the initial screening stage, the model leverages the sensitivity of data-driven approaches, while at the final annotation stage, it incorporates the conservativeness of marker gene-based verification. This layered strategy could significantly enhance the clinical applicability of annotation models. These integrative innovations not only help overcome the technical bottlenecks of individual methods but also pave the way for constructing clinically interpretable intelligent annotation systems, ultimately accelerating the translational application of single-cell analysis technologies in precision medicine.

Conclusion

This review offers a thorough and comprehensive overview of recent advancements in cell type annotation using scRNA-seq technology, emphasizing the transformative new perspectives it brings to understanding cellular heterogeneity. We systematically analyze and categorize various annotation methods, including those based on specific gene expression, correlation-based reference models, data-driven reference models, and large-scale pretrained models, to evaluate the strengths, weaknesses, and applicability of each approach. To address key challenges such as data sparsity, long-tail distributions, and cellular heterogeneity, we explore the potential of integrating multi-omics data and dynamic clustering algorithms to enhance annotation accuracy and robustness. Moreover, future research should focus on continual learning strategies to improve model adaptability in open-world environments, where the identification of emerging cell types is crucial. Such efforts, supported by robust evaluation frameworks and enabled by interdisciplinary collaboration, will provide a solid foundation for advancing single-cell annotation, thereby shedding light on the pivotal role of cellular complexity in biomedical research.

Key Points

- Conduct a comprehensive analysis of various single-cell transcriptome cell type annotation methods, categorizing and elaborating on their characteristics to provide insights for the development of new methods and inspire innovation through cross-method integration.
- Thoroughly examine the experimental evaluation process, covering data processing, preprocessing, evaluation metrics, and performance assessment, optimizing the evaluation framework to assist researchers in improving experimental design and method selection.
- Precisely analyze the challenges in single-cell annotation, focusing on long-tail rare cell identification and the classification of unseen cells in open-world scenarios, while proposing targeted strategies. Highlight the importance of interdisciplinary collaboration, advocate for multi-omics data integration, and encourage the use of dynamic clustering algorithms to enhance continuous learning and foster comprehensive development.

Conflict of interest

None declared.

Funding

This work is supported by the National Natural Science Foundation of China under Grant nos 62272067 and 62131004, the Sichuan Science and Technology Program under Grant no. 2023NSFSC0499, the Scientific Research Foundation of Sichuan Province under Grant no. MZGC20230078, and the Scientific Research Foundation of Chengdu University of Information Technology under Grant no. KYQN202208.

References

1. Szałata A, Hrovatin K, Becker S. *et al*. Transformers in single-cell omics: a review and new perspectives. *Nat Methods* 2024;**21**: 1430–43. https://doi.org/10.1038/s41592-024-02353-z
2. Paik DT, Cho S, Tian L. *et al*. Single-cell RNA sequencing in cardiovascular development, disease and medicine. *Nat Rev Cardiol* 2020;**17**:457–73. https://doi.org/10.1038/s41569-020-0359-y
3. Onda M, Willingham M, Nagata S. *et al*. New monoclonal antibodies to mesothelin useful for immunohistochemistry, fluorescence-activated cell sorting, western blotting, and ELISA. *Clin Cancer Res* 2005;**11**:5840–6. https://doi.org/10.1158/1078-0432.CCR-05-0578
4. Clarkson YL, Weatherall E, Waterfall M. *et al*. Extracellular localisation of the c-terminus of ddx4 confirmed by immunocytochemistry and fluorescence-activated cell sorting. *Cells* 2019;**8**:578. https://doi.org/10.3390/cells8060578
5. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet* 2019;**10**:317.
6. de Sousa Abreu R, Penalva LO, Marcotte EM. *et al*. Global signatures of protein and mrna expression levels. *Mol Biosyst* 2009;**5**: 1512–26. https://doi.org/10.1039/b908315d
7. Macosko EZ, Basu A, Satija R. *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**:1202–14. https://doi.org/10.1016/j.cell.2015.05.002
8. Hie B, Peters J, Nyquist SK. *et al*. Computational methods for single-cell RNA sequencing. *Ann Rev Biomed Data Sci* 2020;**3**: 339–64.
9. Yunjin Li L, Ma DW, Chen G. Advances in bulk and single-cell multi-omics approaches for systems biology and precision medicine. *Brief Bioinform* 2021;**22**:bbab024.
10. De Kanter JK, Lijnzaad P, Candelli T. *et al*. Chetah: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 2019;**47**:e95–5. https://doi.org/10.1093/nar/gkz543
11. Kiselev VY, Yiu A, Hemberg M. Scmap: projection of single-cell rna-seq data across data sets. *Nat Methods* 2018;**15**:359–62. https://doi.org/10.1038/nmeth.4644
12. Li C, Liu B, Kang B. *et al*. Scibet as a portable and fast single cell type identifier. *Nat Commun* 2020;**11**:1818.
13. Chen J, Hao X, Tao W. *et al*. Transformer for one stop interpretable cell type annotation. *Nat Commun* 2023;**14**:223.
14. Zhang AW, Flanagan CO', Chavez EA. *et al*. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods* 2019;**16**:1007–15. https://doi.org/10.1038/s41592-019-0529-1
15. Aran D, Looney AP, Liu L. *et al*. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**:163–72. https://doi.org/10.1038/s41590-018-0276-y
16. Xiong Y-X, Wang M-G, Chen L. *et al*. Cell-type annotation with accurate unseen cell-type identification using multiple references. *PLoS Comput Biol* 2023;**19**:e1011261. https://doi.org/10.1371/journal.pcbi.1011261
17. Yang F, Wang W, Wang F. *et al*. Scbert as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;**4**:852–66. https://doi.org/10.1038/s42256-022-00534-z
18. Pasquini G, Arias JER, Schäfer P. *et al*. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* 2021;**19**:961–9. https://doi.org/10.1016/j.csbj.2021.01.015
19. Cheng C, Chen W, Jin H. *et al*. A review of single-cell rna-seq annotation, integration, and cell–cell communication. *Cells* 2023;**12**:1970. https://doi.org/10.3390/cells12151970
20. Liu Z, Sun D, Wang C. Evaluation of cell-cell interaction methods by integrating single-cell rna sequencing data with spatial information. *Genome Biol* 2022;**23**:218.

21. Han Z, Johnson T, Zhang J. *et al*. Functional virtual flow cytometry: a visual analytic approach for characterizing single-cell gene expression patterns. *Biomed Res Int* 2017;**2017**:3035481. https://doi.org/10.1155/2017/3035481
22. Miao Z, Humphreys BD, McMahon AP. *et al*. Multi-omics integration in the age of million single-cell data. *Nat Rev Nephrol* 2021;**17**:710–24. https://doi.org/10.1038/s41581-021-00463-x
23. Cao J, Packer JS, Ramani V. *et al*. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**:661–7. https://doi.org/10.1126/science.aam8940
24. Franzén O, Gan L-M, Björkegren JLM. Panglaodb: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019;**2019**:baz046.
25. Zhang X, Lan Y, Jinyuan X. *et al*. Cellmarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* 2019;**47**:D721–8. https://doi.org/10.1093/nar/gky900
26. Stuart T, Satija R. *Integrative single-cell analysis Nat Rev Genet* 2019;**20**:257–72. https://doi.org/10.1038/s41576-019-0093-7
27. Yuan L, Zhao L, Jiang Y. *et al*. Scmgatgrn: a multiview graph attention network–based method for inferring gene regulatory networks from single-cell transcriptomic data. *Brief Bioinform* 2024;**25**:bbae526.
28. Amit I, Ardlie K, Arzuaga F. *et al*. The commitment of the human cell atlas to humanity. *Nat Commun* 2024;**15**:10019.
29. Han X, Wang R, Zhou Y. *et al*. Mapping the mouse cell atlas by microwell-seq. *Cell* 2018;**172**:1091–107. https://doi.org/10.1016/j.cell.2018.02.001
30. Schaum N, Karkanias J, Neff NF. *et al*. Single-cell transcriptomics of 20 mouse organs creates a tabula muris: the tabula muris consortium. *Nature* 2018;**562**:367. https://doi.org/10.1038/s41586-018-0590-4
31. Hodge RD, Bakken TE, Miller JA. *et al*. Conserved cell types with divergent features in human versus mouse cortex. *Nature* 2019;**573**:61–8. https://doi.org/10.1038/s41586-019-1506-7
32. Congxue H, Li T, Yingqi X. *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**:D870–6.
33. Yuan H, Yan M, Zhang G. *et al*. Cancersea: a cancer single-cell state atlas. *Nucleic Acids Res* 2019;**47**:D900–8. https://doi.org/10.1093/nar/gky939
34. Rozenblatt-Rosen O, Stubbington M, Regev A. *et al*. The Human Cell Atlas: from vision to reality. *Nature* 2017;**550**:451–53. https://doi.org/10.1038/550451a
35. Han X, Zhou Z, Fei L. *et al*. Construction of a human cell landscape at single-cell level. *Nature* 2020;**581**:303–9. https://doi.org/10.1038/s41586-020-2157-4
36. Barrett T, Wilhite SE, Ledoux P. *et al*. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Res* 2012;**41**:D991–5. https://doi.org/10.1093/nar/gks1193
37. GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30. https://doi.org/10.1126/science.aaz1776
38. Zheng GXY, Terry JM, Belgrader P. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049.
39. Picelli S, Faridani OR, Björklund ÅK. *et al*. Full-length RNA-seq from single cells using smart-seq2. *Nat Protoc* 2014;**9**:171–81. https://doi.org/10.1038/nprot.2014.006
40. Cheng J-X, Liu B-L, Zhang X. How powerful is cd133 as a cancer stem cell marker in brain tumors? *Cancer Treat Rev* 2009;**35**:403–8. https://doi.org/10.1016/j.ctrv.2009.03.002
41. Zeppernick F, Ahmadi R, Campos B. *et al*. Stem cell marker cd133 affects clinical outcome in glioma patients. *Clin Cancer Res* 2008;**14**:123–9. https://doi.org/10.1158/1078-0432.CCR-07-0932
42. Ren F, Sheng W-Q, Xiang D. Cd133: a cancer stem cells marker, is used in colorectal cancers. *World J Gastroenterol: WJG* 2013;**19**:2603. https://doi.org/10.3748/wjg.v19.i17.2603
43. Zdolsek HA, Ernerudh J, Holt PG. *et al*. Expression of the t-cell markers cd3, cd4 and cd8 in healthy and atopic children during the first 18 months of life. *Int Arch Allergy Immunol* 1999;**119**:6–12. https://doi.org/10.1159/000024169
44. Wang K, Wei G, Liu D. Cd19: a biomarker for b cell development, lymphoma diagnosis and therapy. *Exp Hematol Oncol* 2012;**1**:1–7.
45. Vaswani A. Attention is all you need. *Adv Neural Inform Process Syst* 2017;**30**:5998–6008.
46. Lu L, Xue W, Wei X. *et al*. Sctrans: multi-scale scRNA-seq sub-vector completion transformer for gene-selective cell type annotation. In: *33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, pp. 5954–62. Jeju, South Korea: International Joint Conferences on Artificial Intelligence, 2024.
47. Jiang P. Quality control of single-cell RNA-seq. *Comput Methods Single-Cell Data Anal* 2019;**1935**:1–9. https://doi.org/10.1007/978-1-4939-9057-3_1
48. Junru L, Sheng Y, Qian W. *et al*. ScRNA-seq data analysis method to improve analysis performance. *IET Nanobiotechnol* 2023;**17**:246–56.
49. Cole MB, Risso D, Wagner A. *et al*. Performance assessment and selection of normalization procedures for single-cell RNA-seq. *Cell Syst* 2019;**8**:315–28. https://doi.org/10.1016/j.cels.2019.03.010
50. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform* 2019;**20**:1583–9. https://doi.org/10.1093/bib/bby011
51. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biol* 2019;**20**:296.
52. Eraslan G, Simon LM, Mircea M. *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**:390.
53. Moon KR, Van Dijk D, Wang Z. *et al*. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**:1482–92. https://doi.org/10.1038/s41587-019-0336-3
54. Haghverdi L, Lun ATL, Morgan MD. *et al*. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7. https://doi.org/10.1038/nbt.4091
55. Korsunsky I, Millard N, Fan J. *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96. https://doi.org/10.1038/s41592-019-0619-0
56. Stuart T, Butler A, Hoffman P. *et al*. *Comprehensive integration of single-cell data cell*. *Cell* 2019;**177**:1888–902. https://doi.org/10.1016/j.cell.2019.05.031
57. Luecken MD, Büttner M, Chaichoompu K. *et al*. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
58. Zhang Y, Sun H, Zhang W. *et al*. Cellstar: a comprehensive resource for single-cell transcriptomic annotation. *Nucleic Acids Res* 2024;**52**:D859–70. https://doi.org/10.1093/nar/gkad874
59. Kim SH, Cho SY. Single-cell transcriptomics to understand the cellular heterogeneity in toxicology. *Mol Cell Toxicol* 2023;**19**:223–8. https://doi.org/10.1007/s13273-022-00304-3
60. Cao Y, Wang X, Peng G. Scsa: a cell type annotation tool for single-cell RNA-seq data. *Front Genet* 2020;**11**:490.

61. Jia Y, Ma P, Yao Q. Cellmarkerpipe: cell marker identification and evaluation pipeline in single cell transcriptomes. *Sci Rep* 2024;**14**:13151.
62. Yang X, Baumgart SJ, Stegmann CM. *et al.* Maca: marker-based automatic cell-type annotation for single-cell expression data. *Bioinformatics* 2022;**38**:1756–60. https://doi.org/10.1093/bioinformatics/btab840
63. Hedlund E, Deng Q. Single-cell RNA sequencing: technical advancements and biological applications. *Mol Aspects Med* 2018;**59**:36–46. https://doi.org/10.1016/j.mam.2017.07.003
64. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022;**13**:1246.
65. Wang Q, Li C-L, Li W. *et al.* Distinct molecular subtypes of systemic sclerosis and gene signature with diagnostic capability. *Front Immunol* 2023;**14**:1257802.
66. Cheong J-H, Wang SC, Park S. *et al.* Development and validation of a prognostic and predictive 32-gene signature for gastric cancer. *Nat Commun* 2022;**13**:774. https://doi.org/10.1038/s41467-022-28437-y
67. Gupta K, Lalit M, Biswas A. *et al.* Modeling expression ranks for noise-tolerant differential expression analysis of scrna-seq data. *Genome Res* 2021;**31**:689–97. https://doi.org/10.1101/gr.267070.120
68. Guo H, Li J. Scsorter: assigning cells to known cell types according to marker genes. *Genome Biol* 2021;**22**:69.
69. Satija R, Farrell JA, Gennert D. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502. https://doi.org/10.1038/nbt.3192
70. Huang Y, Sanguinetti G. Uncertainty versus variability: Bayesian methods for analysis of scRNA-seq data. *Curr Opin Syst Biol* 2021;**28**:100375. https://doi.org/10.1016/j.coisb.2021.100375
71. Magaña-López G, Calzone L, Zinovyev A. *et al.* Scboolseq: linking scRNA-seq statistics and Boolean dynamics. *PLoS Comput Biol* 2024;**20**:e1011620. https://doi.org/10.1371/journal.pcbi.1011620
72. Moon TK. The expectation-maximization algorithm. *IEEE Signal Process Mag* 1996;**13**:47–60. https://doi.org/10.1109/79.543975
73. Simmons SK, Lithwick-Yanai G, Adiconis X. *et al.* Mostly natural sequencing-by-synthesis for scRNA-seq using ultima sequencing. *Nat Biotechnol* 2023;**41**:204–11. https://doi.org/10.1038/s41587-022-01452-6
74. Song P, Li W, Guo L. *et al.* Identification and validation of a novel signature based on nk cell marker genes to predict prognosis and immunotherapy response in lung adenocarcinoma by integrated analysis of single-cell and bulk RNA-sequencing. *Front Immunol* 2022;**13**:850745. https://doi.org/10.3389/fimmu.2022.1076784
75. Song P, Li W, Xiaoxuan W. *et al.* Integrated analysis of single-cell and bulk RNA-sequencing identifies a signature based on b cell marker genes to predict prognosis and immunotherapy response in lung adenocarcinoma. *Cancer Immunol Immunother* 2022;**71**:2341–54. https://doi.org/10.1007/s00262-022-03143-2
76. Zhang Z, Luo D, Zhong X. *et al.* Scina: a semi-supervised subtyping algorithm of single cells and bulk samples. *Genes* 2019;**10**:531. https://doi.org/10.3390/genes10070531
77. Cortal A, Martignetti L, Six E. *et al.* Gene signature extraction and cell identity recognition at the single-cell level with cell-id. *Nat Biotechnol* 2021;**39**:1095–102. https://doi.org/10.1038/s41587-021-00896-6
78. Abdi H, Valentin D. Multiple correspondence analysis. *Encycl Measure Stat* 2007;**2**:651–7.
79. Shao X, Liao J, Xiaoyan L. *et al.* Sccatch: automatic annotation on cell types of clusters from single-cell RNA sequencing data. *Iscience* 2020;**23**:100882. https://doi.org/10.1016/j.isci.2020.100882
80. Cohen I, Huang Y, Chen J. *et al.* Pearson correlation coefficient. *Noise Reduction in Speech Processing* 2009;**2**:1–4.
81. Zar JH. Significance testing of the spearman rank correlation coefficient. *J Am Stat Assoc* 1972;**67**:578–80.
82. Rahutomo F, Kitasuka T, Aritsugi M. *et al.* Semantic cosine similarity. In: *The 7th International Student Conference on Advanced Science and Technology ICAST, Vol.* 4, p. 1. University of Seoul South Korea, 2012.
83. Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. *IEEE Trans Syst Man Cybern* 1985;**SMC-15**:580–5. https://doi.org/10.1109/TSMC.1985.6313426
84. Cao Z-J, Wei L, Shen L. *et al.* Searching large-scale scRNA-seq databases via unbiased cell embedding with cell blast. *Nat Commun* 2020;**11**:3458.
85. Goodfellow I, Pouget-Abadie J, Mirza M. *et al.* Generative adversarial networks. *Commun ACM* 2020;**63**:139–44. https://doi.org/10.1145/3422622
86. Hou R, Denisenko E, Forrest ARR. Scmatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* 2019;**35**:4688–95. https://doi.org/10.1093/bioinformatics/btz292
87. Lizio M, Abugessaisa I, Noguchi S. *et al.* Update of the fantom web resource: expansion to provide additional transcriptome atlases. *Nucleic Acids Res* 2019;**47**:D752–8. https://doi.org/10.1093/nar/gky1099
88. Ke P, Xiang S, Xie C. *et al.* Unsupervised continual learning of single-cell clustering based on novelty detection and memory replay. In:*2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 3031–8. Piscataway, NJ: IEEE, 2022.
89. Wan H, Yuan M, Yiwei F. *et al.* Continually adapting pre-trained language model to universal annotation of single-cell RNA-seq data. *Brief Bioinform* 2024;**25**:bbae047.
90. Duan B, Zhu C, Chuai G. *et al.* Learning for single-cell assignment. *Sci Adv* 2020;**6**:eabb8340.
91. Rui F, Gillen AE, Sheridan RM. *et al.* Clustifyr: an r package for automated single-cell RNA sequencing cluster classification. *F1000Research* 2020;**9**:223.
92. Zhai Y, Chen L, Deng M. Scbol: a universal cell type identification framework for single-cell and spatial transcriptomics data. *Brief Bioinform* 2024;**25**:bbae188.
93. Yuan L, Sun S, Jiang Y. *et al.* Scrgcl: a cell type annotation method for single-cell RNA-seq data using residual graph convolutional neural network with contrastive learning. *Brief Bioinform* 2025;**26**:bbae662.
94. Auria L, Moro RA. Support vector machines (SVM) as a technique for solvency analysis. *SSRN Electron J* 2008. https://doi.org/10.2139/ssrn.1424949
95. Qi Y. Random forest for bioinformatics. In: Zhou ZH, Dietterich TG (eds.), *Ensemble Machine Learning: Methods and Applications*, pp. 307–23. New York, NY: Springer; 2012.
96. Alquicira-Hernandez J, Sathe A, Ji HP. *et al.* Scpred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* 2019;**20**:1–17.
97. Tan Y, Cahan P. Singlecellnet: a computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst* 2019;**9**:207–13. https://doi.org/10.1016/j.cels.2019.06.004

98. McCarthy DJ, Campbell KR, Lun ATL. *et al.* Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in r. *Bioinformatics* 2017;**33**:1179–86.
99. Abdelaal T, Michielsen L, Cats D. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:1–19.
100. Lopez R, Regier J, Cole MB. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
101. Ding J, Adiconis X, Simmons SK. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat Biotechnol* 2020;**38**:737–46. https://doi.org/10.1038/s41587-020-0465-8
102. Shao X, Yang H, Zhuang X. *et al.* 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**:e122–2. https://doi.org/10.1093/nar/gkab775
103. Chen L, He Q, Zhai Y. *et al.* Single-cell RNA-seq data semi-supervised clustering and annotation via structural regularized domain adaptation. *Bioinformatics* 2021;**37**:775–84. https://doi.org/10.1093/bioinformatics/btaa908
104. Li T, Yugui X, He S. *et al.* Cell-specific highly correlated network for self-supervised distillation in cell type annotation. In:*In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 988–93. Piscataway, NJ: IEEE, 2024.
105. Jiao L, Wang G, Dai H. *et al.* Sctranssort: transformers for intelligent annotation of cell types by gene embeddings. *Biomolecules* 2023;**13**:611. https://doi.org/10.3390/biom13040611
106. Jing X, Zhang A, Liu F. *et al.* Ciform as a transformer-based model for cell-type annotation of large-scale single-cell RNA-seq data. *Brief Bioinform* 2023;**24**:bbad195.
107. Zhai Y, Chen L, Deng M. Scgad: a new task and end-to-end framework for generalized cell type annotation and discovery. *Brief Bioinform* 2023;**24**:bbad045.
108. Krishna K, Narasimha M, Murty. Genetic k-means algorithm. *IEEE Trans Syst Man Cybern B Cybern* 1999;**29**:433–9. https://doi.org/10.1109/3477.764879
109. Lotfollahi M, Naghipourfar M, Luecken MD. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022;**40**:121–30. https://doi.org/10.1038/s41587-021-01001-7
110. Doersch C. Tutorial on variational autoencoders. arXiv [cs.LG]. 2016;arXiv:1606.05908. https://arxiv.org/abs/1606.05908
111. Zhai Y, Liang C, Deng M. Realistic cell type annotation and discovery for single-cell RNA-seq data. *In IJCAI* 2023;4967–74. https://doi.org/10.24963/ijcai.2023/552
112. Peyré G, Cuturi M. *et al.* Computational optimal transport: with applications to data science. *Found Trends Mach Learn* 2019;**11**:355–607.
113. Zhai Y, Chen L, Deng M. Scevolve: cell-type incremental annotation without forgetting for single-cell RNA-seq data. *Brief Bioinform* 2024;**25(2)**:bbae039. https://doi.org/10.1093/bib/bbae039
114. Gepperth A, Hammer B. Incremental learning algorithms and applications. In: Wermter S, Weber C, Barschtipan W (eds.), *Incremental Learning: Towards Human-Like Learning Capabilities*, pp. 1–34. Cham, Switzerland: Springer, 2016. Available from: https://hal.science/hal-01418129
115. Hendrycks D, Mazeika M, Kadavath S. *et al.* Using self-supervised learning can improve model robustness and uncertainty. *Adv Neural Inform Process Syst* 2019;**32**:15637–48.
116. Gou J, Baosheng Y, Maybank SJ. *et al.* Knowledge distillation: a survey. *Int J Comput Vision* 2021;**129**:1789–819. https://doi.org/10.1007/s11263-021-01453-z
117. Zhou S, Li Y, Wenyuan W. *et al.* Scmmt: a multi-use deep learning approach for cell annotation, protein prediction and embedding in single-cell rna-seq data. *Brief Bioinform* 2024;**25**:bbad523.
118. Sun H, Haowen Q, Duan K. *et al.* Scmgcn: a multi-view graph convolutional network for cell type identification in scRNA-seq data. *Int J Mol Sci* 2024;**25**:2234. https://doi.org/10.3390/ijms25042234
119. Li Z, Wang Y, Ganan-Gomez I. *et al.* A machine learning-based method for automatically identifying novel cells in annotating single-cell RNA-seq data. *Bioinformatics* 2022;**38**:4885–92. https://doi.org/10.1093/bioinformatics/btac617
120. Liu H, Li H, Sharma A. *et al.* Scanno: a deconvolution strategy-based automatic cell type annotation tool for single-cell rna-sequencing data sets. *Brief Bioinform* 2023;**24**:bbad179.
121. Liu Y, Wei G, Li C. *et al.* Tripletcell: a deep metric learning framework for accurate annotation of cell types at the single-cell level. *Brief Bioinform* 2023;**24**:bbad132.
122. Wei Z, Zhang S. Callr: a semi-supervised cell-type annotation method for single-cell RNA sequencing data. *Bioinformatics* 2021;**37**:i51–8.
123. Jia S, Lysenko A, Boroevich KA. *et al.* Scdeepinsight: a supervised cell-type identification method for scRNA-seq data with deep learning. *Brief Bioinform* 2023;**24**:bbad266.
124. Jian H, Li X, Gang H. *et al.* Iterative transfer learning with neural network for clustering and cell type classification in single-cell rna-seq analysis. *Nat Mach Intell* 2020;**2**:607–18. https://doi.org/10.1038/s42256-020-00233-7
125. Song Q, Jing S, Zhang W. Scgcn is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat Commun* 2021;**12**:3826.
126. Kimmel JC, Kelley DR. Scnym: semi-supervised adversarial neural networks for single cell classification. *Genome Res* 2021;**31**:1781–93.
127. Ma F, Pellegrini M. Actinn: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* 2020;**36**:533–8. https://doi.org/10.1093/bioinformatics/btz592
128. Wan H, Chen L, Deng M. Scname: neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data. *Bioinformatics* 2022;**38**:1575–83. https://doi.org/10.1093/bioinformatics/btac011
129. Kiselev VY, Kirschner K, Schaub MT. *et al.* Sc3: consensus clustering of single-cell RNA-seq data. *Nat Methods* 2017;**14**:483–6. https://doi.org/10.1038/nmeth.4236
130. Chen L, Wang W, Zhai Y. *et al.* Deep soft k-means clustering with self-training for single-cell RNA sequence data. *NAR Genom Bioinform* 2020;**2**:lqaa039.
131. Fischer F, Fischer DS, Mukhin R. *et al.* Sctab: scaling cross-tissue single-cell annotation models. *Nat Commun* 2024;**15**:6611.
132. Zhai Y, Liang C, Deng M. Distribution-independent cell type identification for single-cell RNA-seq data. In: De Raedt L (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pp. 6143–51. Marina del Rey, CA: International Joint Conferences on Artificial Intelligence Organization, 2024.
133. Cui H, Wang C, Maan H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:529–39.
134. Hao M, Gong J, Zeng X. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91.

135. Park S, Lee H. Robust self-supervised learning strategy to tackle the inherent sparsity in single-cell RNA-seq data. *Brief Bioinform* 2024;**25**:bbae586.
136. Xia Y, Liu Y, Li T. *et al.* Assessing parameter efficient methods for pre-trained language model in annotating scRNA-seq data. *Methods* 2024;**228**:12–21. https://doi.org/10.1016/j.ymeth.2024.05.007
137. Xiao Y, Liu J, Zheng Y et al. CellAgent: an LLM-driven multi-agent framework for automated single-cell data analysis. bioRxiv. 2024. https://doi.org/10.1101/2024.05.13.594090
138. Fan X, Liu J, Yang Y. *et al.* Scgraphformer: unveiling cellular heterogeneity and interactions in scRNA-seq data using a scalable graph transformer network. *Commun Biol* 2024;**7**:1463.
139. Liu Y, Li T, Wang Z. *et al.* Exploring parameter-efficient fine-tuning of a large-scale pre-trained model for scRNA-seq cell type annotation. In:*In 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 580–5. Piscataway, NJ, USA: IEEE, 2023.
140. Abdelaal T, Michielsen L, Cats D. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**:1–19.
141. Anguita D, Ghelardoni L, Ghio A. *et al. The'k'in k-fold cross validation In ESANN* 2012;**102**:441–6.
142. Chiou J, Zeng C, Cheng Z. *et al.* Single-cell chromatin accessibility identifies pancreatic islet cell type -and state-specific regulatory programs of diabetes risk. *Nat Genet* 2021;**53**:455–66. https://doi.org/10.1038/s41588-021-00823-0
143. Miao Z, Moreno P, Huang N. *et al.* Putative cell type discovery from single-cell gene expression data. *Nat Methods* 2020;**17**:621–8. https://doi.org/10.1038/s41592-020-0825-9
144. Boehm KM, Khosravi P, Vanguri R. *et al.* Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 2022;**22**:114–26. https://doi.org/10.1038/s41568-021-00408-3
145. Lin Y, Tung-Yu W, Wan S. *et al.* Scjoint integrates atlas-scale single-cell RNA-seq and atac-seq data with transfer learning. *Nat Biotechnol* 2022;**40**:703–10. https://doi.org/10.1038/s41587-021-01161-6
146. Gayoso A, Steier Z, Lopez R. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nat Methods* 2021;**18**:272–82. https://doi.org/10.1038/s41592-020-01050-x
147. Stoeckius M, Hafemeister C, Stephenson W. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**:865–8. https://doi.org/10.1038/nmeth.4380
148. Argelaguet R, Arnol D, Bredikhin D. *et al.* Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* 2020;**21**:1–17.
149. Mimitou EP, Lareau CA, Chen KY. *et al.* Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol* 2021;**39**:1246–58. https://doi.org/10.1038/s41587-021-00927-2
150. Zhong Z, Fini E, Roy S. *et al.* Neighborhood contrastive learning for novel class discovery. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10867–75. Piscataway, NJ, USA: IEEE, 2021.
151. Kenton J, DM-WC, Bert LKT. 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 (NAACL)*, pp. 1. Stroudsburg, PA, USA: Association for Computational Linguistics (ACL), 2019.
152. Choromanski K, Likhosherstov V, Dohan D. *et al.* Rethinking attention with performers. arXiv preprint, arXiv:2009.14794. 2021.
153. Hospedales T, Antoniou A, Micaelli P. *et al.* Meta-learning in neural networks: a survey. *IEEE Trans Pattern Anal Mach Intell* 2021;**44**:5149–69.
154. Snell J, Swersky K, Zemel R. Prototypical networks for few-shot learning. *Adv Neural Inform Process Syst* 2017;**30**:4080–90.
155. Yang M, Yang Y, Xie C. *et al.* Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nat Mach Intell* 2022;**4**:696–709. https://doi.org/10.1038/s42256-022-00518-z
156. Kirkpatrick J, Pascanu R, Rabinowitz N. *et al.* Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci* 2017;**114**:3521–6. https://doi.org/10.1073/pnas.1611835114
157. Li S, Guo H, Zhang S. *et al.* Attention-based deep clustering method for scRNA-seq cell type identification. *PLoS Comput Biol* 2023;**19**:e1011641. https://doi.org/10.1371/journal.pcbi.1011641

Briefings in Bioinformatics, 2025, **26(3)**, bbaf207
https://doi.org/10.1093/bib/bbaf207
Review