RESEARCH ARTICLE

scFoundation, a deep learning-based single-cell analysis method, is comparable to Scanpy in analyzing scRNA-seq data of liver cirrhosis

Shuya Liu[1], Meijuan Zhang[2, *]

[1]International Division of Shandong Experimental High School, Jinan, Shandong, China. [2]Department of General Practice, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Shandong Engineering Laboratory for Health Management, Jinan, Shandong 250023, China.

Received: March 28, 2025; **accepted:** August 12, 2025.

Single-cell RNA sequencing (scRNA-seq) technology has revolutionized the understanding of cellular heterogeneity by enabling the detailed analysis of individual cell transcriptomes. Traditional analysis methods like Scanpy rely on feature selection and dimensionality reduction, which may introduce bias and overlook subtle biological signals. In contrast, deep learning models like scFoundation provide a data-driven approach, processing the entire gene set to capture complex molecular patterns without manual intervention. This study compared the performance of scFoundation and Scanpy using publicly available liver cirrhosis scRNA-seq data, which profiled over 100,000 single cells from healthy and cirrhotic liver tissues. The results showed that both methods produced comparable cell type annotations with scFoundation showing a slightly higher Adjusted Rand Index (ARI) of 0.977 than that of 0.962 from Scanpy, demonstrating the model's potential to match or surpass classical methods in some cases. The results were additionally confirmed that there were no biases with donor clinical condition or sample ID. The findings of this research suggested that AI-based models like scFoundation could enhance single-cell RNA-seq analysis, offering robust and scalable solutions while uncovering finer biological structures within complex datasets.

Keywords: single-cell RNA sequencing; deep learning; liver cirrhosis; Scanpy.

[*]**Corresponding author:** Meijuan Zhang, Department of General Practice, The First Affiliated Hospital of Shandong First Medical University & Shandong Provincial Qianfoshan Hospital, Shandong Engineering Laboratory for Health Management, Jinan, Shandong 250023, China. Email: xiaohuaquanke@126.com.

Introduction

Single-cell technology has transformed modern biology by enabling the high-resolution analysis of individual cells, thereby revealing heterogeneity hidden in bulk assays [1]. Its applications now extend from plants to human tissues with innovations such as digital polymerase chain reaction (PCR) for single-cell detection and refined microenvironment profiling in oncology [2-4]. Technical advances in precise cell isolation further underpin these gains [5]. The approach has delivered breakthroughs across many disciplines. In cancer research, it uncovers novel biomarkers and lineage trajectories [6], while, in immunology and metabolism, it illuminates emergent cellular properties [7], and in hepatology, it aids early hepatocellular carcinoma diagnosis [8]. Together, these studies underscore the method's

broad utility. The processing and analysis of single-cell data typically involves several key steps including preprocessing, dimension reduction, clustering, annotation, and visualization.

Scanpy is one of the widely used analysis pipelines for this purpose [9]. In Scanpy, feature selection is used to reduce the dimensionality of the data by selecting the most variable genes across cells. This step is essential because working with all genes simultaneously can be computationally expensive and may introduce noise, leading to less accurate downstream analyses. The selection process inherently introduces bias as it prioritizes certain genes over others based on predefined criteria. While this approach simplifies the analysis and focuses on the most informative features, it risks overlooking genes that may be biologically relevant but do not exhibit high variability in the data. This bias can affect downstream results such as clustering or differential expression analysis by potentially missing subtle or rare signals that are crucial for understanding specific biological processes or identifying rare cell types. Deep-learning methods offer a compelling alternative. Convolutional networks can automatically detect patterned single cells in imaging datasets [10], while more general frameworks learn complex, nonlinear gene-expression relationships directly from full transcriptomes [11, 12]. Multimodal architectures integrate multi-omics layers [13], and recent studies highlight how such models are reshaping single-cell analysis [14]. Crucially, models like scVI and the large-scale foundation model scFoundation eliminate manual feature selection by training on complete gene sets [15]. Large-scale pretrained models have revolutionized natural language processing, so is its application in life sciences [16]. To address the challenges imposed by traditional single-cell analysis methods, deep learning models were developed. scFoundation stands as the largest model of its kind as it's trained on over 50 million human single-cell transcriptomics data with more than 100M parameters tuned. Therefore, scFoundation captures complex molecular features across all known cell types and, in theory, should achieve better performance in various downstream tasks.

Liver cirrhosis is a severe condition characterized by extensive fibrosis, leading to significant morbidity and mortality worldwide. This research utilized data from a published study that profiled the transcriptomes of over 100,000 single cells from both healthy and cirrhotic human liver tissue to qualitatively and quantitatively evaluate the performance of the self-supervised learning model scFoundation against the traditional Scanpy pipeline, which required manual input. By comparing cell type annotations generated by scFoundation and the traditional Scanpy pipeline, the performance of both approaches in functional analysis was evaluated to assess how effectively each method identified distinct cellular niches between healthy and diseased liver cirrhosis samples. This research provided deeper insights into the differential cellular environments associated with disease progression.

Materials and methods

Data sources

The single-cell transcriptomic dataset of human liver cirrhosis was downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (Bethesda, MD, USA) under the accession number of GSE136103. A total of about 220 GB of raw FASTQ files and 5.6 GB of processed gene-cell count matrix (HDF5) were retrieved from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc =GSE136103). The matrix contained 101, 370 single cells from ten human liver samples with five healthy donors and five cirrhotic patients covering 33,694 expressed genes. After quality control that the cells demonstrated 100 – 4,000 detected genes, < 5 % mitochondrial reads, and genes detected in ≥ 3 cells, 97,842 cells and 21,317 genes were retained for downstream analysis.

Data preprocessing and quality control

Gene symbols were first annotated against curated dictionaries to identify mitochondrial (MT-), ribosomal (RPL/RPS-) and hemoglobin (HB-) genes, which served as quality-control (QC) markers. For every cell, three QC metrics were computed including the number of detected genes (nGenes), total UMI counts (nCounts) and the percentage of reads mapping to mitochondrial genes (pct-MT). Violin and scatter plots of these metrics guided the empirical thresholds used for filtering. Cells were retained only when $100 \leq nGenes \leq 4,000$ and pct-MT ≤ 5 %, while genes expressed in fewer than three cells were removed, which eliminated low-complexity droplets, potential multiplets, and transcripts dominated by technical noise. The remaining counts were normalized to a library size of 10,000 per cell and log1p-transformed to stabilize variance. Highly variable genes (HVGs) were selected by ranking all genes on mean expression and dispersion, retaining those with $0.0125 \leq mean \leq 3$ and dispersion ≥ 0.5 to provide the most informative feature subset. To minimize technical effects, linear regression was applied to each gene to remove contributions from nCounts and pct-MT, after which residuals were z-scored to unit variance with values exceeding ±10 SD clipped. Principal component analysis (PCA) on the scaled HVG matrix revealed an elbow at the tenth component as PC1-PC5 explained 4.3%, 3.6%, 2.7%, 1.9%, and 1.4% of the total variance, respectively, while the first ten PCs together accounted for 17.8%. These ten PCs were retained for construction of the k-nearest-neighbor graph underlying Leiden clustering and for subsequent uniform manifold approximation and projection (UMAP) visualization.

Selection of highly variable genes and data scaling

Highly variable genes were critical for downstream analysis and were identified based on thresholds for mean expression and dispersion. These genes were visualized and used to refine the dataset for dimensionality reduction. The dataset was then adjusted by regressing out unwanted sources of variation including total counts and mitochondrial gene content followed by scaling to unit variance while clipping extreme values.

Dimensionality reduction and clustering

PCA was performed using Scikit-learn v1.3.0 (https://scikit-learn.org) to reduce dimensionality with the top 40 components selected for neighborhood graph construction. Clustering was carried out using Leidenalg v0.10.2 wrapped around igraph v0.10.8 (https://github.com/vtraag/leidenalg) to identify distinct cell populations. The clustering results were visualized using Umap-learn v0.5.5 (https://github.com/lmcinnes/umap), providing a two-dimensional representation of the data.

Annotation comparison and confusion matrix analysis

To evaluate the clustering results, a confusion matrix was constructed to compare Scanpy-generated clusters using Scanpy v1.10.1 (https://github.com/scverse/scanpy) to pre-existing lineage annotations. This matrix was visualized using heatmaps to assess the correspondence between clusters. Cell types were annotated based on the dominant assignment in each cluster and mapped back to the dataset. Adjusted rand index (ARI) was calculated to quantify the agreement between Scanpy-annotated clusters and original annotations.

Visualization and final annotation

Final UMAP visualizations were created to compare Scanpy-derived cell type annotations with the original lineage annotations. Heatmaps were generated to visualize the confusion matrix, highlighting areas of strong agreement or divergence between methods. This analysis provided a robust framework for benchmarking clustering and annotation accuracy in single-cell data analysis.

Results and discussion

To enable a systematic, like-for-like evaluation,

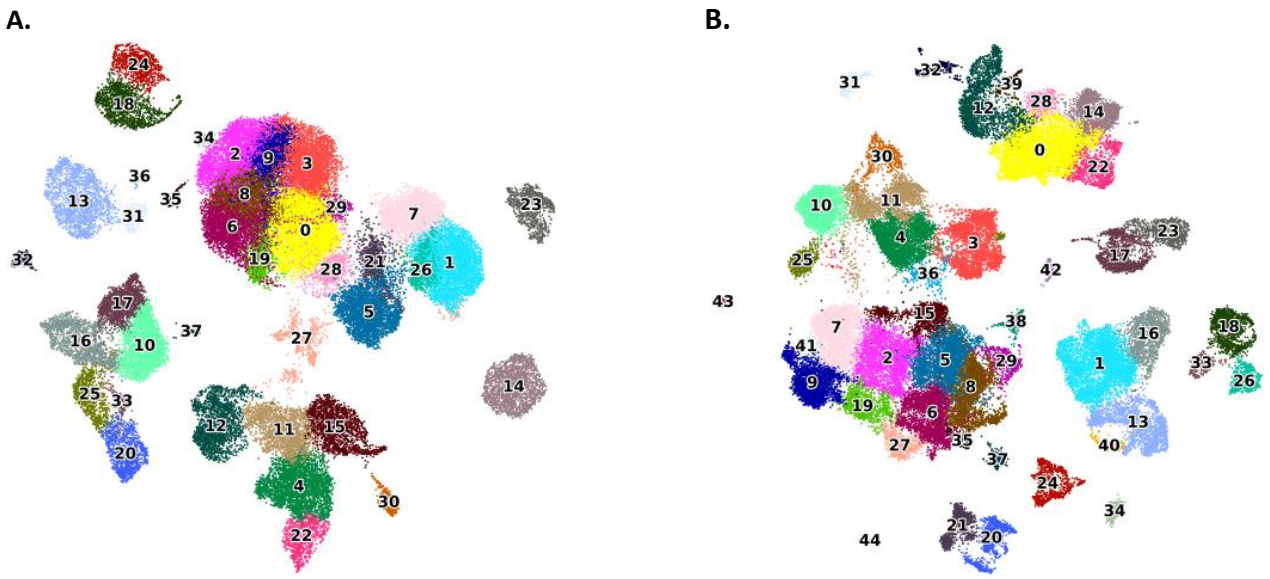

Figure 1. Comparative UMAP visualization of Leiden clustering results. **A.** Scanpy workflow: application of the Leiden algorithm to the PCA-derived embedding delineated 37 transcriptionally distinct clusters, each corresponding to a putative cell type or lineage and forming well-segregated cell-type groupings. **B.** scFoundation workflow: using identical clustering parameters on the scFoundation latent space yielded 43 clusters, revealing additional transcriptional diversity and finer subdivision of cell populations within the same dataset.

results produced with the standard Scanpy pipeline were juxtaposed with those derived from scFoundation embeddings at each stage of the workflow including model-based clustering, marker-guided cluster annotation, and quantitative assessment of the annotated partitions. Identical parameter settings and marker panels were retained across pipelines, so that any divergence in outcome could be ascribed exclusively to the underlying embedding strategy.

Model-based clustering

After applying uniform preprocessed parameters, Leiden clustering was performed on the principal-component manifold generated by Scanpy and on the scFoundation latent space, respectively. The resulting cluster structures were reported together with their dimensionality-reduction maps, providing the baseline on which downstream biological interpretation was built.

(1) Scanpy workflow

By using 10 nearest neighbors and the first 40 principal components, the Leiden algorithm resolved 37 distinct clusters (Figure 1A). The UMAP projection displayed well-separated groups that corresponded to major immune and stromal lineages reported in the source study.

(2) scFoundation workflow

After generating low-dimensional embeddings with the scFoundation API processed in < 1,000-cell batches, an identical Leiden configuration yielded 43 clusters (Figure 1B). The additional clusters represented a finer partitioning of the cellular landscape, suggesting improved sensitivity to subtle transcriptional differences.

Marker-driven cluster annotation

Canonical lineage markers were employed to translate purely computational clusters into biologically interpretable cell types. The same gene panel, detection thresholds and dot-plot visualisation strategy were applied to both embeddings so that any disparity in annotation could be traced exclusively to the upstream clustering step.

(1) Scanpy workflow

Among the 37 Leiden clusters obtained from the

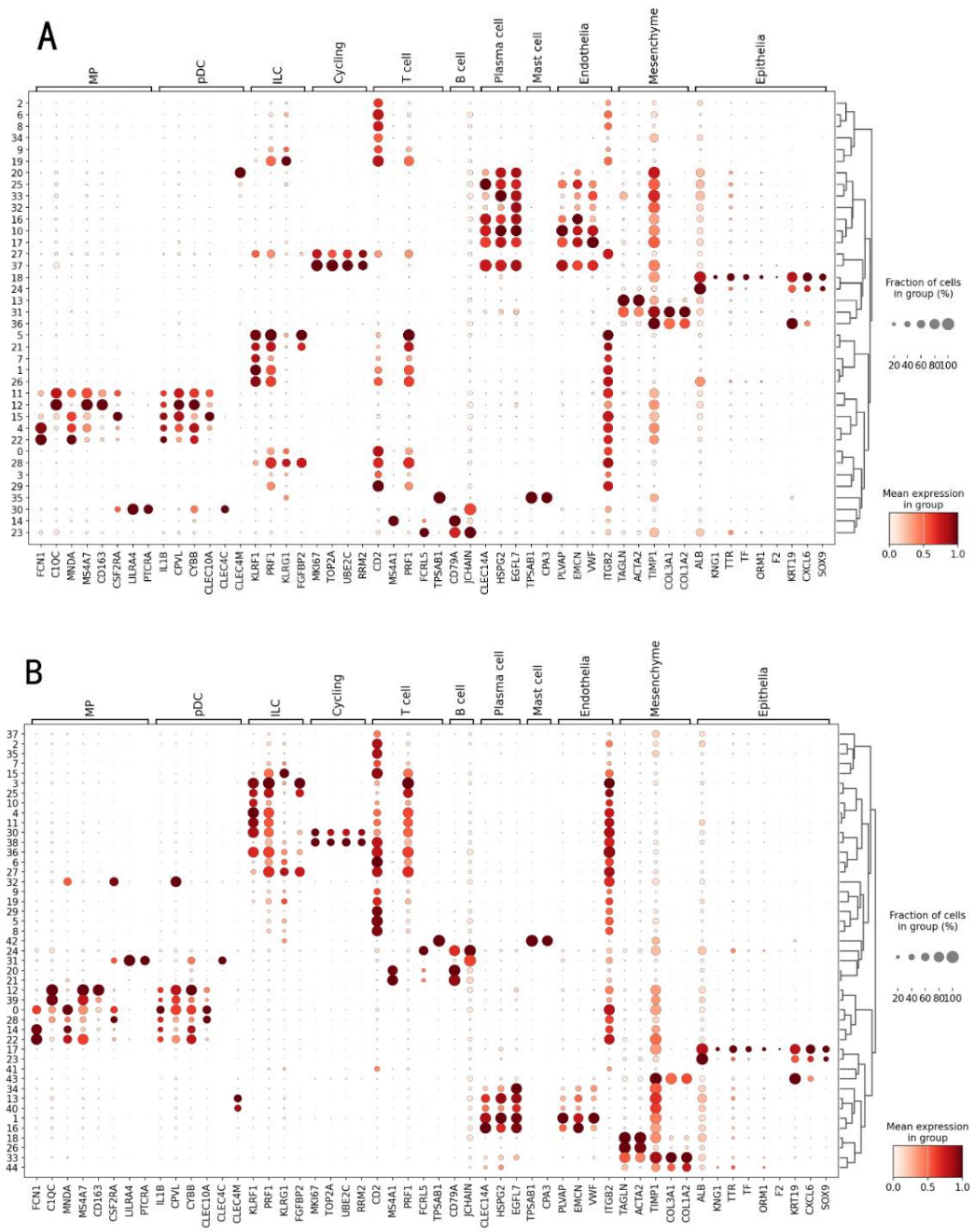

Figure 2. Marker gene landscape of cell clusters. **A.** Expression profiles of canonical marker genes across clusters. The horizontal axis denotes marker genes, and the vertical axis represents cluster IDs. Dual encoding by dot size (fraction of cells expressing the gene) and color intensity (mean expression level) revealed molecular heterogeneity among clusters. **B.** Cell-type annotation based on marker gene signatures. Clusters were assigned to specific cell types by matching expression patterns against established cellular signatures. Dot size and color intensity followed the same encoding rules, collectively providing visual evidence for cell identity assignment.

Scanpy principal-component manifold, 11 major lineages were identified. Macrophage identity was assigned to clusters 4, 11, 12, 15, and 22 because all five groups displayed robust MNDA, CD163, and MS4A7 expression. Of these, cluster 12 showed the strongest MS4A7 and CD163 signals, pointing to a tissue-resident phenotype. Cluster 30 was annotated as plasmacytoid dendritic cells owing to its pronounced CLEC4C and LILRA4 expression. Innate lymphoid cells

were recognized in clusters 1 and 5 on the basis of KLRF1 enrichment, whereas cluster 27 consisted of actively cycling cells marked by high MKI67 and TOP2A. The T-cell compartment comprised several CD2-positive clusters that further segregated into helper and cytotoxic subsets according to CD3D co-expression with CD4 or CD8A. Cluster 13, characterized by elevated CD79A, represented the B-cell lineage, while plasma cells localized to clusters 10, 16, and 17, distinguished by strong EGFL7 and immunoglobulin transcripts. Mast cells were confined to cluster 35, which expressed TPSAB1 and CPA3, and epithelial cells were concentrated in cluster 18, marked by KRT19 and SOX9. Hierarchical clustering of both genes and clusters reinforced these assignments that macrophage clusters formed a tight branch, which was clearly segregated from endothelial and epithelial signatures, yielding a coherent overview of lineage distribution in the Scanpy embedding (Figure 2A).

(2) scFoundation workflow

Applying the identical marker panel to the 43 Leiden clusters derived from the scFoundation latent space not only reproduced every major lineage detected with Scanpy but also exposed additional transcriptional states. Macrophage diversity was partitioned into 8 clusters. Tissue-resident macrophages appeared in clusters 9, 24, and 33, whereas inflammatory macrophages with elevated IL1B, NLRP3, and SPP1 dominated clusters 6 and 19. Two small clusters of 31 and 37 selectively expressed MHC-II genes such as HLA-DRA and HLA-DRB1, suggesting antigen-presenting macrophages, and cluster 41 co-expressed APOE and LGALS3, consistent with a lipid-handling programme. Innate lymphoid cells separated into three transcriptionally distinct groups including cluster 2 showing TBX21 and IFNG expression characteristic of ILC1, cluster 17 carrying the RORC, IL23R, and AHR signature of ILC3, and cluster 29 retaining high KLRF1 with low cytokine levels, corresponding to a quiescent ILC pool. Epithelial heterogeneity was similarly refined. Basal epithelial cells in cluster 28 displayed KRT5 and TP63, secretory epithelial cells in cluster 38 up-regulated MUC1, KRT8, and EPCAM, and a rare proliferative subset in cluster 42 co-expressed KRT19 with MKI67 and TOP2A, revealing an actively cycling epithelial compartment that had not been resolved in the Scanpy analysis. The T-cell compartment expanded to 12 clusters including a CCR7-high naïve subset (cluster 4) and a CD8A-positive, GZMB-positive cytotoxic subset (cluster 12). B cells were found in clusters 5 and 20, while three plasma-cell clusters of 11, 26, and 36 exhibited a gradient of XBP1 and PRDM1 expression indicative of progressive plasmablast maturation. A plasmacytoid dendritic-cell cluster (cluster 23), a mast-cell cluster (cluster 35), and an endothelial cluster (cluster 30), which expressed PECAM1 and VWF, were recovered with high fidelity (Figure 2B). Hierarchical clustering of the gene-by-cluster matrix corroborated these stratifications, where inflammatory and resident macrophage clusters segregated from one another, basal and secretory epithelial clusters formed separate branches, and naïve versus effector T-cell subsets were clearly resolved. Thus, the marker-guided annotation based on scFoundation embeddings not only recapitulated all principal cell lineages but also revealed previously hidden transcriptional sub-states, underscoring the superior resolving power of the learned representation.

Quantitative assessment of annotated partitions

The accuracy of the annotated cluster solutions was evaluated against the published ground-truth labels through both visual inspection of overlapping heat-maps and calculation of the adjusted rand index. These metrics quantified the degree to which each workflow reproduced established cell-type boundaries, thereby providing an objective measure of biological fidelity.

(1) Scanpy workflow

The UMAP annotated with Scanpy labels (Figure 3A-a) was visually concordant with the reference map reported in the original publication (Figure 3A-b). Major lineages occupied comparable

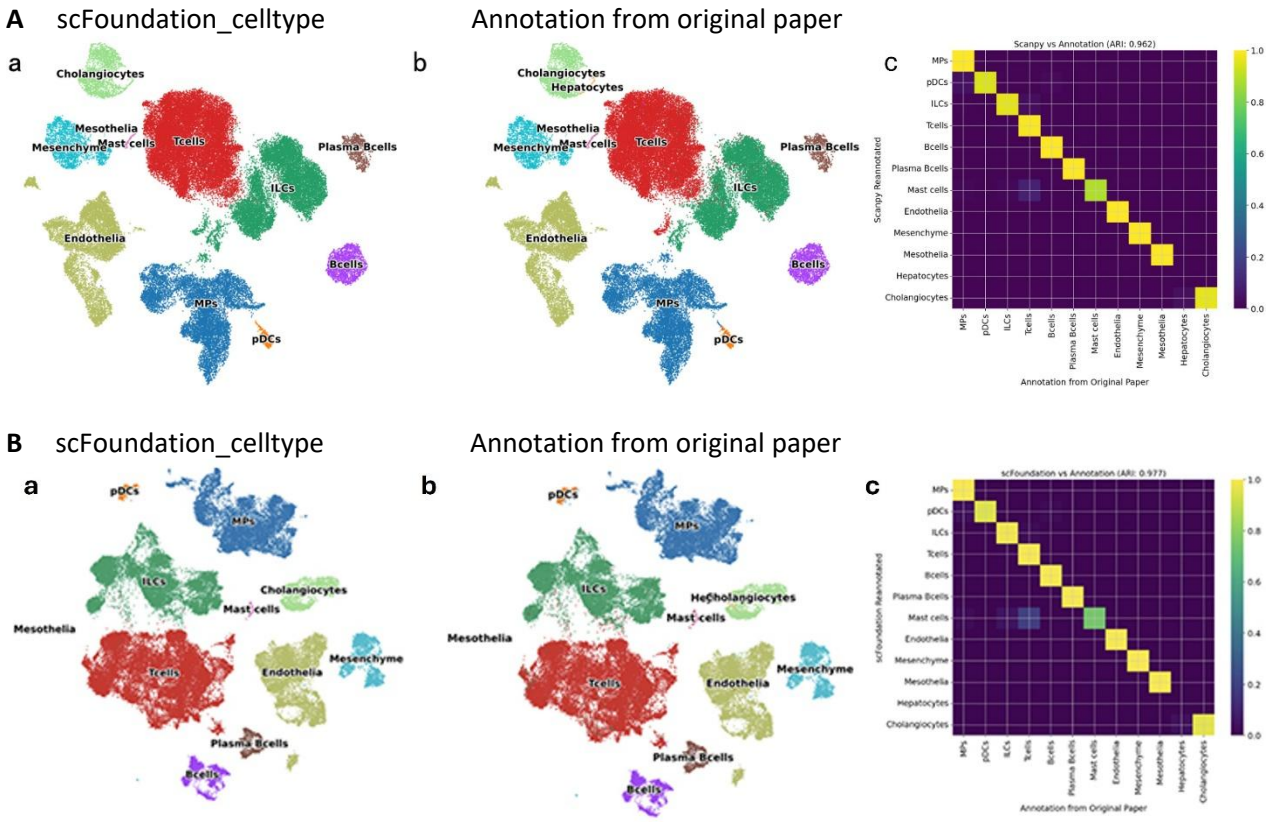

Figure 3. Concordance between cell-type annotations of this study and the original study. **A.** Side-by-side evaluation of the first dataset. (**a**) UMAP embedding colored by the cell-type labels assigned in this study, revealing well-segregated clusters of T cells, macrophages (MPs), B cells, ILCs, plasmacytoid dendritic cells (pDCs), and others. (**b**) Corresponding UMAP from the original publication with its initial annotations, visually mirroring the cluster structure in panel A. (**c**) Heatmap of cell-type overlap between the two label sets, quantified by ARI of 0.962. **B.** Benchmarking on the second dataset. (**a**) UMAP plot displaying the cell-type assignments produced in this study. (**b**) UMAP from the original work showing the author-provided labels. (**c**) Overlap heatmap between the two annotations, yielding an ARI of 0.977.

regions, and inter-cluster boundaries were largely preserved. The quantitative overlap matrix (Figure 3A-c) exhibited a pronounced diagonal, and the resulting ARI of 0.962 confirmed near-perfect agreement. Off-diagonal elements were confined to small macrophage and epithelial subclusters, indicating that minor over-merging rather than misclassification accounted for most discrepancies. Together, the qualitative and quantitative evidence validated the robustness of the Scanpy-based preprocessing and marker-driven annotation protocol.

(2) scFoundation workflow

The UMAP generated from scFoundation embeddings and annotated with the same marker schema (Figure 3B-a) not only recapitulated the global structure of the reference map (Figure 3B-b) but also delineated additional intra-lineage variation, particularly within the macrophage and epithelial compartments. The corresponding overlap heatmap (Figure 3B-c) displayed an even sharper diagonal than that of the Scanpy analysis, yielding an ARI of 0.977. This increase reflected two complementary improvements including a reduction in off-diagonal discordance for previously merged macrophage subsets and finer subdivision of epithelial states without generating spurious cross-lineage assignments. Importantly, every lineage present in the reference annotation was recovered with ≥ 95 % cell-wise fidelity, underscoring that the enhanced granularity did not compromise biological accuracy. While both pipelines achieved high

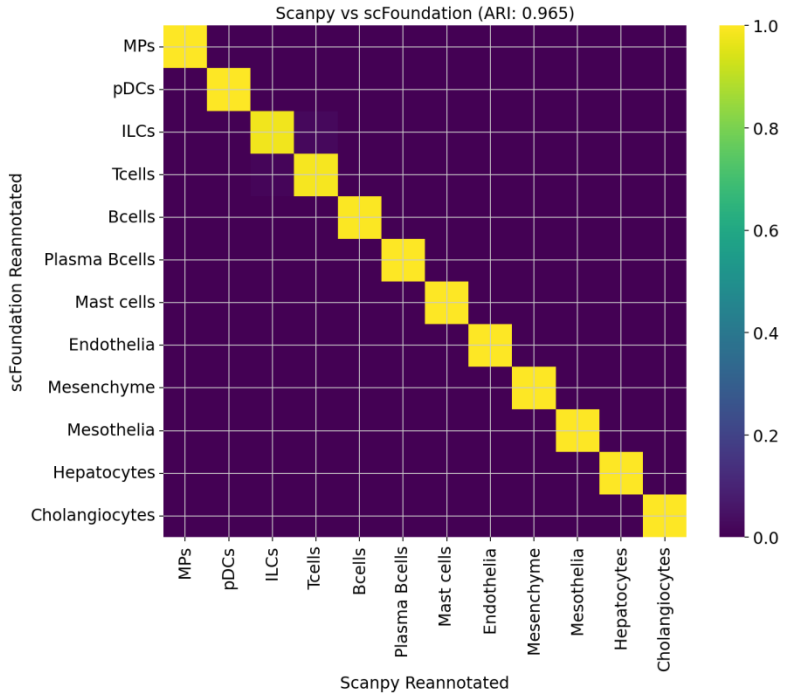

Figure 4. Comparison of cell-type annotations between Scanpy and scFoundation. The heatmap illustrated the overlap between cell-type annotations generated by the two methods with an ARI of 0.965, indicating strong agreement and consistency in clustering results.

concordance with the published labels, the scFoundation-based workflow attained superior alignment and resolved additional biologically meaningful sub-clusters, highlighting the value of learned embeddings for precise single-cell stratification.

Resolution of fine-grained transcriptional states

The extent to which each embedding could unveil sub-lineage heterogeneity was investigated. Differential gene-expression analyses and hierarchical clustering were used to determine whether additional, biologically interpretable states such as inflammatory versus resident macrophages or proliferative epithelial subsets were selectively resolved. The findings illustrated the practical implications of embedding choice for discovering previously obscured cellular programs. To determine whether the two analytical pipelines yielded comparable biological conclusions, the cell-type labels produced by Scanpy were directly contrasted with those inferred from the scFoundation embedding (Figure 4). The resulting confusion heat-map was dominated by a near-perfect diagonal with each of the twelve major lineages including macrophages, plasmacytoid dendritic cells, innate lymphoid cells, T cells, B cells, plasma cells, mast cells, endothelial cells, mesenchymal cells, mesothelial cells, hepatocytes, and cholangiocytes being mapped almost exclusively to its counterpart in the alternative analysis. An ARI of 0.965 placed the concordance within the “almost perfect” range. Off-diagonal signal remained confined to minor sub-populations that scFoundation subdivided more finely-principally inflammatory versus resident macrophage states and proliferative epithelial subsets, whereas Scanpy merged these into single clusters. No lineage detected by one method was absent from the other, indicating that observed differences were restricted to cluster granularity rather than lineage presence or absence. The effect of inter-individual variability and clinical status on this concordance was subsequently examined. When the scFoundation UMAP was colored by donor condition (healthy versus diseased) or by sample

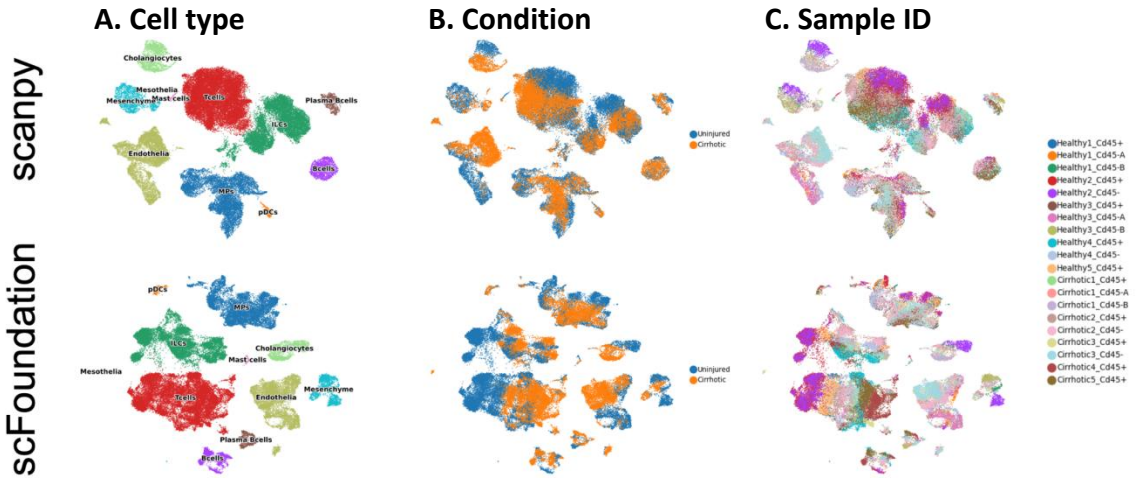

Figure 5. Robustness of cell-type annotations across clinical conditions and sample IDs. **A.** UMAP plot showing cell-type annotations generated by scFoundation. **B.** Annotations stratified by donor clinical condition. **C.** Annotations stratified by sample ID.

identifier, all clusters remained compositionally mixed as no cluster being dominated (> 75 %) by a single condition or sample. A χ^2 test of independence confirmed that cell-type assignment was not significantly associated with either metadata variable with $P > 0.05$ for all lineages after Benjamini-Hochberg correction (Figure 5). The results demonstrated that the consistency in annotations between Scanpy and scFoundation was not influenced by donor clinical condition or sample ID, confirming the robustness of the conclusions across different embedding methods and patient characteristics. Accordingly, the high Scanpy-scFoundation agreement was retained across heterogeneous patient backgrounds, ruling out confounding by donor-specific batch effects. The heat-map overlap, the ARI of 0.965, and the condition agnostic distribution of cells collectively demonstrated that the two embedding strategies were effectively interchangeable at the level of broad cell-type identification with scFoundation providing additional resolution without compromising cross-sample robustness.

Conclusion

The benchmark of the deep-learning framework scFoundation against the classical Scanpy pipeline on liver-cirrhosis scRNA-seq data showed that AI-based single-cell models were already capable of matching, and in some respects, surpassing conventional analyses. Both workflows recovered nearly identical cell-type annotations, yet scFoundation achieved a slightly higher concordance with the original labels (ARI ≈ 0.97), indicating that a large pretrained model could reproduce established biological insights without manual feature engineering. Moreover, the UMAP embeddings derived from scFoundation revealed finer transcriptomic structures that were not apparent in the Scanpy results, hinting at previously unrecognized cellular states linked to cirrhotic remodeling. While these additional patterns might represent bona-fide biology, they could also reflect amplified batch effects, therefore, orthogonal validation such as spatial transcriptomics or protein level assays remained essential to confirm their biological relevance. The advantages of scFoundation extended beyond accuracy. Its end-to-end architecture eliminated dataset-specific preprocessing, thereby lowering the analytical barrier for non-computational laboratories and improving reproducibility across studies. Nonetheless, three practical limitations currently restrict wider adoption, which include that the public API caps the number of cells that can be processed in a single request, limiting

scalability for large consortia datasets. Local deployment demands high-end GPUs and specialized expertise that many wet-lab environments lack. The model exhibits sensitivity to batch effects, underscoring the need for intrinsic correction mechanisms. Until these challenges are resolved, Scanpy remains a cost-effective and accessible alternative for many users. scFoundation exemplifies the transformative potential of deep-learning foundation models in single-cell genomics, offering streamlined workflows and enhanced resolution, yet its full promise will be realized only after improvements in batch robustness, computational efficiency, and cloud-native accessibility broaden its reach to the wider biomedical community.

References

1. Wu X, Yang X, Dai Y, Zhao Z, Zhu J, Guo H, *et al*. 2024. Single-cell sequencing to multi-omics: Technologies and applications. Biomark Res. 12(1):110.
2. Rhaman MS, Ali M, Ye W, Li B. 2024. Opportunities and challenges in advancing plant research with single-cell omics. Genomics Proteomics Bioinformatics. 22(2):qzae026.
3. Fang W, Liu X, Maiga M, Cao W, Mu Y, Yan Q, *et al*. 2024. Digital PCR for single-cell analysis. Biosensors (Basel). 14(2):64.
4. Caligola S, De Sanctis F, Canè S, Ugel S. 2022. Breaking the immune complexity of the tumor microenvironment using single-cell technologies. Front Genet. 13:1013701.
5. Chen F, Liu K, Shang L, Wang Y, Tang X, Liang P, *et al*. 2024. Precision isolation and cultivation of single cells by vortex and flat-top laser ejection. Front Microbiol. 15:1234567.
6. Lei Y, Tang R, Xu J, Wang W, Zhang B, Liu J, *et al*. 2021. Applications of single-cell sequencing in cancer research: Progress and perspectives. J Hematol Oncol. 14(1):91.
7. Wei D, Xu M, Wang Z, Tong J. 2022. The development of single-cell metabolism and its role in studying cancer emergent properties. Front Oncol. 11:812345.
8. Aliya S, Lee H, Alhammadi M, Umapathi R, Huh YS. 2022. An overview on single-cell technology for hepatocellular carcinoma diagnosis. Int J Mol Sci. 23(3):1402.
9. Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: A tutorial. Mol Syst Biol. 15(6):e8746.
10. Debnath T, Hattori R, Okamoto S, Shibata T, Santra TS, Nagai M. 2022. Automated detection of patterned single cells within hydrogel using deep learning. Sci Rep. 12(1):18343.
11. Molho D, Ding J, Tang W, Li Z, Wen H, Wang Y, *et al.* 2024. Deep learning in single-cell analysis. ACM Trans Intell Syst Technol. 15(3):Article 40.
12. Weiskittel TM, Correia C, Yu GT, Ung CY, Kaufmann SH, Billadeau DD, *et al*. 2021. The trifecta of single-cell, systems-biology, and machine-learning approaches. Genes (Basel). 12(7):1098.
13. Lin X, Tian T, Wei Z, Hakonarson H. 2022. Clustering of single-cell multi-omics data with a multimodal deep learning method. Nat Commun. 13(1):7705.
14. Ma Q, Xu D. 2022. Deep learning shapes single-cell data analysis. Nat Rev Mol Cell Biol. 23(5):303-304.
15. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, *et al*. 2024. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 21(8):1481-1491.
16. Ramachandran P, Dobie R, Wilson-Kanamori JR, Dora EF, Henderson BEP, Luu NT, *et al*. 2019. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. Nature. 575(7783):512-518.