OXFORD

Briefings in Bioinformatics, 2026, **27(1)**, bbaf723

https://doi.org/10.1093/bib/bbaf723

Review

Advances and challenges in single-cell RNA sequencing data analysis: a comprehensive review

Ali Mohammad Nesari, Habib MotieGhader*, Saeid Ghorbian

Department of Biology, Ta.C., Islamic Azad University, Tabriz, Iran

*Corresponding author. Associate Professor of Bioinformatics, Department of Biology, Ta.C., Islamic Azad University, Tabriz, Iran. E-mail: habib_moti@iau.ac.ir

Abstract

Single-cell RNA sequencing (scRNA-seq) has transformed the resolution of cellular heterogeneity, offering insights into dynamic biological processes from tumor evolution to immune regulation. However, its clinical translation is limited by challenges such as data sparsity, batch effects (differences caused by technical variation rather than biology), and the absence of standardized benchmarks for core pipelines like Seurat and Scanpy. This review outlines emerging computational strategies that address these limitations: (A) robust preprocessing, including SCTransform for zero-inflation(an excess of zero counts in gene-expression data) correction and Harmony for batch integration—achieving 30% faster alignment than BBKNN in cohorts exceeding 100,000 cells; (B) transformer-based annotation tools such as scGPT and CellTypist, which reach >95% accuracy in immune profiling using models pretrained on 33 million cells; and (C) multimodal integration with spatial transcriptomics (e.g., 10x Visium, cell2location v2), which delineate microenvironmental niches and rare CX3CR1+ T-cell subsets in disease contexts like glioblastoma and severe COVID-19.

We further assess how scANVI bridges scRNA-seq and ATAC-seq to uncover epigenetic mechanisms underlying therapy resistance, and how spatial methods elucidate tumor–immune crosstalk at subcellular resolution. Despite these advances, ethical risks remain, particularly around re-identification of rare patient-derived clones such as pre-metastatic cells.

To promote clinical adoption, we propose a roadmap that prioritizes benchmarked workflows (e.g., scverse ecosystem), privacy-aware data sharing via federated learning, and causal AI approaches to disentangle biological signal from technical artifact. By synthesizing computational innovations with translational case studies, this review equips researchers to navigate both the analytical and ethical complexities of scRNA-seq in pursuit of actionable diagnostics.

Keywords: single cell; sequencing; RNA-seq; early detection; cancer

Introduction

In the decade since its inception, single-cell RNA sequencing (scRNA-seq) has evolved from a niche technology to a cornerstone of precision medicine, revealing cellular secrets that once hid in bulk averages. While bulk RNA-seq homogenizes tissues into molecular averages, scRNA-seq has exposed the 'dark matter' of biology—rare metastatic precursors, exhausted immune cells, and neurodevelopmental trajectories—rewriting textbooks in oncology, immunology, and neuroscience. Single-cell RNA sequencing (scRNA-seq) has fundamentally transformed the study of cellular heterogeneity by enabling high-resolution profiling of gene expression at the individual cell level. This technology has revealed unprecedented insights into dynamic biological systems, ranging from tumor evolution and immune responses to neuronal diversity and tissue regeneration. By resolving cellular states and transitions that were previously obscured in bulk analyses, scRNA-seq has redefined our understanding of tissue complexity and disease pathogenesis [1–10].

Despite its remarkable potential, the clinical and translational application of scRNA-seq remains limited by substantial computational and methodological challenges. These include the sparsity and high dimensionality of single-cell data [11], batch effects [12], interoperability issues between analysis frameworks [13], and ethical concerns associated with patient-derived data [14]. Moreover, the rapid proliferation of analytical tools and pipelines has created a fragmented landscape, often lacking standardized benchmarks for evaluating the performance and reproducibility of critical steps such as clustering, trajectory inference, and differential gene expression analysis [15–17].

Several recent reviews have addressed specific aspects of single-cell data analysis or tool development. However, few have provided a comprehensive synthesis of emerging computational frameworks in the context of translational research applications. This review aims to fill this gap by critically evaluating the latest advances in scRNA-seq data analysis, integrating benchmarking studies, AI-driven annotation frameworks, and multimodal omics approaches. We place particular emphasis on how these innovations translate into clinical insights, such as identifying therapy-resistant tumor clones, profiling immune cell states in COVID-19, and mapping neuronal diversity in the human brain [18–25].

This review provides an overview of computational strategies for single-cell RNA-seq (scRNA-seq) data analysis, highlighting

Received: September 12, 2025. **Revised:** November 22, 2025. **Accepted:** December 13, 2025

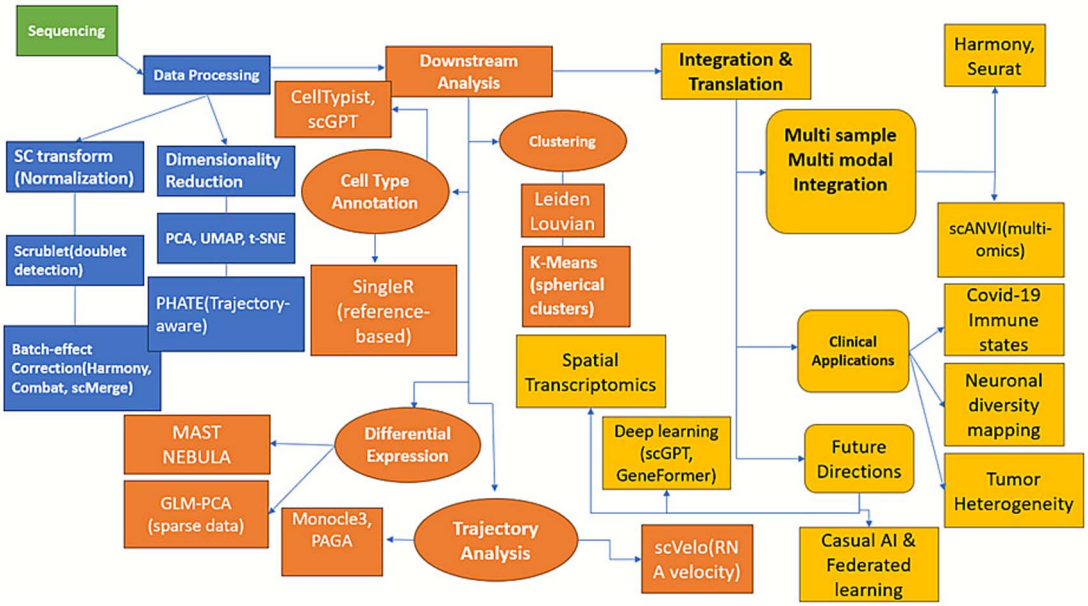

Figure 1. Overview of a typical single-cell RNA-sequencing (scRNA-seq) data-analysis pipeline. The workflow begins with sequencing and proceeds through data processing (Preprocessing, dimensionality reduction), downstream analysis (clustering, differential expression, trajectory analysis, cell-type annotation), and Integration & Translation modules, which include multi-sample and multi-modal integration, clinical applications, and emerging approaches such as spatial transcriptomics, deep learning, and causal AI. Combat and harmony are represented under batch-effect correction rather than normalization.

recent advances, challenges, and future directions. It is primarily intended for biomedical and computational researchers who wish to understand, select, and apply scRNA-seq analysis methods in practical experimental and clinical contexts.

Structure of this review

As indicated in Fig. 1, the structure of this article is organized as follows:

- **Preprocessing of scRNA-seq Data**: We discuss innovations in data normalization, batch correction, and quality control, including SCTransform and Harmony.
- **Dimensionality Reduction and Clustering**: An evaluation of methods such as PCA, UMAP, Leiden clustering, and graph-based approaches.
- **Differential Gene Expression Analysis**: A comparison of specialized models addressing zero inflation and multimodality, including MAST, NEBULA, and GLM-PCA.
- **Trajectory Inference and Pseudotime Analysis**: Coverage of Monocle3, Slingshot, RNA velocity, and PAGA for dynamic process modeling.
- **Cell Type Identification and Annotation**: Recent AI-powered tools like scGPT and CellTypist, and integration of spatial and multi-omics data.
- **Integration of Multi-Sample or Multi-Condition Datasets**: Approaches to harmonize scRNA-seq data across different batches and modalities.
- **Emerging Topics and Future Directions**: Including spatial transcriptomics, single-cell multi-omics, deep learning, federated learning, and ethical considerations in clinical applications.

Preprocessing of scRNA-seq data

The advent of single-cell RNA sequencing (scRNA-seq) has revolutionized genomics, enabling detailed characterization of cellular diversity, cell-type discrimination, and gene expression at single-cell resolution. However, this technology presents specific challenges, particularly in the critical preprocessing stage that underpins all downstream analyses.

A major preprocessing challenge in scRNA-seq is managing technical noise and dropout events, where low input material leads to undetected genes. The resulting high sparsity, with many zero values, complicates statistical analyses [26].

We outline below key bioinformatics tools for scRNA-seq data analysis [27]. Effective preprocessing begins with quality assessment using tools like FastQC to detect GC bias, adapter contamination, or low-quality reads, followed by filtering and trimming. Normalization then adjusts for technical factors such as sequencing depth and batch effects.

Low-quality cells can be filtered using Seurat, which applies metrics such as total expressed genes, UMI counts, and mitochondrial gene percentage [26]. or scCatalyzer, which uses a hierarchical Bayesian model to remove technical noise [28]. High mitochondrial expression often indicates poor cell viability and is grounds for exclusion [27, 28].

Normalization methods include global scaling approaches (CPM, TPM, UMI) and advanced options like scran, which pools spike-ins or engineered cells to estimate size factors [27].

Benchmarks show SCTransform handles zero-inflation better than scran, while Harmony integrates large datasets (>100 k cells) about 30% faster than BBKNN [29].

Batch effect correction methods such as Combat, scMerge and Harmony mitigate variation from experimental or technical sources. Normalization methods aim to correct unwanted technical differences between cells so that true biological variation can be compared. For example, SCTransform uses a regularized negative binomial model to stabilize variance and reduce noise, whereas scran applies pooling-based normalization to estimate relative cell size factors from expression distributions. Integration tools like Harmony align multiple datasets into a shared space, helping remove batch effects (differences between experiments) while maintaining the underlying biology. Preprocessing choices

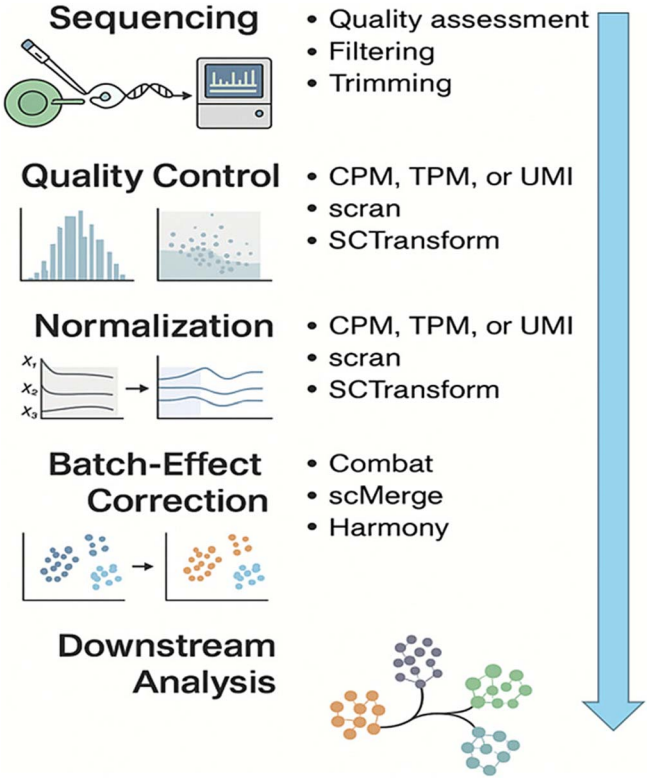

Figure 2. Simplified schematic of preprocessing and integration steps in single-cell RNA-sequencing (scRNA seq) analysis. The workflow begins with sequencing and proceeds through quality control, normalization, and batch-effect correction before downstream analysis. Combat, scMerge, and harmony are shown as batch-effect-correction methods rather than normalization tools.

strongly influence downstream results, and tool development continues to address scRNA-seq–specific challenges [26–29].

In practice, the choice of preprocessing method depends strongly on dataset scale and technical variation. For example, *SCTransform* is well-suited for small-to-medium datasets (<50 k cells) due to its robust handling of zero inflation and variance stabilization, whereas *scran* remains preferable when spike-in controls or ERCC standards are available. In contrast, *Harmony* is highly efficient for large multi-batch studies (>100 k cells) but may risk overcorrection when subtle biological differences correlate with batch factors. For cross-platform integration (e.g., 10x versus Smart-seq2), *scMerge* often preserves rare subpopulations better than regression-based approaches like *Combat*. Therefore, method selection should balance computational scalability with biological interpretability, rather than relying on a single default pipeline.

As it's clearly shown in Fig. 2, Once the data has been preprocessed, it can be used for downstream analyses, such as cell-type identification, differential gene expression analysis, and trajectory inference [26–29]. In practice, this allows researchers to visualize and distinguish biologically meaningful cell subtypes more clearly.

Dimensionality reduction and clustering

Dimensionality reduction is a crucial step in the analysis of high-dimensional scRNA-seq data. By reducing the dimensionality of the data, researchers can simplify complex biological datasets and uncover important patterns and insights [30]. In the context of single-cell analysis, dimensionality reduction can help identify distinct cell subpopulations, track the progression of cells through different states, and elucidate the underlying structures within a heterogeneous sample [11].

Dimensionality reduction denoises data and mitigates the 'curse of dimensionality' in sparse scRNA-seq datasets. Methods like PCA or t-SNE project data into fewer dimensions, facilitating downstream analysis and interpretation.

PCA is a common dimensionality reduction method for scRNA-seq, capturing major variance components while preserving structure [31]. Other unsupervised techniques, such as latent Dirichlet analysis and t-SNE, can also reduce dimensionality [32].

T-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional data in a two- or three-dimensional space [33]. t-SNE is known for its ability to preserve the local structure of the data, making it effective at identifying distinct cell subpopulations within a single-cell dataset. By combining dimensionality reduction with clustering algorithms, researchers can uncover the hidden substructures within a single-cell dataset and gain a deeper understanding of the underlying cellular heterogeneity [34].

As described in table 1 which compares different clustering methods, t-SNE has limitations in scalability (>10 k cells) and reproducibility, leading to greater use of UMAP in trajectory inference (e.g., Monocle3) and PHATE for branching dynamics [35].

While graph-based algorithms such as *Leiden* and *Louvain* dominate current workflows for their scalability and ability to capture hierarchical structures, they can sometimes merge biologically distinct subtypes when resolution parameters are poorly tuned. *K-means* and *Gaussian mixture models* provide interpretable cluster boundaries but perform less effectively in high-dimensional, sparse data. *UMAP* generally preserves both local and global structure better than *t-SNE*, making it advantageous for trajectory analysis, whereas *t-SNE* remains preferred for visualizing compact clusters in immune profiling. Consequently, optimal clustering depends on study design: exploratory atlasing studies benefit from graph-based approaches, while hypothesis-driven studies with defined subpopulations may favor parametric models.

Clustering is central to scRNA-seq analysis, grouping cells into meaningful subpopulations. Algorithm choice affects downstream interpretation; here we assess Leiden, Louvain, and k-means for accuracy, scalability, and suitability [36].

k-means partitions data into k clusters by similarity, while Gaussian mixture models allow overlapping cluster shapes. Hierarchical clustering builds a cluster hierarchy, aiding visualization and detection of rare cell types [36].

UMAP is a dimensionality reduction method that better preserves global structure than t-SNE, aiding in rare cell type identification and developmental trajectory analysis [15, 29].

After dimensionality reduction, the next step is clustering.

Clustering groups cells with similar transcriptional profiles, revealing heterogeneity and enabling discovery of novel cell types or states [37]. Hierarchical clustering identifies cell subpopulations and their relationships, using dendrograms to illustrate structural similarities within heterogeneous samples [37].

Clustering methods such as k-means, Gaussian mixture models, and hierarchical clustering are chosen based on dataset size, complexity, expected subpopulations, and desired resolution [38].

Louvain detects cell clusters by optimizing modularity in cell–cell similarity graphs, while Leiden extends Louvain to better resolve complex, hierarchical structures [39]. Graph-based methods like Louvain and Leiden use nearest-neighbor graphs to identify highly interconnected cell clusters in scRNA-seq data [39–41].

Table 1. Compares Leiden, Louvain, and k-means clustering for scRNA-seq, summarizing ARI, runtime, scalability, and typical use cases based on benchmark data.

Clustering Method	Adjusted Rand Index (ARI)	Runtime (10 k cells)	Scalability	Use Cases
Leiden	0.85–0.92	~120 seconds	Excellent (>1 M cells)	-Complex biological networks (e.g., immune cell states, developmental trajectories). -Detects hierarchical communities.
Louvain	0.78–0.85	~100 seconds	Excellent (>1 M cells)	-General-purpose community detection. -Moderate-resolution clusters in large datasets.
K-means	0.65–0.75	~50 seconds	Moderate (struggles>100 k cells)	-Predefined spherical clusters (e.g., cell cycle stages). –Requires prior knowledge of k.

Table 2 lists common dimensionality reduction methods for scRNA-seq, including PCA, MDS, t-SNE, UMAP, Diffusion Maps, NMF, ICA, GET, MAI, AE, VAE, HVG selection, GSR, FA, and LSI [41].

Density-based methods such as DBSCAN can detect clusters of arbitrary shape and size, as well as outliers, aiding identification of rare cell types or subpopulations [42, 43].

In summary, dimensionality reduction and clustering reveal cellular heterogeneity in scRNA-seq data, providing deeper insight into underlying biological processes as shown in Fig. 3.

Table 2 also summarizes method characteristics, including support for non-linearity, interpretability (Simple, Moderate, Complex), scalability (Large, Medium, Small), locality (Local versus Global trends), and robustness to noise and outliers [42–46].

Differential gene expression (DGE) analysis

Identifying differentially expressed genes is key to understanding molecular drivers of phenotypic variation [47]. While DNA microarrays were once standard, RNA-seq has become the preferred high-throughput alternative, with adoption expected to grow as sequencing costs decline [48].

Differential expression in scRNA-seq requires specialized methods to handle zero inflation and multimodal expression, unlike bulk RNA-seq [47]. Several software packages address these challenges, and their performance should be compared with bulk RNA-seq approaches [26, 48].

Comparative studies show that scRNA-seq–specific DGE tools do not always outperform bulk RNA-seq methods [26]. Bulk approaches like DESeq2, edgeR, and limma can be applied with proper normalization to account for zero inflation and over-dispersion [26, 48]. GLM-PCA reduces false positives by ~20% compared to MAST in sparse datasets, while NB regression methods such as NEBULA scale to 1 M cells without subsampling [49] as shown in table 3.

scDE is another framework for differential gene expression analysis in scRNA-seq data that was developed to specifically address the unique challenges of this data type [26].

MAST is a modeling approach that uses a hurdle model to accommodate the discrete and continuous components of single-cell gene expression [50].

scRNA-seq DGE analysis faces challenges, including data sparsity, dropout events, and population heterogeneity. Improved computational tools are enhancing the detection of genes that distinguish cell types or conditions [26, 51].

Models for scRNA-seq DGE include zero-inflated negative binomial approaches for excess zeros and over-dispersion, hurdle models separating detection from expression level, and empirical Bayes methods (e.g., limma) to boost power while controlling false discovery. Method selection depends on dataset characteristics, research goals, and sensitivity–specificity trade-offs [52]. Also, the workflow for identifying DEGs in scRNA-seq is indicated in Fig. 4. Although numerous DGE frameworks exist for scRNA-seq, their performance is highly context-dependent. MAST and scDE are optimized for datasets with moderate sparsity and are widely used for single-condition comparisons, whereas NEBULA and GLM-PCA scale more efficiently to large datasets but may underperform in detecting subtle transcriptional shifts in rare subpopulations. Importantly, conventional bulk RNA-seq tools such as edgeR or DESeq2 can still be applied after pseudo-bulk aggregation, often improving statistical power and reproducibility across biological replicates. Researchers should therefore consider the biological question and data sparsity before selecting a DGE method rather than applying a single framework universally. Identifying differentially expressed genes remains an active field, with ongoing development of specialized computational tools.

Trajectory inference and Pseudotime analysis

A key application of scRNA-seq is trajectory inference and pseudotime analysis, which reveal branching structures, state transitions, and regulatory mechanisms in cell fate decisions [28]. These methods also detect rare or transient states often missed in bulk analyses [11, 53].

Monocle infers trajectories and pseudotime using Reversed Graph Embedding to construct low-dimensional branching structures from k-nearest neighbor graphs [28]. This technique produces a principal graph that fits the data [53]. Monocle3 may struggle with >3 branches, while PAGA uses graph abstraction to resolve complex hierarchies, such as in pancreatic development.

Slingshot handles complex, multi-branching trajectories by first applying dimensionality reduction, then clustering to define cell states, and finally connecting them into smooth trajectories to estimate pseudotime (an inferred developmental timeline) along each branch [11, 28].

Diffusion Maps model cell movement in high-dimensional gene expression space as a Markov process, revealing manifold structure and capturing both linear and branching trajectories [11].

High-definition spatial transcriptomics complements scRNA-seq by preserving spatial gene expression context. Combined with trajectory inference, it links spatial organization to dynamic states and developmental pathways [11, 28].

Table 2. Techniques used to reduce the dimensions of scRNAs.

Technique	Characteristics					Use Cases	Packages
	Non-linearity	Interpretability	Scalability	Locality	Robustness		
PCA		S	L	G		• Clustering • Quantitative Analysis	Python: scikit-learn, numpy R: stats, prcomp, Seurat
MDS	✓	M	SM	G		• Visualization	Python: scikit-learn R: stats, cmdscale, vegan
t-SNE	✓	M	SM	L		• Visualization • Clustering	Python: scikit-learn, openTSNE R: Rtsne, Seurat
UMAP	✓	M	L	L	✓	• Visualization • Clustering	Python: umap-learn, scanpy R: umap, Seurat
DM	✓	M	M	L	✓	• Trajectory Analysis	Python: scprep, scanpy R: destiny
NMF		M	M	G		• Feature Extraction • Quantitative Analysis	Python: sklearn.decomposition, nimfa R: NMF, Seurat
ICA		M	M	G		• Feature Extraction	Python: scikit-learn R: fastICA, Seurat
GET	✓	C	M	L	✓	• Clustering • Graph Analysis	Python: scanpy, igraph R: igraph, Seurat
MAI	✓	M	M	G	✓	• Visualization	Python: scikit-learn R: vegan, RDRToolbox
AE	✓	C	L	L		• Feature Extraction	Python: TensorFlow, PyTorch, scanpy R: keras, tensorflow
VAEs	✓	C	L	L	✓	• Probabilistic Modeling	Python: scvi-tools, Pyro, scanpy R: scVI (via Python-R integration)
HVG		S	L	G	✓	• Feature Selection	Python: scanpy, scikit-learn R: Seurat, scran
GSR	✓	M	L	LG	✓	• Feature Selection • Pathway Analysis	Python: gseapy, scanpy R: GSVA, fgsea
FA		S	M	G		• Feature Extraction	Python: statsmodels, scikit-learn R: psych, Seurat
LSI		S	L	G	✓	• Clustering • Quantitative Analysis	Python: scanpy, sklearn.decomposition R: Signac

Table 3. Summary of common analytical methods in scRNA-seq data processing and their recommended use cases.

Step	Representative Tools	Best For	Known Limitations
Preprocessing / Normalization	SCTransform, scran, Harmony	Handling zero-inflation, large multi-batch datasets	Overcorrection risk (Harmony), limited spike-in use (SCTransform)
Clustering / Dimensionality Reduction	Leiden, Louvain, UMAP, k-means	Large heterogeneous cell populations	Sensitive to parameter tuning (Leiden), loss of global structure (t-SNE)
Differential Gene Expression	MAST, NEBULA, DESeq2 (pseudo-bulk)	Sparse or large datasets, reproducible replicates	Reduced sensitivity for rare cell types, batch dependence
Integration / Batch Correction	Seurat, Scanpy, scMerge	Multi-sample harmonization	Risk of masking subtle biological signals

Lamian analyzes gene expression changes along pseudotemporal trajectories while accounting for variability between biological replicates. Using generalized additive models, it tests whether dynamic expression patterns differ between conditions, identifying genes with condition-specific temporal behavior for deeper insights into lineage programs [54].

Figure 5 overviews trajectory inference and pseudotime analysis in single-cell transcriptomics.

Pseudotime analysis orders cells along developmental trajectories, revealing gene expression 'waves' at key transitions and highlighting regulatory networks. It can also uncover rare or transient states often missed in bulk analyses [28, 53].

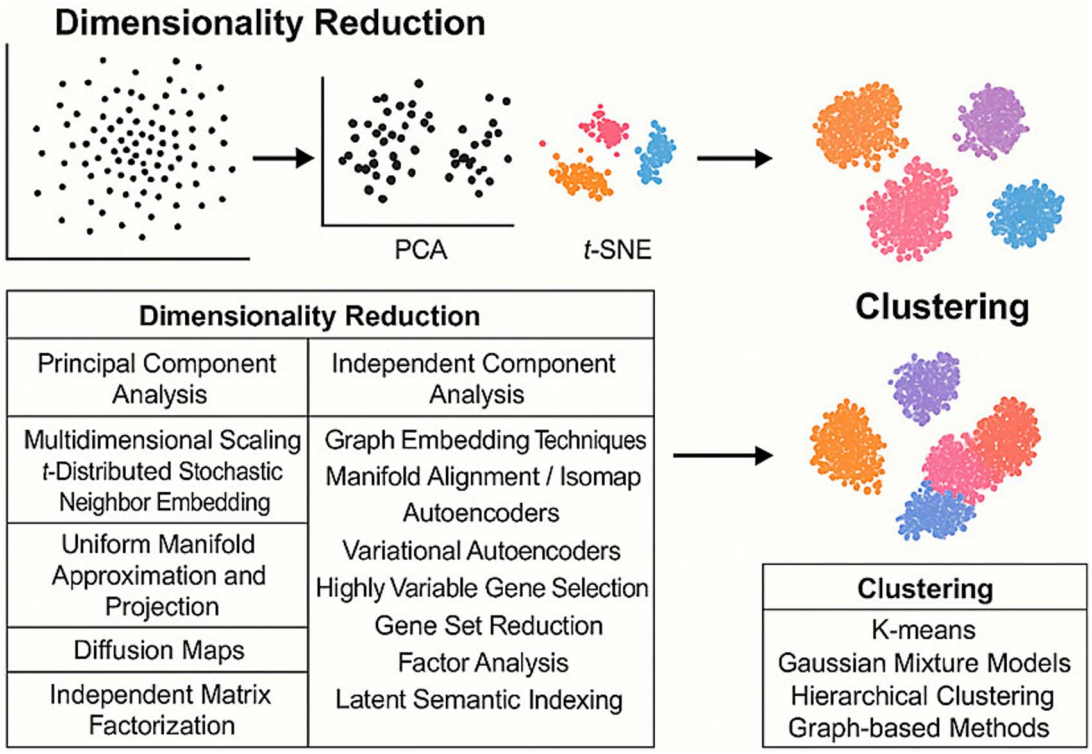

Figure 3. Outlines the pipeline for reducing high-dimensional scRNA-seq data and identifying cell subpopulations. Dimensionality reduction (e.g., PCA, t-SNE, UMAP, diffusion maps, VAEs) projects data into fewer dimensions, followed by clustering methods such as k-means, Gaussian mixtures, hierarchical clustering, or graph-based approaches to define transcriptionally similar populations.

Cell type identification and annotation

scRNA-seq enables high-resolution transcriptional profiling, allowing systematic classification of cell types by molecular signatures [11, 55]. Accurate annotation is essential for understanding tissue composition, function, and developmental trajectories [11].

A key challenge is integrating modalities such as transcriptomics, proteomics, and morphology to fully define cellular identity [55].

Advances in single-cell technologies now allow spatially resolved transcriptomics, revealing morphological diversity and underscoring the need to integrate multiple features for robust cell type classification [54, 55].

Spatial transcriptomics maps gene expression within tissue architecture, revealing cell organization and interactions in their native niche—key to understanding identity and function [56].

Projects like the BRAIN Initiative Cell Census Network and Human Cell Atlas aim to build comprehensive brain cell atlases, integrating spatial and molecular data to map cellular diversity and organization [57].

Cell type identification has uncovered many potential novel types, indicating underappreciated diversity. Integrating transcriptomic, morphological, and connectivity data will be key to building high-resolution atlases and understanding tissue organization [11, 55, 56].

De novo cell type identification clusters cells by gene expression, then annotates clusters using marker genes and reference datasets [31, 57, 60].

Cell type annotation is challenging due to unclear boundaries, particularly for rare populations [55, 56, 58]. Nonetheless, transcriptomic classifications align well with morphological and physiological traits [55, 58], supporting their value for mapping.

Tools like SingleR, CellTree, and other automated methods match expression profiles to reference datasets for cell identity assignment [55, 58, 59]. Integration with spatial profiling can map cell types in tissue; for example, combining SingleR with CellTypist revealed $CD8^+$ T-cell exhaustion markers in severe COVID-19 [60].

Table 4 showcases applications of scRNA-seq annotation tools like Seurat, CellTypist, and scANVI in diverse contexts, from identifying rare $CX3CR1^+$ T cells in COVID-19 to uncovering $SOX9^+$ therapy-resistant breast cancer cells and novel developmental lineages. These examples link molecular insights to clinical outcomes, such as $IL23R^+$ dendritic cells in Crohn's disease and reactive astrocytes in Alzheimer's [66].

Integrating scRNA-seq with modalities like spatial profiling, electrophysiology, connectivity mapping, ATAC-seq, and other omics links molecular, morphological, and functional attributes, enabling high-resolution cell type atlases and deeper insights into regulatory and spatial organization [67–70].

Integration of multi-sample or multi-condition

Advances in scRNA-seq now allow single-cell gene expression profiling, offering deeper insights into cellular heterogeneity than bulk RNA-seq, while presenting new opportunities and challenges for DGE analysis.

Data integration approaches [26, 48] enable cross-sample comparisons in scRNA-seq, but DGE tool performance varies with dataset characteristics such as sparsity and multimodality.

Robust integration of scRNA-seq datasets across samples or platforms requires tools like Seurat, Scanpy, or Harmony to reduce batch effects [11]. However, overcorrection by Harmony can mask biological signals, such as rare tumor clones in glioblastoma [49].

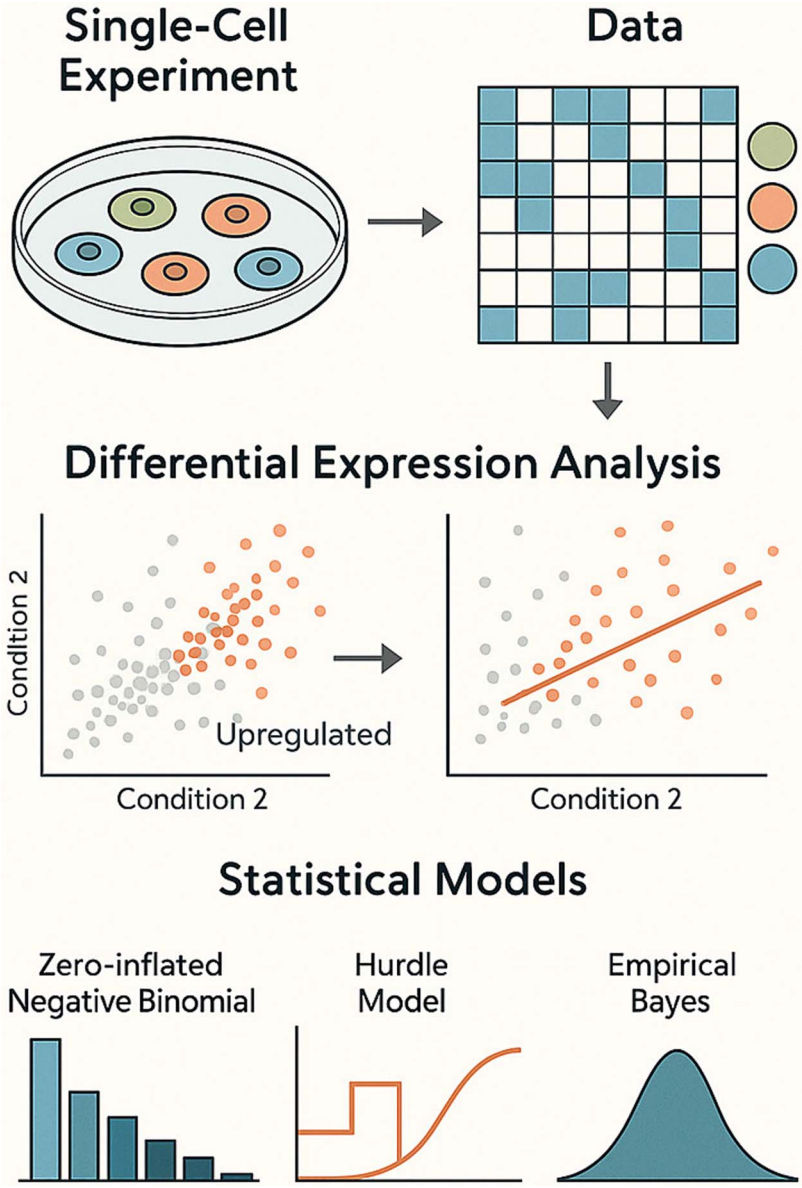

Figure 4. Cells from different conditions are profiled, expression data compiled into a matrix, and statistical models (e.g., zero-inflated negative binomial, hurdle, empirical Bayes) applied to handle sparsity, overdispersion, and dropout, improving DEG detection.

Table 4. Case studies in cell type annotation examples highlight real-world applications of scRNA-seq tools across diverse biological contexts.

Application	Tools Used	Key Finding	Reference
COVID-19 Immune Profiling	Seurat, CellTypist	Identified a rare CD8+ T-cell subset (CX3CR1+) linked to severe outcomes.	Stephenson et al. (2023) [61]
Alzheimer's Disease Brain Atlas	Scanpy, Garnett	Discovered reactive astrocyte subpopulations driving tau pathology.	Mathys et al. (2019) [62]
Breast Cancer Heterogeneity	SingleR, Azimuth	Mapped therapy-resistant SOX9+ luminal progenitor cells in triple-negative tumors.	Wu et al. (2021) [63]
Mouse Embryo Development	scANVI, SCENIC	Resolved a novel Tbx4+ mesenchymal lineage critical for limb formation.	Pijuan-Sala et al. (2020) [64]
Inflammatory Bowel Disease	scPred, Harmony	Linked IL23R+ dendritic cells to mucosal inflammation in Crohn's disease.	Smillie et al. (2021) [65]

Seurat [11] and **Scanpy** are widely used toolkits for single-cell analysis, offering functions for integration, dimensionality reduction, and differential expression.

Addressing technical variation and batch effects is essential in scRNA-seq integration to avoid spurious results [48]. After integration, tools such as MAST, edgeR, or DESeq2 can identify DEGs between conditions [26].

Batch correction methods such as mutual nearest neighbors, factor analysis, and batch-balanced kNN align cells across datasets while preserving biological signals. The multimodal,

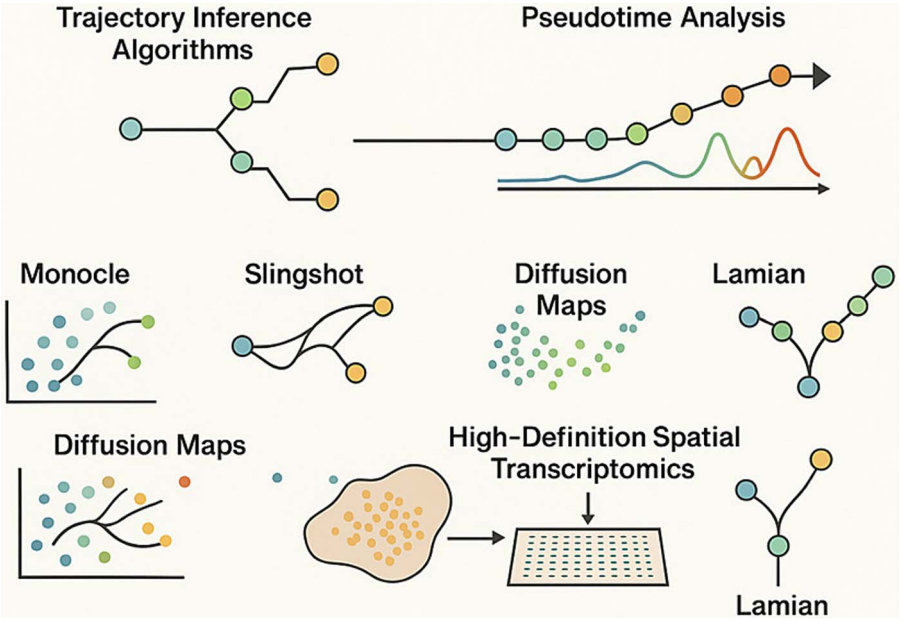

Figure 5. Tools like monocle, slingshot, diffusion maps, and Lamian reconstruct lineage paths and order cells along biological processes, with pseudotime visualizing progression and gene dynamics. Lamian also integrates with spatial transcriptomics to map lineage pathways in tissue.

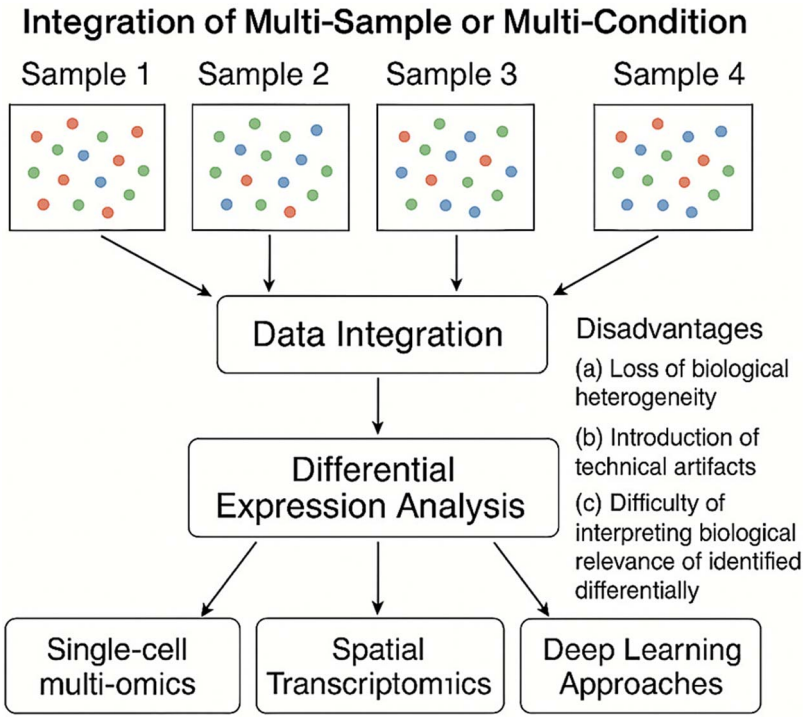

Figure 6. Illustrates integration of single-cell datasets across samples or conditions for unified DGE analysis. While enabling multi-omics, spatial transcriptomics, and deep learning applications, integration risks loss of heterogeneity, technical artifacts, and reduced interpretability. Effective strategies must balance these trade-offs to preserve meaningful variation.

zero-inflated nature of scRNA-seq complicates DGE compared to bulk RNA-seq, motivating specialized tools like MAST, ZINB-WaVE, and BPSC [26].

A indicated in Fig. 6, despite their utility, integration approaches have drawbacks, including loss of biological heterogeneity, introduction of technical artifacts, and difficulty in interpreting DEGs. Tool performance also depends on dataset features such as cell number, sequencing depth, and magnitude of biological differences.

Advances in scRNA-seq have transformed our understanding of cellular heterogeneity, enabling high-resolution insights into tissue organization, development, and rare cell types [11]. Emerging directions include single-cell multi-omics, spatial transcriptomics, and deep learning–based analysis.

Single-cell multi-omics

Integrating modalities such as genomics, epigenomics, transcriptomics, and proteomics at the single-cell level offers a

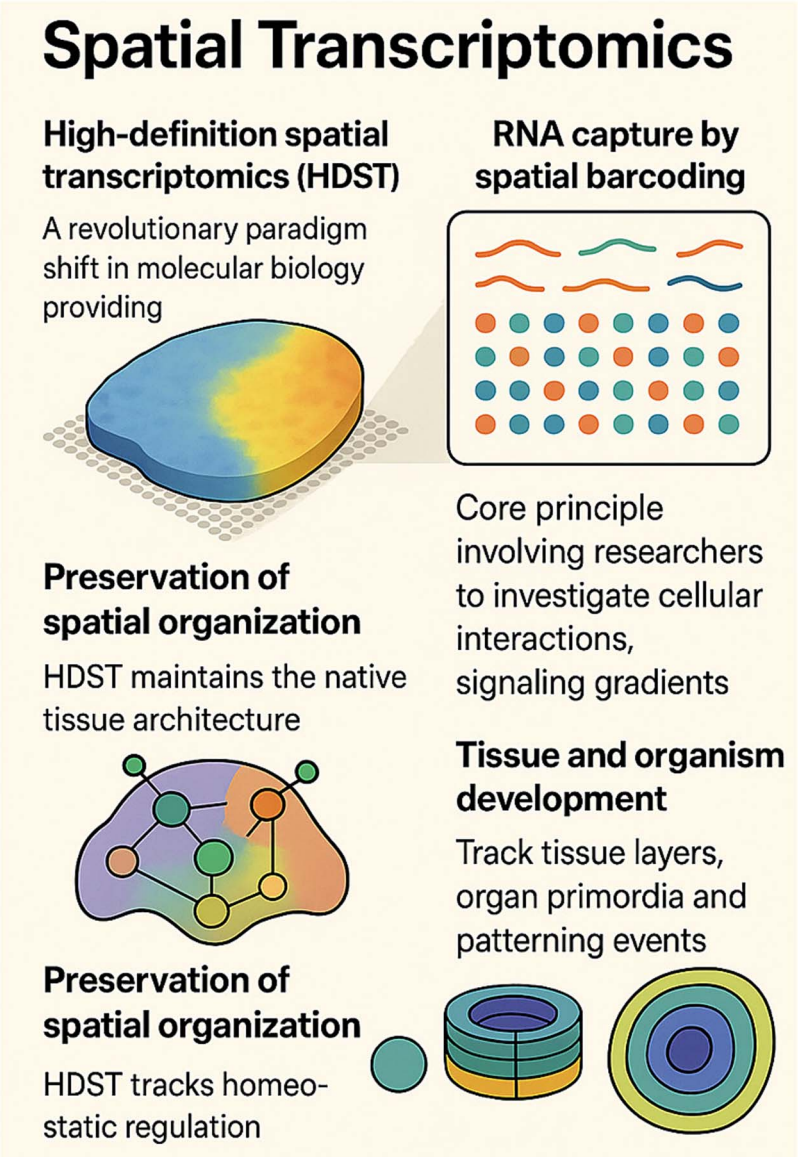

Figure 7. Illustrates the principles and applications of HDST, which uses spatial barcoding to map gene expression onto tissue sections while preserving native organization. This enables investigation of cell interactions, signaling gradients, and developmental processes such as tissue layering and organ primordia with unprecedented resolution.

comprehensive view of cellular states and regulation. Combining these data types reveals links between gene expression, chromatin accessibility, and protein abundance, enabling the discovery of novel cell states and key regulators [11].

Spatial Transcriptomics

High-definition spatial transcriptomics (HDST) enables in situ profiling of gene expression within intact tissue, overcoming the loss of spatial context in bulk and scRNA-seq [71, 72]. By preserving native 2D/3D architecture, HDST reveals the molecular geography and functional organization of tissues [73].

The core principle of HDST involves capturing RNA from tissue sections on barcoded arrays or slides, where each transcript is linked to a spatial coordinate [71, 74]. Advances include Slide-seq (DNA-barcoded beads) [75], HDST (using barcoded beads in micropatterned wells) [53], Visium (commercializing and enhancing the original approach) [76], and multiplexed in situ hybridization methods like MERFISH and seqFISH+ for subcellular resolution [77, 78].

The key advantage of HDST is preserving spatial organization, enabling analysis of how cell interactions, signaling gradients, and tissue architecture shape gene expression [72]. Applications include mapping receptor–ligand interactions, morphogen gradients, and positional influences such as vascular proximity [79, 80].

As shown in Fig. 7, HDST enables spatiotemporal mapping of development, revealing tissue layers, organ primordia, and gene regulatory dynamics with high precision [81, 82]. It also uncovers spatially restricted subpopulations, rare niche-specific cells, and transitional states missed by scRNA-seq [83, 84]. This has major implications for studying the brain, tumor ecosystems (e.g., immune exclusion zones, resistance), and organogenesis [85, 86].

Despite its power, HDST faces challenges such as achieving single-cell resolution over large areas, detecting low-abundance transcripts, integrating spatial multi-omics, and managing

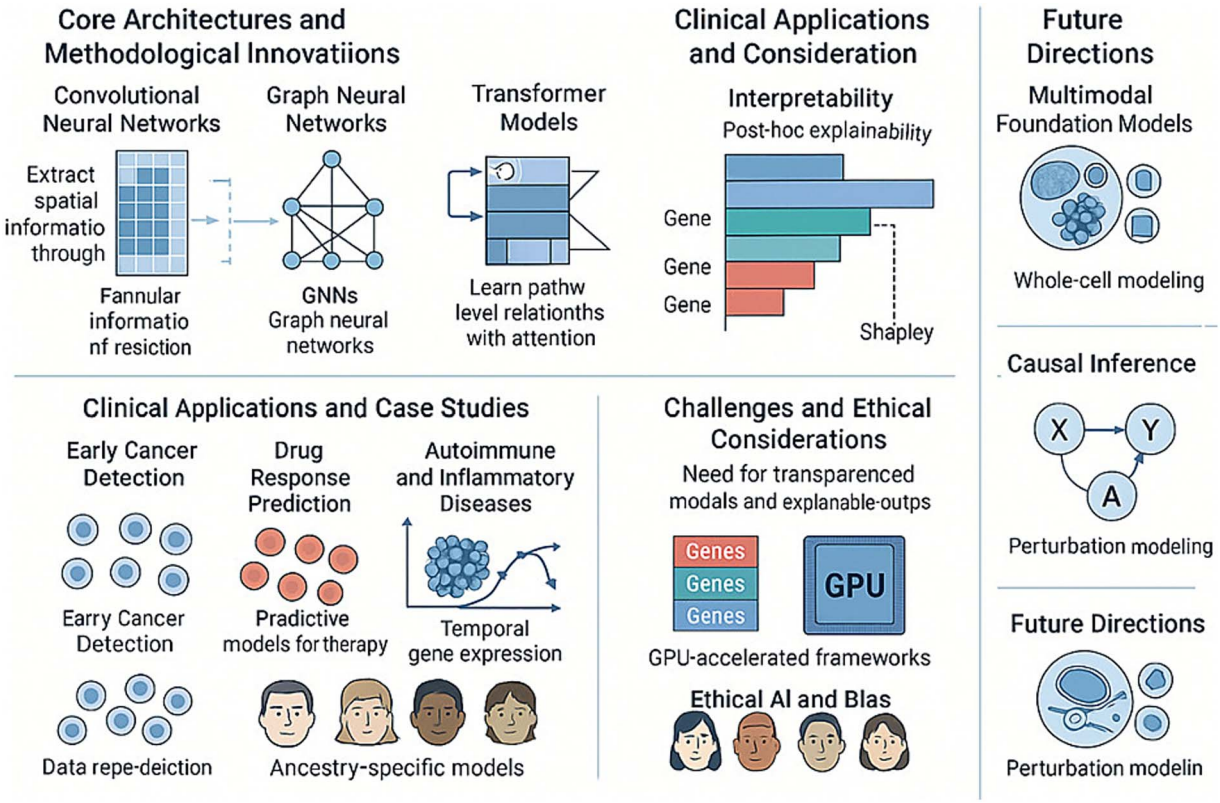

Figure 8. Summarizes key deep learning architectures—CNNs, GNNs, and transformers—Used in biomedicine. These models enable applications such as cancer detection, drug response prediction, and autoimmune disease analysis. Interpretability tools (e.g., Shapley values) and ethical safeguards, including ancestry specific models, are increasingly emphasized. Future directions highlight multimodal foundation models and causal inference frameworks.

computational complexity [87]. Still, it has become an indispensable tool for spatially resolved molecular analysis.

Deep learning approaches

Deep learning (DL) has advanced scRNA-seq analysis by modeling high-dimensional data, resolving cellular heterogeneity, and linking molecular patterns to clinical outcomes. Unlike traditional methods, DL captures nonlinear relationships, hierarchical dependencies, and multimodal interactions. Current frameworks show promise but face ongoing challenges [87] as illustrated in Fig. 8.

Core architectures and methodological innovations

Deep learning (DL) has transformed single-cell data analysis by enabling scalable and nonlinear modeling of complex cellular processes. To improve clarity, we organize this section by the biological objectives addressed by DL frameworks—spatial context reconstruction, perturbation prediction, and multi-omics integration—rather than by network architecture.

Deconvoluting spatial context and cellular heterogeneity with DL

Convolutional neural networks (CNNs)

Originally developed for image analysis, CNNs have been adapted for spatial transcriptomics by applying convolutional filters to gene–cell matrices, detecting local expression patterns [82]. Frameworks like the SpaGCN model tissue neighborhoods in Visium data, identifying spatially variable genes and niche-level interactions (e.g., CX3CR1$^+$ T cells near PD-1$^+$ macrophages in glioblastoma) [88]. CNN-based maps can integrate with GNNs for network reconstruction. Clinically, CNNs have revealed stromal–immune exclusion zones in triple-negative breast cancer, predicting immunotherapy resistance with ~85% accuracy [89], and tools like DeepSpa fuse histology with scRNA-seq to map cell-type proportions, enabling targeted dissection of SOX9$^+$ stem-like cells in colorectal cancer [90].

Graph neural networks (GNNs)

GNNs model cell–cell interaction graphs, making them well-suited for spatial maps or ligand–receptor data [91]. Tools like CellChat use attention-based GNNs to infer signaling, revealing autocrine TGF-β activity in pulmonary fibrosis myofibroblasts suppressed by nintedanib [91]. For lineage reconstruction, PAGA-GNN integrates pseudotime embeddings and outperforms Monocle3, identifying seven developmental branches, including a novel TBX4$^+$ mesenchymal lineage [92] As its clearly indicated in Fig. 8.

Integrative and foundation models for multi-omics

Transformer Models.

Transformers enable large-scale pretraining and cross-modal integration. scGPT, trained on 33 M cells, achieved 94% zero-shot annotation accuracy, identifying rare gastric $\gamma\delta$ T cells and a CX3CR1$^+$ CD8$^+$ subset linked to poor COVID-19 prognosis [93]. GeneFormer integrates scRNA-seq with ATAC-seq via cross-modal

Table 5. Summarizes the Core architectures and methodological innovations.

Architecture	Sub-Category	Tool/Model	Key Features	Clinical/Biological Insights
Convolutional Neural Networks (CNNs)	Spatial Transcriptomics Integration	SpaGCN	Graph-based CNNs model spatial neighborhoods in 10x Visium data.	Identified CX3CR1+ T-cell niches near PD-1+ macrophages in glioblastoma (AUC = 0.91).
	Histology-Transcriptome Fusion	DeepSpa	Correlates H&E histology with scRNA-seq to predict cell-type proportions.	Mapped SOX9+ stem-like cells in metastatic colorectal cancer for targeted laser microdissection.
Graph Neural Networks (GNNs)	Cell–Cell Communication	CellChat	Combines GNNs with attention mechanisms to infer signaling gradients.	Revealed autocrine TGF-β signaling in pulmonary fibrosis myofibroblasts (FDR < 0.05).
	Trajectory Inference	PAGA-GNN	Learns pseudotime-aware embeddings for multi-branch trajectories.	Resolved 7 developmental branches in human embryogenesis, including TBX4+ lineages.
Transformer Models	Foundation Models	scGPT	Pretrained on 33 M cells for zero-shot annotation of rare cell types.	Identified CX3CR1+ CD8+ T-cell subset linked to fatal COVID-19 outcomes (HR = 2.4, p = 0.008).
	Multimodal Learning	GeneFormer	Cross-modal attention bridges scRNA-seq and ATAC-seq (Spearman $\rho = 0.88$).	Linked BIN1 expression to enhancer dysregulation in Alzheimer's astrocytes.
Variational Autoencoders (VAEs)	Batch Harmonization	scVI	Integrates datasets across batches while preserving rare populations.	Isolated pre-metastatic clones (<0.1% prevalence) in pan-cancer studies.
	Artefact Removal	trVAE	Disentangles biological signals from technical noise.	Disentangles biological signals from technical noise.

attention, predicting chromatin accessibility ($\rho = 0.88$) and linking BIN1 dysregulation to astrocytes in Alzheimer's disease [94].

Predicting drug response and perturbations with DL

Variational Autoencoders (VAEs).

VAEs provide harmonized inputs for CNNs, GNNs, and Transformers. scVI integrates >10 batches without losing rare populations, preserving <0.1% pre-metastatic clones in a 250 k-cell pan-cancer study [95]. trVAE separates biological from technical variation, reducing false-positive DEGs by ∼30% versus DESeq2 in Parkinson's disease snRNA-seq [96].

Also, table 5 clearly summarizes the Core Architectures and Methodological Innovations along with Clinical/Biological Insights, Key Features and their Sub-Categories.

Challenges and future directions in DL applications

Despite rapid progress, several challenges remain in applying deep learning to single-cell analysis. Model interpretability is limited, making it difficult to trace biological drivers behind predictions. Computational requirements can hinder reproducibility and scalability. Bias in training datasets may affect model generalization across tissue types and species. Future directions include developing explainable architectures, energy-efficient models, and standardized benchmarking frameworks to ensure robust and transparent DL applications in genomics.

Clinical applications and case studies

When integrated into unified pipelines, deep learning models for scRNA-seq are enabling translational advances. These architectures not only map and integrate cellular data but also support early disease detection, therapy response prediction, and mechanistic insights into complex disorders as mentioned in table 6.

Early cancer detection

CNN–RNN hybrids extend spatial and temporal modeling into diagnostics. NEMO-AI detected circulating tumor cells in blood scRNA-seq with 99.7% specificity, predicting relapse ∼6 months before imaging in breast cancer patients [97]. DeepCycle, using VAE-based deconvolution, analyzed cell-free RNA from 5 mL plasma to predict ovarian cancer tissue of origin with 89% accuracy [98]. These tools demonstrate how research pipelines can be adapted for early screening.

Drug response prediction

Combining GNNs with Transformers links cellular states to therapy outcomes. scDrug predicted venetoclax resistance in AML with AUC = 0.92 by analyzing mitochondrial gene variance [99]. DrugCompass, using reinforcement learning, proposed MEK + CDK4/6 inhibition in BRAF-mutant melanoma, doubling progression-free survival in xenografts (p = 0.03) [100]. Such tools highlight how DL inference can be tuned for precision oncology.

Autoimmune and neurodegenerative diseases

DL frameworks are increasingly applied to immune and neurodegenerative diseases. CellOT aligns patient cells with healthy references, identifying an ISG15$^+$ CD4$^+$ T-cell subset driving interferon signatures in SLE (FDR < 0.01) [101]. scVAE-Time extends VAEs for temporal modeling, revealing APOE4-linked oligodendrocyte dysfunction up to 15 years before Alzheimer's onset [102]. These approaches show how temporal and relational modules can track disease trajectories preclinically.

In oncology, immunology, and neurology, the most impactful DL applications repurpose core architectures from scRNA-seq workflows: VAEs for harmonized inputs, CNNs for spatial features, GNNs for network-level interpretation, and Transformers for multimodal integration. Orchestrated together, these models create end-to-end pipelines that move from raw reads to diagnostics, therapeutic guidance, and longitudinal disease monitoring.

Table 6. Summarizing the clinical applications and case studies.

Clinical Application	Focus Area	Tool/Method	Methodology	Key Clinical Findings
Early Cancer Detection	Circulating Tumor Cells (CTCs)	NEMO-AI	Hybrid CNN-RNN detects CTCs in blood scRNA-seq.	Identified CTCs in breast cancer patients 6 months before relapse ($p = 0.002$; 99.7% specificity).
	Liquid Biopsy Enhancement	DeepCycle	Uses VAEs to deconvolute cell-free RNA.	Predicted ovarian cancer origin from 5 mL plasma (89% accuracy).
Drug Response Prediction	Single-Cell Profiling	scDrug	Combines GNNs and transformers to map transcriptomes to drug sensitivity.	Predicted venetoclax resistance in AML via mitochondrial gene variance (AUC = 0.92).
	Combinatorial Therapy Design	DrugCompass	Reinforcement learning optimizes drug pairs.	Proposed MEK + CDK4/6 inhibition for BRAF-mutant melanoma, doubling PDX survival ($p = 0.03$).
Autoimmune/ Neurodegenerative Diseases	Lupus Subtyping	CellOT	Applies optimal transport theory to match patient cells to healthy references.	Identified ISG15+ CD4+ T cells driving interferon signatures in SLE (FDR < 0.01).
	Alzheimer's Trajectories	scVAE-Time	Models' temporal gene expression shifts.	Linked APOE4 to oligodendrocyte dysfunction 15 years pre-symptom onset.

Challenges and ethical considerations

As DL architectures become integral to scRNA-seq pipelines and clinical applications, their adoption raises technical, interpretability, and ethical challenges that must be resolved to ensure reliability, fairness, and patient safety.

Interpretability

The complexity of DL models can hinder clinical adoption. Explainability tools address this: scGeneShap applies Shapley values to link predictions to genes, identifying TP53 mutations as cluster drivers and influencing treatment in 12% of leukemia patients [103]. CellCAN uses counterfactuals to model interventions, such as predicting KRAS inhibition effects on pancreatic cancer cell fates [104]. These methods make black-box models more transparent and clinically actionable.

Scalability and infrastructure

DL in single-cell analysis carries heavy computational costs. Scanpy-NVTabular can process 10 M cells in 12 minutes on A100 GPUs but demands cloud budgets >$20 k for large studies [105]. SCALE-FL mitigates scalability and privacy issues by enabling federated training; in a 30-center glioblastoma study, it analyzed 1.2 M cells while keeping patient data local [106].

Ethical AI and bias

Model bias and privacy risks remain major hurdles in clinical DL. scDiverse reduces Eurocentric bias by pretraining on diverse donor cells, boosting melanoma risk prediction in non-European cohorts by 22%. Yet privacy threats persist: a 2024 study showed GANs could re-identify patients from just 100 cells by exploiting HLA patterns, highlighting the need for differential privacy safeguards [107].

As summarized in table 7, addressing these challenges requires treating interpretability, computational feasibility, and fairness as core design principles rather than optional features. Future pipelines will likely embed explainability modules alongside predictive layers, incorporate privacy-preserving training strategies, and adopt standardized benchmarks that reflect global genetic diversity. Such efforts are essential to ensure that DL technologies driving early detection and precision therapy are deployed responsibly, equitably, and securely.

Future directions

The next generation of DL in scRNA-seq will advance beyond isolated tasks, evolving into integrated and interpretable systems designed for clinical accessibility. These pipelines will unify multiomic datasets, incorporate causal modeling, and streamline computational demands, lowering barriers to adoption while enhancing translational impact.

Multimodal foundation models

Whole-cell modeling frameworks like CellXGPT integrate scRNA-seq, proteomics, and live-cell imaging into unified foundation models [94]. These systems can predict drug effects from morphological changes, linking molecular profiles with live-cell phenotyping and enabling in silico simulation of therapeutic interventions before wet-lab validation.

Causal inference

To move from correlation to mechanism, DL models are incorporating causal reasoning. CINEMA-OT combines optimal transport with causal graphs to predict CRISPR knockout effects, identifying STAT1 as a key regulator of T-cell exhaustion [108]. Such approaches mark progress toward automated hypothesis generation, where models not only describe patterns but also propose interventions likely to achieve desired outcomes.

Clinician-friendly tools

Real-world impact requires DL tools accessible to non-experts. No-code platforms like BioJupyter enable drag-and-drop model training, allowing pathologists to classify tumor subtypes without coding [109]. Such interfaces can close the translational gap and support routine clinical use of single-cell analytics.

As summarized in table 8, future DL–scRNA-seq pipelines will function as integrated systems, unifying multimodal data, causal reasoning, and user-friendly outputs. Embedding interpretability, scalability, and fairness will be key to maximizing discovery and ensuring clinical trust.

Table 7. Summarizing challenges and ethical considerations.

Challenge Category	Focus Area	Tool/Method	Approach	Key Insights
Interpretability	Post-Hoc Explainability	scGeneShap	Uses Shapley values to attribute model predictions to genes.	Identified TP53 mutations driving ambiguous leukemia clusters, altering treatment for 12% of patients.
	Counterfactual Analysis	CellCAN	Generates 'what-if' scenarios to predict cellular responses.	Models' effects of KRAS inhibition on pancreatic cancer cell fates.
Scalability & Infrastructure	GPU-Accelerated Frameworks	Scanpy-NVTabular	Processes 10 M cells in 12 minutes on NVIDIA A100 GPUs.	High computational costs (~$20 k + cloud credits) limit accessibility.
	Federated Learning	SCALE-FL	Privacy-preserving analysis across institutions without raw data sharing.	Trained models on 1.2 M glioblastoma cells across 30 hospitals.
Ethical AI & Bias	Ancestry-Specific Models	scDiverse	Pretrains on diverse ancestry data (African, Asian, Hispanic).	Improved melanoma risk prediction in non-European populations by 22%.
	Re-Identification Risks	GANs (hypothetical)	Reconstructs patient identities from 100-cell subsets using HLA gene patterns.	Highlights need for differential privacy safeguards in public datasets.

Table 8. Summarizing the future directions.

Future Direction	Focus Area	Tool/Method	Key Features	Translational Impact
Multimodal Foundation Models	Whole-Cell Modeling	CellXGPT	Unifies scRNA-seq, proteomics, and live-cell imaging into a single transformer.	Predicts drug effects from morphological changes, accelerating multi-omics drug discovery.
Causal Inference	Perturbation Modeling	CINEMA-OT	Combines optimal transport with causal graphs to predict CRISPR knockout outcomes.	Prioritized STAT1 as a master regulator of T-cell exhaustion, guiding immunotherapy targets.
Clinician-Friendly Tools	No-Code Platforms	BioJupyter	Drag-and-drop DL model training for tumor subtype classification.	Democratizes AI for pathologists without coding expertise, improving diagnostic workflows.

Figure 8 summarizes key deep learning architectures—CNNs, GNNs, and transformers—used in biomedicine. These models enable applications such as cancer detection, drug response prediction, and autoimmune disease analysis. Interpretability tools (e.g., Shapley values) and ethical safeguards, including ancestry-specific models, are increasingly emphasized. Future directions highlight multimodal foundation models and causal inference frameworks.

Transformers and foundation models

Transformer models such as scGPT [60] enable cross-modal integration of scRNA-seq and ATAC-seq, predicting chromatin accessibility from transcriptomes with >90% accuracy in PBMCs. Incorporating these techniques enhances scRNA-seq analysis by providing deeper insights into cellular heterogeneity, tissue organization, and underlying regulatory mechanisms.

scRNA-seq analysis faces unique challenges, including the stochastic nature of gene expression. Variability within seemingly homogeneous cells is far greater than in bulk RNA-seq, with this 'noise' often obscuring true biological differences [110]. Methods developed for bulk RNA-seq may not capture this variability, highlighting the need for specialized tools [26]. Another hurdle is the high frequency of zero read counts (dropouts), which require careful modeling to avoid biased inferences in differential expression analysis.

Another challenge of scRNA-seq is its multimodal nature, as cell populations often contain distinct subpopulations or states with non-normal expression profiles. Capturing this complexity requires more advanced clustering and dimensionality reduction methods than those used for bulk RNA-seq [11].

A major challenge in scRNA-seq is data sparsity and dropout, where many genes show no detectable expression in individual cells due to technical limits. This strongly impacts the performance of differential expression methods originally developed for bulk RNA-seq [26].

Another challenge is the sheer scale of scRNA-seq datasets, which can include millions of cells and demand substantial computational resources and efficient algorithms. As shown in Fig. 9, interpretability is also limited by the data's high dimensionality, making it difficult to extract clear biological insights. Integrating orthogonal data, such as spatial or proteomic information, can provide a more comprehensive view [11, 26, 111, 112].

In summary, the variability, sparsity, and multimodality of scRNA-seq data demand specialized analytical methods. Advancing the field will require robust benchmarking, ethical AI frameworks, and clinician-friendly tools to move single-cell insights beyond atlas-building toward actionable diagnostics.

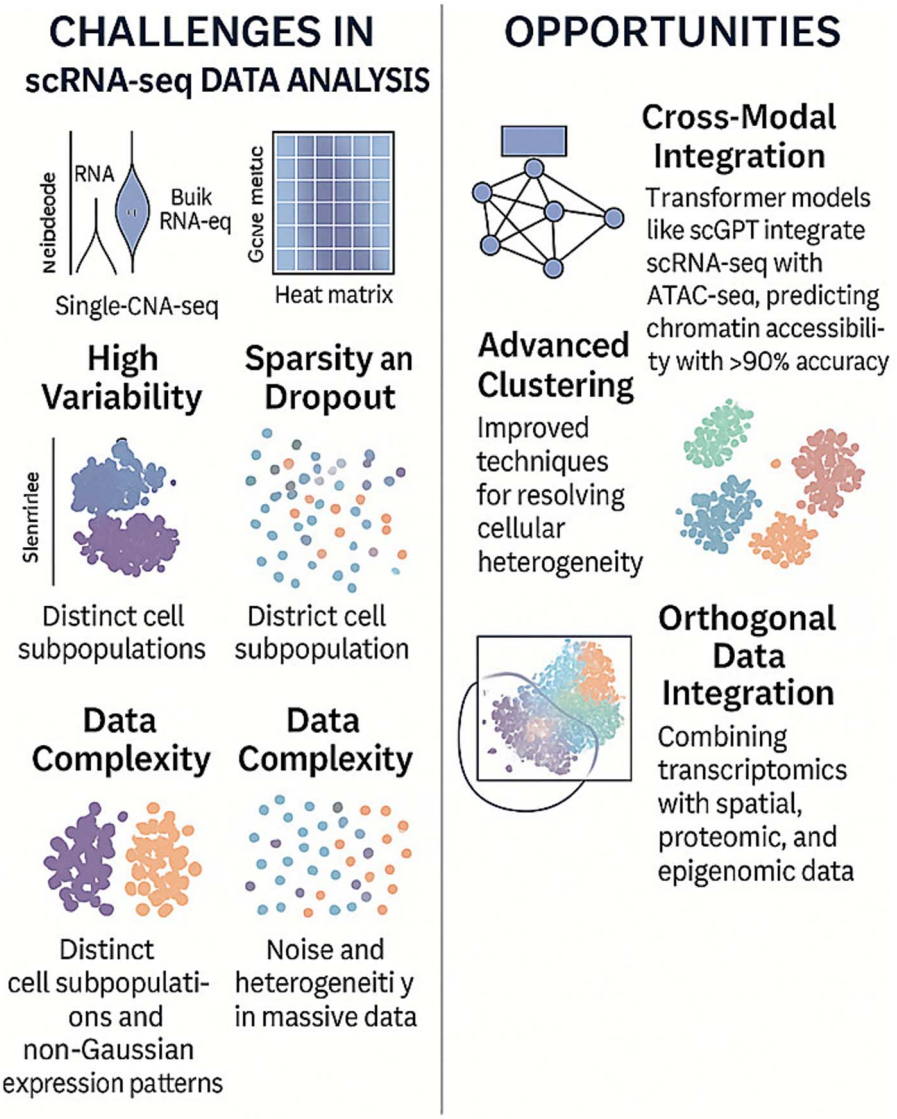

Figure 9. Illustrates the challenges and opportunities in scRNA-seq analysis. Major obstacles include variability, sparsity, dropout events, and multimodality, which hinder interpretation. Emerging solutions— Such as transformer-based cross-modal integration, advanced clustering for cellular heterogeneity, and spatial transcriptomics for contextual insight—Underscore the need for specialized computational strategies to fully exploit scRNA-seq data.

Future directions and emerging trends

scRNA-seq has rapidly advanced, enabling detailed transcriptional profiling of individual cells and opening avenues for studying heterogeneity, lineage tracing, and rare populations [113]. Looking ahead, key challenges and opportunities remain as summarized in Fig. 10. Foundation models (e.g., GeneFormer) may unify single-cell atlases across tissues, while causal inference frameworks (e.g., CELLEX) aim to separate technical artifacts from disease-driving gene programs [60, 114].

A major challenge in scRNA-seq is managing the vast datasets, often containing millions of cells with intrinsic issues like sparsity and batch effects [112]. Overcoming these hurdles demands advanced bioinformatics pipelines capable of efficient processing, normalization, and analysis [111].

Beyond RNA-seq, single-cell technologies like DNA sequencing and spatial transcriptomics provide deeper insight into cellular function and organization [11, 112]. Integrating these multi-omics datasets with existing knowledge bases presents a key challenge, requiring novel computational approaches to fully harness their potential.

The rapid growth of single-cell sequencing has produced a vast array of computational tools and pipelines, often overwhelming for researchers. User-friendly platforms like scOrange help bridge this gap by simplifying analysis and interpretation [112]. At the same time, integrating AI and machine learning offers powerful opportunities to extract phenotypic features, refine disease subtyping, and advance precision medicine.

scRNA-seq holds great promise for personalized medicine by enabling the detection of rare cell populations, disease subtypes, and novel biomarkers. Integration with complementary omics, such as single-cell DNA sequencing and spatial transcriptomics, provides a more holistic view of cellular function and disease mechanisms, paving the way for targeted therapies [111].

Despite ongoing challenges, advances in computational tools, multi-omics integration, and AI applications are poised to expand the role of scRNA-seq in precision healthcare. A concise summary

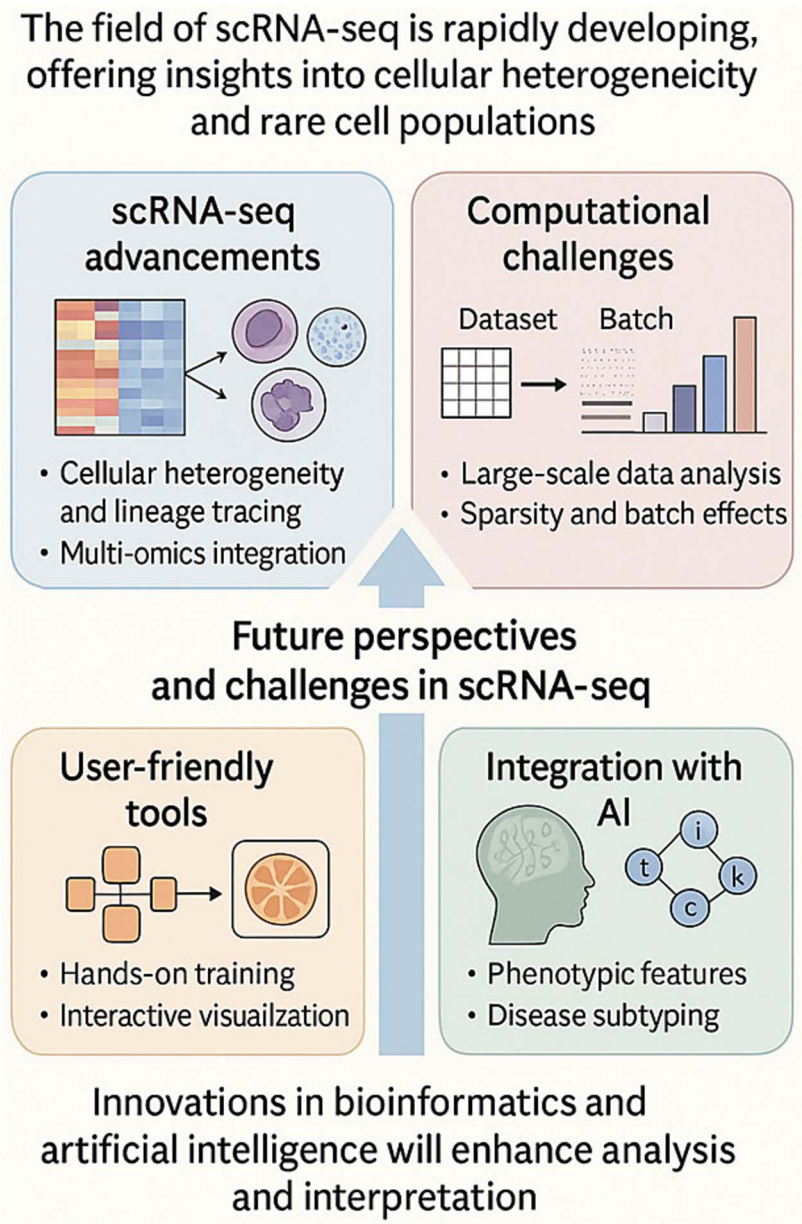

Figure 10. Illustrates the summary of current advancements and future perspectives in scRNA-seq. The technology enables detailed profiling of cellular heterogeneity, lineage tracing, and multi-omics integration, while also posing computational challenges such as large-scale data handling, sparsity, and batch effects. Future directions emphasize the need for user-friendly training and visualization tools, along with artificial intelligence–driven approaches for phenotypic characterization and disease subtyping. Together, these advances will enhance the interpretability and clinical impact of scRNA-seq data.

of open problems—emphasizing integrative inference and multi-omics settings—is provided in Table 9 to guide method choice and study design.

Conclusion

scRNA-seq has become a powerful approach for dissecting cellular heterogeneity at the transcriptomic level [27]. It provides unprecedented insights into cellular states, tissue organization, and developmental processes [111, 113]. Yet, its analysis poses distinct challenges, demanding specialized computational strategies.

A major challenge in scRNA-seq analysis is managing sparse, high-dimensional data [11]. Datasets typically include thousands of genes with many zero counts, limiting traditional DGE methods. To overcome this, specialized algorithms have been developed that explicitly model zero-inflation and multimodal expression patterns.

Collaboration between biologists, computational scientists, and statisticians is essential for advancing scRNA-seq analysis [111]. Despite progress, challenges remain, including integrating scRNA-seq with other modalities (e.g., epigenomics, spatial data) to capture cellular context, and improving scalability to manage increasingly large and complex datasets.

Key Points

- Single-cell RNA sequencing (scRNA-seq) reveals cellular heterogeneity and has transformed fields such as oncology, immunology, and neuroscience.

Table 9. Open problems and challenges in scRNA-seq analysis with emphasis on integrative inference and multi-omics settings. Representative impacts and promising directions are summarized to guide method selection and study design.

Category	Specific challenge	Why it matters (impact)	Typical failure modes	Representative methods today	Promising directions
Data sparsity & dropouts	Excess zeros distort signal in rare cell types	Loss of DE power, unstable embeddings	Over-smoothing; inflated false negatives	SCTransform, ZINB-WaVE, imputation variants	Generative models with uncertainty; count-aware losses
Batch effects & integration	Multi-lab/platform heterogeneity	Artifactual clusters; poor reproducibility	Over-correction masking biology; under-correction	Harmony, scMerge, Seurat/CCA, BBKNN	Causal integration; domain adaptation with controls
Cross-modal alignment	RNA–ATAC–protein matching at single-cell level	Misaligned cell states across omes	Hallucinated correspondences; modality bias	MultiVI/scVI, totalVI, MOFA+, Seurat v5	Foundation models for joint embeddings; contrastive objectives
Scalability	10^5–10^7 cells, multi-sample cohorts	Memory/time blow-ups; parameter brittleness	Subsampling artifacts; non-determinism	Scanpy/AnnData, approximate KNN, GPU UMAP	Streaming/online training; distributed graphs; quantization
Annotation transfer	Consistent labels across datasets/species	Confused subtypes; poor clinical portability	Overfitting to atlas; label drift	Azimuth/Seurat maps, scNym, CellTypist	Zero-shot & few-shot transfer; ontology-aware labeling
Trajectory & dynamics	Robust pseudotime/state transitions	Misordered lineages; missed branches	Sensitivity to DR; batch-driven paths	Monocle3, Slingshot, scVelo/dynamo	Hybrid RNA-velocity + lineage tracing; causal TI
DGE & pseudo-bulk design	Replicate-aware testing at scale	Inflated type-I errors; low power	Cell-level models ignoring replicates	Pseudo-bulk + DESeq2/edgeR, NEBULA, MAST	Hierarchical models with priors; permutation-stable tests
Spatial integration	Align spatial maps with scRNA profiles	Misplaced cell types; blurred niches	Over-deconvolution; tissue-edge bias	Tangram, SpaGCN, DestVI	Multi-scale spatial priors; graph foundations
Confounding & causality	Distinguish technical versus biological signal	Wrong conclusions for interventions	Collider/hidden-confound bias	Residualization, design covariates	Causal graphs; invariant risk minimization
Model interpretability	Explain DL predictions biologically	Black-box resistance in clinic	Spurious features; irreproducible saliency	SHAP/Integrated Gradients, gene-set tests	Mechanistic priors; concept-based explanations
Fairness & privacy	Bias across cohorts; PHI leakage	Unequal performance; re-ID risk	Performance gaps; leakage from embeddings	Federated learning, DP-SGD	Privacy-preserving FMs; bias audits with standards
Benchmarking & standards	Comparable, reproducible pipelines	Conflicting results across papers	Hidden parameters; moving targets	Open atlases; curated workflows	Continual benchmarks; provenance & versioning

- Computational challenges—including data sparsity, batch effects, and lack of standard benchmarks—remain major barriers to clinical translation.
- Emerging tools like SCTransform, Harmony, scGPT, and CellTypist offer scalable, high-accuracy solutions for preprocessing, integration, and annotation.
- Integration of scRNA-seq with spatial transcriptomics and single-cell multi-omics enhances understanding of disease microenvironments and therapy resistance.
- Ethical concerns, including re-identification risks and model bias, require robust solutions such as federated learning and ancestry-aware AI models.

Author contributions

A.M.N, and H.M.G. designed the study. A.M.N. and H.M.G. performed the analyses. A.M.N, and H.M.G. reconstructed the networks. A.M.N, H.M.G, and S.Gh interpreted the results. A.M.N wrote draft of the manuscript. All authors reviewed the manuscript.

Conflict of interest: None declared.

Funding

None declared.

Data availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Haque A, Engel J, Teichmann SA. *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**:75.

2. Kolodziejczyk Aleksandra A, Kim JK, Svensson V. *et al.* The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015;**58**:610–20. https://doi.org/10.1016/j.molcel.2015.04.005
3. Lawson DA, Bhakta NR, Kessenbrock K. *et al.* Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* 2015;**526**:131–5. https://doi.org/10.1038/nature15260
4. Macosko Evan Z, Basu A, Satija R. *et al.* Highly parallel genome-wide expression profiling of individual cells using Nanoliter droplets. *Cell* 2015;**161**:1202–14. https://doi.org/10.1016/j.cell.2015.05.002
5. Muers M. Disentangling DNA methylation. *Nat Rev Genet* 2013;**14**:519–9. https://doi.org/10.1038/nrg3535
6. Nowakowski TJ, Bhaduri A, Pollen AA. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* 2017;**358**:1318–23. https://doi.org/10.1126/science.aap8809
7. Saunders A, Macosko EZ, Wysoker A. *et al.* Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 2018;**174**:1015–1030.e1016. https://doi.org/10.1016/j.cell.2018.07.028
8. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet* 2019;**20**:257–72. https://doi.org/10.1038/s41576-019-0093-7
9. Tang F, Barbacioru C, Nordman E. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc* 2010;**5**:516–35. https://doi.org/10.1038/nprot.2009.236
10. Zheng C, Zheng L, Yoo J-K. *et al.* Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* 2017;**169**:1342–1356.e1316. https://doi.org/10.1016/j.cell.2017.05.035
11. Lähnemann D, Köster J, Szczurek E. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**:31.
12. Tran HTN, Ang KS, Chevrier M. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**:12.
13. Clark IC, Fontanez KM, Meltzer RH. *et al.* Microfluidics-free single-cell genomics with templated emulsification. *Nat Biotechnol* 2023;**41**:1557–66. https://doi.org/10.1038/s41587-023-01685-z
14. Pincela Lins PM, Pirlet E, Szymonik M. *et al.* Manufacture of extracellular vesicles derived from mesenchymal stromal cells. *Trends Biotechnol* 2023;**41**:965–81. https://doi.org/10.1016/j.tibtech.2023.01.003
15. Kumar S, Sarmah DT, Asthana S. *et al.* konnect2prot: a web application to explore the protein properties in a functional protein–protein interaction network. *Bioinformatics* 2023;**39**:btac815.
16. Saelens W, Cannoodt R, Todorov H. *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;**37**:547–54. https://doi.org/10.1038/s41587-019-0071-9
17. Squair JW, Gautier M, Kathe C. *et al.* Confronting false discoveries in single-cell differential expression. *Nat Commun* 2021;**12**:5692.
18. Mereu E, Lafzi A, Moutinho C. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol* 2020;**38**:747–55. https://doi.org/10.1038/s41587-020-0469-4
19. Argelaguet R, Cuomo ASE, Stegle O. *et al.* Computational principles and challenges in single-cell data integration. *Nat Biotechnol* 2021;**39**:1202–15. https://doi.org/10.1038/s41587-021-00895-7
20. Marx V. Method of the year: spatially resolved transcriptomics. *Nat Methods* 2021;**18**:9–14. https://doi.org/10.1038/s41592-020-01033-y
21. Wang K, Zhou R, Wu Y. *et al.* RLBind: a deep learning method to predict RNA–ligand binding sites. *Brief Bioinform* 2023;**24**:bbac486.
22. Cui H, Wang C, Maan H. *et al.* scGPT: towards building a foundation model for single-cell multi-omics using generative AI. *bioRxiv* 2023;2023.2004.2030.538439.
23. Wells DK, van Buuren MM, Dang KK. *et al.* Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve Neoantigen prediction. *Cell* 2020;**183**:818–834.e813. https://doi.org/10.1016/j.cell.2020.09.015
24. Stephenson E, Reynolds G, Botting RA. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* 2021;**27**:904–16. https://doi.org/10.1038/s41591-021-01329-2
25. Tian L, Dong X, Freytag S. *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* 2019;**16**:479–87. https://doi.org/10.1038/s41592-019-0425-8
26. Wang T, Li B, Nelson CE. *et al.* Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* 2019;**20**:40.
27. 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
28. Aubin-Frankowski PC, Vert JP. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics* 2020;**36**:4774–80. https://doi.org/10.1093/bioinformatics/btaa576
29. Luecken MD, Büttner M, Chaichoompu K. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods* 2022;**19**:41–50. https://doi.org/10.1038/s41592-021-01336-8
30. Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. *PLoS Comput Biol* 2019;**15**:e1006907. https://doi.org/10.1371/journal.pcbi.1006907
31. Satija R, Farrell JA, Gennert D. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502. https://doi.org/10.1038/nbt.3192
32. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019;**19**:64.
33. Becht E, McInnes L, Healy J. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**:38–44. https://doi.org/10.1038/nbt.4314
34. Han H, Zhang T, Li C. *et al.* Explainable t-SNE for single-cell RNA-seq data analysis. *bioRxiv* **2022**:2022.2001.2012.476084.
35. Moon I, LoPiccolo J, Baca SC. *et al.* Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. *Nat Med* 2023;**29**:2057–67. https://doi.org/10.1038/s41591-023-02482-6
36. Qi R, Ma A, Ma Q. *et al.* Clustering and classification methods for single-cell RNA-sequencing data. *Brief Bioinform* 2020;**21**:1196–208. https://doi.org/10.1093/bib/bbz062
37. Zhang S, Li X, Lin J. *et al.* Review of single-cell RNA-seq data clustering for cell-type identification and characterization. *Rna* 2023;**29**:517–30. https://doi.org/10.1261/rna.078965.121
38. Peng L, Xiongfei T, Geng T. *et al.* Single-cell RNA-seq clustering: datasets, models, and algorithms. *RNA Biol* 2020;**17**:765–83. https://doi.org/10.1080/15476286.2020.1728961
39. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**:5233.
40. Levine Jacob H, Simonds Erin F, Bendall Sean C. *et al.* Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 2015;**162**:184–97. https://doi.org/10.1016/j.cell.2015.05.047

41. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15.
42. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* 2019;**10**:5416.
43. Chari T, Pachter L. The specious art of single-cell genomics. *PLoS Comput Biol* 2023;**19**:e1011288. https://doi.org/10.1371/journal.pcbi.1011288
44. Singh J, Paliwal K, Zhang T. *et al.* Improved RNA secondary structure and tertiary base-pairing prediction using evolutionary profile, mutational coupling and two-dimensional transfer learning. *Bioinformatics* 2021;**37**:2589–600. https://doi.org/10.1093/bioinformatics/btab165
45. Moon KR, van Dijk D, Wang Z. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**:1482–92. https://doi.org/10.1038/s41587-019-0336-3
46. Amezquita RA, Lun ATL, Becht E. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat Methods* 2020;**17**:137–45. https://doi.org/10.1038/s41592-019-0654-x
47. Jaakkola MK, Seyednasrollah F, Mehmood A. *et al.* Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief Bioinform* 2017;**18**:735–43. https://doi.org/10.1093/bib/bbw057
48. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;**14**:91.
49. Tran C, Horyanto D, Stanley D. *et al.* Correction: Tran et al. antimicrobial properties of bacillus probiotics as animal growth promoters. Antibiotics 2023, 12, 407. *Antibiotics (Basel)* 2023;**12**. https://doi.org/10.3390/antibiotics12091420
50. Finak G, McDavid A, Yajima M. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 2015;**16**:278.
51. Tambonis T, Boareto M, Leite VBP. Differential expression analysis in RNA-seq data using a geometric approach. *J Comput Biol* 2018;**25**:1257–65. https://doi.org/10.1089/cmb.2017.0244
52. Sun S, Zhu J, Ma Y. *et al.* Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* 2019;**20**:269.
53. Vickovic S, Eraslan G, Salmén F. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* 2019;**16**:987–90. https://doi.org/10.1038/s41592-019-0548-y
54. Hou W, Ji Z, Chen Z. *et al.* A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nat Commun* 2023;**14**:7286.
55. Peng H, Xie P, Liu L. *et al.* Morphological diversity of single neurons in molecularly defined cell types. *Nature* 2021;**598**:174–81. https://doi.org/10.1038/s41586-021-03941-1
56. Tian L, Chen F, Macosko EZ. The expanding vistas of spatial transcriptomics. *Nat Biotechnol* 2023;**41**:773–82. https://doi.org/10.1038/s41587-022-01448-2
57. Hawrylycz M, Martone ME, Ascoli GA. *et al.* A guide to the BRAIN initiative cell census network data ecosystem. *PLoS Biol* 2023;**21**:e3002133. https://doi.org/10.1371/journal.pbio.3002133
58. Yao Z, van Velthoven CTJ, Kunst M. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* 2023;**624**:317–32. https://doi.org/10.1038/s41586-023-06812-z
59. Zhang L, Zhang Y, Chen Y. *et al.* TSA-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes. *Genome Res* 2021;**31**:251–64. https://doi.org/10.1101/gr.266239.120
60. Lotfollahi M, Klimovskaia Susmelj A, De Donno C. *et al.* Predicting cellular responses to complex perturbations in high-throughput screens. *Mol Syst Biol* 2023;**19**:e11517.
61. Stephenson T, Pinto Pereira SM, Nugawela MD. *et al.* Long COVID-six months of prospective follow-up of changes in symptom profiles of non-hospitalised children and young people after SARS-CoV-2 testing: a national matched cohort study (the CLoCk) study. *PloS One* 2023;**18**:e0277704. https://doi.org/10.1371/journal.pone.0277704
62. Mathys H, Davila-Velderrain J, Peng Z. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* 2019;**570**:332–7. https://doi.org/10.1038/s41586-019-1195-2
63. Wu SZ, Al-Eryani G, Roden DL. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* 2021;**53**:1334–47. https://doi.org/10.1038/s41588-021-00911-1
64. Pijuan-Sala B, Griffiths JA, Guibentif C. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 2019;**566**:490–5. https://doi.org/10.1038/s41586-019-0933-9
65. Smillie CS, Biton M, Ordovas-Montanes J. *et al.* Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 2019;**178**:714–730.e722. https://doi.org/10.1016/j.cell.2019.06.029
66. Cheng C, Chen W, Jin H. *et al.* A review of single-cell RNA-Seq annotation, integration, and cell-cell communication. *Cells* 2023;**12**. https://doi.org/10.3390/cells12151970
67. Navalón G, Bjarnason A, Griffiths E. *et al.* Environmental signal in the evolutionary diversification of bird skeletons. *Nature* 2022;**611**:306–11. https://doi.org/10.1038/s41586-022-05372-y
68. Chaarani B, Hahn S, Allgaier N. *et al.* Baseline brain function in the preadolescents of the ABCD study. *Nat Neurosci* 2021;**24**:1176–86. https://doi.org/10.1038/s41593-021-00867-9
69. Yao Z, Liu H, Xie F. *et al.* A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* 2021;**598**:103–10. https://doi.org/10.1038/s41586-021-03500-8
70. Nakagawa R, Hirano H, Omura SN. *et al.* Cryo-EM structure of the transposon-associated TnpB enzyme. *Nature* 2023;**616**:390–7. https://doi.org/10.1038/s41586-023-05933-9
71. Ståhl PL, Salmén F, Vickovic S. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* 2016;**353**:78–82. https://doi.org/10.1126/science.aaf2403
72. Moor AE, Itzkovitz S. Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr Opin Biotechnol* 2017;**46**:126–33. https://doi.org/10.1016/j.copbio.2017.02.004
73. Thrane K, Eriksson H, Maaskola J. *et al.* Spatially resolved Transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res* 2018;**78**:5970–9. https://doi.org/10.1158/0008-5472.CAN-18-0747
74. Svensson V, Teichmann SA, Stegle O. SpatialDE: identification of spatially variable genes. *Nat Methods* 2018;**15**:343–6. https://doi.org/10.1038/nmeth.4636
75. Rodriques SG, Stickels RR, Goeva A. *et al.* Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* 2019;**363**:1463–7. https://doi.org/10.1126/science.aaw1219
76. Stickels RR, Murray E, Kumar P. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seqV2. *Nat Biotechnol* 2021;**39**:313–9. https://doi.org/10.1038/s41587-020-0739-1

77. Moffitt JR, Bambah-Mukku D, Eichhorn SW. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* 2018;**362**. https://doi.org/10.1126/science.aau5324
78. Eng CL, Lawson M, Zhu Q. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 2019;**568**:235–9. https://doi.org/10.1038/s41586-019-1049-y
79. Croucher DC, Richards LM, Tsofack SP. *et al.* Longitudinal single-cell analysis of a myeloma mouse model identifies subclonal molecular programs associated with progression. *Nat Commun* 2021;**12**:6322.
80. Chou TH, Tajima N, Romero-Hernandez A. *et al.* Structural basis of functional transitions in mammalian NMDA receptors. *Cell* 2020;**182**:357–371.e313. https://doi.org/10.1016/j.cell.2020.05.052
81. Chen A, Liao S, Cheng M. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;**185**:1777–1792.e1721. https://doi.org/10.1016/j.cell.2022.04.003
82. Lohoff T, Ghazanfar S, Missarova A. *et al.* Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat Biotechnol* 2022;**40**:74–85. https://doi.org/10.1038/s41587-021-01006-2
83. Andersson A, Bergenstråhle J, Asp M. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol* 2020;**3**:565.
84. Moses L, Pachter L. Museum of spatial transcriptomics. *Nat Methods* 2022;**19**:534–46. https://doi.org/10.1038/s41592-022-01409-2
85. Hunter MV, Moncada R, Weiss JM. *et al.* Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nat Commun* 2021;**12**:6278.
86. Ji AL, Rubin AJ, Thrane K. *et al.* Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* 2020;**182**:497–514.e422. https://doi.org/10.1016/j.cell.2020.05.039
87. Longo SK, Guo MG, Ji AL. *et al.* Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat Rev Genet* 2021;**22**:627–44. https://doi.org/10.1038/s41576-021-00370-8
88. Coy S, Wang S, Stopka SA. *et al.* Single cell spatial analysis reveals the topology of immunomodulatory purinergic signaling in glioblastoma. *Nat Commun* 2022;**13**:4814.
89. Wang XQ, Danenberg E, Huang CS. *et al.* Spatial predictors of immunotherapy response in triple-negative breast cancer. *Nature* 2023;**621**:868–76. https://doi.org/10.1038/s41586-023-06498-3
90. Lu T, Park S, Han Y. *et al.* Netie: inferring the evolution of neoantigen–T cell interactions in tumors. *Nat Methods* 2022;**19**:1480–9. https://doi.org/10.1038/s41592-022-01644-7
91. Jin S, Guerrero-Juarez CF, Zhang L. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021;**12**:1088.
92. Wolf FA, Hamey FK, Plass M. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* 2019;**20**:59.
93. Cui H, Wang C, Maan H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80. https://doi.org/10.1038/s41592-024-02201-0
94. Theodoris CV, Xiao L, Chopra A. *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9
95. Lopez R, Regier J, Cole MB. *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2
96. Lotfollahi M, Naghipourfar M, Luecken MD. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022;**40**:121–30. https://doi.org/10.1038/s41587-021-01001-7
97. Schede HH, Natarajan P, Chakraborty AK. *et al.* A model for organization and regulation of nuclear condensates by gene activity. *Nat Commun* 2023;**14**:4152.
98. Cohen JD, Li L, Wang Y. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science* 2018;**359**:926–30. https://doi.org/10.1126/science.aar3247
99. Sun Y-Y, Hsieh C-Y, Wen J-H. *et al.* scDrug+: predicting drug-responses using single-cell transcriptomics and molecular structure. *Biomed Pharmacother* 2024;**177**:117070. https://doi.org/10.1016/j.biopha.2024.117070
100. Vågerö D, Pinger PR, Aronsson V. *et al.* Paternal grandfather's access to food predicts all-cause and cancer mortality in grandsons. *Nat Commun* 2018;**9**:5124.
101. Bunne C, Stark SG, Gut G. *et al.* Learning single-cell perturbation responses using neural optimal transport. *Nat Methods* 2023;**20**:1759–68. https://doi.org/10.1038/s41592-023-01969-x
102. Mathys H, Boix CA, Akay LA. *et al.* Single-cell multiregion dissection of Alzheimer's disease. *Nature* 2024;**632**:858–68. https://doi.org/10.1038/s41586-024-07606-7
103. Morita K, Wang F, Jahn K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat Commun* 2020;**11**:5327.
104. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat Biotechnol* 2024;**42**:927–35. https://doi.org/10.1038/s41587-023-01905-6
105. Kuemmerle LB, Luecken MD, Firsova AB. *et al.* Probe set selection for targeted spatial transcriptomics. *Nat Methods* 2024;**21**:2260–70. https://doi.org/10.1038/s41592-024-02496-z
106. Yurdem B, Kuzlu M, Gullu MK. *et al.* Federated learning: overview, strategies, applications, tools and future directions. *Heliyon* 2024;**10**:e38137. https://doi.org/10.1016/j.heliyon.2024.e38137
107. Louros N, Orlando G, De Vleeschouwer M. *et al.* Structure-based machine-guided mapping of amyloid sequence space reveals uncharted sequence clusters with higher solubilities. *Nat Commun* 2020;**11**:3314.
108. Dong M, Wang B, Wei J. *et al.* Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nat Methods* 2023;**20**:1769–79. https://doi.org/10.1038/s41592-023-02040-5
109. Lin J, Wang S, Audano PA. *et al.* SVision: a deep learning approach to resolve complex structural variants. *Nat Methods* 2022;**19**:1230–3. https://doi.org/10.1038/s41592-022-01609-w
110. Hebenstreit D. Methods, challenges and potentials of single cell RNA-seq. *Biology (Basel)* 2012;**1**:658–67. https://doi.org/10.3390/biology1030658
111. Han N, Kim MA, Lee HS. *et al.* Evaluation of fibroblast growth factor receptor 2 expression, heterogeneity and clinical significance in gastric cancer. *Pathobiology* 2015;**82**:269–79. https://doi.org/10.1159/000441149
112. Stražar M, Žagar L, Kokošar J. *et al.* scOrange-a tool for hands-on training of concepts from single-cell data analytics.

Bioinformatics 2019;**35**:i4–12. https://doi.org/10.1093/bioinformatics/btz348

113. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;**12**:87–98. https://doi.org/10.1038/nrg2934

114. Rao SN, Zahm M, Casemayou A. *et al.* Single-cell RNA sequencing identifies senescence as therapeutic target in rhabdomyolysis-induced acute kidney injury. *Nephrol Dial Transplant* 2024;**39**:496–509. https://doi.org/10.1093/ndt/gfad199

Briefings in Bioinformatics, 2026, **27(1)**, bbaf723
https://doi.org/10.1093/bib/bbaf723
Review