



# Immune cell annotation in the single-cell studies: technologies, challenges, and integrative solutions

Sabrina George<sup>1</sup> · Nor Adzimah Johdi<sup>1</sup>

Received: 17 October 2025 / Accepted: 20 April 2026  
© The Author(s) 2026

## Abstract

Single-cell RNA sequencing has transformed immunological research by enabling high-resolution transcriptional profiling of individual immune cells. Despite its transformative impact, annotating immune cells based solely on transcriptomic data remains challenging. These difficulties arise from biological factors, including gene expression heterogeneity and post-transcriptional regulation, as well as technical limitations that contribute to mismatches between mRNA and protein expression. Such discrepancies may lead to cell misclassification and obscure functional insights, particularly in heterogeneous populations such as peripheral blood mononuclear cells. This review highlights the major challenges in immune cell annotation by detailing the mechanisms underlying mRNA-protein discrepancies, examining both the biological factors and technical artifacts that drive this divergence, and emphasizing their implications for accurate cell classification. A critical overview of current single-cell profiling technologies follows, with evaluation of the respective advantages and limitations of transcriptomic, proteomic, and multimodal approaches. In particular, technologies such as Cellular Indexing of Transcriptomes and Epitopes by Sequencing integrate transcriptomic and proteomic data, addressing the shortcomings of single-modality analyses. Further examination focuses on computational strategies for immune cell annotation, with emphasis on automated methods and bioinformatics frameworks tailored to multi-omics datasets. The unique computational challenges of integrating mRNA and protein data, together with solutions for improved annotation accuracy, are discussed. This review integrates key challenges, technologies, and computational tools, highlighting the need for standardized multimodal profiling of immune cells. Such integration enhances annotation reliability and advances disease understanding and therapy discovery.

**Keywords** Single-cell RNA sequencing · Annotation · Immune cells profiling · Transcriptomic · Proteomic · Multimodal approaches

## Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized the field of immunology by enabling detailed transcriptional profiling at single-cell resolution. This technology allows researchers to dissect immune cell heterogeneity, trace developmental trajectories, and study dynamic immune responses in both health and disease [1, 2]. However, one of the primary limitations of scRNA-seq is its reliance on transcriptomic data alone for cell-type annotation.

This poses particular challenges in studies of peripheral blood mononuclear cells (PBMCs), where closely related immune subsets often exhibit overlapping gene expression profiles [3, 4]. Accurate annotation is essential for identifying immune responses, uncovering disease-associated signatures, and guiding therapeutic interventions. For example, precise classification of monocyte subsets in COVID-19 and influenza has revealed immune signatures linked to disease severity and treatment targets [5].

A key obstacle to accurately annotating immune cells is the discordance between transcript and protein abundance. Because immune-cell identity and function are chiefly governed at the protein stage via post-transcriptional regulation, translational control, and protein degradation [6], mRNA readouts can diverge markedly from the corresponding protein expression, leading to misclassification of cell types

---

✉ Nor Adzimah Johdi  
adzimah@hctm.ukm.edu.my

<sup>1</sup> UKM Medical Molecular Biology Institute (UMBI), National University Malaysia, Cheras, Malaysia

[7]. For instance, CD15 surface marker, encoded by the gene *FUT4*, is highly expressed on myeloid-derived suppressor cells (MDSCs), yet its mRNA is often undetectable in scRNA-seq, highlighting how transcript-based annotation may fail to capture immunosuppressive populations that are clearly defined at the protein level [6]. Similarly, immune checkpoint protein PD-1, encoded by *PDCDI*, may be abundantly expressed on the surface of exhausted T cells while showing minimal mRNA expression, resulting in the underestimation of immunoregulatory states when relying solely on transcriptomic data [8].

This discordance is further compounded by technical limitations inherent to scRNA-seq. For example, dropout events, caused by low capture efficiency or stochastic gene expression, can result in undetected transcripts, especially those expressed at low abundance [9]. Additionally, transcripts with short half-lives may degrade rapidly, leading to insufficient mRNA for sequencing [3]. Furthermore, shallow sequencing depth further exacerbates these issues by reducing the likelihood of detecting key markers essential for accurate cell-type identification [10]. Collectively, these factors impair the sensitivity of mRNAs encoding critical surface proteins, such as CD25, CD69, or PD-1, which are central to defining T cell activation and immune regulation.

Traditionally, immune cell annotation has been based on surface protein expression, with flow cytometry remaining the gold standard for immunophenotyping. Protein-based methods enable precise classification using well-characterized markers such as CD3, CD4, CD14, PD-1, and HLA-DR, and have been pivotal in studies of hematopoiesis and tumor immunology [11, 12]. However, these approaches lack of transcriptomic resolution and are limited in multiplexing capacity. The development of multimodal technologies such as cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) has bridged this gap, enabling the simultaneous measurement of gene expression and surface protein expression in single cells. Despite their potential, widespread adoption of such technologies remains limited due to the technical barriers and costs involved.

Accurate immune cell annotation does not depend on a single data modality, but rather emerges from the interaction between biological regulation, technological measurement, and computational interpretation. In practice, annotation accuracy is shaped by three interdependent factors: (i) biological mechanisms governing the relationship between mRNA and protein expression, (ii) technological constraints that determine which cellular features can be reliably captured, and (iii) computational frameworks used to integrate and interpret heterogeneous single-cell data. Failure to account for any one of these dimensions can lead to incomplete or misleading immune cell classification.

This mini-review explores the challenges of immune cell annotation in single-cell studies through this integrated lens. We first outline the key biological and technical challenges underlying immune cell annotation, with particular emphasis on mRNA-protein discordance. We then review transcriptomic, proteomic, spatial, and multimodal technologies in terms of how they enable or constrain immune cell annotation, rather than as standalone platforms. Finally, we discuss computational approaches for integrating multimodal data and highlight unresolved challenges and future directions toward more robust and biologically grounded immune cell annotation strategies.

## Importance and key challenges in immune cell annotation

Accurate immune cell annotation is fundamental to single-cell transcriptomics and essential for uncovering biological insights in immunology. In the context of cancer, infection, and autoimmune diseases, the ability to reliably identify and characterize immune cell types and states is crucial for mapping cellular heterogeneity, tracing differentiation trajectories, and understanding disease progression at single-cell resolution. As scRNA-seq and multimodal omics technologies evolve, the demand for precise annotation has become increasingly vital for both basic research and clinical applications [13, 14].

In practice, immune cell annotation follows a structured analytical pipeline that integrates biological interpretation with technical preprocessing. Single-cell data are first generated using transcriptomic, proteomic, or multimodal platforms, followed by stringent quality control to remove low-quality cells and technical artifacts. Normalization and batch correction are then applied to mitigate technical variability across samples and experimental conditions. Dimensionality reduction and clustering enable the identification of transcriptionally or phenotypically similar cell populations, which are subsequently annotated using marker-based, reference-based, or integrative computational approaches. At each stage, biological regulation and technical constraints jointly influence annotation accuracy, underscoring the importance of considering mRNA–protein relationships and platform-specific limitations when interpreting immune cell identities.

Cell annotation transforms raw sequencing data into biologically interpretable information. By assigning cellular identities based on molecular signatures, researchers can deconstruct complex immune landscapes and identify rare or transitional populations such as tissue-resident memory T cells (Trm), regulatory dendritic cells, and T peripheral helper (Tph) cells [15, 16]. Without accurate annotation,

such populations may be misclassified or overlooked, obscuring their functional relevance in immune surveillance, regulation, or tumor suppression.

Despite its importance, immune cell annotation from scRNA-seq data remains inherently challenging due to substantial gene expression heterogeneity, technical noise, and sparsity [17]. These issues are especially pronounced in immune populations, where closely related cell types often share overlapping transcriptional profiles. For instance, both CD8<sup>+</sup> cytotoxic T cells and natural killer (NK) cells express granzymes (*GZMB*) and perforin (*PRF1*), making it difficult to distinguish between them using transcriptomic data alone [18]. Similarly, transitional stages between B cells and plasma cells can display ambiguous expression patterns, with both expressing immunoglobulin genes [19]. Further complexity arises in dynamic states such as activation, exhaustion, or tissue residency, where conventional markers such as *PDCD1* and *CD39* are expressed at low or fluctuating levels, complicating annotation [20, 21].

The most common annotation strategy involves unsupervised clustering via dimensionality reduction techniques such as t-distributed stochastic neighbour embedding (t-SNE) or uniform manifold approximation and projection (UMAP), followed by manual inspection of canonical marker gene expression. While widely adopted, this approach is time-consuming, subjective, and lacks scalability for large datasets. Critically, it risks misclassifying or missing rare and functionally distinct populations, particularly when marker genes are absent, downregulated, or when mRNA and protein expression are discordant. These limitations are especially evident in disease contexts. In the context of COVID-19, precise delineation of monocyte and dendritic cell subsets was crucial for correlating immune responses with disease severity and patient outcomes [22]. In cancer immunology, poor annotation may lead to underestimation of immunosuppressive populations such as myeloid-derived suppressor cells (MDSCs), which often express low levels of lineage-specific transcripts, or to the misclassification of exhausted T cells when checkpoint receptor transcripts are absent [23].

Transcript–protein discordance represents a major barrier to accurate cell annotation. The mRNA levels do not always correlate with surface protein abundance due to post-transcriptional regulation, protein degradation, and translational control (Liu et al., 2016). For instance, exhausted T cells frequently exhibit high surface expression of immune checkpoint proteins such as CTLA-4, even when their corresponding transcripts, CTLA-4, are undetectable in scRNA-seq data. This discrepancy can lead to the misclassification of these cells as active effector populations [24]. The observed mRNA–protein discordance likely stems from

tight post-transcriptional regulation, where checkpoint-encoding transcripts undergo rapid degradation following translation, resulting in transient mRNA visibility despite sustained protein expression [25]. Thus, the integration of surface protein data via multimodal approaches like CITE-seq and RNA Expression and Protein sequencing (REAP-seq) has proven instrumental in mitigating mRNA–protein discordance. These techniques employ DNA-barcoded antibodies to simultaneously quantify transcripts and surface proteins in single cells, enabling direct detection of stable surface markers, even when their cognate mRNAs are absent or degraded [26, 27].

As summarized in Table 1, current computational tools have significantly advanced automated immune cell annotation, yet their generalizability across disease contexts remains a topic of active investigation. While methods like Collaborative Agent System for Single-cell Interpretable Annotation (CASSIA) and Area Under the Curve (AUC) cell-based approaches show improved performance, persistent limitations, including difficulty in annotating novel populations, capturing transitional states, and ensuring cross-platform reproducibility, still emerge as common challenges across these tools (Table 1). These gaps highlight the need for next-generation annotation frameworks that integrate multimodal data such as transcriptomic, proteomic, and epigenomic layers, which represent a promising avenue but will require rigorous benchmarking and biological validation to ensure both interpretability and cross-study consistency.

The consequences of poor annotation are far-reaching, with misclassification potentially distorting biological interpretations, misguiding therapeutic predictions, and obscuring disease mechanisms. For instance, mislabelling exhausted T cells as activated effectors could lead to erroneous predictions of patient responses to immune checkpoint inhibitors [23]. Similarly, failure to detect tissue-resident macrophages due to weak transcript signals may underestimate their contributions to inflammation or tissue remodelling. Inconsistent annotation practices further impede the development of reliable immune atlases and complicate cross-study meta-analyses [28]. To address these challenges, initiatives like the Human Cell Atlas are advancing standardized guidelines, curated reference datasets, and benchmarking protocols to enhance reproducibility and data harmonization [29].

Collectively, immune cell annotation is not merely a technical step but a critical determinant of rigor in single-cell immunology. Emerging multimodal technologies, coupled with improved computational tools and community-driven reference frameworks, now empower researchers to decipher complex immune landscapes with unprecedented resolution.

**Table 1** Computational tools for automated immune cell annotation

| Tool                | Strengths                                                                                                                       | Limitations                                                                                                                                    | Reference |
|---------------------|---------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| ImmunIC             | High accuracy of up to 98% for major immune lineages                                                                            | Limited resolution for fine-grained subtypes                                                                                                   | [17]      |
| ImmClassifier       | Improved distinction of nuanced subsets (e.g., CD8+ effector memory T cells)                                                    | Limited generalizability to novel/intermediate states, as it is trained on reference datasets                                                  | [30]      |
| sc-ImmuCC           | Uses a hierarchical framework which achieves 71–90% accuracy                                                                    | Struggles with rare subsets (e.g., $\gamma\delta$ T cells, MAIT cells)                                                                         | [5]       |
| CASSIA              | Superior performance (>25% improvement over others); excels in functionally relevant states                                     | Not explicitly discussed, as it may require benchmarking for novel states                                                                      | [31]      |
| AUCell-based method | Consistently outperforms Azimuth; automated label assignment using marker gene sets                                             | Reliance on predefined marker lists limits its adaptability when novel or context-specific markers emerge.                                     | [32]      |
| Garnett             | Robust and accurate when classifying immune cells even with rare or missing cell types, and when data quality is low.           | Limitation in identifying novel cells that are not included in the original marker file                                                        | [33]      |
| scBalance           | Scales effectively to million-cell datasets and offer a fast and stable computation speeds                                      | Relies on quality reference atlases, potentially struggling with novel or highly divergent immune cell states                                  | [34]      |
| TCAT                | Derives a catalogue of gene expression programs (GEPs) from multiple T-cell scRNA-seq references spanning tissues               | Reliance on predefined GEPs, potentially missing novel states, and focus primarily on T-cells rather than broad immune annotation              | [35]      |
| CellTypist          | Provides fast annotation using curated references, achieving high accuracy for immune cell subtypes like naïve/memory T/B cells | Struggles distinguishing fine-grained states such as naïve vs. central memory T-cells without CITE-seq                                         | [36]      |
| Symphony            | Robustly handling noisy or incomplete labels in immune scRNA-seq while preserving batch-corrected structures                    | High computational demands for training on large references and potential overfitting risks                                                    | [37]      |
| TACIT               | Uncovers rare populations effectively in spatial/multi-omics contexts adaptable to scRNA-seq markers                            | Needs for abundant background signals, risk of unassigned cells from poor intensities or novel types, and unsuitability for small cell regions | [38]      |

Abbreviation: *ImmunIC* Immune Cell Identifier and Classifier, *MAIT* Mucosal-associated invariant T, *CASSIA* Collective Agent System for Single-Cell Interpretable Annotation, *TCAT* T-CellAnnoTator, *TACIT* Threshold-based Assignment of Cell Types from Multiplexed Imaging DaTa

## Mechanisms underlying mrna-protein discrepancies in immune cells

The correlation between mRNA abundance and protein expression is a central tenet in molecular biology, yet this relationship is frequently weak or inconsistent in immune cells, a phenomenon with significant implications for both immunological research and clinical applications. While mRNA provides the transcriptional template for protein synthesis, it is the proteins themselves that mediate critical immune functions, including antigen presentation, cytokine signalling, and cytotoxic activity. Given the dynamic nature and tight regulation of immune responses, understanding the discordance between transcript and protein levels is essential for accurate immune cell characterization and functional annotation [39, 40]. This ensures coverage of the full spectrum of immune responses, capturing phenotypic and functional diversity, reducing cell-state misclassification, and advancing immunological insight.

Although a direct, linear mRNA-protein relationship is often assumed, extensive empirical evidence contradicts this presumption [41, 42]. For instance, key immune mediators such as interferons (IFN- $\alpha$ ), tumor necrosis factor (TNF) receptor superfamily agonists, and chemokines routinely exhibit non-linear transcript-protein relationships due to stringent post-transcriptional and translational control [43]. This is particularly apparent in the melanoma microenvironment, where distinct immune subtypes display divergent expression patterns of checkpoint proteins and cytokine regulators such as CXCL10 and IFNAR1 [44]. These disparities directly influence immune phenotypes and therapeutic responses, including outcomes to mRNA-based vaccines.

Such discordances arise from interconnected biological and technical factors, reflecting the multilayered regulation spanning transcription to functional protein production [7]. In immune cells, where rapid post-transcriptional and protein turnover are essential for responding to environmental stimuli, these disconnects are particularly pronounced.

Consequently, elucidating the mechanisms underlying mRNA-protein uncoupling not only refines the interpretation of single-cell multiomics data but also provides fundamental insights into immune physiology and pathology.

### Biological factors influencing mRNA-protein expression

The disparity between mRNA and protein levels in immune cells is governed by three key biological processes, which are post-transcriptional regulation, translation efficiency, and protein degradation (Fig. 1) [45–47]. These mechanisms operate in a cell-type and context-specific manner, with particular relevance in immune cells due to their need for rapid functional shifts during immune surveillance and activation.

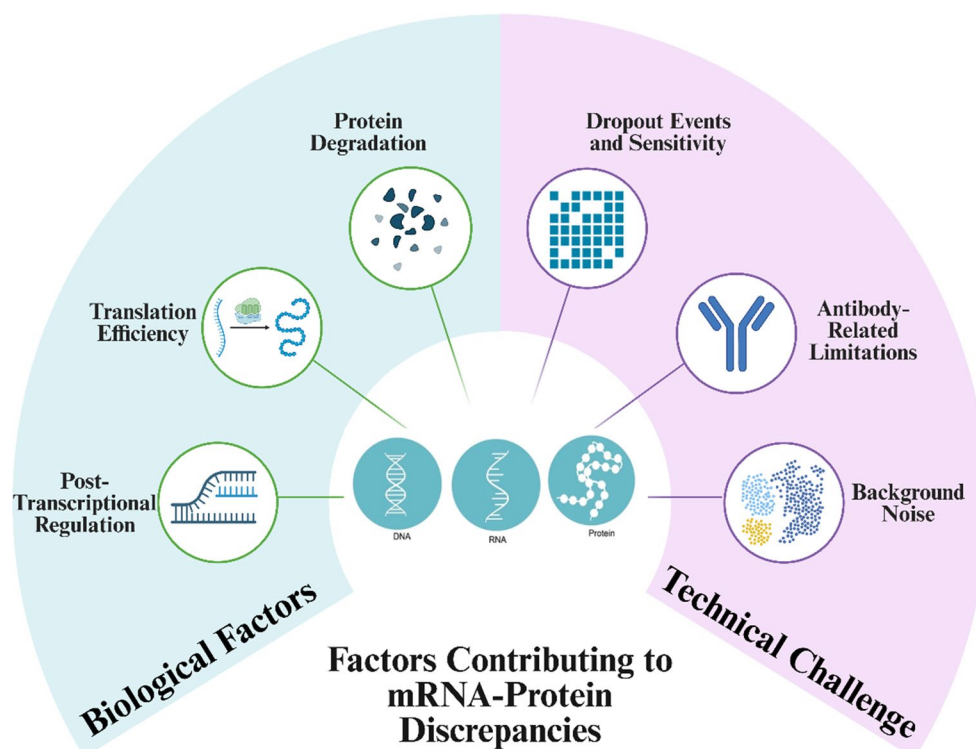
Post-transcriptional mechanisms significantly influence transcript stability and translational capacity [48]. Central to this process are RNA-binding proteins (RBPs) and microRNAs (miRNAs), which modulate mRNA degradation or inhibit translation, thereby decoupling transcript levels from protein synthesis [49, 50]. In immune cells such as macrophages and T cells, inflammatory cytokine mRNAs such as IL-6 and TNF are tightly controlled by RBPs, including tristetraprolin (TTP), Regnase-1, AUF1, and Roquin [51]. These regulators bind to AU-rich elements (AREs) in the 3' UTRs of target mRNAs, destabilizing transcripts or suppressing their translation. This fine-tuning is crucial for preventing excessive cytokine production during immune

activation while still enabling rapid responses. For example, TLR-stimulated macrophages transiently transcribe cytokine mRNAs, but their short half-lives, which are dictated by RBP-mediated decay, ensure tight control over protein output, independent of transcriptional bursts [52].

Even when mRNA is abundant, its conversion into protein depends heavily on translation efficiency, which varies based on multiple factors such as ribosome binding site (RBS), codon composition, tRNA availability, and ribosomal dynamics [53, 54]. A strong RBS enhances ribosome recruitment and increases translation initiation, while optimal codon usage increases translation speed [55]. Additionally, the initiation and elongation phases of translation are modulated by various cis-regulatory elements. For example, sequences such as polyproline motifs can cause ribosome stalling during elongation, reducing overall protein output [56]. In immune cells, where rapid protein production is essential, these bottlenecks significantly impact functional output. If elongation stalls exceed initiation rates, protein synthesis is drastically impaired [57, 58].

Finally, protein abundance is dynamically regulated by degradation pathways, which operate independently of mRNA turnover [59]. Key mechanisms include the ubiquitin-proteasome system, lysosomal autophagy, and non-specific proteolysis [60–62]. These pathways selectively eliminate damaged, misfolded, or excess proteins and can contribute to reduced protein levels even in the presence of abundant transcripts. Interestingly, studies have demonstrated that mRNA and protein degradation can be

**Fig. 1** Biological and technical factors contributing to mRNA-protein discordance in a single-cell study. Illustration created using BioRender (<https://biorender.com>)



uncoupled. In prostate tissue, mRNA turnover occurs at a high rate, yet protein stability remains unaffected, suggesting these processes are independently regulated [63, 64]. Moreover, many proteins exhibit longer half-lives than their corresponding transcripts, allowing for sustained protein function despite transcriptional downregulation. This temporal buffering is particularly relevant in central and effector memory T cells [25], macrophage subsets [65], and exhausted T cell [66] where persistent protein expression is needed for prolonged signalling or effector functions.

### Technical factors influencing mRNA-protein expression

In addition to biological regulation, various technical factors inherent to single cell multiomics platforms significantly affect the observed correlation between mRNA and protein expression. These include issues such as dropout events, antibody-related limitations, and background noise, all of which can distort true cellular expression profiles and complicate data interpretation (Fig. 1).

A primary technical challenge in scRNA-seq is the occurrence of dropout events, where mRNA transcripts fail to be detected despite their presence in the cell [67]. These events stem from both the stochastic nature of gene expression and technical limitations such as inefficient mRNA capture and reverse transcription, particularly for low-abundance transcripts [68]. Dropouts disproportionately affect genes with transient or low expression, including critical immune regulatory genes activated under specific stimuli or stress conditions [67]. As a result, the incomplete detection of these transcripts can lead to misinterpretation of cellular states and obscure the detection of key functional pathways. To address this, various imputation algorithms have been developed, each employing distinct computational strategies.

These include DrImpute, which utilizes a clustering-based approach that aggregates similar cells to improve local expression patterns, making it particularly effective for enhancing cell type clustering [69]. In contrast, scDoc implements a probabilistic model that specifically distinguishes biological absence of expression from technical dropouts, providing more accurate recovery of true zeros in low-coverage datasets [68]. The scHinter algorithm takes a different approach by applying hierarchical network regularization to maintain gene-cell relationships during imputation, offering robust performance for noisy or sparse data [70].

More advanced methods like I-Impute integrate multiple data modalities by incorporating biological prior knowledge alongside expression data, enabling context-aware imputation, which is particularly valuable for immune cell states [71]. Most recently, CCI employs deep learning through

diffusion models to capture complex global expression patterns, showing superior performance for challenging datasets such as those capturing immune cell activation dynamics [72]. While these tools have significantly improved the resolution of single-cell transcriptomic analyses, dropout-related inaccuracies still impact downstream interpretation. The choice of algorithm depends heavily on experimental context, where DrImpute may suffice for basic cell type identification, CCI's sophisticated pattern recognition becomes crucial when analyzing dynamic immune responses. All methods nevertheless share inherent limitations in completely resolving dropout artifacts, which can still lead to biased biological interpretations in tasks like differential expression analysis and cell state annotation.

In multimodal single-cell technologies, antibody-based detection techniques such as CITE-seq, REAP-seq and AbSeq enable simultaneous profiling of surface protein expression alongside transcriptomic data. These techniques are widely implemented on platforms like BD Rhapsody™ and 10x Genomics™ Chromium, facilitating integrated multi-omics analysis at single-cell resolution. However, several antibody-related issues may introduce artifacts that compromise the accuracy of mRNA–protein correlations. One critical limitation is antibody specificity. Off-target binding or failure to distinguish between closely related protein isoforms can result in false-positive or ambiguous signals [73]. In complex immune environments, where closely related proteins are often co-expressed, cross-reactivity poses a significant challenge, leading to non-specific binding and distorted protein measurements [74].

Additionally, antibody affinity and saturation further complicate the protein quantification. At high antigen concentrations, antibody saturation can occur, limiting the dynamic range of detection and potentially masking subtle but biologically meaningful differences [75]. Conversely, low-affinity antibodies may fail to detect proteins present at low abundance. A more fundamental constraint is the limited availability of well-validated antibodies for many immune targets. Commercially available antibodies often lack comprehensive validation for single-cell applications, particularly for emerging immune markers or unconventional protein isoforms. To mitigate these effects, researchers must employ rigorous antibody validation, panel optimization, and high-affinity reagents when available. However, even these measures are constrained by the commercial antibody landscape. Alternative solutions include custom antibody development, though this is time-consuming and costly, epitope tagging strategies for recombinant proteins, and computational correction methods to account for antibody noise.

Another major technical challenge is background noise, particularly in droplet-based platforms such as 10x Genomics™ Chromium systems. Protein measurements via

DNA-barcoded antibodies (CITE-seq) are prone to contamination from unbound antibody carryover, non-specific binding, and ambient mRNA/protein in empty droplets [76]. In some cases, antibody-derived UMIs in empty droplets even exceeded those in droplets containing cells, emphasizing the magnitude of the ambient signal [77]. This necessitates normalization approaches such as “dsb” (denoised and scaled by background), which leverage empty droplet profiles for denoising. Comparative analyses of 10x Chromium and BD Rhapsody™ further highlight platform-specific noise profiles in which BD Rhapsody™ tends to display a more uniform but generally higher baseline of ambient noise, while 10x Chromium™ shows lower overall background but exhibits stronger cell type-specific biases, particularly from neutrophil-derived transcripts in damaged samples [78]. These differences, rooted in the platforms’ distinct molecular designs—microwell versus droplet systems, highlight the importance of computational correction tools such as dsb for reliable immune cell profiling. Finally, the major biological and technical sources of mRNA–protein discordance and their implications for immune cell annotation are summarized in Table 2.

Technologies for immune cell profiling

Advances in transcriptomic and proteomic technologies have fundamentally reshaped our understanding of immune cell diversity, function, and regulation. Early immune profiling methods relied heavily on microarrays, which enabled gene expression analysis across predefined probes [79]. While instrumental in biomarker discovery and disease stratification, microarrays were limited by probe dependency, hybridization artifacts, narrow dynamic range, and

high RNA input requirements, rendering them suboptimal for analyzing rare or low-yield immune samples [80, 81]. The advent of bulk RNA sequencing (RNA-seq) overcame these limitations by enabling transcriptome-wide, probe-independent detection of gene expression with enhanced sensitivity and dynamic range [82–85]. This approach uncovered novel transcripts, alternative splicing events, and non-coding RNAs. However, bulk RNA-seq averages gene expression across mixed populations, thereby masking cellular heterogeneity and limiting its utility in resolving rare or transient immune states [86].

The introduction of scRNA-seq marked a transformative step, enabling transcriptomic profiling at single-cell resolution. Unlike bulk RNA-seq, which averages signals across cell populations, scRNA-seq isolates individual cells via microfluidics, droplet-based encapsulation, or microwell plates. The individual cells are then lysed, and their RNA molecules are uniquely barcoded during reverse transcription. By quantifying gene expression per cell, this technology has revealed rare subsets, delineated developmental trajectories, and captured context-specific immune responses in cancer, infection, and autoimmunity that are otherwise obscured in bulk analyses [87–89]. Nonetheless, scRNA-seq poses challenges, including loss of spatial context, variable capture efficiency, and limited transcript recovery, particularly from small immune cells. These limitations are addressed by spatial transcriptomics technologies such as Multiplexed Error-Robust Fluorescence In Situ Hybridization (MERFISH) and Sequential Fluorescence In Situ Hybridization (seqFISH), which preserve tissue architecture while mapping transcripts in situ [90, 91].

MERFISH employs combinatorial binary barcoding and sequential imaging to detect hundreds to tens of thousands of genes simultaneously with nanometre-scale spatial localization, offering robust error correction quantification of preselected gene panels [90]. In contrast, seqFISH uses color-coded barcodes combined over multiple cycles to achieve super-resolved imaging, excelling in detecting low-copy transcripts like transcription factors [91]. Both methods surpass scRNA-seq in spatial mapping and sensitivity for targeted genes, including lower dropout than scRNA-seq but are restricted to predefined gene sets. Complementing these approaches, single-cell ATAC-seq (scATAC-seq) profiles chromatin accessibility at the single-cell level using a hyperactive Tn5 transposase to insert sequencing adapters into open chromatin regions [92]. This reveals active regulatory elements like enhancers or promoters, and infers transcription factor binding, thereby linking epigenetic states to immune cell function. When integrated with scRNA-seq, scATAC-seq provides a multi-omics framework to decipher gene regulatory mechanisms underlying immune responses.

**Table 2** Biological and technical sources of mRNA–protein discordance and their implications for immune cell annotation

| Source of mRNA–protein discordance                                  | Implications for immune cell annotation                                                                         |
|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Post-transcriptional regulation (e.g. miRNA control, RNA stability) | Protein-defined immune states may be underrepresented or misclassified when relying on transcript-based markers |
| Translational control and protein stability                         | Functional immune phenotypes can persist despite low or transient mRNA expression, leading to annotation bias   |
| Transcript dropout in scRNA-seq                                     | False-negative marker detection increases annotation uncertainty and ambiguous cell identities                  |
| Limited sequencing depth                                            | Low-abundance marker genes essential for immune subset definition may be missed                                 |
| Stochastic capture efficiency                                       | Inconsistent marker detection across cells reduces annotation robustness and reproducibility                    |

While transcriptomics provides information on gene activity, proteomic technologies are essential for functional immune profiling, since proteins are the effectors of immune responses. Flow cytometry has long served as the cornerstone for immunophenotyping, allowing multiparametric quantification of surface and intracellular markers, functional states, and cytokine production [93, 94]. Clinically, it has been extensively used to validate immune signatures in cancer, autoimmune diseases, and infectious conditions such as COVID-19 [95, 96]. However, conventional flow cytometry is constrained by spectral overlap (typically limiting multiplexing to ~15–20 markers), the need for meticulous spillover compensation of antibody–fluorophore conjugates, laser-line restrictions on concurrent excitation, and diminishing signal-to-noise in larger panels. Consequently, Spectral flow cytometry partially overcomes these constraints by capturing full emission spectra rather than discrete channels, enabling simultaneous detection of 30–50 markers with reduced compensation artifacts [97, 98]. This approach enhances resolution for dim markers in cytokines and improves the unmixing of fluorophores with overlapping spectra. However, spectral systems demand extensive pre-panel optimization, reference controls for spectral unmixing, and higher computational costs for data deconvolution. Additionally, they remain constrained by antibody quality and cellular autofluorescence in certain types of samples.

Mass cytometry (CyTOF) further extended immune profiling capabilities by using metal-conjugated antibodies and time-of-flight mass spectrometry, enabling simultaneous detection of over 40 proteins per cell without spectral interference [99]. The CyTOF has been instrumental in identifying rare immune populations and intracellular signalling events in cancer and autoimmunity, although it lacks spatial resolution and requires specialized instrumentation [100, 101]. Thus, Imaging mass cytometry (IMC) overcomes CyTOF's spatial limitations by combining high-dimensional proteomics with tissue imaging. In IMC, tissue sections stained with metal-tagged antibodies are ablated pixel-by-pixel using a laser, and the vaporized metal ions are quantified by mass spectrometry [102, 103]. This preserves tissue architecture while enabling multiplexed detection of more than 40 markers, revealing immune-tumor interactions, tissue-resident niches, and cellular neighborhoods *in situ*. However, IMC suffers from lower throughput, complex data analysis pipelines, and higher costs compared to suspension CyTOF.

Thus, the growing complexity of immune responses has driven innovation in multimodal profiling platforms. Building on their ability to resolve mRNA–protein discrepancies, as discussed earlier, CITE-seq and REAP-seq have become particularly valuable for identifying immune subsets where protein expression diverges from transcriptional profiles. For

example, they are reliable in detecting CD103+ tissue-resident memory T cells and FCεRIα+ mast cells that frequently escape transcriptomic identification [104]. Furthermore, these methods enable more accurate discrimination of functionally distinct states, such as differentiating activated (CD69+CD25+) from exhausted (PD-1+TIM-3+) T cell populations within tumor microenvironments [105].

Building upon the CITE-seq multimodal framework, Extended CITE-seq (ECCITE-seq), an advanced version of the original CITE-seq, significantly expands single-cell profiling capabilities by integrating four key modalities, which are transcriptome-wide mRNA sequencing, surface protein quantification via antibody-derived tags (ADTs), paired T-cell receptor/B-cell receptor (TCR/BCR) sequencing for immune clonotype tracking, and detection of CRISPR single-guide RNAs (sgRNAs) for perturbation screening [106]. Unlike conventional CITE-seq or REAP-seq which combine only transcriptomics and proteomics, ECCITE-seq's unique value lies in its ability to simultaneously capture genetic perturbations, via sgRNA detection, while maintaining immune receptive specificity. Therefore, making it particularly powerful for functional immunology studies investigating gene knockout effects on immune cell phenotypes [106, 107].

While multimodal approaches such as ECCITE-seq provide high-dimensional mRNA–protein profiling, they lack spatial context, which is critical for immune cell annotation in tissues where cellular identity and function depend on local microenvironmental cues. Spatial transcriptomic platforms address this limitation by preserving tissue architecture, cellular morphology, and spatial co-localization, thereby enabling immune cell annotation that integrates molecular identity with positional information.

Spatial approaches such as CosMx<sup>®</sup>, Visium, and Xenium improve immune cell annotation by resolving immune populations within their native tissue niches, facilitating discrimination of transcriptionally similar but spatially distinct immune states. This is particularly relevant for immune subsets such as tumor-associated macrophages (TAMs) [108], tissue-resident myeloid cells, and neutrophils, which are often underrepresented or ambiguously annotated in dissociation-based scRNA-seq due to transcriptional sparsity or loss during tissue processing. Among spatial platforms, Visium supports discovery-driven annotation through transcriptome-wide spatial mapping integrated with histological context, albeit at multi-cellular resolution [109], whereas Xenium and CosMx<sup>®</sup> enable higher-resolution, targeted spatial profiling with the capacity to integrate RNA and protein signals [110]. Such multimodal spatial integration enhances annotation robustness by cross-validating immune identities across molecular expression, morphology, and spatial co-localization, allowing more confident resolution of immune cell states within complex tissue microenvironments.

For studies prioritizing throughput over spatial context, the Chromium GEM-X platform (10x Genomics™), an advanced generation of droplet-based scRNA-seq, leverages microfluidics and optimized GEM-X chemistry to achieve high-throughput single-cell multiomics [111]. Its gel bead-in-emulsion (GEM) system supports simultaneous transcriptome-wide RNA-seq, surface protein detection via ADTs, and multiplexed analysis of up to 128 samples per run, making it well-suited for large-scale cohort studies [112]. Nonetheless, GEM-X shares common drawbacks of scRNA-seq platforms, including the absence of spatial information, high costs for expansive cohorts, and reduced sensitivity for rare immune subsets.

To address these limitations, emerging technologies like Evercode™ Immune Profiling (Parse Biosciences) employ a microfluidics-free, combinatorial indexing approach that enables ultra-high-throughput analysis of up to 1 million fixed cells per experiment, which is a 10-fold increase over Chromium GEM-X's typical throughput. Unlike Chromium's droplet-based GEM system, Evercode™ utilizes split-pool barcoding to progressively tag transcripts and paired immune receptors across multiple rounds of indexing, eliminating the need for specialized instrumentation [113, 114]. This approach offers a deep clonotype resolution and full-length immune receptor sequencing with matched transcriptomes. However, Evercode™ currently lacks Chromium GEM-X's integrated protein detection capability. Where Chromium excels in multimodal profiling with standardized workflows, Evercode™ prioritizes massive-scale immune repertoire studies, making them complementary rather than competing solutions.

More recently, Beacon Discovery™ (Bruker) introduced an optofluidic platform designed for real-time, functional single-cell analysis, ideal for immunology, oncology, and therapeutic discovery. The system utilizes Opto-Electrical Positioning (OEP) paired with OptoSelect™ microfluidic chips, where light-actuated pixels precisely isolate live cells into NanoPen chambers for dynamic profiling. Unlike Evercode's sequencing focus or Chromium GEM-X's high-throughput omics, Beacon platform prioritizes functional immune metrics, which is demonstrated in adoptive T-cell therapy studies where it quantified single-cell cytotoxicity via caspase-3 activation, cytokine secretion, and phenotype switching during immune synapse formation [115]. These different techniques are summarized in Table 3.

## Bioinformatic obstacles in integrating mrna and protein datasets

The field of single-cell multimodal analysis has evolved significantly since the introduction of pioneering technologies like CITE-seq and REAP-seq, which first enabled

**Table 3** Conceptual overview of single-cell and spatial platforms and their implications for immune cell annotation

| Conceptual Category    | Platform Examples          | Data Type                      | Key Advantages for Immune Annotation                                                                             | Key Limitations for Immune Annotation                                 |
|------------------------|----------------------------|--------------------------------|------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------|
| Transcriptomic         | scRNA-seq                  | mRNA                           | Enables discovery of immune cell states and transcriptional programs                                             | Limited sensitivity for surface markers; dropout affects annotation   |
| Spatial transcriptomic | Visium, Xenium, CosMx      | Spatial mRNA (±protein)        | Preserves tissue context and cell–cell interactions; improves annotation of spatially defined immune populations | Limited gene panels or resolution constraints depending on platform   |
| Proteomic              | Flow cytometry, CyTOF, IMC | Surface/intracellular proteins | High sensitivity for canonical immune markers; robust phenotyping                                                | Restricted to predefined antibody panels; limited discovery potential |
| Multimodal             | CITE-seq, ECCITE-seq       | mRNA + protein                 | Improves annotation confidence by integrating transcriptomic and proteomic signals                               | Antibody-dependent; increased analytical complexity                   |

Abbreviation: *scRNA-seq* Single-cell RNA sequencing, *CyTOF* Cytometry by Time-of-Flight (Mass Cytometry), *IMC* Imaging Mass Cytometry, *CITE-seq* Cellular Indexing of Transcriptomes and Epitopes by sequencing, *ECCITE-seq* Expanded CRISPR-compatible Cellular Indexing of Transcriptomes and Epitopes

simultaneous measurement of mRNA and protein expression within individual cells [26, 27]. These methods have been widely implemented on high-throughput platforms such as 10x Genomics™ Chromium and BD Rhapsody™, establishing them as standard workflows in immunology research. Modern iterations like 10x Genomics™ Chromium and BD Rhapsody™ now support even larger-scale multimodal profiling while addressing early limitations in sensitivity and throughput [89, 116]. Despite these technological advances, computational integration of mRNA and protein datasets remains challenging due to fundamental differences in the data structure of mRNA and protein.

One of the primary barriers to effective integration is the disparity in data sparsity and distribution between transcriptomic and proteomic measurements. The scRNA-seq generates high-dimensional count matrices exhibiting extreme

sparsity, with 60–90% zero values per cell due to transcriptional stochasticity and technical dropout events [117]. This contrasts sharply with bulk RNA-seq, where the proportion of zeros typically ranges between 10% and 40%, necessitating distinct statistical modelling approaches. While scRNA-seq requires zero-inflated or negative binomial distributions to account for this sparsity, ADT-based protein data, although less sparse, are constrained by smaller panel sizes of typically 30–200 markers [118] and confounded by technical artifacts, including non-specific binding, cross-reactivity, and antibody saturation [119, 120]. As a result, normalization approaches developed for transcriptomic data, such as log-transformation or library size scaling, are often inappropriate for protein data, which may require distinct strategies for batch correction and scaling. Moreover, differences in the dynamic range and detection sensitivity between RNA and protein further complicate joint analysis. Specifically, RNA data require normalization methods that address high variability and dropout rates, as previously mentioned, while protein data, with their broader dynamic range and more stable signal, require scaling approaches such as centered log-ratio normalization or quantile normalization that can preserve relative abundance across cells.

Another significant challenge lies in the computational scalability and infrastructure requirements. As multimodal platforms now routinely profile millions of cells across thousands of features, traditional computational pipelines often fail to manage the memory and processing demands of such large datasets [121]. Single-cell data's inherent sparsity demands optimized algorithms capable of performing high-dimensional analyses at scale and memory-efficient storage formats, which reduces memory requirement by one to two orders of magnitude. Methods such as variational autoencoders (VAEs), graph neural networks, and other machine learning-based techniques often require high-performance computing (HPC) environments or graphics processing unit (GPU) acceleration due to the complexity of their deep neural network architecture, volume of data and the need for fast, large-scale matrix computations which exceed the capacity of typical central processing unit (CPU) for efficient training and inference [122]. When RNA and protein data are generated from different platforms such as 10x Genomics™ Chromium, BD Rhapsody™, or Parse Bioscience™ or across distinct laboratory batches or time points, platform-specific technical noise and batch effects are introduced. These sources of variability must be carefully addressed using robust harmonization techniques such as HarmonizR [123] and Harmony [124] to remove unwanted technical noise while retaining true biological signals.

A key computational challenge is modality alignment, which accurately matches each cell's RNA and protein readouts despite data gaps. Dropout events, which are common

in both modalities, often lead to partial observations where low-abundance features are undetected due to biological or technical factors. Although imputation methods, such as nearest-neighbour averaging or deep learning-based models, can estimate missing values, they risk introducing artifacts and obscuring true biological variation if not properly validated [125, 126]. Furthermore, mRNA-protein relationships are frequently nonlinear and context-dependent, rendering classical linear approaches such as canonical correlation analysis (CCA) inadequate for capturing complex regulatory mechanisms, as CCA models linear relationships between two sets of variables. This limitation has spurred demand for machine-learning frameworks capable of modelling nonlinear dependencies, which are an essential capability for accurately interpreting CITE-seq data and deriving biologically meaningful insights that linear methods cannot provide.

To address these computational challenges, a growing number of integrative analysis tools have been developed, each with unique strengths and limitations (Table 4). For instance, Seurat aligns modalities through canonical correlation analysis (CCA) coupled with mutual nearest-neighbors (MNN) matching, yet its reliance on linear models imposes limitations [127]. The Multi-Omics Factor Analysis Plus (MOFA+) uses factor analysis to identify shared and modality-specific sources of variation, yet performs suboptimally on extremely sparse datasets [128]. Total Variational Inference (TotalVI), a deep generative model built on the VAE framework, jointly denoises and imputes RNA and protein data while accounting for technical covariates, though its high computational cost limits scalability [126]. More recent approaches, such as the Single-cell Multi-omics Transformer-based Framework (scmFormer) [129], Single-cell Protein Reconstruction via Cross-Attention (scProca) [130], Single-cell Network-Enhanced Transformer (scNET) [131], and Single-cell Transformer Encoder for Labeling (scTEL) [132] utilize transformer architectures or graph neural networks for improved modelling of cross-modal relationships. While these methods offer improved performance and flexibility, they often come at the expense of interpretability, requiring careful benchmarking and biological validation. Table 4 summarizes the overview of the mentioned computational tools.

Looking forward, several key directions must be pursued to address the current limitations in multimodal data integration. First, algorithmic innovation is essential to develop methods that balance nonlinear modelling capacity with scalability and interpretability. Such innovation will enable the creation of integration frameworks that are not only accurate and scalable but also interpretable, capable of handling the multifaceted technical complexities inherent in modern single-cell multimodal datasets. Second, the field

**Table 4** Overview of computational tools for integrating single-cell transcriptomic and proteomic data

| Tool              | Modelling Approach                                                                                      | Integration Mode                                                                                                             | Strengths                                                                                                                                                                                                                                                                                                            | Limitations                                                                                                                                                                                                                                                                                              |
|-------------------|---------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Seurat (v4) [127] | Statistical integration using Canonical Correlation Analysis (CCA) and Weighted Nearest Neighbors (WNN) | Integrates RNA and ADT data; supports batch correction via anchors                                                           | <ul style="list-style-type: none"> <li>Widely used and well-documented</li> <li>Flexible for various single-cell omics</li> <li>Good batch correction</li> </ul>                                                                                                                                                     | <ul style="list-style-type: none"> <li>Challenging multi-sample integration</li> <li>Requires separate normalization for RNA and protein</li> <li>Sensitive to batch effects and rare cell types</li> </ul>                                                                                              |
| MOFA+ [128]       | Bayesian factor analysis                                                                                | Integrates matched multi-modal data; RNA and protein from same cells                                                         | <ul style="list-style-type: none"> <li>Unsupervised and interpretable factor discovery</li> <li>Supports missing data</li> <li>Differentiates shared and modality-specific variation</li> </ul>                                                                                                                      | <ul style="list-style-type: none"> <li>Cannot model non-linear relationships</li> <li>Assumes feature independence</li> <li>Less effective at batch correction</li> </ul>                                                                                                                                |
| TotalVI [126]     | Deep generative model based on Variational Auto-encoder (VAE)                                           | Joint latent space for RNA and surface protein                                                                               | <ul style="list-style-type: none"> <li>Denoises and integrates modalities simultaneously</li> <li>Captures non-linear dependencies</li> <li>Handles technical noise</li> </ul>                                                                                                                                       | <ul style="list-style-type: none"> <li>Requires GPU for efficient training</li> <li>Difficulties in interpreting latent space</li> <li>May be difficult for users without ML background</li> </ul>                                                                                                       |
| sciPENN [133]     | Deep learning framework using repeated neural network blocks                                            | Integrates CITE-seq and scRNA-seq data into a common latent embedding space                                                  | <ul style="list-style-type: none"> <li>Handles missing data using censored loss</li> <li>Outperforms Seurat v4 and TotalVI in benchmarks</li> <li>Provides uncertainty estimates for protein predictions</li> <li>Robust to protocol and platform variability</li> </ul>                                             | <ul style="list-style-type: none"> <li>Requires high-quality matched RNA–protein reference datasets</li> <li>Computationally intensive for large-scale training</li> <li>Latent features may be less interpretable</li> <li>May underperform in unseen biological contexts or rare cell types</li> </ul> |
| scMHNN [134]      | Hypergraph Neural Network with dual-contrastive self-supervised learning                                | Integrates tri-modal single-cell data (RNA, protein, epigenome) from same cells                                              | <ul style="list-style-type: none"> <li>Captures high-order relationships across modalities</li> <li>Effective with few labeled cells via self-supervised learning</li> <li>Strong performance in clustering and annotation</li> </ul>                                                                                | <ul style="list-style-type: none"> <li>Requires matched tri-modal data</li> <li>Computationally intensive</li> <li>Complex architecture demands expertise to tune</li> </ul>                                                                                                                             |
| scmFormer [129]   | Transformer-based architecture with multi-task learning                                                 | Integrates scRNA-seq and CITE-seq/ASAP-seq data across batches and tissues; supports spatial datasets                        | <ul style="list-style-type: none"> <li>Scales to &gt; 1.48 million cells on a personal computer</li> <li>High accuracy in label transfer</li> <li>Robust to batch effects</li> <li>Can generate unmeasured modalities</li> </ul>                                                                                     | <ul style="list-style-type: none"> <li>Requires significant resources for training</li> <li>Complex attention mechanisms may obscure biological insights and interpretability</li> </ul>                                                                                                                 |
| scProca [130]     | Deep generative model combining VAE with cross-attention mechanisms.                                    | Integrates scRNA-seq and CITE-seq, including partially overlapping ADT panels; can infer protein from RNA.                   | <ul style="list-style-type: none"> <li>Outperforms Seurat/TotalVI in integration metrics</li> <li>Effectively removes batch effects while preserving biological variation</li> <li>Supports large, multi-omics mosaic datasets</li> </ul>                                                                            | <ul style="list-style-type: none"> <li>Requires paired RNA-protein data for optimal performance</li> <li>Complex training and tuning</li> <li>Limited to ADTs, not full proteomes</li> </ul>                                                                                                             |
| scNET [131]       | Dual-view Graph Neural Network (GNN) using gene–gene and cell–cell relationships                        | Integrates scRNA-seq with protein–protein interaction (PPI) networks                                                         | <ul style="list-style-type: none"> <li>Enhances identification of differentially enriched pathways and gene functional annotation.</li> <li>Reduces noise and zero inflation typical of scRNA-seq data through graph-based smoothing.</li> </ul>                                                                     | <ul style="list-style-type: none"> <li>Focuses on integration of scRNA-seq with PPI networks rather than direct protein abundance measurements</li> <li>Relies on availability and quality of PPI data</li> <li>Model interpretability is limited due to GNN complexity</li> </ul>                       |
| scTEL [132]       | Deep learning framework using Transformer encoder layers                                                | Predicts protein expression from scRNA-seq data, enabling joint analysis of RNA and protein in the same cell computationally | <ul style="list-style-type: none"> <li>Reduces experimental cost by computationally imputing protein expression instead of measuring it directly.</li> <li>Integrates multiple CITE-seq datasets with partial protein overlap</li> <li>Performs prediction, cell typing, and integration in one framework</li> </ul> | <ul style="list-style-type: none"> <li>Relies on availability of high-quality paired RNA-protein training data</li> <li>Prediction accuracy depends on the quality and representativeness of training data.</li> </ul>                                                                                   |

Abbreviation: *CCA* Canonical Correlation Analysis, *WNN* Weighted Nearest Neighbors, *ADT* Antibody-Derived Tag, *MOFA+* Multi-Omics Factor Analysis Plus, *TotalVI* Total Variational Inference, *sciPENN* Single Cell imputation Protein Embedding Neural Network, *VAE* Variational Autoencoder, *RNA* Ribonucleic Acid, *GPU* Graphics Processing Unit, *ML* Machine Learning, *scMHNN* Single-cell Multi-modal Hypergraph Neural Network, *scmFormer* Single-cell Multi-omics Transformer-based Framework, *CITE-seq* Cellular Indexing of Transcriptomes and Epitopes by Sequencing, *ASAP-seq* Assay for Single-cell Accessibility and Protein Sequencing, *scProca* Single-cell Protein Reconstruction via Cross-Attention, *scNET* Single-cell Network-Enhanced Transformer, *GNN* Graph Neural Network, *PPI* Protein–Protein Interaction, *scTEL* Single-cell Transformer Encoder for Labeling

urgently needs standardized benchmarking frameworks to evaluate tool performance across diverse datasets and varying levels of technical noise, using biologically meaningful metrics. Without such standardized benchmarks, it becomes challenging to objectively compare methods, reproduce findings, or ensure methodological robustness in the face of highly variable and heterogeneous single-cell data. Finally, with the growing clinical adoption of single-cell technologies, transitioning from research settings into routine clinical applications for disease diagnosis, there is a growing need for robust, reproducible, and scalable pipelines. These pipelines must bridge the gap between complex single-cell assays and real-world clinical practice by delivering consistent, interpretable, and timely data to support personalized medicine. In particular, automated and clinically relevant pipelines that incorporate quality control, data integration, and biological interpretation are essential for reliably processing large volumes of diverse patient samples while maintaining consistent analytical quality and reproducibility.

## Current challenges and future directions

Despite the transformative potential of single-cell multi-omics in immunology, critical challenges persist in achieving precise immune cell annotation. First, it is still unclear whether multimodal single-cell data can fully resolve ambiguous intermediate immune states, which often display overlapping or transitional features. Second, limitations in antibody panels, such as incomplete coverage of emerging or rare markers, constrain the ability to accurately profile diverse immune populations at the protein level. Finally, zero-shot annotation of novel immune cell types remains a significant challenge, as current tools depend heavily on existing reference data and struggle to classify previously uncharacterized populations. Addressing these issues will be crucial for improving annotation accuracy and advancing our understanding of immune heterogeneity.

Additionally, current benchmarking remains difficult in part due to the absence of standardized metrics and comprehensive frameworks to reliably evaluate annotation accuracy across tools and datasets. Several web-based annotation platforms such as SCSA [135], Annotation of Cell Types (ACT) [136], CellHint [36] and curated resources like CellMarker database [137] have expanded accessibility for researchers by providing automated or semi-automated annotation pipelines. While these platforms accelerate cell-type identification, their accuracy remains dependent on the quality and completeness of reference marker sets or training datasets. As such, they may misclassify rare or novel

immune populations or fail to capture subtle functional states, underscoring the continued need for well-curated, gold-standard reference datasets to ensure reliable benchmarking across methods.

Moreover, existing multimodal reference atlases primarily derive from healthy PBMCs or normal tissues, limiting their utility in disease contexts. Comprehensive atlases integrating matched mRNA, protein, and epigenomic data from tumor tissues remain scarce, hindering the development of computational models for precise annotation in cancer. While current integrative tools have advanced scalability and nonlinear modelling, they face limitations in interpretability and computational demands, often requiring GPUs and extensive training, which restricts their accessibility for routine analyses.

To address these gaps, future efforts may need to prioritize the development of a standardized, disease-inclusive multimodal atlas with matched molecular layers, coupled with community-driven benchmarks for tool validation. Scalable computational methods, such as lightweight neural networks or hybrid models integrating biological priors, must be optimized for usability without sacrificing interpretability. Collaborative initiatives, such as the Human Cell Atlas (HCA) [138], Single Cell Portal [139], scImmO-mics [140] should expand to incorporate dynamic immune states such as activation or exhaustion, and clinical samples, enabling translation to therapeutic discovery and precision immunology.

**Acknowledgements** This work was supported by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme (FRGS) (grant number: FRGS/1/2021/SKK0/UKM/02/7).

**Author contributions** SG and NAJ developed the study concept and design. SG prepared the initial draft, compiled and synthesized materials, and generated all visualizations. NAJ performed final editing and contributed additional ideas. Both authors reviewed and approved the submitted manuscript.

**Funding** Open access funding provided by The Ministry of Higher Education Malaysia and Universiti Kebangsaan Malaysia. This work was supported by the Ministry of Higher Education, Malaysia, under the Fundamental Research Grant Scheme (FRGS) (grant number: FRGS/1/2021/SKK0/UKM/02/7).

**Data availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing interests** The authors declare no competing interests.

**Conflict of interest** The authors declare no conflict of interest, financial or otherwise.

**Clinical Trial Number** Not applicable.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol*. 2018;18:35–45. <https://doi.org/10.1038/nri.2017.76>.
- Wang G, Tao H, Zhou L, Zhang J, Pu W, Xu T, et al. scRNA-seq reveals involvement of monocytes in immune response in SLE patients. *Genomics*. 2025;117:110994. <https://doi.org/10.1016/j.ygeno.2025.110994>.
- Reimegård J, Tarbier M, Danielsson M, Schuster J, Baskaran S, Panagiotou S, et al. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun Biol*. 2021;4:624. <https://doi.org/10.1038/s42003-021-02142-w>.
- Mullan KA, de Vrij N, Valkiers S, Meysman P. Current annotation strategies for T cell phenotyping of single-cell RNA-seq data. *Front Immunol*. 2023;14:1306169. <https://doi.org/10.3389/fimmu.2023.1306169>.
- Jiang Y, Chen Z, Han N, Shang J, Wu A. sc-ImmuCC: hierarchical annotation for immune cell types in single-cell RNA-seq. *Front Immunol*. 2023;14:1–12. <https://doi.org/10.3389/fimmu.2023.1223471>.
- Li J, Zhang Y, Yang C, Rong R. Discrepant mRNA and Protein Expression in Immune Cells. *Curr Genomics*. 2020;21:560–3. <https://doi.org/10.2174/1389202921999200716103758>.
- Prabakar A, Zamora R, Barclay D, Yin J, Ramamoorthy M, Bagheri A, et al. Unraveling the complex relationship between mRNA and protein abundances: a machine learning-based approach for imputing protein levels from RNA-seq data. *NAR Genomics Bioinforma*. 2024;6:lqae019. <https://doi.org/10.1093/nargab/lqae019>.
- De Biasi S, Meschiari M, Gibellini L, Bellinazzi C, Borella R, Fidanza L, et al. Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nat Commun*. 2020;11:3434. <https://doi.org/10.1038/s41467-020-17292-4>.
- Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun*. 2018;9:997. <https://doi.org/10.1038/s41467-018-03405-7>.
- Wani SA, Quadri S, Mir MS, Gulzar Y. A Comparative Study of Machine Learning Techniques for Cell Annotation of scRNA-Seq Data. *Algorithms*. 2025;18:232. <https://doi.org/10.3390/a18040232>.
- Preglej T, Brinkmann M, Steiner G, Aletaha D, Göschl L, Bonelli M. Advanced immunophenotyping: A powerful tool for immune profiling, drug screening, and a personalized treatment approach. *Front Immunol*. 2023;14:1096096. <https://doi.org/10.3389/fimmu.2023.1096096>.
- Zhou Z, Ye C, Wang J, Zhang NR. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat Commun*. 2020;11:1–10. <https://doi.org/10.1038/s41467-020-14391-0>.
- Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184:3573–e358729. <https://doi.org/10.1016/j.cell.2021.04.048>.
- Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet*. 2019;20:257–72. <https://doi.org/10.1038/s41576-019-0093-7>.
- Mueller SN, Mackay LK. Tissue-resident memory T cells: local specialists in immune defence. *Nat Rev Immunol*. 2016;16:79–89. <https://doi.org/10.1038/nri.2015.3>.
- Khan O, Giles JR, McDonald S, Manne S, Ngiew SF, Patel KP, et al. TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion. *Nature*. 2019;571:211–8. <https://doi.org/10.1038/s41586-019-1325-x>.
- Park SY, Ter-Saakyan S, Faraci G, Lee HY. Immune cell identifier and classifier (ImmunIC) for single cell transcriptomic readouts. *Sci Rep*. 2023;13:12093. <https://doi.org/10.1038/s41598-023-39282-4>.
- Artyomov MN, Van den Bossche J. Immunometabolism in the Single-Cell Era. *Cell Metab*. 2020;32:710–25. <https://doi.org/10.1016/j.cmet.2020.09.013>.
- Nutt SL, Hodgkin PD, Tarlinton DM, Corcoran LM. The generation of antibody-secreting plasma cells. *Nat Rev Immunol*. 2015;15:160–71. <https://doi.org/10.1038/nri3795>.
- Bengsch B, Ohtani T, Khan O, Setty M, Manne S, O'Brien S, et al. Epigenomic-Guided Mass Cytometry Profiling Reveals Disease-Specific Features of Exhausted CD8 T Cells. *Immunity*. 2018;48:1029–e10455. <https://doi.org/10.1016/j.immuni.2018.04.026>.
- Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat Med*. 2018;24:978–85. <https://doi.org/10.1038/s41591-018-0045-3>.
- Tian Y, Carpp LN, Miller HER, Zager M, Newell EW, Gottardo R. Single-cell immunology of SARS-CoV-2 infection. *Nat Biotechnol*. 2022;40:30–41. <https://doi.org/10.1038/s41587-021-01131-y>.
- Li H, van der Leun AM, Yofe I, Lubling Y, Gelbard-Solodkin D, van Akkooi ACJ, et al. Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. *Cell*. 2020;181:747. <https://doi.org/10.1016/j.cell.2020.04.017>.
- Corselli M, Saksena S, Nakamoto M, Lomas WE, Taylor I, Chattopadhyay PK. Single cell multiomic analysis of T cell exhaustion in vitro. *Cytom Part A*. 2022;101:27–44. <https://doi.org/10.1002/cyto.a.24496>.
- Nicolet BP, Wolkers MC. The relationship of mRNA with protein expression in CD8+ T cells associates with gene class and gene characteristics. *PLoS ONE*. 2022;17:e0276294. <https://doi.org/10.1371/journal.pone.0276294>.
- Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14:865–8. <https://doi.org/10.1038/nmeth.4380>.
- Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol*. 2017;35:936–9. <https://doi.org/10.1038/nbt.3973>.
- Mabbott NA, Baillie J, Brown H, Freeman TC, Hume DA. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*. 2013;14:632. <https://doi.org/10.1186/1471-2164-14-632>.
- Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The Human Cell Atlas: from vision to reality. *Nature*. 2017;550:451–3. <https://doi.org/10.1038/550451a>.
- Liu X, Gosline SJC, Pflieger LT, Wallet P, Iyer A, Guinney J, et al. Knowledge-based classification of fine-grained immune cell

- types in single-cell RNA-Seq data. *Brief Bioinform.* 2021;22:1–11. <https://doi.org/10.1093/bib/bbab039>.
31. Xie E, Cheng L, Shireman J, Cai Y, Liu J, Mohanty C et al. CASSIA: a multi-agent large language model for reference free, interpretable, and automated cell annotation of single-cell RNA-sequencing data. *BioRxiv* 2024:2012–24. <https://doi.org/10.1101/2024.12.04.626476>
  32. Truong L, Truong T, Nguyen H. Automated cell annotation in scRNA-seq data using unique marker gene sets. *BioRxiv.* 2024;2005–24. <https://doi.org/10.1101/2024.05.24.595477>.
  33. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods.* 2019;16:983–6.
  34. Cheng Y, Fan X, Zhang J, Li Y. A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data. *Commun Biol.* 2023;6:545.
  35. Kotliar D, Raychaudhu S. Reproducible annotation of T cell subsets and activation states with gene expression programs. *Nat Methods.* 2025;22:1774–5.
  36. Xu C, Prete M, Webb S, Jardine L, Stewart BJ, Hoo R, et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell.* 2023;186:5876–e589120. <https://doi.org/10.1016/j.cell.2023.11.026>.
  37. Kang JB, Nathan A, Weinand K, Zhang F, Millard N, Rumker L, et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nat Commun.* 2021;12:5890.
  38. Huynh KLA, Tyc KM, Matuck BF, Easter QT, Pratapa A, Kumar NV, et al. Deconvolution of cell types and states in spatial multiomics utilizing TACIT. *Nat Commun.* 2025;16:3747.
  39. Ponomarenko EA, Krasnov GS, Kiseleva OI, Kryukova PA, Arzumanyan VA, Dolgalev GV, et al. Workability of mRNA Sequencing for Predicting Protein Abundance. *Genes (Basel).* 2023;14:2065. <https://doi.org/10.3390/genes14112065>.
  40. Srivastava H, Lippincott MJ, Currie J, Canfield R, Lam MPY, Lau E. Protein prediction models support widespread post-transcriptional regulation of protein abundance by interacting partners. *PLOS Comput Biol.* 2022;18:e1010702. <https://doi.org/10.1371/journal.pcbi.1010702>.
  41. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet.* 2020;21:630–44. <https://doi.org/10.1038/s41576-020-0258-4>.
  42. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet.* 2012;13:227–32. <https://doi.org/10.1038/nrg3185>.
  43. Shi Y, Shi M, Wang Y, You J. Progress and prospects of mRNA-based drugs in pre-clinical and clinical applications. *Signal Transduct Target Ther.* 2024;9:322. <https://doi.org/10.1038/s41392-024-02002-z>.
  44. Yang R, He J, Kang D, Chen Y, Huang J, Li J, et al. Bioinformatics analysis reveals novel tumor antigens and immune subtypes of skin cutaneous melanoma contributing to mRNA vaccine development. *Front Immunol.* 2025;16:1520505. <https://doi.org/10.3389/fimmu.2025.1520505>.
  45. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA. *Abundance Cell.* 2016;165:535–50. <https://doi.org/10.1016/j.cell.2016.03.014>.
  46. Salovska B, Zhu H, Gandhi T, Frank M, Li W, Rosenberger G, et al. Isoform-resolved correlation analysis between <sc>mRNA abundance regulation and protein level degradation. *Mol Syst Biol.* 2020;16:e9170. <https://doi.org/10.15252/msb.20199170>.
  47. Kershaw CJ, Nelson MG, Castelli LM, Jennings MD, Lui J, Talavera D, et al. Translation factor and RNA binding protein mRNA interactomes support broader RNA regulons for posttranscriptional control. *J Biol Chem.* 2023;299:105195. <https://doi.org/10.1016/j.jbc.2023.105195>.
  48. Franks A, Airolidi E, Slavov N. Post-transcriptional regulation across human tissues. *PLOS Comput Biol.* 2017;13:e1005535. <https://doi.org/10.1371/journal.pcbi.1005535>.
  49. HafezQorani S, Lafzi A, de Bruin RG, van Zonneveld AJ, van der Veer EP, Son YA, et al. Modeling the combined effect of RNA-binding proteins and microRNAs in post-transcriptional regulation. *Nucleic Acids Res.* 2016;44:e83–83. <https://doi.org/10.1093/nar/gkw048>.
  50. Kim S, Kim S, Chang HR, Kim D, Park J, Son N, et al. The regulatory impact of RNA-binding proteins on microRNA targeting. *Nat Commun.* 2021;12:5057. <https://doi.org/10.1038/s41467-021-25078-5>.
  51. Yoshinaga M, Takeuchi O. Post-transcriptional control of immune responses and its potential application. *Clin Transl Immunol.* 2019;8:e1063. <https://doi.org/10.1002/cti2.1063>.
  52. Lai TH-C. Post-transcriptional Regulation of Gene Expression in Innate Immunity 2021. <https://doi.org/https://doi.org/10.26190/unswworks/2048>
  53. Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, et al. A short translational ramp determines the efficiency of protein synthesis. *Nat Commun.* 2019;10:5774. <https://doi.org/10.1038/s41467-019-13810-1>.
  54. Wu Q, Bazzini AA. Translation and mRNA Stability Control. *Annu Rev Biochem.* 2023;92:227–45. <https://doi.org/10.1146/annurev-biochem-052621-091808>.
  55. Nedialkova DD, Leidel SA. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell.* 2015;161:1606–18. <https://doi.org/10.1016/j.cell.2015.05.022>.
  56. Lyu X, Yang Q, Zhao F, Liu Y. Codon usage and protein length-dependent feedback from translation elongation regulates translation initiation and elongation speed. *Nucleic Acids Res.* 2021;49:9404–23. <https://doi.org/10.1093/nar/gkab729>.
  57. Yip MCJ, Shao S. Detecting and Rescuing Stalled Ribosomes. *Trends Biochem Sci.* 2021;46:731–43. <https://doi.org/10.1016/j.tibs.2021.03.008>.
  58. Park J, Park J, Lee J, Lim C. The trinity of ribosome-associated quality control and stress signaling for proteostasis and neuronal physiology. *BMB Rep.* 2021;54:439–50. <https://doi.org/10.5483/BMBRep.2021.54.9.097>.
  59. Payne SH. The utility of protein and mRNA correlation. *Trends Biochem Sci.* 2015;40:1–3. <https://doi.org/10.1016/j.tibs.2014.10.010>.
  60. Rusilowicz-Jones EV, Urbé S, Clague MJ. Protein degradation on the global scale. *Mol Cell.* 2022;82:1414–23. <https://doi.org/10.1016/j.molcel.2022.02.027>.
  61. Tsai JM, Nowak RP, Ebert BL, Fischer ES. Targeted protein degradation: from mechanisms to clinic. *Nat Rev Mol Cell Biol.* 2024;25:740–57. <https://doi.org/10.1038/s41580-024-00729-9>.
  62. Varshavsky A. The Ubiquitin System, Autophagy, and Regulated Protein Degradation. *Annu Rev Biochem.* 2017;86:123–8. <https://doi.org/10.1146/annurev-biochem-061516-044859>.
  63. Csárdi G, Franks A, Choi DS, Airolidi EM, Drummond DA. *PLOS Genet.* 2015;11:e1005206. <https://doi.org/10.1371/journal.pgen.1005206>. Accounting for Experimental Noise Reveals That mRNA Levels, Amplified by Post-Transcriptional Processes, Largely Determine Steady-State Protein Levels in Yeast.
  64. Shao W, Guo T, Toussaint NC, Xue P, Wagner U, Li L, et al. Comparative analysis of mRNA and protein degradation in prostate tissues indicates high stability of proteins. *Nat Commun.* 2019;10:2524. <https://doi.org/10.1038/s41467-019-10513-5>.
  65. Koh Y-C, Yang G, Lai C-S, Weerawatanakorn M, Pan M-H. Chemopreventive Effects of Phytochemicals and Medicines on M1/M2 Polarized Macrophage Role in Inflammation-Related Diseases. *Int J Mol Sci.* 2018;19:2208. <https://doi.org/10.3390/ijms19082208>.

66. Blank CU, Haining WN, Held W, Hogan PG, Kallies A, Lugli E, et al. Defining 'T cell exhaustion'. *Nat Rev Immunol*. 2019;19:665–74. <https://doi.org/10.1038/s41577-019-0221-9>.
67. Chen G, Ning B, Shi T, Single-Cell RNA-S. *Front Genet*. 2019;10:1–13. <https://doi.org/10.3389/fgene.2019.00317>. Technologies and Related Computational Data Analysis.
68. Ran D, Zhang S, Lytal N, An L. scDoc: correcting drop-out events in single-cell RNA-seq data. *Bioinformatics*. 2020;36:4233–9. <https://doi.org/10.1093/bioinformatics/btaa283>.
69. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics*. 2018;19:220. <https://doi.org/10.1186/s12859-018-2226-y>.
70. Ye P, Ye W, Ye C, Li S, Ye L, Ji G, et al. scHinter: imputing dropout events for single-cell RNA-seq data with limited sample size. *Bioinformatics*. 2020;36:789–97. <https://doi.org/10.1093/bioinformatics/btz627>.
71. Feng X, Chen L, Wang Z, Li SC. I-Impute: a self-consistent method to impute single cell RNA sequencing data. *BMC Genomics*. 2020;21:618. <https://doi.org/10.1186/s12864-020-07007-w>.
72. Juan W, Ahn KW, Chen Y-G, Lin C-W. A Consensus Clustering-Based Imputation Method for Addressing Dropout Events in scRNA-Seq Data. *Bioengineering*. 2025;12:31. <https://doi.org/10.3390/bioengineering12010031>.
73. Upadhyay SR, Ryan CJ. Antibody reliability influences observed mRNA–protein correlations in tumour samples. *Life Sci Alliance*. 2023;6:e202201885. <https://doi.org/10.26508/lsa.202201885>.
74. Ausserwöger H, Schneider MM, Herling TW, Arosio P, Invernizzi G, Knowles TPJ, et al. Non-specificity as the sticky problem in therapeutic antibody development. *Nat Rev Chem*. 2022;6:844–61. <https://doi.org/10.1038/s41570-022-00438-x>.
75. Dempsey ME, Woodford-Berry O, Darling EM. Quantification of Antibody Persistence for Cell Surface Protein Labeling. *Cell Mol Bioeng*. 2021;14:267–77. <https://doi.org/10.1007/s12195-021-00670-3>.
76. Janssen P, Kliesmete Z, Vieth B, Adiconis X, Simmons S, Marshall J, et al. The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biol*. 2023;24:1–22. <https://doi.org/10.1186/s13059-023-02978-x>.
77. Mulè MP, Martins AJ, Tsang JS. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nat Commun*. 2022;13:2099. <https://doi.org/10.1038/s41467-022-29356-8>.
78. Colino-Sanguino Y, Rodriguez de la Fuente L, Gloss B, Law AMK, Handler K, Pajic M, et al. Performance comparison of high throughput single-cell RNA-Seq platforms in complex tissues. *Heliyon*. 2024;10:e37185. <https://doi.org/10.1016/j.heliyon.2024.e37185>.
79. Hyatt G, Melamed R, Park R, Seguritan R, Laplace C, Poirot L, et al. Gene expression microarrays: glimpses of the immunological genome. *Nat Immunol*. 2006;7:686–91. <https://doi.org/10.1038/nri0706-686>.
80. Pulumati A, Pulumati A, Dwarakanath BS, Verma A, Papineni RVL. Technological advancements in cancer diagnostics: Improvements and limitations. *Cancer Rep*. 2023;6:e1764. <https://doi.org/10.1002/cnr2.1764>.
81. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*. 2014;9:e78644. <https://doi.org/10.1371/journal.pone.0078644>.
82. Ermann J, Rao DA, Teslovich NC, Brenner MB, Raychaudhuri S. Immune cell profiling to guide therapeutic decisions in rheumatic diseases. *Nat Rev Rheumatol*. 2015;11:541–51. <https://doi.org/10.1038/nrrheum.2015.71>.
83. Shendure J. The beginning of the end for microarrays? *Nat Methods*. 2008;5:585–7. <https://doi.org/10.1038/nmeth0708-585>.
84. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harb Protoc*. 2015;2015:top084970. <https://doi.org/10.1101/pdb.top084970>.
85. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10:57–63. <https://doi.org/10.1038/nrg2484>.
86. Li X, Wang C-Y. From bulk, single-cell to spatial RNA sequencing. *Int J Oral Sci*. 2021;13:36. <https://doi.org/10.1038/s41368-021-00146-0>.
87. Huang K, Xu Y, Feng T, Lan H, Ling F, Xiang H, et al. The Advancement and Application of the Single-Cell Transcriptome in Biological and Medical Research. *Biology (Basel)*. 2024;13:451. <https://doi.org/10.3390/biology13060451>.
88. Molla Desta G, Birhanu AG. Advancements in single-cell RNA sequencing and spatial transcriptomics: transforming biomedical research. *Acta Biochim Pol*. 2025;72:13922. <https://doi.org/10.3389/abp.2025.13922>.
89. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. 2017;9:75. <https://doi.org/10.1186/s13073-017-0467-4>.
90. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Sci* (80–). 2015;348:aaa6090. <https://doi.org/10.1126/science.aaa6090>.
91. Xia C, Fan J, Emanuel G, Hao J, Zhuang X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc Natl Acad Sci*. 2019;116:19490–9. <https://doi.org/10.1073/pnas.1912459116>.
92. Zeng Y, Luo M, Shangguan N, Shi P, Feng J, Xu J, et al. Deciphering cell types by integrating scATAC-seq data with genome sequences. *Nat Comput Sci*. 2024;4:285–98. <https://doi.org/10.1038/s43588-024-00622-7>.
93. Blundell MP, Sanderson SL, Long TA. Flow Cytometry as an Important Tool in Proteomic Profiling. *Proteomic Profiling Methods Protoc.*, Springer; 2021, pp. 213–27. [https://doi.org/10.1007/978-1-0716-1186-9\\_13](https://doi.org/10.1007/978-1-0716-1186-9_13).
94. Maes E, Cools N, Willems H, Baggerman G. FACS-Based Proteomics Enables Profiling of Proteins in Rare Cell Populations. *Int J Mol Sci*. 2020;21:6557. <https://doi.org/10.3390/ijms21186557>.
95. Potts M, Fletcher-Etherington A, Nightingale K, Mescia F, Bergamaschi L, Calero-Nieto FJ, et al. Proteomic analysis of circulating immune cells identifies cellular phenotypes associated with COVID-19 severity. *Cell Rep*. 2023;42:112613. <https://doi.org/10.1016/j.celrep.2023.112613>.
96. Lyons YA, Wu SY, Overwijk WW, Baggerly KA, Sood AK. Immune cell profiling in cancer: molecular approaches to cell-specific identification. *Npj Precis Oncol*. 2017;1:26. <https://doi.org/10.1038/s41698-017-0031-0>.
97. Andersen MN, Al-Karradi SNH, Kragstrup TW, Hokland M. Elimination of erroneous results in flow cytometry caused by antibody binding to Fc receptors on human monocytes and macrophages. *Cytom Part A*. 2016;89:1001–9. <https://doi.org/10.1002/cyto.a.22995>.
98. Fadziruddin NFZ, Abd Hamid IJ, Hashim IF, Zainudeen ZT. Advantages and Limitations of Immunophenotyping and Lymphocyte Proliferation Assay in Clinical Flow Cytometry in the Field of Primary Immunodeficiency Diseases. *J Heal Transl Med*. 2022;1:80–9. <https://doi.org/10.22452/jumec.sp2022no1.10>.
99. Tracey LJ, An Y, Justice MJ. CyTOF: An Emerging Technology for Single-Cell Proteomics in the Mouse. *Curr Protoc*. 2021;1:e118. <https://doi.org/10.1002/cpz1.118>.

100. Zhang Q, Ye M, Lin C, Hu M, Wang Y, Lou Y, et al. Mass cytometry-based peripheral blood analysis as a novel tool for early detection of solid tumours: a multicentre study. *Gut*. 2023;72:996–1006. <https://doi.org/10.1136/gutjnl-2022-327496>.
101. Castellano E, Samba C, Estes G, Simpson L, Vendrame E, Garcia-Cuesta EM, et al. CyTOF analysis identifies unusual immune cells in urine of BCG-treated bladder cancer patients. *Front Immunol*. 2022;13:970931. <https://doi.org/10.3389/fimmu.2022.970931>.
102. Xiao X, Guo Q, Cui C, Lin Y, Zhang L, Ding X, et al. Multiplexed imaging mass cytometry reveals distinct tumor-immune microenvironments linked to immunotherapy responses in melanoma. *Commun Med*. 2022;2:131. <https://doi.org/10.1038/s43856-022-00197-2>.
103. Elaldi R, Hemon P, Petti L, Cosson E, Desrues B, Sudaka A, et al. High Dimensional Imaging Mass Cytometry Panel to Visualize the Tumor Immune Microenvironment Contexture. *Front Immunol*. 2021;12:666233. <https://doi.org/10.3389/fimmu.2021.666233>.
104. Colpitts SJ, Budd MA, Monajemi M, Reid KT, Murphy JM, Ivison S, et al. Strategies for optimizing CITE-seq for human islets and other tissues. *Front Immunol*. 2023;14:1107582. <https://doi.org/10.3389/fimmu.2023.1107582>.
105. Todorovic V. Single-cell RNA-seq—now with protein. *Nat Methods*. 2017;14:1028–9. <https://doi.org/10.1038/nmeth.4488>.
106. Smibert P, Mimitou E, Stoeckius M. ECCITE-seq. *Protoc Exch*. 2019. <https://doi.org/10.1038/protex.2019.025>.
107. Papalexli E, Mimitou EP, Butler AW, Foster S, Bracken B, Mauck WM, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet*. 2021;53:322–31. <https://doi.org/10.1038/s41588-021-00778-2>.
108. Matusiak M, Hickey JW, van IJendoorn DGP, Lu G, Kidziński L, Zhu S, et al. Spatially Segregated Macrophage Populations Predict Distinct Outcomes in Colon Cancer. *Cancer Discov*. 2024;14:1418–39. <https://doi.org/10.1158/2159-8290.CD-23-1300>.
109. de Oliveira MF, Romero JP, Chung M, Williams SR, Gottschö AD, Gupta A, et al. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nat Genet*. 2025;57:1512–23. <https://doi.org/10.1038/s41588-025-02193-3>.
110. Cook DP, Jensen KB, Wise K, Roach MJ, Dezem FS, Ryan NK et al. A Comparative Analysis of Imaging-Based Spatial Transcriptomics Platforms. *BioRxiv* 2023:2012–23. <https://doi.org/10.1101/2023.12.13.571385>.
111. Otolano N. The neXt Generation of Single Cell RNA-Seq: An Introduction to GEM-X Technology. 10x Genomics 2024.
112. Chen J-S, Yang S, Shuga J, Hui W, Santiago A, Akhremichev I, et al. Abstract 1440: GEM-X Universal Multiplex: a cost-effective, sample agnostic multiplexing solution for scRNA-seq in cancer research. *Cancer Res*. 2025;85:1440–1440. <https://doi.org/10.1158/1538-7445.AM2025-1440>.
113. Hatje K, Schneider K, Danilin S, Koechl F, Giroud N, Juglair L et al. Comparison of Fixed Single Cell RNA-seq Methods to Enable Transcriptome Profiling of Neutrophils in Clinical Samples. *BioRxiv* 2024:2008–24. <https://doi.org/10.1101/2024.08.14.607767>.
114. Filippov I, Philip CS, Schauser L, Peterson P. Comparative transcriptomic analyses of thymocytes using 10x Genomics and Parse scRNA-seq technologies. *BMC Genomics*. 2024;25:1069. <https://doi.org/10.1186/s12864-024-10976-x>.
115. Ingram PN, Gao Q, Liu L, Perreau J, Kukreja R, Gadish O, et al. Abstract 3188: Functional interrogation of human T cells using the accessible Beacon DiscoveryTM optofluidic system. *Cancer Res*. 2025;85:3188–3188. <https://doi.org/10.1158/1538-7445.AM2025-3188>.
116. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50:1–14. <https://doi.org/10.1038/s12276-018-0071-8>.
117. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol*. 2022;23:31. <https://doi.org/10.1186/s13059-022-02601-5>.
118. Reimegård J, Tarbier M, Danielsson M, Schuster J, Baskaran S, Panagiotou S, et al. A combined approach for single-cell mRNA and intracellular protein expression analysis. *Commun Biol*. 2021;4:624. <https://doi.org/10.1038/s42003-021-02142-w>.
119. Wang F, Liu C, Li J, Yang F, Song J, Zang T, et al. SPDB: a comprehensive resource and knowledgebase for proteomic data at the single-cell resolution. *Nucleic Acids Res*. 2024;52:D562–71. <http://doi.org/10.1093/nar/gkad1018>.
120. Tian J, Bai X, Quek C. Single-Cell Informatics for Tumor Microenvironment and Immunotherapy. *Int J Mol Sci*. 2024;25:4485. <https://doi.org/10.3390/ijms25084485>.
121. Zhang K, Zemke NR, Armand EJ, Ren B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat Methods*. 2024;21:217–27. <https://doi.org/10.1038/s41592-023-02139-9>.
122. Gronbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*. 2020;36:4415–22. <https://doi.org/10.1093/bioinformatics/btaa293>.
123. Voß H, Schlumbohm S, Barwikowski P, Wurlitzer M, Dottermusch M, Neumann P, et al. HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nat Commun*. 2022;13:3523. <https://doi.org/10.1038/s41467-022-31007-x>.
124. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, 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>.
125. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21:31. <https://doi.org/10.1186/s13059-020-1926-6>.
126. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, 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>.
127. Butler A, Hoffman P, Smibert P, Papalexli E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36:411–20. <https://doi.org/10.1038/nbt.4096>.
128. Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Mariotti JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 2020;21:111. <https://doi.org/10.1186/s13059-020-02015-1>.
129. Xu J, Huang D, Zhang X. Transformer Adv Sci. 2024;11:1–14. <https://doi.org/10.1002/adv.202307835>. scmFormer Integrates Large-Scale Single-Cell Proteomics and Transcriptomics Data by Multi-Task.
130. Xiong J, Zheng S, Gong F, Ma L, Wan L. Integrate and generate single-cell proteomics from transcriptomics with cross-attention. 2025:1–7. <https://doi.org/10.1101/2025.01.28.635217>.
131. Sheinin R, Sharan R, Madi A. scNET: learning context-specific gene and cell embeddings by integrating single-cell gene expression data with protein–protein interactions. *Nat Methods*. 2025;22. <https://doi.org/10.1038/s41592-025-02627-0>.
132. Chen Y, Fan X, Shi C, Shi Z, Wang C. A joint analysis of single cell transcriptomics and proteomics using transformer. *NPJ Syst Biol Appl*. 2025;11:1. <https://doi.org/10.1038/s41540-024-00484-9>.
133. Lakkis J, Schroeder A, Su K, Lee MYY, Bashore AC, Reilly MP, et al. A multi-use deep learning method for CITE-seq and

- single-cell RNA-seq data integration with cell surface protein prediction and imputation. *Nat Mach Intell.* 2022;4:940–52. <https://doi.org/10.1038/s42256-022-00545-w>.
134. Li W, Xiang B, Yang F, Rong Y, Yin Y, Yao J, et al. scMHNN: a novel hypergraph neural network for integrative analysis of single-cell epigenomic, transcriptomic and proteomic data. *Brief Bioinform.* 2023;24:1–15. <https://doi.org/10.1093/bib/bbad391>.
135. Cao Y, Wang X, Peng G. SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data. *Front Genet.* 2020;11:490. <https://doi.org/10.3389/fgene.2020.00490>.
136. Quan F, Liang X, Cheng M, Yang H, Liu K, He S, et al. Annotation of cell types (ACT): a convenient web server for cell type annotation. *Genome Med.* 2023;15:91. <https://doi.org/10.1186/s13073-023-01249-5>.
137. [137] Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, 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. <https://doi.org/10.1093/nar/gkac947>.
138. [138] Atlas HC. Human cell atlas. *Hum Cell Atlas Seq First 250K Dev Cells.* 2018.
139. [139] Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an interactive home for single-cell genomics data. *BioRxiv* 2023. <https://doi.org/10.1101/2023.07.13.548886>.
140. [140] Li Y-Y, Zhou L-W, Qian F-C, Fang Q-L, Yu Z-M, Cui T, et al. scImmOmics: a manually curated resource of single-cell multi-omics immune data. *Nucleic Acids Res.* 2025;53:D1162–72. <https://doi.org/10.1093/nar/gkae985>.

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