

Review article

Advancing automated cell type annotation with large language models and single-cell isoform sequencing

Hettiarachchige Wijewardena ^a, Saloni Bhatia ^a, Namrata Bhattacharya ^{b,c,d},
Debarika Sengupta ^{c,e,f}, Siyuan Wu ^{a,g,h,*}, Ulf Schmitz ^{a,h,i,*}

^a Computational Biomedicine Lab, College of Science and Engineering, James Cook University, Townsville, QLD, Australia

^b Australian Prostate Cancer Research Centre-Queensland, Faculty of Health, School of Biomedical Sciences, Centre for Genomics and Personalised Health, Queensland University of Technology, Brisbane, QLD 4000, Australia

^c Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi 110020, India

^d Translational Research Institute, Princess Alexandra Hospital, Woolloongabba, QLD 4102, Australia

^e Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi 110020, India

^f Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi 110020, India

^g School of Mathematics, Monash University, Melbourne, Victoria 3800, Australia

^h Centre for Tropical Bioinformatics and Molecular Biology, Australian Institute of Tropical Health and Medicine, James Cook University, Cairns, Australia

ⁱ Centenary Institute, The University of Sydney, Camperdown, Australia

ARTICLE INFO

Keywords:

Single-cell RNA sequencing
Automatic cell type annotation
Machine learning
Transcript isoforms
Alternative splicing
Large language models
Natural language processing

ABSTRACT

Accurate cell type identification is critical for interpreting single-cell transcriptomic data and understanding complex biological systems. In this review, we discuss how natural language processing and large language models can enhance the accuracy and scalability of cell type annotation. We also highlight how emerging single-cell long-read sequencing technologies enable isoform-level transcriptomic profiling, offering higher resolution than conventional gene expression-based methods and providing opportunities to redefine cell types. By integrating the insights of key technical and algorithmic advances across sequencing and computational approaches, we provide a unified overview of recent developments that are reshaping automated cell type annotation and improving the precision of biological interpretation.

1. Introduction

Cells are the fundamental units of life. The ensemble of expressed genes and proteins in a cell defines its identity and function, or phenotype, at any specific point in time. It is therefore no surprise that anomalies in gene expression are an observable manifestation of many diseases. Likewise, while orchestrated transcript isoforms expression enables the regulation of critical biological processes, abnormal isoform expression has been linked to illnesses such as cancer [1].

Consequently, knowing the canonical transcriptome profile of cell types can facilitate the identification of early signs of diseases and the discovery of therapeutic targets. Conversely, precise identification of cell types is essential for cellular engineering, including somatic cell reprogramming [2], guided differentiation of pluripotent stem cells [3], and direct conversions among differentiated cell lineages [4], such as therapeutic T-cell engineering [5].

Single-cell sequencing enables high-throughput transcriptomic profiling of thousands of individual cells from a tissue or organism in parallel. This is achieved using droplet-, plate-, or microwell-based approaches combined with second- or third-generation sequencing technologies, allowing transcriptomic analysis at the gene, isoform, or even spatial resolution [6, 7].

1.1. Cell type annotation

Cell type annotation facilitates the assignment of an identity to cellular transcriptomic profiles, which can be grouped to allow for inter-cell-type or cross-sample comparison. Conventional cell type annotation in single-cell transcriptomic data relies on expert manual labelling of cell clusters using known marker genes and biological insight [8]. It typically involves two steps: clustering cells based on transcriptomic similarity, followed by assigning cell type labels by comparing marker gene

* Corresponding authors at: Computational Biomedicine Lab, College of Science and Engineering, James Cook University Townsville, QLD, Australia.

E-mail addresses: thaddeus.wu@jcu.edu.au (S. Wu), ulf.schmitz@jcu.edu.au (U. Schmitz).

<https://doi.org/10.1016/j.csbj.2025.11.008>

Received 10 September 2025; Received in revised form 2 November 2025; Accepted 3 November 2025

Available online 6 November 2025

2001-0370/© 2025 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

expression to references [9]. Annotators often need to consult literature and mine existing data to identify context-specific markers, especially when canonical markers are insufficient or ambiguous [8, 9]. Although effective, this process is slow, labour-intensive, and requires both computational and domain expertise, and because it is often not based on standardized cell label ontologies, it is difficult to reproduce [8, 10].

As a consequence, automatic cell type annotation systems have advanced in recent years. In this context, machine learning (ML) can significantly improve the understanding of the cellular composition of tissues from single-cell RNA sequencing data [11]. This entails analysing the data to understand the structure and composition of tissue at the cellular level, capturing the interrelationships of cell markers, and developing models of cell populations. ML enables the identification of diverse cell types within complex tissues across different species, provided that suitable training datasets are available, thereby facilitating the reconstruction of cellular networks.

1.2. Natural language processing

ML has witnessed the remarkable impact of transformers, which were initially developed for natural language processing (NLP) tasks. NLP encompasses computational approaches that enable the representation and interpretation of human language, converting unstructured text into analysable forms that ML algorithms can process more effectively. These have now enabled NLP-based cell characterisation methods that significantly expand the scope of automated cell type annotation, allowing for more nuanced identification of cell subpopulations and rare cell types [12]. Building on these methods, the development of large language models (LLMs) has further enhanced this capability. LLMs are accessed via provider-specific interfaces or open-source implementations and serve as the core reasoning component of AI systems, supported by mechanisms such as memory and tool use [13, 14]. By utilizing extensive training datasets, LLMs can assist in the identification of complex cell types and their characteristics [15]. However, challenges remain, particularly with reproducibility, and the fact that LLMs are not specifically designed for cell type annotation [16].

1.3. Third-generation sequencing

A recent uptake of single-cell long-read sequencing studies have allowed studying single-cell transcriptomes at finer granularity. Long-read sequencing or isoform sequencing (a.k.a. third-generation sequencing, TGS) technologies facilitate the investigation of full-length isoforms and transcriptomic complexity patterns [17, 18]. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are two prominent platforms for TGS that enable sequencing of fragments exceeding 10 kb in length.

However, cell type annotation methods have not yet fully leveraged the enhanced transcriptomic resolution to improve classification performance and enable the identification of novel or rare cell types. This review examines recent developments in automated cell type classification, single-cell RNA sequencing technologies, and the potential for LLMs and long-read sequencing to further advance the field.

See major milestone in the evolution of cell type annotation in Fig. 1.

2. Machine-learning models for automated cell type classification

Automation of cell type identification eliminates the need for manual annotation, making it accessible to those unfamiliar with cell markers while also saving resources when working with large datasets [19, 20]. As a result, there has been a surge in the development of computational tools over the past decade, specifically designed to automate cell type identification from single-cell RNA sequencing data (Supplementary Table 1).

These tools can be broadly classified into two main categories: (1) Reference-based methods utilise existing datasets or models to classify new cells. These methods include supervised approaches that require explicit model training on labelled datasets, unsupervised approaches that compare new cells to reference profiles based on patterns learned from unlabelled datasets, and pretrained classifiers adapted from NLP or LLMs. (2) Marker-based methods rely on predefined marker genes or cell type signatures from biological databases and ontologies, assigning cell

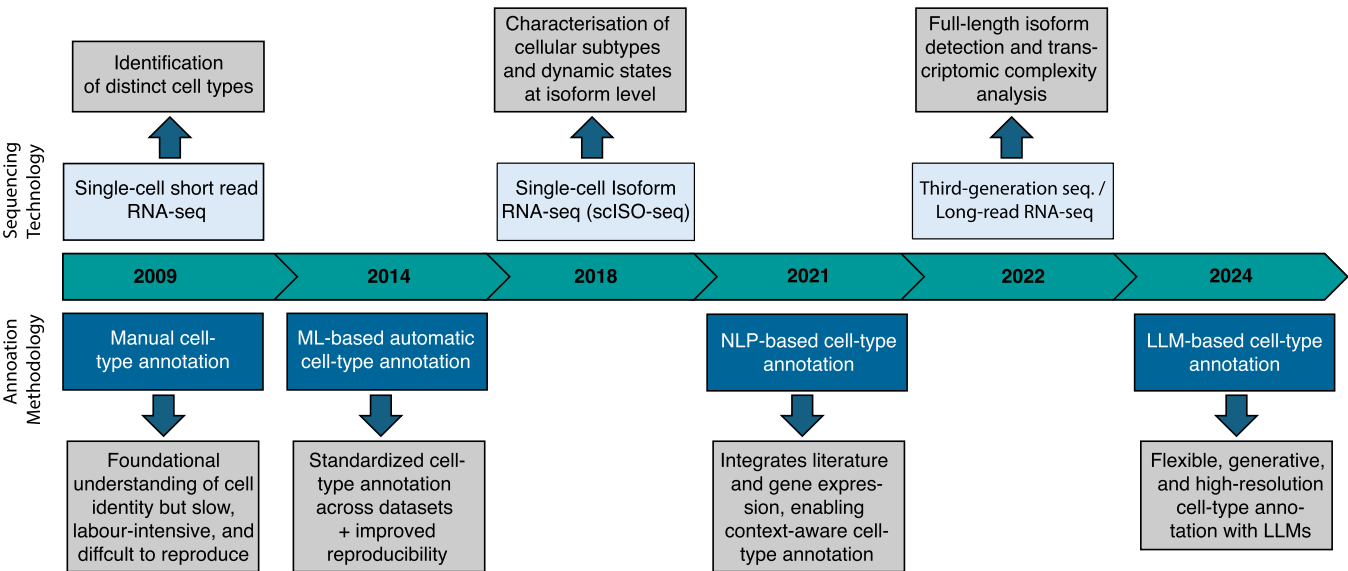


Fig. 1. Timeline of key milestones in the evolution of single-cell RNA-seq-based cell type annotation. The timeline highlights major technological and algorithmic advances shaping the field of cell type annotation. Single-cell RNA sequencing (scRNA-seq) was first introduced in 2009 [116], enabling the detection of cellular heterogeneity previously masked in bulk RNA-seq, followed by manual annotation approaches based on marker genes. ML-based automatic annotation tools were subsequently developed [117], enhancing reproducibility and scalability across datasets. The introduction of single-cell isoform sequencing [82] allowed the characterization of cellular subtypes and dynamic states at the isoform level. Later, natural language processing (NLP)-based methods such as *scETM* [56] and *CellMeSH* [57] integrated literature and gene expression information for context-aware annotation. Advances in third-generation long-read sequencing technologies (PacBio, Nanopore) [118] further enabled full-length isoform detection. Most recently, the emergence of large language model (LLM)-based frameworks [59, 63] has facilitated flexible, generative, and high-resolution approaches to cell type annotation.

identities based on marker expression patterns without large annotated references.

Within the reference-based category, most tools integrate ML techniques, ranging from traditional classifiers such as neural networks (NN), random forests (RF), and support vector machines (SVM), to ensemble and ML-based hybrid approaches. A recent trend involves composite models combining multiple ML algorithms to enhance performance, robustness, and interpretability (e.g., Moana [21], *scRe-Classify* [22], or *ImmClassifier* [23]). While many non-ML methods still rely on statistical or rule-based frameworks, some hybrid tools like Digital Cell Sorter [24] incorporate ML within marker-based approaches.

To assess the effectiveness of these diverse approaches, multiple benchmark studies have compared the performance of cell type annotation tools and ML models (Table 1). Despite the diversity of strategies, marker-based methods do not consistently outperform others, underscoring the importance of robust marker gene selection [10]. Abdelaal et al., Xie et al., and Hung et al. benchmarked both reference-based and marker-based methods, whereas Zhao et al. evaluated only reference-based approaches. In contrast, Huang and Zhang et al. and Tortelote et al. focused on benchmarking general-purpose ML models for cell type annotation.

Benchmark studies show that cell type annotation tools vary widely in performance depending on dataset complexity, tissue type, and cell-type similarity. Tools such as *scPred* [29], *SingleCellNet* [30], *ACTINN* [31], *CaSTLe* [32], *scmapcell* [33], *SingleR* [34], and *SCINA* [35] achieve near-perfect accuracy with well-structured datasets but struggle with heterogeneous or complex datasets [10, 27], while *Cell BLAST* [36], *CHETAH* [37], and *scVI* [38] often underperform in these settings. General-purpose ML models such as SVM and RF perform well with low-complexity datasets but decline in performance when faced with large or overlapping cell populations [10, 25, 39]. Across all benchmarks, no single tool consistently outperforms others, and runtime differences reflect trade-offs between computational cost and algorithmic complexity [27, 33, 34, 40].

Notably, most automated cell type classification tools are optimised for short-read sequencing data. Future efforts should focus more on adapting models for isoform-resolved annotation using long-read RNA sequencing data.

Detailed descriptions of the datasets used in the cell type annotation benchmark studies discussed here are provided in Supplementary Table 2.

Even though traditional ML-based approaches play a crucial role in

automating single-cell type annotation, they often face dataset-specific challenges that limit their generalisability and interpretability, which was also the observation in recent cell type annotation benchmark studies [15]. ML-based models typically rely on reference datasets, making their performance highly dependent on the quality and representativeness of the training data. Consequently, ML-based methods tend to perform well within specific tissues or experimental conditions but exhibit reduced accuracy when applied to novel or heterogeneous datasets [14, 15]. Moreover, the high dimensionality and nonlinear relationships inherent in mRNA expression data pose additional challenges. Most ML pipelines rely on highly variable gene (HVG) selection and dimensionality reduction techniques such as PCA, which risk discarding biologically relevant information, especially for rare cell types, and introduce artificial biases due to parameter sensitivity [29, 31, 41–44]. The selected HVGs are often dataset-specific and may not capture co-expression or biological interactions of genes which are critical for accurate annotation [45]. These limitations collectively restrict their ability to generalise beyond the training datasets and discover new cell types. In contrast, LLMs have been introduced to address these issues by reframing transcriptomic data within a natural language context, allowing models to capture complex gene–gene relationships, contextual dependencies, and nonlinear patterns [46]. By leveraging pre-training on large, diverse biological corpora, LLMs can overcome the narrow scope and limited scalability of ML-based annotation methods [43, 47, 48].

3. Large language model-based cell type annotation

LLMs, particularly transformer-based foundation models, offer a unified framework capable of learning from massive, heterogeneous datasets, enabling broader applicability across tissues and conditions [47].

Foundation models, is a category of large scale pre-trained models that are characterised by their scalability and adaptability. Trained on vast datasets, they capture complex biological patterns and can be reused across diverse tasks with minimal modification. However, this power comes with high computational demands, often requiring GPU or TPU clusters and involving billions of parameters [49].

LLMs evolved from traditional ML methods through breakthroughs in deep learning, specifically the transformer architecture, which uses self-attention, an attention mechanism that relates different positions of a single sequence to compute a context-aware representation of the sequence, to capture long-range dependencies in sequential data [50–52]. Even though transformers were originally developed for NLP, they have now underlined the foundation models in biology for LLMs, where genes are treated as tokens in a sequence, much like words in a sentence [49, 53].

In the context of single-cell biology, LLMs facilitate automated cell type annotation by embedding gene expression data into meaningful representations. Drawing parallels to language, where texts comprise words, cells are defined by genes, though this remains an analogy rather than a direct equivalence - it simply explains the complex relationship [47]. This enables the possibility of classification of cell subtypes and rare populations, even across batches or modalities [49, 54]. By modelling genes as textual tokens, LLMs can continuously improve with new data, easily incorporate additional features and metadata, and uncover subtle functional relationships and developmental trajectories [47].

LLMs are trained in two stages: pretraining and fine-tuning. In pre-training, the model learns general patterns from large-scale unlabelled single-cell datasets using self-supervised objectives. Cell representations are often learned via special tokens or via a cell embedding matrix derived from the model’s output. During fine-tuning, these pretrained embeddings are adapted to specific downstream tasks including cell type annotation, perturbation prediction, or integration. This two-phase training enables LLMs to generalise well while retaining task-specific

Table 1
Benchmark studies comparing the performance of cell type annotation tools and ML models.

Study	Notes	Reference
Abdelaal et al.	Evaluated the efficacy of 22 different approaches for automated cell identification with 27 single-cell RNA sequencing datasets.	[10]
Xie et al.	Assessed 32 existing automated cell type identification techniques for scRNA-seq, examining their efficacy across various datasets.	[19]
Huang and Zhang	Performed an extensive assessment of 10 ML models for the automatic assignment of cell phenotypes utilizing 20 publically accessible scRNA-seq datasets.	[25]
Hung et al.	Evaluated eleven cell type annotation methodologies, accessible as R packages, examining their efficacy across a diverse array of public scRNA-seq datasets and simulated data.	[26]
Zhao et al.	Assessed nine classification algorithms tailored for scRNA-seq datasets utilising three distinct sources which contains highly credible cell type labels.	[27]
Tortelote	In a preprint article, Tortelote assessed the efficacy of eight ML models for cell annotation in scRNA-seq versus single-nucleus RNA sequencing (snRNA-seq) datasets.	[28]

precision [49, 53].

Pretrained on comprehensive atlas-level datasets such as PanglaoDB (panglao.db.se) [55] and the Human Cell Atlas (www.humancellatlas.org), models like *scBERT* [43] integrate database-driven annotation tools with generative artificial intelligence (AI). These models combine multiple annotation techniques and apply self-iterative optimisation, selecting the best strategies for each case and improving accuracy and interpretability [14, 56]. As a result, LLMs facilitate precise annotation of rare cell types and reveal complex differentiation trajectories previously difficult to resolve [54].

Recent advancements in NLP- and LLM-based cell type annotation have led to the development of a growing number of software tools. Among them, *MarkerGeneBERT*, a direct application of NLP, systematically extracts cell type and marker gene information from scientific literature to enhance the accuracy and efficiency of scRNA-seq based cell type annotation [12]. *CellMeSH*, on the other hand, automatically constructs a comprehensive gene–cell-type association database from indexed biomedical literature and uses a probabilistic querying approach to predict cell types from scRNA-seq clusters, enabling scalable and literature-informed annotation [57]. Although tools such as *SCellBOW* [58] and *scETM* [56] do not directly annotate cell types, they provide a crucial supportive role. *SCellBOW*, an unsupervised transfer learning method, treats genes as words and cells as documents to enable clustering, phenotypic analysis, and the detection of malignant subpopulations [58]. *scETM* on the other hand employs embedded topic modeling and NNs to evaluate scRNA-seq data, providing interpretable gene embeddings and zero-shot transfer learning across many tissues and species [56]. Table 2 provides a comparative summary of LLM-based tools developed for cell type annotation. Supplementary Table 3 presents detailed information on each LLM based cell type annotation tool, including input modalities, training data sources and sizes, model interpretability, computational cost, and biological validation strategies. Supplementary Table 4 lists the datasets used to evaluate annotation accuracy and specifies the corresponding cell types analysed in each study.

Besides dedicated cell type annotation tools, several specialized single-cell foundation models also enable cell type annotation, even though it is not their primary purpose. *CellPLM* is a pre-trained single-cell language model capturing cell–cell relationships and spatial transcriptomic patterns, supporting downstream tasks including clustering, perturbation prediction, and cell type annotation [65]. *Geneformer* models gene network dynamics from large-scale single-cell transcriptomic data and identifies candidate therapeutic targets, while also facilitating cell type annotation [66]. *GenePT* leverages LLM-derived gene embeddings to represent gene–cell relationships, making it applicable for cell type identification [67]. *Cell2Text* is a multimodal generative model that produces interpretable natural-language descriptions from single-cell RNA-seq data, with the additional capability of annotating cell types [68].

3.1. Benchmarks of LLM-based cell type annotation algorithms

SOAR is a large-scale benchmarking study that has evaluated the performance of eight instruction-tuned LLMs (*DeepSeek-LLM-67B*, *Qwen2-72B*, *Llama-3-70B*, *Mixtral-8 × 7B*, *Mixtral-8 × 22B*, *Cell2Sentence*, *GPT-4o mini*, *GPT-4o*) across eleven datasets for cell type annotation in single-cell genomics. Their findings highlighted that LLMs exhibit strong interpretive capabilities in scRNA-seq data without extensive fine-tuning, while also showing promising potential for cross-modality analysis in multi-omics contexts [69]. Additionally, AnnDictionary has been utilised to assess commercially available LLMs for cell type annotation using the Tabula Sapiens datasets [13, 70]. This study demonstrated that LLM annotation of most primary cell types achieves an accuracy exceeding 80–90 %. A leaderboard for LLM cell type annotation, based on evaluations using Tabula Sapiens data, is available at <https://singlecellgpt.com/celltype-annotation-leaderboard> [13].

Table 2
LLM-based cell type annotation tools.

Software	Description	Reference
<i>LICT</i>	A software package that employs a multi-model fusion and "talk-to-machine" technique to enhance annotation reliability, particularly in datasets characterised by low cellular heterogeneity.	[15]
<i>GPTCelltype</i>	An R package integrating <i>ChatGPT</i> powered by <i>GPT-4</i> , which enables LLMs to execute cell-type annotations independently, without extensive domain knowledge or reference datasets.	[16]
<i>CellAgent</i>	A multi-agent framework leveraging LLMs to automate single-cell RNA sequencing data analysis and cell type annotation, delivering high-quality results without human intervention.	[14]
<i>CASSIA</i>	A multi-agent LLM-based tool that enhances annotation accuracy while improving interpretability by providing reasoning and quality scores for each prediction.	[59]
<i>Celler</i>	<i>Celler</i> is a transformer-based annotation tool that applies LLM concepts to improve the identification of rare and disease-relevant cell types. Using innovations like Glnf Loss and Hard Data Mining, it effectively handles long-tailed distributions in single-cell data.	[46]
<i>CellReasoner</i>	A lightweight, open-source LLM for single-cell type annotation that maps gene expression profiles to cell types with strong generalisation and interpretable, marker-level reasoning	[60]
<i>CellTypeAgent</i>	LLM-based annotation tool that uses marker genes and integrates GPT models with curated databases to improve accuracy and reduce hallucinations, enabling efficient and reliable cell type identification.	[9]
<i>ReCellTy</i>	A retrieval-augmented LLM framework for single-cell annotation that leverages a structured knowledge graph built from a refined CellMarker2.0 database.	[61]
<i>scBERT</i>	A transformer-based model inspired by LLMs such as <i>BERT</i> , repurposed for single-cell RNA-seq data. It leverages pretraining on large-scale unlabelled gene expression data followed by fine-tuning to accurately annotate cell types.	[43]
<i>scExtract</i>	An automated framework for single-cell RNA-seq that uses LLMs to extract insights from research articles, guiding data processing, integration, and annotation for large-scale meta-analysis with minimal manual effort.	[62]
<i>scInterpreter</i>	Harnesses the broad biological knowledge and reasoning capabilities of LLMs to interpret and classify cell types from gene expression data, demonstrating the value of integrating general-domain knowledge into single-cell analysis.	[63]
<i>scGPT</i>	Transformer-based foundation model that learns gene and cell embeddings from over 33 million cells, enabling cell type annotation, data integration, and perturbation prediction, offering high accuracy and interpretability.	[47, 64]

3.2. Advantages and challenges of LLM-based models for cell type annotation

Transformer-based LLMs show strong generalisation to unseen data and support user-guided annotation via chatbot interfaces [53]. Pre-trained on large-scale datasets, they leverage attention mechanisms to encode prior biological knowledge, enabling batch-insensitive annotations across diverse tissues, species, and technologies [49, 53]. LLMs facilitate reference-query data integration without needing explicit batch labels and can handle intra- and inter-dataset predictions [49]. They offer improved annotation consistency compared to manual methods and enable end-to-end analysis without additional fine-tuning, enhancing scalability and automation in single-cell genomics [48].

LLM-based models offer several advantages over traditional ML-based methods by enhancing scalability, interpretability, and accessibility in cell type annotation. For instance, *GPT-4* demonstrates both

cost-efficiency and seamless integration with existing single-cell analysis frameworks such as Seurat, removing the need for separate pipelines or curated reference datasets [16]. Its broad pretraining across extensive datasets enables robust generalisation across tissues and species, while its interactive chatbot nature allows for dynamic, user-guided annotation refinement [9, 16]. Beyond single-agent models, multi-agent frameworks like CASSIA extend these benefits by providing automated, accurate, and interpretable annotations, along with annotation-specific quality scores that flag uncertain predictions for manual review or model-based correction. Benchmarking studies have shown that such frameworks outperform traditional ML and semi-automated methods, particularly in complex datasets from cancer, immunology, and rare species [59]. Moreover, the capacity of LLMs to capture long-range dependencies and contextual associations between genes through transformer architectures enables improved modelling compared to ML methods that rely on linear dimensionality reduction and dataset-specific feature selection approaches [43]. Collectively, these advances illustrate the transformative potential of LLM-based annotation systems to deliver higher accuracy, broader applicability, and reduced reliance on expert intervention compared to conventional ML-based tools.

Despite their strengths, LLMs face several limitations. General-purpose LLMs such as GPT remain limited in biological applications [59]. Their lack of domain-specific pretraining often results in biologically irrelevant outputs, limited interpretability, and suboptimal performance in tasks like cell type annotation [10, 71]. High-quality and context-diverse pretraining datasets, spanning different cell types, disease states, tissues, genders, and species, are essential for improving the generalisability and biological relevance of single-cell LLMs [72, 73]. Incorporating quality verification strategies, such as online learning frameworks, can further refine dataset selection and ensure robust model performance across downstream tasks [74]. These models are also difficult to deploy due to their scale and reliance on proprietary application programming interfaces (APIs) [60]. For example, while *GPT-4* performs well on peripheral blood mononuclear cell (PBMC) and gastric cancer datasets, it struggles with more complex data like human embryonic cells. Similarly, *ERNIE 4.0* shows high accuracy in low-heterogeneity datasets, such as stromal cells, but lacks generalisability [15]. These shortcomings have prompted the development of domain-adapted solutions. Agent-based frameworks, such as *CellAgent* [14], integrate LLMs with bioinformatics tools in structured workflows, enabling context-aware, interpretable analysis of scRNA-seq data [75]. Fine-tuning LLMs on curated marker gene databases further improves their biological relevance and task-specific accuracy [16].

Integrating language-based models with omics data, structurally distinct from natural language, is inherently complex [53]. Evaluation is further complicated by non-deterministic outputs and the evolving nature of proprietary models like *GPT-4*. A lack of transparency in training data, biases, and hallucinations, plausible yet factually incorrect outputs that can skew scientific conclusions and undermine accuracy, pose significant concerns [48]. LLMs also struggle with rare or novel cell types, low-heterogeneity data, skewed cell type distributions, and rigid input formats. Annotation quality is often influenced by training data bias, and outputs typically require expert validation [16, 48]. High computational costs, environmental impact, and the need for extensive tuning limit reproducibility and accessibility. Overreliance on LLMs without rigorous human oversight increases the risk of misinterpretation in downstream analyses. [48, 53].

While ethical issues are relatively limited at this stage, as most LLM-based cell type annotation tools operate as standalone systems, reproducibility remains an ongoing challenge [15]. However, many recent tools have begun integrating mechanisms to improve transparency and consistency of results [14, 46, 60]. The opacity often associated with proprietary LLMs has been partially mitigated by developers openly disclosing training data sources, model architectures, and fine-tuning strategies, as summarized in [Supplementary Table 3](#). For instance,

CASSIA enhances interpretability by providing validator cross-checks, quantitative quality scores, and HTML reports documenting each decision [59], *CellReasoner* employs reasoning-augmented annotation through chain-of-thought (CoT) supervision that combines AI- and human-curated reasoning paths [60], and *ReCellTy* offers intermediate reasoning steps and transparent annotation tracing linked to selected features [61]. Moreover, most tools explicitly report their training datasets, improving data provenance and traceability [43, 46]. Nevertheless, challenges persist when these models rely on general-purpose LLMs trained on undisclosed or heterogeneous biomedical corpora, where complete data transparency and ethical assurance remain difficult to guarantee [14, 16].

4. Cell type annotation with isoform-resolved transcriptomics

Alternative splicing (AS) is a fundamental post-transcriptional mechanism that enables a single gene to produce multiple mRNA isoforms by selectively including or excluding specific exons during pre-mRNA processing [1, 76, 77] (Fig. 2A). By allowing genes to produce several transcript isoforms, transcriptomics complexity, protein diversity, and ultimately cellular complexity and their functional versatility can be increased, e.g., in species with fewer protein-coding genes [77–80].

Isoform diversity rises in proportion to the number of potential combinations of AS events [81] (Fig. 2B). Studies indicate that the fraction of genes that exhibit AS has progressively increased over the past 1.4 billion years of eukaryotic evolution, and is significantly correlated with organismal complexity (e.g. quantified by the number of unique cell types; Fig. 2B). Therefore, isoforms resulting from AS should also be a relevant consideration in cell type annotation, which could ultimately yield more profound insights.

4.1. Single-cell isoform sequencing

The power of single-cell RNA sequencing (scRNA-seq) to distinguish distinct cell types is often illustrated by the well-known fruit smoothie versus fruit salad analogy: Bulk RNA sequencing is like a smoothie, i.e. a blended mix of fruits, because gene expression is averaged across all cell types present, obscuring individual contributions. In contrast, scRNA-seq resembles a fruit salad, where each fruit remains identifiable, as reads are barcoded to indicate their cell of origin, enabling cell-specific gene expression analysis. While gene-level annotation captures the fruit types, single-cell isoform sequencing provides even finer resolution, distinguishing subtle differences such as ripeness or variety, analogous to detecting isoform-level diversity within a cell type (Fig. 2C). In the context of cell-type identification, single-cell isoform sequencing enables the discrimination of cellular subtypes and dynamic states [82–85], offering an unprecedented level of granularity particularly valuable for tissues with high cell-type diversity.

Single-cell isoform sequencing facilitates the elucidation of cell-specific gene and isoform expression, hence allowing for the identification of biological processes and molecular activities associated with both established and novel cell types. TGS platforms have been tested with single-cell platforms including droplet-, plate-, and microwell-based approaches (Fig. 3A) to advance transcriptome studies by capturing full-length transcripts at single-cell level and increase precision in cellular isoform characterisation. This approach has facilitated applications in human health research, including the identification of novel isoforms, fusion events, and potential neoepitopes for cancer vaccine development [7, 86, 87].

Pacific Bioscience's (PacBio's) Single-Molecule Real-Time (SMRT) sequencers utilise a silicon chip comprising 8 million (Sequel IIe) or 25 million (Revio) nanometre-scale wells, each acting as a chamber for high-fidelity cDNA sequencing reactions that occur in real time and produce fluorescent signals that must be base-called. Oxford Nanopore Technology (ONT) flowcells contain nanopores embedded in an electro-

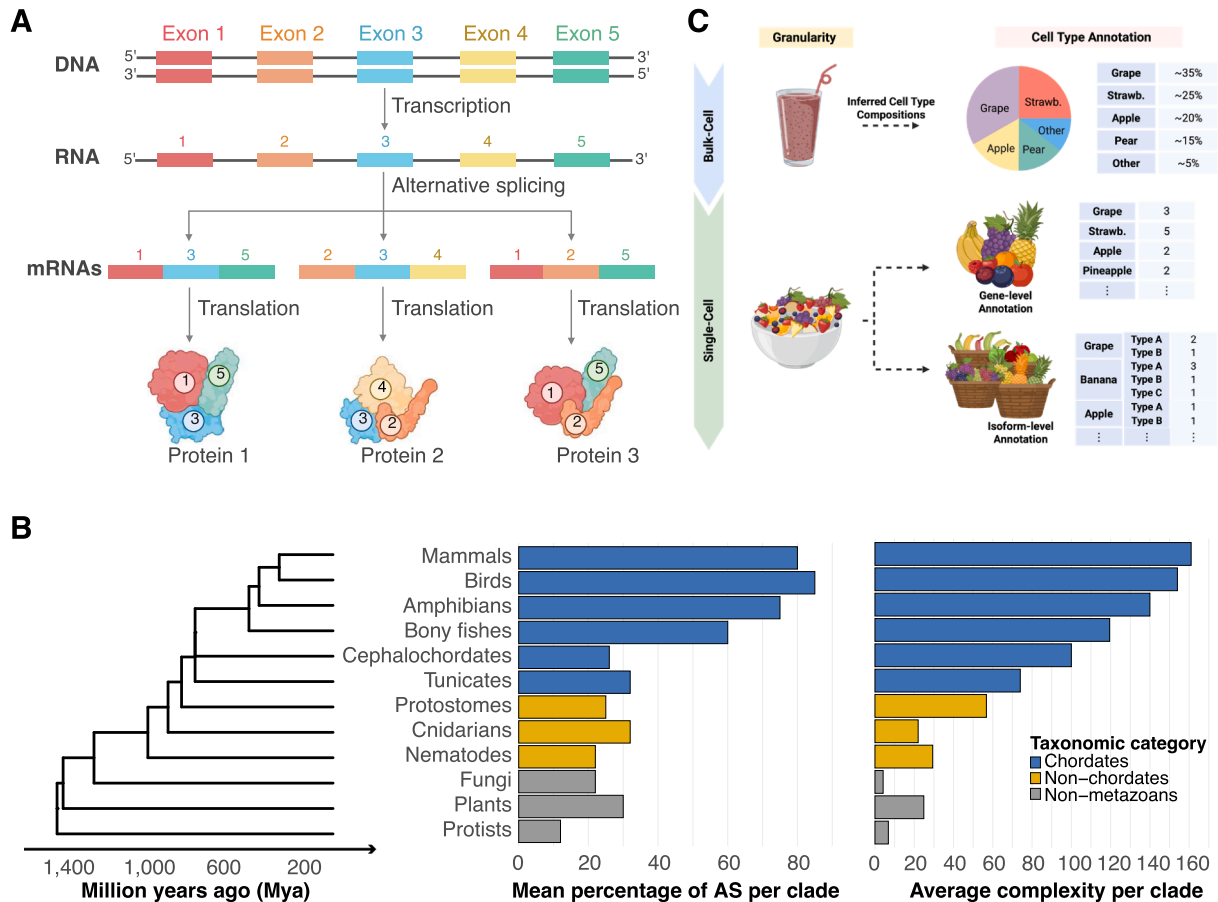


Fig. 2. Transcriptomic complexity in evolution and cell type annotation. (A) Process of AS and generation of transcript isoforms. Protein-coding genes are transcribed into precursor mRNA (pre-mRNA) containing exons, introns, and 5' and 3' untranslated regions (UTRs). During AS, introns are removed and exons are joined to produce mature mRNA transcripts. The selective inclusion or exclusion of specific exons generates multiple isoforms from a single gene, contributing to proteomic diversity. (B) The phylogenetic tree depicts the evolutionary relationships and divergence timelines of species groups, including mammals, spanning from the present to 1400 million years ago (MYA). The bar charts illustrate the average percentage of alternatively spliced genes for each animal group (middle) and the average complexity of organisms within each taxonomic group (right), quantified by the number of unique cell types as an indicator of organismal complexity. The colours of the bars are categorized based on taxonomic groups (chordates, nonchordate metazoans, or nonmetazoans) [1, 2]. (C) Extension of the fruit salad analogy to illustrate cell type annotation. Bulk RNA sequencing averages gene expression across all cells (smoothie), whereas single-cell RNA sequencing identifies individual cell types (fruit types in the salad). Single-cell isoform sequencing further enhances this resolution by capturing cellular subtypes and fine differences (ripeness or variety of each fruit in the salad).

resistant membrane allowing the identification of change in electric current as nucleotide fragments transfers the pores in real time.

PacBio's *Iso-seq* workflow and ONT's *wf-single-cell* workflow provide end-to-end analysis, from preprocessing raw reads to plotting Uniform Manifold Approximation and Projections (UMAPs). Bioinformatics tools such as *FLAMES* [88] and *SiCeLoRe* [6] implement sequential algorithms to extract barcodes, correct them, assemble transcripts, and generate transcript count matrices. Users can also call variants from their single-cell long-read sequencing data using these tools. After extracting barcodes, reads can be aligned using long-read aligner, such as *minimap2* [89], which retains barcode tags and is widely adopted for splice-aware long-read alignment.

For transcript identification and quantification, tools like *SCOTCH* [90], *Isosceles* [91] and *IsoQuant* [92] can be used. The resulting isoform count matrix can be used with downstream tools like *Seurat* [93] and *SCANPY* [94] for cell clustering. However, due to a lack of isoform-resolution markers for classification, clustering is typically performed based on gene expression. Isoforms can be classified and filtered using the *SQANTI* [95] workflow to improve cell types annotation (Fig. 3B).

4.2. Cell type identification in studies using single-cell long read sequencing

Recently, an increasing number of single-cell long-read sequencing studies have emerged. Here is how they have tackled cell type annotation for their data. Yang et al., (2023) performed parallel single-cell short (Illumina) and long-read sequencing (PacBio) in induced pluripotent stem cell-derived cerebral organoids. After the standard preprocessing of the short-read data and clustering using Seurat, marker genes for each cluster were identified using the FindMarker function [96]. The principal marker genes were then utilised for determining the cell types for each cluster. The long-read data was processed using the PacBio's *Iso-seq3* pipeline, followed by isoform annotation using *SQANTI* [95]. However, no refinement of cell type annotations was attempted with the help of the isoform-resolved data.

Dondi et al., (2023) performed single-cell short and long-read sequencing on ovarian cancer patient samples [87]. Cell types in short-read and long-read data were both annotated with *scROSHI* [71], which relies on a priori-defined cell type-specific genes [87]. The high similarity (Jaccard distance) between short and long-read-based cell clusters (exceeding 94 % for most cell types) demonstrated that long-read sequencing aligns well with short-read sequencing for cell

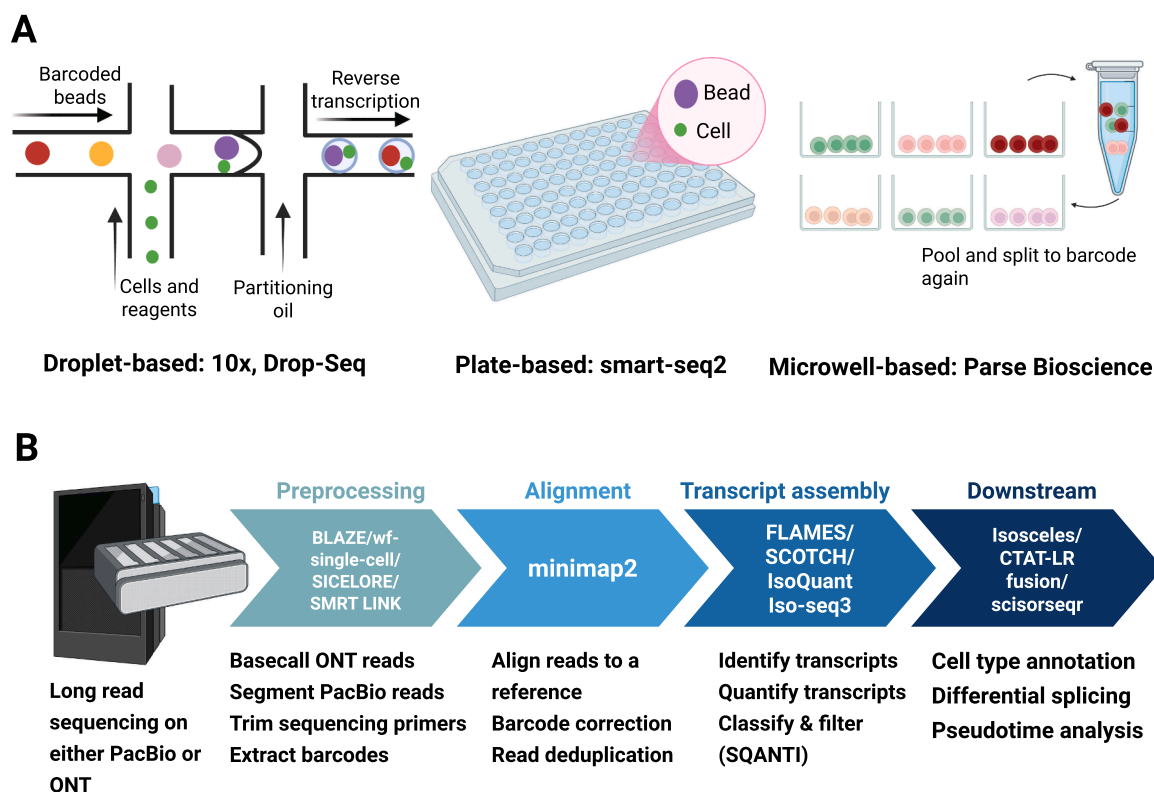


Fig. 3. Single-cell Long read RNA-seq workflow. (A) Different methods to barcode single cells include droplet based, plate based, and combinatorial indexing. (B) Basic bioinformatics analysis workflow with examples of tools available for each step.

type identification. Isoform expression analysis with *SQANTI* [95] revealed an increased transcriptomic diversification in cancer cells indicating that further subclassification of cell types could be achieved via long-read data.

Shiau et al., (2023) developed *scNanoGPS*, a tool that analyses single-cell nanopore sequencing data to capture both cellular genotypes (ie. mutations) and phenotypes (ie. gene isoform expression) for each cell. While the authors performed cell-type-specific isoform analysis, further refinement of cell types based on these patterns was not attempted. In contrast, tumour cells were identified using *CopyKAT*, which inferred chromosomal copy number alteration (CNA) profiles from the UMI count matrix, labelling cells with genome-wide CNAs as tumour cells [97].

In this context, Penter et al., (2024), developed *nanoranger*, a long-read sequencing workflow that leverages single-cell cDNA libraries to identify cell lineage-defining “natural barcodes,” such as single-nucleotide variants, fusion genes, isoforms, and sequences of chimeric antigen receptors and TCRs [84]. These barcodes are then used for amplicon-based long-read sequencing. Initial cell type annotation is performed using either canonical marker gene expression or alignment of single-cell profiles with a healthy reference dataset. Long-read sequencing then enables the detection of complex variants beyond the reach of short-read sequencing, allowing the tracking of single-cell tumour-immune co-evolution.

Byrne et al., (2024) developed *scTaILoR-seq*, a targeted long-read sequencing method that enhances isoform detection at single-cell resolution. Applied to ovarian cancer samples, it enabled isoform quantification, variant analysis, and allelic imbalance detection across cell populations. In single-cell data analysis, following preprocessing, cells were clustered using the Leiden algorithm, and the resulting clusters were annotated based on marker gene expression [86].

Li et al., (2024), combined short- and long-read single-cell RNA sequencing to build an isoform-resolution colorectal cancer atlas, identifying dysregulated and tumour-specific isoforms and recurrent

neopeptides for potential cancer vaccines. Cell types were annotated based on gene expression profiles via Seurat’s transfer learning from the Human Colon Cancer Atlas, with tumour epithelial cells refined using Xgboost and copy number variations inferred from the gene expression data. [7].

Collectively, these studies reveal that current approaches to cell type identification in single-cell long-read sequencing still rely heavily on gene-level expression and clustering-based methods. Studies such as Yang et al., (2023), Byrne et al., (2024), Penter et al., (2024), and Li et al., (2024), adopt reference-based annotation strategies, where cell clustering and marker gene identification are performed using tools like Seurat or external reference atlases, while long-read data are primarily used for isoform characterization and validation. Similarly, Dondi et al. [87] and Shiau et al. [97] combine clustering of single-cell profiles with classifiers based on predefined gene signatures or genomic features, such as *scROSHI* and *CopyKAT*, to assign cell identities, maintaining a gene-level focus while exploring isoform diversity within each cell type. Across these studies, clustering remains the foundational step for defining cellular populations, with annotation decisions guided predominantly by marker gene expression rather than isoform-level signatures, highlighting the continued reliance on short-read-based references and the need for computational models that can fully leverage isoform-resolved expression for finer cellular discrimination.

4.3. Challenges and opportunities in modelling cell types with single-cell long-read sequencing data

While short-read and long-read sequencing generally produce concordant gene expression estimates for highly expressed genes, this agreement weakens for low-abundance transcripts and rare cell types [98]. Moreover, short-read data often obscure isoform diversity, limiting the ability to distinguish between functionally or developmentally distinct cellular states. The diversity of transcript isoforms, which reflects cell-type-specific splicing programs, introduces a new and

biologically meaningful feature space for automated cell type annotation [17, 18].

Long-read sequencing technologies address many of the inherent limitations of short-read methods. In short-read sequencing, RNA molecules are fragmented into short segments (typically ~150 bp), making it challenging to accurately assign reads to similar isoforms of the same gene [99]. These methods frequently exhibit 3' bias, insufficient coverage across splice junctions, and difficulty detecting alternative polyadenylation, RNA editing, or fusion transcripts. In contrast, long-read approaches such as PacBio and Oxford Nanopore Technologies (ONT) can capture full-length transcripts in single reads, providing direct insights into complex RNA processing events and isoform coordination. By enabling comprehensive isoform reconstruction, long-read sequencing enhances the resolution of transcriptomic landscapes and offers a powerful foundation for improving the precision and interpretability of automated cell type annotation [17].

Single-cell long-read RNA sequencing generates substantial amounts of data, presenting both opportunities and challenges. Computational techniques must be fast and effective to handle these datasets [31]. ML advancements have enabled the development of fast and accurate computational models [100]. However, high-throughput biological data pose risks of overfitting due to their complexity and small sample sizes, particularly for rare cell types. Additionally, systematic sequencing biases may limit the application of classic learning models [101].

Due to the large dataset size, visualising and interpreting clustering results is challenging. Linear transformation techniques like PCA struggle to capture cellular relationships accurately due to high dropout rates and noise levels. Nonlinear techniques, such as t-Distributed Stochastic Neighbor Embedding (tSNE) and UMAP, offer more flexibility but require careful parameter selections, which significantly impacts visualisation outcomes [20].

The immense number of unannotated isoforms is another major limitation. Understanding the splicing mechanisms responsible for transcriptome diversity is crucial for enhancing the precision and efficiency of cell-fate determination modelling as well as refining model assumptions [17]. Since not all genes are relevant for cell type identification, classification models may suffer from overfitting, leading to suboptimal performances [102, 103]. Furthermore, the high frequency of zero reads (dropouts) in single-cell data complicates preprocessing and filtering [20].

Benchmarking of isoform-aware cell-type annotation tools is challenging since there is a lack of datasets combining long-read sequencing with well-curated cell-type labels. One exception is a murine hematopoietic development dataset that provides experimentally validated labels from FACS-isolated embryonic cells [104]. Another potential benchmark dataset is the LongBench cross-platform reference dataset, which was created using 8 cell lines from 3 different cancer types and profiled across 3 long-read sequencing technologies, including ONT, PacBio, and Illumina short-read sequencing. The dataset includes genotype-based cell line annotations [105].

5. Discussion

Traditional cell type classification methods heavily depend on prior knowledge and human input. However, since 2018, substantial progress has been made in automated cell type classification technologies. Current models, trained on publicly available single-cell RNA sequencing datasets, can now generate direct predictions of cell types without requiring extensive knowledge of cell markers [19]. This automation has significantly reduced the dependency on manual annotation.

Automatic cell type annotation technologies generally fall into two categories: supervised approaches and prior knowledge-based methods. Prior knowledge-based methods have not consistently outperformed other classifiers, as their effectiveness often hinges on the careful selection of marker genes. While most tools rely on ML methodologies, several non-ML models leveraging statistical techniques have also

demonstrated efficacy. The majority of these models have been developed for single-cell RNA sequencing, though a few have been adapted for bulk RNA sequencing data. However, to classify cell types effectively based on isoform expression data, specialised models designed explicitly for long-read RNA sequencing data are essential.

Based on published automatic cell type annotation benchmark studies, traditional classifiers such as SVM, RF, NNs, and Logistic Regression consistently demonstrate strong performance, particularly in well-annotated datasets like Tabula Muris [106] and PBMC3K (www.10xgenomics.com/datasets), where they often achieve F1 scores exceeding 90 %. However, these models struggle in more complex datasets, such as PBMC10K (www.10xgenomics.com/datasets) and snHeart [107], where closely related cell types are more challenging to distinguish.

Among annotation tools, scmap-cell [33], scPred [29], ACTINN [31], and Cell BLAST [36] have performed well in large datasets, while *SingleR* [34], *CP*, and *RPC* [108] have excelled in simulated datasets. However, certain methods, such as *scID* [109] and *Garnett* [110], have underperformed, particularly in self-projection tasks and datasets with heterogeneous cell populations [26]. Notably, while *scVI* [38] has demonstrated strong performance in certain datasets, it has faced challenges in highly annotated datasets such as Tabula Muris and AMB92 (Allan Mouse Brain, <http://celltypes.brain-map.org/rnaseq>).

Computational efficiency is another critical consideration in cell type annotation. Based on computational time analysis, scmap-cluster [33] has been identified as an efficient model [27]. Many automatic cell type annotation systems incorporate feature selection procedure, which significantly impact model performance. Tools such as *Clustifyr* [111], *SCENIC* [112], and prior knowledge-based methods can be particularly useful for the detection of tumour cells [19].

5.1. Emerging trends in cell type annotation

LLMs have emerged as a promising tool for enhancing cell type annotation. By integrating gene expression data with generative AI, LLMs improve the accuracy of identifying rare and complex cell types. Nevertheless, challenges such as data variability and potential inconsistencies persist. Benchmarking studies, such as SOAR [69] and AnnDictionary [13], have demonstrated that LLMs can achieve high accuracy levels (80 – 90 %) in principal cell-type annotations. Emerging tools, such as *LICT* and *GPTCelltype*, offer further advancements in the field.

Owing to technological advancements, particularly in read length, accuracy, and the decreasing cost of sequencing, single-cell long-read RNA sequencing is increasingly being utilised in biological research [84, 95, 96]. This method enables comprehensive molecular profiling and precise annotation of diverse cell types across nearly all tissues of an organism. Accurate identification of cell types is crucial for researchers utilising single-cell long-read sequencing data, as it facilitates a deeper understanding of transcriptome complexity and isoform diversity concerning cell identity, fate, and state transitions [17]. However, two major limitations remain: the relatively high error rate of long-read sequencing compared to short-read sequencing and the uncertainty surrounding the 5' end of transcripts.

As single-cell long-read RNA sequencing continues to advance, it is imperative to develop specialized annotation models tailored to isoform-level resolution. By addressing current limitations, researchers can enhance the precision and reliability of automated cell type classification, ultimately driving further discoveries in cellular biology and disease research.

The continued advancement of single-cell LLMs will benefit from integrating multi-modal datasets, combining scRNA-seq with isoform-resolved and epigenetic information to capture a more comprehensive view of cellular identity and regulatory mechanisms. Self-supervised pretraining on RNA and single-cell foundation models, leveraging diverse RNA types, including coding, non-coding, and UTR sequences,

across multiple organisms, can enhance the ability of models to learn robust, context-aware representations for downstream tasks such as cell type annotation, trajectory inference, and functional prediction [43, 49]. While current LLM applications to isoform-level transcriptomics remain limited, emerging frameworks like *IsoFormer* demonstrate the potential of language-model architectures to integrate multi-modal biological information and capture isoform-level diversity [113]. Complementary tools like *IsoDiffR*, reveals cell-type-specific isoform usage and functional divergence [114].

Training models on continuous-valued, high-dimensional RNA isoform expression matrices poses significant challenges due to the large number of isoforms per gene, sparsity of single-cell measurements, and the need to distinguish isoform-driven from gene-driven expression patterns [82]. Transformer-based approaches such as *IsoFormer* and *IsoDiffR* overcome these challenges by leveraging multi-modal encoders, attention mechanisms, and robust aggregation strategies to capture isoform-specific patterns across cells while integrating complementary DNA, RNA, and protein information [113, 114]. Similarly, these strategies could be extended to isoform-aware LLM architectures, where isoform embeddings and attention-based modeling may enable accurate, context-aware cell type annotation from high-dimensional isoform matrices.

Tokenization and embedding are fundamental steps in adapting LLMs for biological data, enabling the conversion of raw RNA sequences or expression profiles into formats suitable for computational analysis. In RNA sequence data, tokenization methods such as one-hot encoding and k-mer segmentation translate nucleotide strings (A, U, C, G) into numerical representations [49, 115]. For single-cell data, tokenization can be based on gene ranking, expression binning, pathway grouping, or patch-based segmentation of expression matrices [43, 47, 64]. Embedding then maps these tokens into continuous vector spaces, capturing semantic and positional relationships among biological features [49]. Extending these strategies to isoform-level modelling could allow the representation of transcript variants and splicing events, creating isoform-aware embeddings that capture transcript-specific regulatory information. While current models primarily rely on gene-aware embeddings to represent cell states, the development of isoform-level tokenization and embeddings remains an open and promising research direction for improving automated cell type annotation and understanding splicing-driven cellular diversity.

Additionally, explainable AI plays a crucial role in validating LLM-based biological predictions by linking model outputs to interpretable biological features such as genes, pathways, or isoforms. By making tokenization and embedding steps biologically transparent, explainable AI enables researchers to trace how specific molecular patterns influence model decisions, thereby increasing trust and interpretability in LLM-driven single-cell analyses [49].

Authors' contributions

HW surveyed the literature and wrote the first draft. SB provided substantial technical advice on single-cell isoform sequencing technologies. NB and DS contributed paragraphs on NLP and LLMs. SW and US supervised the work, supported data visualisation, and helped with reviewing and revising the manuscript. All authors read and approved the final manuscript.

CRedit authorship contribution statement

Debarka Sengupta: Writing – review & editing, Supervision. **Siyuan Wu:** Writing – review & editing, Supervision, Conceptualization. **Ulf Schmitz:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Hettiarachchi Wijewardena:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Saloni Bhatia:** Writing – original draft, Visualization, Investigation. **Namrata Bhattacharya:** Writing – original

draft, Conceptualization.

Funding

This work was supported by the National Health and Medical Research Council (Grant #1196405 to US); the Tropical Australian Academic Health Centre (Grant #SF01124); the Townsville University Hospital (Grant #THSSERTA_RPG05_2024, THSSERTA_RPG15_2024, and THSSERTA_RCG05_2024). HW is supported by a James Cook University International Higher Degree Research Fellowship.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.11.008.

References

- [1] Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res* 2018; 46:582–92.
- [2] Wang H, Yang Y, Liu J, Qian L. Direct cell reprogramming: approaches, mechanisms and progress. *Nat Rev Mol Cell Biol* 2021;22:410–24.
- [3] McIntire E, Barr KA, Gonzales NM, Gilad Y. Guided differentiation of pluripotent stem cells for cardiac cell diversity. *bioRxiv Prepr* 2025. <https://doi.org/10.1101/2023.07.21.550072>.
- [4] Xu J, Du Y, Deng H. Direct lineage reprogramming: strategies, mechanisms, and applications. *Cell Stem Cell* 2015;16:119–34.
- [5] Sadelain M, Rivière I, Riddell S. Therapeutic T cell engineering. *Nature* 2017;545: 423–31.
- [6] Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat Commun* 2020;11: 4025.
- [7] Li Z, Zhang B, Chan JJ, Tabatabaiean H, Tong QY, Chew XH, Fan X, Driguez P, Chan C, Cheong F, et al. An isoform-resolution transcriptomic atlas of colorectal cancer from long-read single-cell sequencing. *Cell Genom* 2024;4:100641.
- [8] Clarke ZA, Andrews TS, Atif J, Pouyababar D, Innes BT, MacParland SA, Bader GD. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nat Protoc* 2021;16:2749–64.
- [9] Chen J, Zhang J, Yao H, Li Y. CellTypeAgent: trustworthy cell type annotation with large language models. *arXiv Prepr* 2025. <https://doi.org/10.48550/arXiv.2505.08844>.
- [10] Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;20:194.
- [11] Pouyan MB, Nourani M. Clustering single-cell expression data using random forest graphs. *IEEE J Biomed Health Inf* 2017;21:1172–81.
- [12] Cheng P, Peng Y, Zhang XL, Chen S, Fang BBin, Li YZ, Sun YM. A natural language processing system for the efficient extraction of cell markers. *Sci Rep* 2024;14: 21183.
- [13] Crowley G, Quake SR. Benchmarking cell type annotation by large language models with anndictionary. *bioRxiv Prepr* 2024. <https://doi.org/10.1101/2024.10.10.617605>.
- [14] Xiao Y, Liu J, Zheng Y, Xie X, Hao J, Li M, Wang R, Ni F, Li Y, Luo J, et al. CellAgent: an LLM-driven multi-agent framework for automated single-cell data analysis. *arXiv Prepr* 2024. <https://doi.org/10.48550/arXiv.2407.09811>.
- [15] Ye W, Xiang J, Ma Y, Liang H, Wang T, Xiang Q, Peng Xiang A, Li W, Huang W. Objectively evaluating the reliability of cell type annotation using LLM-based strategies. *arXiv Prepr* 2024. <https://doi.org/10.48550/arXiv.2409.15678>.
- [16] Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;21:1462–5.
- [17] Wu S, Schmitz U. Single-cell and long-read sequencing to enhance modelling of splicing and cell-fate determination. *Comput Struct Biotechnol J* 2023;21: 2373–80.
- [18] Wu S, Schmitz U. ScsoX: a multidimensional framework for measuring isoform-level transcriptomic complexity in single cells. *Genome Biol* 2025;26:289.
- [19] Xie B, Jiang Q, Mora A, Li X. Automatic cell type identification methods for single-cell RNA sequencing. *Comput Struct Biotechnol J* 2021;19:5874–87.
- [20] Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;20:273–82.

- [21] Wagner F, Yanai I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. *bioRxiv Prepr* 2018. <https://doi.org/10.1101/456129>.
- [22] Kim T, Lo K, Geddes TA, Kim HJ, Yang JYH, Yang P. ScReClassify: Post hoc cell type classification of single-cell RNA-seq data. *BMC Genom* 2019;20:913.
- [23] Liu X, Gosline SJC, Pflieger LT, Wallet P, Iyer A, Guinney J, Bild AH, Chang JT. Knowledge-based classification of fine-grained immune cell types in single-cell RNA-Seq data. *Brief Bioinform* 2021;22:bbab039.
- [24] Domanski S, Hakansson A, Bertus TJ, Paternostro G, Piermarocchi C. Digital Cell Sorter (DCS): a cell type identification, anomaly detection, and Hopfield landscapes toolkit for single-cell transcriptomics. *BMC Bioinforma* 2019;9:10670.
- [25] Huang Y, Zhang P. Evaluation of machine learning approaches for cell-type identification from single-cell transcriptomics data. *Brief Bioinform* 2021;22:bbab217.
- [26] Huang Q, Liu Y, Du Y, Garmire LX. Evaluation of cell type annotation R packages on single-cell RNA-seq Data. *Genom Proteom Bioinforma* 2021;19:267–81.
- [27] Zhao X, Wu S, Fang N, Sun X, Fan J. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Brief Bioinform* 2020;21:1581–95.
- [28] Tortelote G. Benchmarking machine learning models for cell type annotation in single-cell vs single-nucleus RNA-seq data. *Res Sq Prepr* 2025. <https://doi.org/10.21203/rs.3.rs-5754289/v1>.
- [29] Alquicira-Hernandez J, Sathie A, Ji HP, Nguyen Q, Powell JE. ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* 2019;20:264.
- [30] Tan Y, Cahan P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Syst* 2019;9:207–213.e2.
- [31] Ma F, Pellegrini M. ACTINN: Automated identification of cell types in single cell RNA sequencing. *Bioinformatics* 2020;36:533–8.
- [32] Lieberman Y, Rokach L, Shay T. CaSTLe - Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* 2018;13:e0205499.
- [33] Kiselev VYu, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 2018;15:359–62.
- [34] Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;20:163–72.
- [35] Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, Mahrt E, Guo W, Stawiski EW, Modrusan Z, et al. Scina: Semi-supervised analysis of single cells in silico. *Genes (Basel)* 2019;10:531.
- [36] Cao ZJ, Wei L, Lu S, Yang DC, Gao G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat Commun* 2020;11:3458.
- [37] de Kanter JK, Lijnzaad P, Candelli T, Margaritis T, Holstege FCP. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 2019;47:E95.
- [38] Ergen C, Pour Amir VV, Kim M, Streets A, Gayoso A, Yosef N. Scvi-hub: an actionable repository for model-driven single cell analysis. *bioRxiv Prepr* 2024. <https://doi.org/10.1101/2024.03.01.582887>.
- [39] Lakkis H. (2023) Assessment of Cell type Annotation Tools of Single Cell RNA Sequencing Data. *McGill University, Montreal*.
- [40] Sun H, Zhou Y, Fei L, Chen H, Guo G. scMCA: A tool to define mouse cell types based on single-cell digital expression. In *Methods in Molecular Biology*, 1935. Humana Press Inc.; 2019. p. 91–6.
- [41] Pasquini G, Rojo Arias JE, Schäfer P, Busskamp V. Automated methods for cell type annotation on scRNA-seq data. *Comput Struct Biotechnol J* 2021;19:961–9.
- [42] Xie P, Gao M, Wang C, Zhang J, Noel P, Yang C, Von Hoff D, Han H, Zhang MQ, Lin W. SuperCT: A supervised-learning framework for enhanced characterization of single-cell transcriptomic profiles. *Nucleic Acids Res* 2019;47:e48.
- [43] Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, Lu H, Yao J. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat Mach Intell* 2022;4:852–66.
- [44] Li C, Liu B, Kang B, Liu Z, Liu Y, Chen C, Ren X, Zhang Z. SciBet as a portable and fast single cell type identifier. *Nat Commun* 2020;11:1818.
- [45] Qiu P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat Commun* 2020;11:1169.
- [46] Zhao H, Liu Y, Yao J, Xiong L, Zhou Z, Zhang Z. Celler: A Genomic Language Model for Long-Tailed Single-Cell Annotation. *arXiv Prepr* 2025. <https://doi.org/10.48550/arXiv.2504.00020>.
- [47] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;21:1470–80.
- [48] Kazmi A, Singh D, Jatav S, Luthra S. Beyond the Hype: The Complexity of Automated Cell Type Annotations with GPT-4. *bioRxiv Prepr* 2025. <https://doi.org/10.1101/2025.02.11.637659>.
- [49] Liu J, Yang M, Yu Y, Xu H, Wang T, Li K, Zhou X. Advancing bioinformatics with large language models: components, applications and perspectives. *arXiv Prepr* 2025. <https://doi.org/10.48550/arXiv.2401.04155>.
- [50] Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *arXiv Prepr* 2023. <https://doi.org/10.48550/arXiv.1706.03762>.
- [51] Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv Prepr* 2019. <https://doi.org/10.48550/arXiv.1810.04805>.
- [52] Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al. On the Opportunities and Risks of Foundation Models. *arXiv Prepr* 2022. <https://doi.org/10.48550/arXiv.2108.07258>.
- [53] Szalata A, Hrovatin K, Becker S, Tejada-Lapueta A, Cui H, Wang B, Theis FJ. Transformers in single-cell omics: a review and new perspectives. *Nat Methods* 2024;21:1430–43.
- [54] Zeng Z, Du H. Revolutionizing Single Cell Analysis: The Power of Large Language Models for Cell Type Annotation. *arXiv Prepr* 2023. <https://doi.org/10.48550/arXiv.2304.02697>.
- [55] Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxf)* 2019. <https://doi.org/10.1093/database/baz046>.
- [56] Zhao Y, Cai H, Zhang Z, Tang J, Li Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat Commun* 2021;12:5261.
- [57] Mao S, Zhang Y, Seelig G, Kannan S. CellMeSH: Probabilistic cell-type identification using indexed literature. *Bioinformatics* 2022;38:1393–402.
- [58] Bhattacharya N, Rockstroh A, Deshpande SS, Thomas SK, Yadav A, Goswami C, Chawla S, Solomon P, Fourgeux C, Ahuja G, et al. Pseudo-grading of tumor subpopulations from single-cell transcriptomic data using Phenotype Algebra. *eLife* 2024:RP98469. <https://doi.org/10.7554/eLife.98469.1>.
- [59] Xie E, Cheng L, Shireman J, Cai Y, Liu J, Mohanty C, Dey M, Kendziorzi C. CASSIA: a multi-agent large language model for reference free, interpretable, and automated cell annotation of single-cell RNA-sequencing data. *bioRxiv Prepr* 2024. <https://doi.org/10.1101/2024.12.04.626476>.
- [60] Cao G, Shen Y, Wu J, Chao H, Chen M, Chen D. CellReasoner: A reasoning-enhanced large language model for cell type annotation. *bioRxiv Prepr* 2025. <https://doi.org/10.1101/2025.05.20.655112>.
- [61] Han D, Jia Y, Chen R, Han W, Guo S, Wang J. ReCellTy: domain-specific knowledge graph retrieval-augmented LLMs workflow for single-cell annotation. *arXiv Prepr* 2025. <https://doi.org/10.48550/arXiv.2505.00017>.
- [62] Wu Y, Tang F. scExtract: leveraging large language models for fully automated single-cell RNA-seq data annotation and prior-informed multi-dataset integration. *Genome Biol* 2025;26:174.
- [63] Li C, Xiao M, Wang P, Feng G, Li X, Zhou Y. scInterpreter: Training Large Language Models to Interpret scRNA-seq Data for Cell Type Annotation. *arXiv Prepr* 2024. <https://doi.org/10.48550/arXiv.2402.12405>.
- [64] Ding S, Li J, Luo R, Cui H, Wang B, Chen R. scGPT: end-to-end protocol for fine-tuned retinal cell type annotation. *Nat Protoc* 2025. <https://doi.org/10.1038/s41596-025-01220-1>.
- [65] Wen H, Tang W, Dai X, Ding J, Jin W, Xie Y, Tang J. CellPLM: pre-training cell language model beyond single cells. *bioRxiv Prepr* 2023. <https://doi.org/10.1101/2023.10.03.560734>.
- [66] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, Mantineo H, Brydon EM, Zeng Z, Liu XS, et al. Transfer learning enables predictions in network biology. *Nature* 2023;618:616–24.
- [67] Chen Y, Zou J. GenePT: a simple but effective foundation model for genes and cells built from ChatGPT. *bioRxiv Prepr* 2024. <https://doi.org/10.1101/2023.10.16.562533>.
- [68] Kharouiche O, Markogiannakis A, Fei X, Chatzianastasis M, Vazirgiannis M. Cell2Text: Multimodal LLM for Generating Single-Cell Descriptions from RNA-Seq Data. *arXiv Prepr* 2025. <https://doi.org/10.48550/arXiv.2509.24840>.
- [69] Liu J, Xu S, Zhang L, Zhang J. Single-Cell Omics Arena: A Benchmark Study for Large Language Models on Cell Type Annotation Using Single-Cell Data. *arXiv Prepr* 2024. <https://doi.org/10.48550/arXiv.2412.02915>.
- [70] Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, Salzman J, Yosef N, Bulthaupt B, Brown P, Harper W, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (1979)* 2022;376:eab14896.
- [71] Prummer M, Bertolini A, Bosshard L, Barkmann F, Yates J, Boeva V, Tumor Profiler Consortium T, Stekhoven D, Singer F. scROSHI: robust supervised hierarchical identification of single cells. *NAR Genom Bioinform* 2023;5:lqad058.
- [72] Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *arXiv Prepr* 2023. <https://doi.org/10.48550/arXiv.2307.09009>.
- [73] Ji Y, Deng Y, Gong Y, Peng Y, Niu Q, Zhang L, Ma B, Li X. Exploring the Impact of Instruction Data Scaling on Large Language Models: An Empirical Study on Real-World Use Cases. *arXiv Prepr* 2023. <https://doi.org/10.48550/arXiv.2303.14742>.
- [74] Carta T, Romac C, Wolf T, Lamprier S, Sigaud O, Oudeyer P-Y. Grounding large language models in interactive environments with online reinforcement learning. *arXiv Prepr* 2024. <https://doi.org/10.48550/arXiv.2302.02662>.
- [75] Wang X, Chai Z, Li S, Liu Y, Li C, Jiang Y, Liu Q. CTISL: a dynamic stacking multi-class classification approach for identifying cell types from single-cell RNA-seq data. *Bioinformatics* 2024;40:btac063.
- [76] Baralle FE, Giudice J. Alternative splicing as a regulator of development and tissue identity. *Nat Rev Mol Cell Biol* 2017;18:437–51.
- [77] Tao Y, Zhang Q, Wang H, Yang X, Mu H. Alternative splicing and related RNA binding proteins in human health and disease. *Signal Transduct Target Ther* 2024;9:26.
- [78] Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley MC, Shini S, Lieschke GJ, Wong JLL, Rasko JEJ. Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol* 2017;18:216.
- [79] Sinitcyn P, Richards AL, Weatheritt RJ, Brademan DR, Marx H, Shishkova E, Meyer JG, Hebert AS, Westphall MS, Blencowe BJ, et al. Global detection of human variants and isoforms by deep proteome sequencing. *Nat Biotechnol* 2023;41:1776–86.
- [80] Roth JF, Braunschweig U, Wu M, Li JD, Lin ZY, Larsen B, Weatheritt RJ, Gingras AC, Blencowe BJ. Systematic analysis of alternative exon-dependent

- interactome remodeling reveals multitasking functions of gene regulatory factors. *Mol Cell* 2023;83:4222–38. [e10.](#)
- [81] Arzalluz-Luqueángeles, Conesa A. Single-cell RNAseq for the study of isoforms-how is that possible? *Genome Biol* 2018;19:110.
- [82] Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, Koopmans F, Barres B, Smit AB, Sloan SA, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* 2018;36:1197–202.
- [83] Xiang X, He Y, Zhang Z, Yang X. Interrogations of single-cell RNA splicing landscapes with SCASL define new cell identities with physiological relevance. *Nat Commun* 2024;15:2164.
- [84] Penter L, Borji M, Nagler A, Lyu H, Lu WS, Cieri N, Maurer K, Oliveira G, Al Khafaji AM, Garimella KV, et al. Integrative genotyping of cancer and immune phenotypes by long-read sequencing. *Nat Commun* 2024;15:32.
- [85] Mincarelli L, Uzun V, Wright D, Scoones A, Rushworth SA, Haerty W, Macaulay IC. Single-cell gene and isoform expression analysis reveals signatures of ageing in haematopoietic stem and progenitor cells. *Commun Biol* 2023;6:558.
- [86] Byrne A, Le D, Sereti K, Menon H, Vaidya S, Patel N, Lund J, Xavier-Magalhães A, Shi M, Liang Y, et al. Single-cell long-read targeted sequencing reveals transcriptional variation in ovarian cancer. *Nat Commun* 2024;15:6916.
- [87] Dondi A, Lischetti U, Jacob F, Singer F, Borgsmüller N, Coelho R, Heinzelmann-Schwarz V, Beisel C, Beerenwinkel N. Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer. *Nat Commun* 2023;14:7780.
- [88] Tian L, Jabbari JS, Thijssen R, Gouli Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MRM, Schuster J, Wang C, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biol* 2021;22:310.
- [89] Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100.
- [90] Xu Z, Qu H-Q, Chan J, Kao C, Hakonarson H, Wang K. Single-Cell Omics for Transcriptome Characterization (SCOTCH): isoform-level characterization of gene expression through long-read single-cell RNA sequencing. *bioRxiv Prepr* 2025. <https://doi.org/10.1101/2024.04.29.590597>.
- [91] Kabza M, Ritter A, Byrne A, Sereti K, Le D, Stephenson W, Sterne-Weiler T. Accurate long-read transcript discovery and quantification at single-cell, pseudo-bulk and bulk resolution with IsoSees. *Nat Commun* 2024;15:7316.
- [92] Pribelski AD, Mikheenko A, Joglekar A, Smetanin A, Jarroux J, Lapidus AL, Tilgner HU. Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* 2023;41:915–8.
- [93] Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;42:293–304.
- [94] Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:15.
- [95] Tardaguila M, De La Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K, et al. SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* 2018;28:396–411.
- [96] Yang Y, Yang R, Kang B, Qian S, He X, Zhang X. Single-cell long-read sequencing in human cerebral organoids uncovers cell-type-specific and autism-associated exons. *Cell Rep* 2023;42:113335.
- [97] Shiau CK, Lu L, Kieser R, Fukumura K, Pan T, Lin HY, Yang J, Tong EL, Lee GH, Yan Y, et al. High throughput single cell long-read sequencing analyses of same-cell genotypes and phenotypes in human tumors. *Nat Commun* 2023;14:4124.
- [98] You Y, Solano A, Lancaster J, David M, Wang C, Su S, Zeglinski K, Ghamsari R, Chauhan M, Gleeson J, et al. Benchmarking long-read RNA-sequencing technologies with LongBench: a cross-platform reference dataset profiling cancer cell lines with bulk and single-cell approaches. *bioRxiv Prepr* 2025. <https://doi.org/10.1101/2025.09.11.675724>.
- [99] Monzó C, Liu T, Conesa A. Transcriptomics in the era of long-read sequencing. *Nat Rev Genet* 2025;26:681–701.
- [100] Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ. Single cells make big data: New challenges and opportunities in transcriptomics. *Curr Opin Syst Biol* 2017;4:85–91.
- [101] Guo Y, Liu S, Li Z, Shang X. BCDForest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC Bioinforma* 2018;19:118.
- [102] Hia NT, Sumon Ahmed. Automatic cell type annotation using supervised classification: A systematic literature review. *Syst Lit Rev MetaAnal J* 2022;3:99–108.
- [103] Qi R, Ma A, Ma Q, Zou Q. Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinform* 2019;21:1196–208.
- [104] Wang F, Tan P, Zhang P, Ren Y, Zhou J, Li Y, Hou S, Li S, Zhang L, Ma Y, et al. Single-cell architecture and functional requirement of alternative splicing during hematopoietic stem cell formation. *Sci Adv* 2022;8:5369.
- [105] You Y, Solano A, Lancaster J, David M, Wang C, Su S, Zeglinski K, Su R, Chauhan M, Gleeson J, et al. Benchmarking long-read RNA-sequencing technologies with LongBench: a cross-platform reference dataset profiling cancer cell lines with bulk single-cell approaches. *bioRxiv Prepr* 2025. <https://doi.org/10.1101/2025.09.11.675724>.
- [106] Schaum N, Karkanas J, Neff NF, May AP, Quake SR, Wyss-Coray T, Darmanis S, Batson J, Botvinnik O, Chen MB, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 2018;562:367–72.
- [107] Selewa A, Dohn R, Eckart H, Lozano S, Xie B, Gauchat E, Elorbany R, Rhodes K, Burnett J, Gilad Y, et al. Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation. *bioRxiv Prepr* 2019. <https://doi.org/10.1101/585901>.
- [108] Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinforma* 2017;18:105.
- [109] Boufea K, Seth S, Batada NN. scID Uses Discriminant Analysis to Identify Transcriptionally Equivalent Cell Types across Single-Cell RNA-Seq Data with Batch Effect. *iScience* 2020;23:100914.
- [110] Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* 2019;16:983–6.
- [111] Fu R, Gillen AE, Sheridan RM, Tian C, Daya M, Hao Y, Hesselberth JR, Riemondy KA. clustifyr: an R package for automated single-cell RNA sequencing cluster classification. *F1000Res* 2020;9:223.
- [112] Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, Poovathingal S, Wouters J, Aibar S, Aerts S. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;20:1355–67.
- [113] Garau-Luis JJ, Bordes P, Gonzalez L, Roller M, de Almeida BP, Hexemer L, Blum C, Laurent S, Grzegorzewski J, Lang M, et al. Multi-modal Transfer Learning between Biological Foundation Models. *arXiv Prepr* 2024. <https://doi.org/10.48550/arXiv.2406.14150>.
- [114] Shi,Z.-X., Qian,Y., Zheng,Z.-C., Zheng,Y. and Dai,Q. (2025) Analyzing cell-type-specific isoform expression using IsoDiffR and long-read single-cell RNA sequencing.
- [115] Akiyama M, Sakakibara Y. Informative RNA base embedding for RNA structural alignment and clustering by deep representation learning. *NAR Genom Bioinform* 2022;4:lqac012.
- [116] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;6:377–82.
- [117] Cahan P, Li H, Morris SA, Lummertz Da Rocha E, Daley GQ, Collins JJ. CellNet: Network biology applied to stem cell engineering. *Cell* 2014;158:903–15.
- [118] Method of the Year 2022. long-read sequencing. *Nat Methods* 2023;20(1).