PLOS Computational Biology

METHODS

scMagnifier: Resolving fine-grained cell subtypes via GRN-informed perturbations and consensus clustering

Zhenhui He, Kangning Dong*

School of Mathematics, Renmin University of China, Beijing China

* dongkangning@ruc.edu.cn

OPEN ACCESS

Citation: He Z, Dong K (2026) scMagnifier: Resolving fine-grained cell subtypes via GRN-informed perturbations and consensus clustering. PLoS Comput Biol 22(6): e1014167. https://doi.org/10.1371/journal.pcbi.1014167

Editor: Suoqin Jin, Wuhan University, CHINA

Received: March 24, 2026

Accepted: June 7, 2026

Published: June 18, 2026

Data availability statement: Data availability All data analyzed in this paper are available in raw form from their original authors. Specifically, the lung adenocarcinoma datasets are available at NCBI Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907). The pancreas dataset is collected from NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.

Abstract

Resolving fine-grained cell subtypes in single-cell RNA sequencing (scRNA-seq) data remains challenging, as their subtle transcriptional differences are often obscured by technical noise and data sparsity. Here, we present scMagnifier, a consensus clustering framework that leverages gene regulatory network (GRN)-informed in silico perturbations to amplify subtle transcriptional differences and uncover latent cell subpopulations. scMagnifier perturbs candidate transcription factors (TFs), propagates perturbation effects through cluster-specific GRNs to simulate post-perturbation expression profiles, and integrates clustering results across multiple perturbations into stable subtype assignments. Additionally, scMagnifier introduces regulatory perturbation consensus UMAP (rpcUMAP), a perturbation-aware visualization that provides clearer separation between cell subtypes and guides the selection of the optimal number of clusters. In both single-batch and multi-batch benchmarks, scMagnifier consistently improves the resolution and accuracy of fine-grained cell type identification. Notably, when integrated with spatial clustering methods such as STAGATE, scMagnifier is compatible with spatial transcriptomics workflows and effectively reveals tumor cell subtypes and their spatial organization in ovarian cancer.

Author summary

Understanding the diversity of cells in tissues is key to studying health and disease, but identifying subtle differences between closely related cell types is often challenging. Small molecular variations and technical noise can hide rare immune cells, transitional states, or tumor subtypes. We developed scMagnifier, a computational tool that amplifies these subtle differences by simulating genetic perturbations, allowing cells to be grouped into fine-grained subtypes more accurately. We also created a visualization method to clearly separate cell groups and guide the identification of meaningful cell types. When tested on

cgi?acc=GSE84133). The BMMC datasets is available at NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122). The UPN19_pre dataset is accessible on Human Antigen Receptor Database (huARdb) website (https://huarc.net/v2/database/browse/UPN19_pre). The ovarian cancer dataset and H&E images are available at SPAtial Transcriptomics resource for subCellular and High-throughput platforms (SPATCH) website (https://spatch.pku-genomics.org/#/dataset/xenium). Code availability The scMagnifier algorithm is implemented in Python and is available on Github (https://github.com/RucDongLab/scMagnifier).

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) [grant number 62402498 to K.D.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

single-cell, multi-batch, and spatial transcriptomics data, scMagnifier consistently outperformed existing methods, revealing rare cell populations and mapping tumor subgroups in ovarian cancer. This approach provides researchers with an accessible and reliable way to uncover hidden cellular diversity, with potential applications in immunology, cancer biology, and studies of disease-related cell states.

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by enabling transcriptome-wide profiling at single-cell resolution [1]. Unsupervised clustering is a standard and essential step in the scRNA-seq analysis workflow, routinely used to identify discrete cell types and continuous cell states [2,3]. With the rapid development of clustering algorithms, current tools can consistently and accurately resolve major cell types—such as T and B lymphocytes in immune tissues, neurons and glia in the brain, or epithelial and stromal cells in solid tumors—across a wide range of biological settings. However, resolving fine-grained cell subtypes and accurately delineating their boundaries remains challenging—particularly for transcriptionally similar cell states such as activated versus resting immune cells, malignant epithelial subclones within tumors, or rare cell populations [2,3]. The high dimensionality, sparsity, and noise inherent in scRNA-seq data often obscure subtle but biologically meaningful transcriptional differences, limiting the resolution of clustering methods for fine-grained heterogeneity [4].

Notably, transcriptionally similar cell populations may respond differently to regulatory perturbations due to variations in their underlying gene regulatory networks (GRNs), offering a powerful means to uncover hidden heterogeneity [5,6]. Several studies have employed in silico gene perturbations to simulate key biological processes, such as cell fate transitions or drug-induced state changes. For example, CellOracle simulates the effects of transcription factor (TF) perturbations on cell identity dynamics by modeling GRN influences, providing interpretable and quantitative insights into developmental trajectories [7]. Similarly, scRank perturbs inferred GRNs to simulate drug-induced regulatory changes, enabling the identification of drug-responsive cell types [8]. In addition, scTenifoldKnk applies GRN perturbation to predict the functional consequences of gene knockouts [9]. Collectively, these studies demonstrate that GRN-informed perturbations provide a powerful approach for revealing functionally relevant cellular heterogeneity. Motivated by these studies, we reasoned that perturbing candidate TFs or genes within GRNs could provide a way to amplify subtle transcriptional differences and thereby facilitate the identification of fine-grained cell subtypes.

Furthermore, different perturbations of regulatory networks can yield distinct patterns of cellular responses, making it necessary to integrate multiple perturbation outcomes to obtain stable clustering results and well-defined subpopulation boundaries. Consensus clustering provides an effective strategy to integrate multiple

results and produce stable cluster assignments. Existing consensus clustering approaches typically generate ensemble diversity by repeatedly applying clustering with varying parameters or initializations to the same expression matrix [10–13], which improves robustness to stochastic variability but fails to enhance the underlying biological signal used to distinguish cell subtypes or states. In contrast, integrating GRN-informed perturbation-derived clustering results into a consensus clustering framework captures the differential responses of subpopulations to distinct regulatory perturbations. This perturbation-aware consensus strategy leads to robust and interpretable boundaries between fine-grained cell types.

Building on these ideas, we develop a GRN-informed perturbation-driven consensus clustering framework named scMagnifier. By integrating with standard clustering algorithms, batch integration methods, or spatial clustering tools, scMagnifier is readily applicable to diverse scenarios—including single-batch, multi-batch, and spatial transcriptomic datasets. Extensive benchmarking demonstrates that scMagnifier consistently improves the resolution and accuracy of fine-grained cell type identification and enhances the detection of rare cell populations. Additionally, scMagnifier integrates perturbation-induced clustering information into the UMAP algorithm, yielding a perturbation-aware visualization, named regulatory perturbation consensus UMAP (rpcUMAP), that provides clearer separation between cell subtypes and guides the selection of the optimal number of cell types.

Results

Overview of scMagnifier

The inputs to scMagnifier are the raw gene expression matrix (GEM) and a basic GRN, defined as a set of regulatory interactions from TFs to their target genes (Fig 1A). An initial clustering is obtained using a standard scRNA-seq preprocessing and clustering pipeline implemented in Scanpy [14]. Cluster-specific GRNs are then constructed by pruning the basic GRN based on the expression levels of TFs and their target genes within each cluster (Fig 1D and see "Cluster-specific GRN construction" section of the Methods).

The core of scMagnifier lies in performing GRN-informed in silico perturbations and yielding an ensemble of clustering results across distinct perturbation conditions (Fig 1B). For each candidate TF, a cell-specific perturbation term is first defined relative to its original expression level. This perturbation is then propagated through the corresponding cluster-specific GRN to model downstream regulatory effects (Fig 1E **and see "TF genes perturbation on the level of gene expression" section of the Methods**). The resulting propagated expression changes are integrated with the original GEM to generate a post-perturbation GEM, from which a perturbation-driven clustering result is obtained.

To derive stable cell subtype assignments, scMagnifier performs consensus clustering on the perturbation-driven ensemble of cluster results (Fig 1C and 1F). To integrate these clustering results, each clustering outcome is converted into a one-hot matrix, enabling the computation of perturbation-informed cell-cell distances across the ensemble. These perturbation-informed cell-cell distances are further combined with expression-derived distances computed from the embedding matrix of the GEM, resulting in a combined distance matrix that captures both transcriptional similarity and regulatory perturbation-driven differences. Based on this combined distance matrix, a k-nearest neighbor (KNN) graph is constructed for consensus clustering and for generating a regulatory perturbation consensus UMAP (rpcUMAP) (**See "Consensus clustering" section of the Methods**).

To avoid missing subtle but biologically meaningful differences, consensus clustering is initially performed at high clustering resolution. These preliminary clusters are then merged based on inter-cluster centroid distances and a minimum cluster-size threshold to merge closely related or very small clusters into their nearest neighbor clusters and produce the final stable consensus clusters (S1 Fig **and see "Cluster merging" section of the Methods**).

More generally, scMagnifier is extensible and can be applied to multi-batch single-cell datasets. Because both the clustering and the computation of expression-derived distances are performed on low-dimensional embeddings, scMagnifier can directly make use of batch-corrected embeddings (e.g., from Harmony [15], Scanorama [16] or scVI [17]), enabling

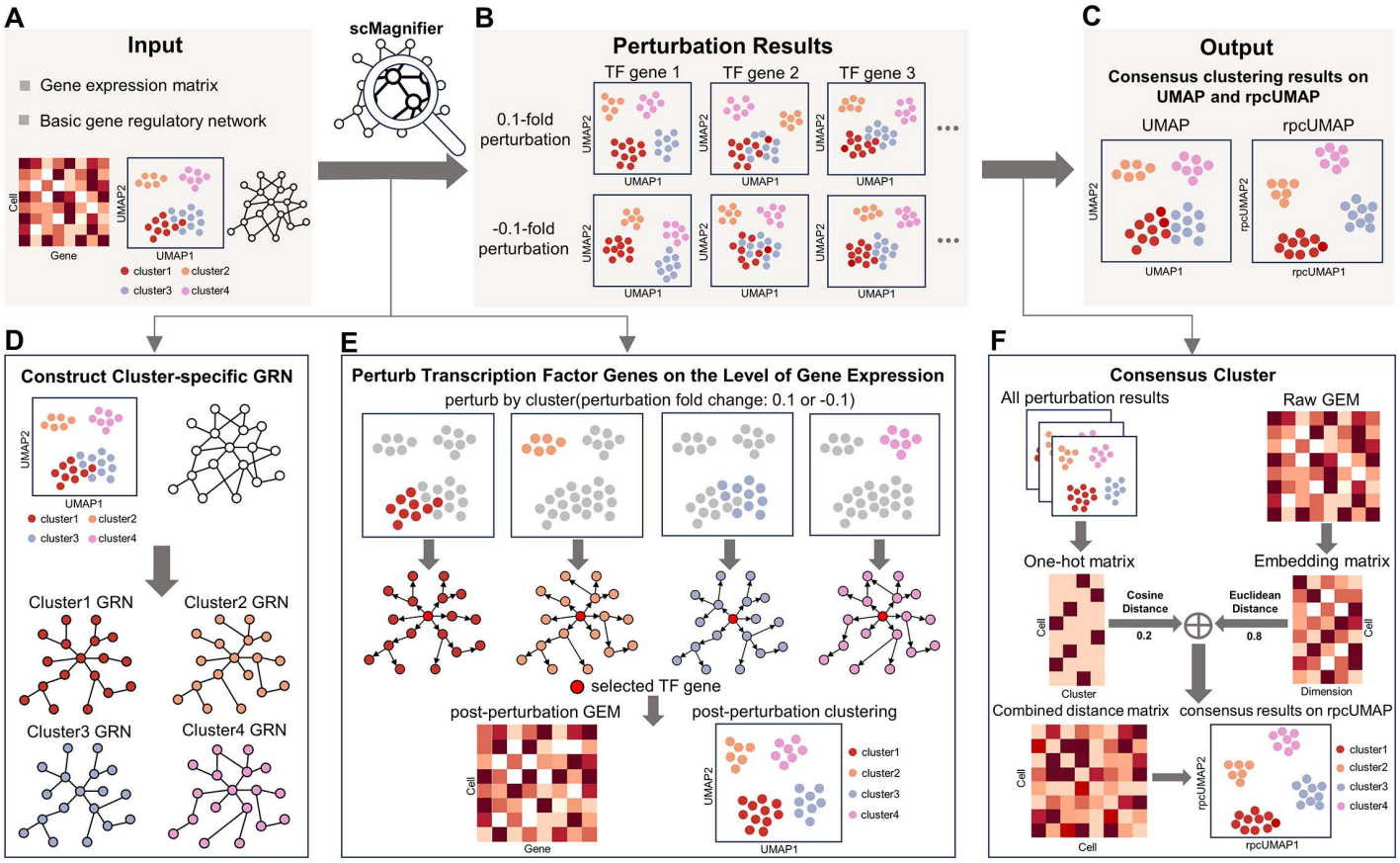

Fig 1. Overview of scMagnifier. (A) scMagnifier takes the gene expression matrix and a basic gene regulatory network as inputs, and derives the initial clustering results using a standard clustering pipeline. (B) scMagnifier obtains distinct clustering results by performing TF perturbations on cluster-specific gene regulatory networks. (C) The output of scMagnifier is a consensus clustering result visualized on UMAP and rpcUMAP embeddings. (D) scMagnifier first constructs cluster-specific GRN based on the basic GRN and initial clustering results. (E) scMagnifier systematically perturbs candidate TF genes and propagates their effects through cluster-specific GRN to generate an ensemble of post-perturbation clustering results. (F) scMagnifier integrates perturbation-driven clustering results with expression-based similarities to construct a combined cell–cell distance matrix, enabling consensus clustering and the generation of a rpcUMAP.

https://doi.org/10.1371/journal.pcbi.1014167.g001

straightforward application to multi-batch datasets (**See "Extension of scMagnifier for multi-batch datasets" section of the Methods**).

Benchmarking scMagnifier in real datasets

To evaluate scMagnifier's ability to resolve fine-grained cell subtypes, we benchmarked its performance on multiple publicly available single-cell datasets. For single-batch evaluation, we selected four lung adenocarcinoma datasets that contain rich cellular substructure and published annotations [18]. In each dataset, we fixed the number of output clusters to match the annotation and compared nine methods (Leiden [19], Louvain [20], scVI(Leiden) [17], scVI(Louvain) [17], SC3s [21], DBSCAN [22], Hierarchical [23], scMagnifier(Leiden) and scMagnifier(Louvain))(**See "Benchmarking setup and parameters" section of the Methods**). To provide a comprehensive evaluation of clustering performance, we assessed both agreement with annotations using adjusted Rand index (ARI) and normalized mutual information (NMI), as well as silhouette score and entropy of cell type mixing. Across these single-batch benchmarks, scMagnifier consistently achieved the highest ARI and NMI, with a substantial increase in Silhouette score (Figs 2A, S2A and S2C; Tables C-F in S1 Text).

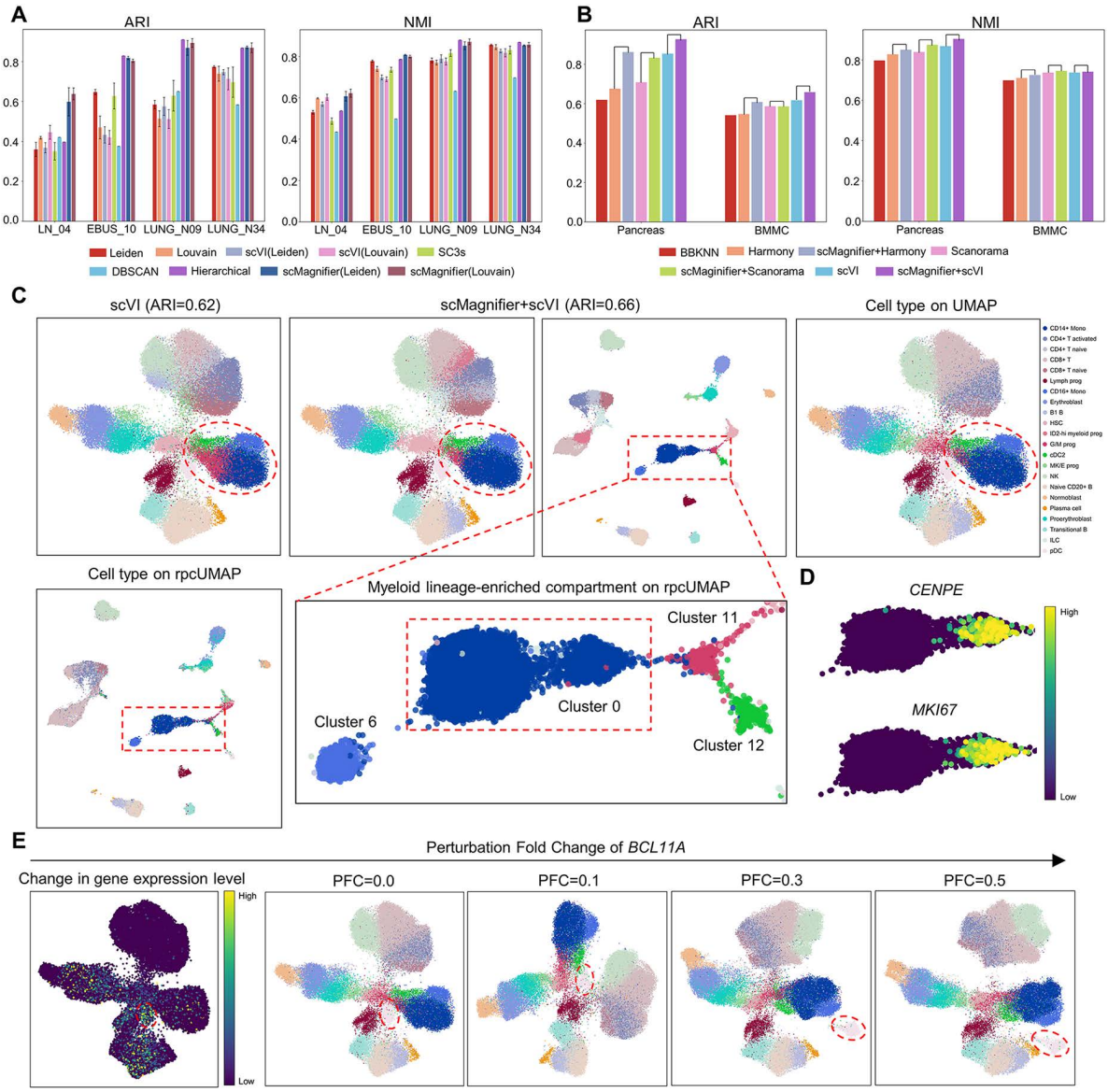

Fig 2. Benchmarking of scMagnifier. (A) Bar plots illustrating the performance of nine algorithms on four single-batch datasets, evaluated using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). (B) Bar plots of clustering performance on two multi-batch datasets for seven batch-correction methods, alone or combined with scMagnifier. (C) UMAP and rpcUMAP visualization of BMMC dataset clustering via scVI (ARI = 0.62) and scMagnifier+scVI (ARI = 0.66). The red circle highlights a myeloid lineage-enriched compartment containing CD14 + Mono, CD16 + Mono, cDC2 and G/M prog cells. A zoomed view contrasts this compartment on standard UMAP versus rpcUMAP. (D) rpcUMAP visualization of *CENPE* and *MKI67* gene expression in Cluster 0 of the BMMC dataset, obtained using the scVI + scMagnifier workflow. (E) UMAP visualization of BCL11A perturbation in the BMMC dataset using the scMagnifier+scVI workflow. The leftmost panel shows the sum of absolute expression changes for *BCL11A* and its target genes across cells. The four panels on the right display UMAP visualizations of post-perturbation clustering results as the perturbation fold change increases.

https://doi.org/10.1371/journal.pcbi.1014167.g002

We next tested scMagnifier's applicability in multi-batch samples. We benchmarked scMagnifier on two public multi-batch datasets (an integrated pancreas dataset [24] and the BMMC dataset [25]). For each dataset, we compared several commonly used batch-correction methods alone and in combination with scMagnifier (BBKNN [26], Harmony [15], scMagnifier+Harmony, Scanorama [16], scMagnifier+Scanorama, scVI [17], scMagnifier+scVI). To comprehensively assess

clustering performance, we quantified ARI, NMI, silhouette score and entropy of batch mixing. Across both datasets, combining scMagnifier with batch-correction methods not only improved ARI and NMI but also substantially increased silhouette scores relative to the corresponding batch-correction methods alone (Figs 2B, S2B and S2D; Tables G-J in S1 Text).

We also performed ablation experiments by removing cluster merging, GEM-derived distances or perturbation-derived distances (**See "Ablation experiments" section of the Methods**). The results (S11 Fig) indicate that each module contributes complementarily to scMagnifier's performance.

Focusing on the myeloid lineage-enriched compartment in the BMMC dataset, we found that scMagnifier+scVI precisely delineated subcluster boundaries, as validated against reference cell type annotations (Fig 2C). In contrast, scVI drew an inaccurate boundary between granulocyte-monocyte progenitors (G/M prog) and $CD14^+$ monocytes, leading to significant cross-contamination of their identities. Notably, the rpcUMAP visualization generated by scMagnifier+scVI outperformed conventional UMAP by achieving clearer separation of distinct cellular clusters. We further found that rpcUMAP identified two morphologically distinct subpopulations within Cluster 0, which were merged into a single cluster in our benchmarking analyses to align with the predefined cell-type number. Differential gene expression analysis confirmed that these subpopulations exhibited divergent expression of proliferation-associated genes (e.g., *CENPE* and *MKI67*), supporting their classification as distinct subclusters (Fig 2D). Thus, rpcUMAP-based evidence supports revising the optimal number of cell types in this myeloid lineage-enriched compartment from four to five. Collectively, these findings demonstrate that rpcUMAP not only enhances the separation of cell subtypes but also guides the selection of the optimal number of cell types. Moreover, by exploring different resolutions, our approach can assist in identifying meaningful cluster numbers without prior knowledge (S13 Fig).

Finally, we explored whether scMagnifier amplifies biologically meaningful differences between cells. In the BMMC dataset, using the scMagnifier+scVI workflow, we perturbed the TF gene *BCL11A* and propagated its regulatory effects through cluster-specific GRNs. We observed that plasmacytoid dendritic cell (pDC) clusters exhibited progressively stronger transcriptional changes compared to neighboring cell populations and, as the perturbation fold change increased, gradually dissociated from surrounding clusters (Fig 2E). Importantly, *BCL11A* has been reported as an essential regulator of pDC development and lineage specification [27]. Its knockout leads to impaired pDC development, aberrant lineage specification and a marked decrease in the number and functional maturation of pDCs [27]. These observations demonstrate that gene perturbations can amplify biologically meaningful transcriptional differences, providing a mechanistic explanation for why scMagnifier can resolve fine-grained cell subtypes.

scMagnifier reveals hidden heterogeneity within MAIT/Th1-Th17 populations

We further evaluate whether scMagnifier can reveal fine-grained cellular heterogeneity that is obscured by conventional clustering. In UPN19_pre dataset [28], using standard Leiden clustering on the original embedding, Mucosal-associated invariant T (MAIT) cells and a T helper 1/T helper 17 (Th1/Th17)-MAIT mixed population were grouped into a single cluster (Cluster 3) and appeared continuously connected on the original UMAP, indicating limited separability in the original embedding space (Fig 3A and 3B). Even when the initial resolution was increased, the two cell populations were still recognized as a single cluster (S4A Fig). This observation is expected, as MAIT cells exhibit transcriptional programs associated with both Th1 and Th17 effector functions [29]. In contrast, applying scMagnifier to the same cells clearly separated this previously merged population into two distinct clusters (Cluster 2 and Cluster 16) in the rpcUMAP space (Fig 3A and 3B).

Differential expression analysis reveals that Cluster 2 is characterized by high expression of cytotoxic-associated genes including *CD8A*, *NKG7*, *NCR3* and *PRF1*, while Cluster 16 shows relatively higher expression of several genes associated with Th1/Th17 programs (Fig 3C). Visualization of gene expression on the rpcUMAP provides a more intuitive view of these differences, showing that cytotoxic-associated genes are strongly enriched in Cluster 2 (Fig 3D).

We next performed KEGG pathway enrichment analysis to further validate the functional heterogeneity of these two scMagnifier-resolved clusters (Fig 3E). For Cluster 2, it was significantly enriched in natural killer cell mediated cytotoxicity

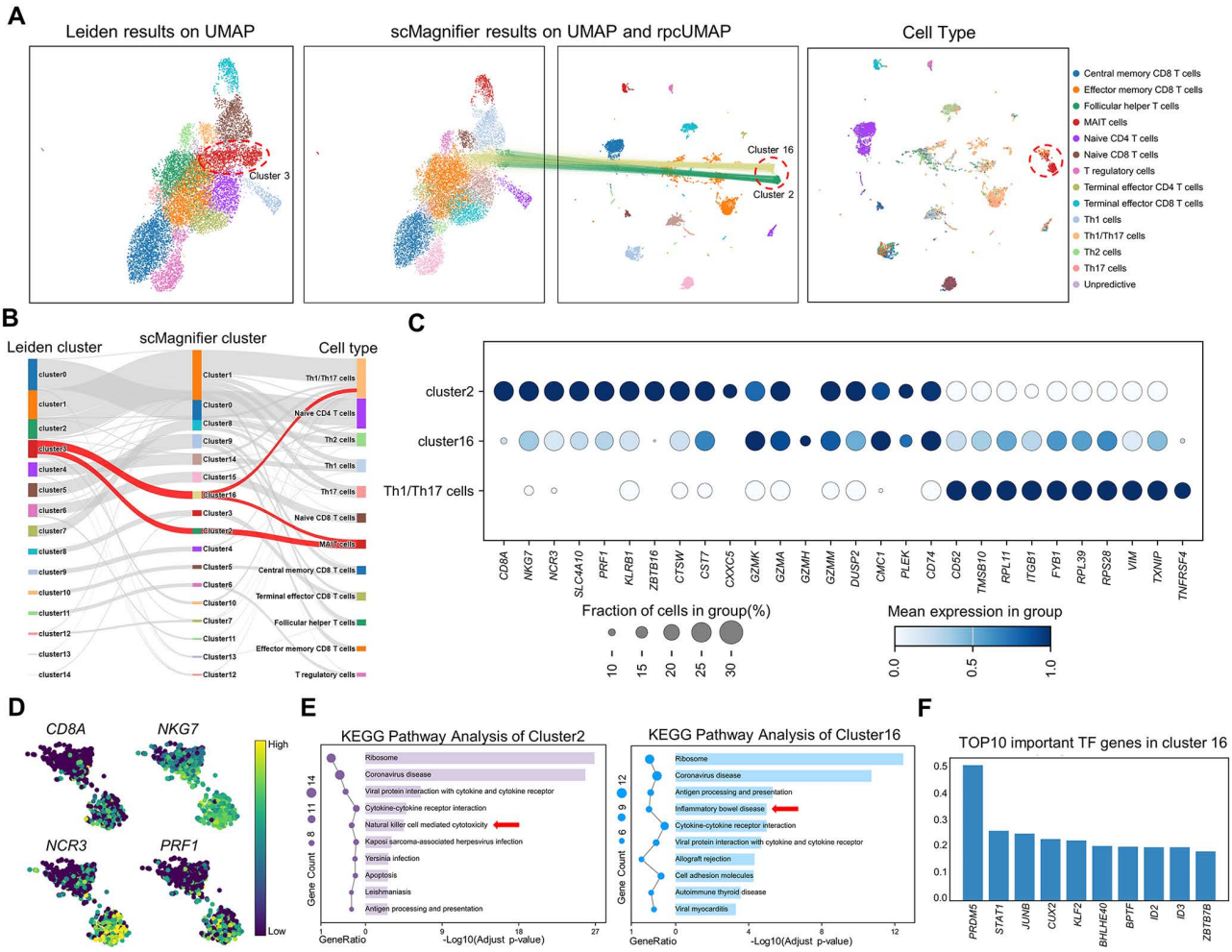

Fig 3. scMagnifier uncovers hidden heterogeneity among MAIT/Th1-Th17 cell populations. (A) Comparison of original Leiden clustering and scMagnifier results. Left: UMAP visualization of standard Leiden clustering (resolution = 0.75). Middle: UMAP and rpcUMAP visualization of scMagnifier clustering. The connecting lines denote the positional correspondence of clusters identified by scMagnifier between the UMAP and rpcUMAP embedding spaces. The red circle highlights the focal clusters (Cluster 2 and Cluster 16). Right: rpcUMAP visualization of cell type distributions in the dataset. (B) Sankey diagram showing the correspondence among original Leiden clusters, scMagnifier clusters and cell-type annotations. (C) Bubble plot of the top 10 differentially expressed genes (DEGs) for each cluster, among cluster 2, cluster 16 and the remaining Th1/Th17 cells. (D) UMAP heatmaps showing the expression of *CD8A*, *NKG7*, *NCR3*, and *PRF1* in cluster 2 and cluster 16. (E) KEGG pathway enrichment analysis of cluster 2 and cluster 16, showing the top 10 most significantly enriched pathways. (F) Bar plot of the top 10 TF genes ranked by importance scores in cluster 16.

https://doi.org/10.1371/journal.pcbi.1014167.g003

pathway (Adjusted p-value = 3.2×10^{-5}), which aligns perfectly with its high expression of cytotoxic genes. Conversely, Cluster 16 was prominently enriched in inflammatory bowel disease pathway, which has been well-documented to be functionally consistent with the molecular features of Th1/Th17 cells [30].

Meanwhile, we derived the TF importance scores based on the changes induced by the perturbed genes (**See "TF gene importance score calculation" section of the Methods**). We observed *STAT1* ranked as the second most important TF genes in Cluster 16 and it did not appear among the top-ranked TF genes in Cluster 2 (Figs 3F and S4B). S*TAT1* has been reported as a key regulator involved in interferon signaling and Th1/Th17-related immune responses [31].

Collectively, analyses at the gene expression, functional pathway and regulatory levels consistently demonstrate that the two scMagnifier-resolved clusters exhibit biological heterogeneity, which is obscured under conventional clustering but

effectively revealed by scMagnifier. Notably, the biological differences observed between the two clusters are consistent with the known plasticity of MAIT cells toward cytotoxic or Th1/Th17-like programs under different contexts [32]. Together, these results highlight scMagnifier's ability to amplify subtle yet meaningful cellular differences and to facilitate the discovery of fine-grained heterogeneity.

scMagnifier enables the identification of rare immune cell types with distinct regulatory programs

Rare cells often represent a very small fraction of the total population and are therefore susceptible to technical noise, dropout events and batch effects, while conventional clustering workflows tend to favor dominant transcriptional programs and can merge low-abundance subpopulations into larger clusters [33,34]. To test whether scMagnifier can effectively identify rare cell populations, we applied it to two independent scRNA-seq datasets, EBUS_10 and LUNG_N30 [18]. When exploring rare cell states, the minimum cluster-size threshold used during the cluster-merge step can be appropriately reduced based on the size of the dataset and the underlying biological context, so as to avoid merging small but potentially meaningful clusters and to retain candidate rare cell populations for downstream evaluation (**See "Consensus clustering" section of the Methods**).

In the EBUS_10 dataset, scMagnifier identified two small cell clusters, R1 (18 cells, 0.40%) and R2 (16 cells, 0.36%), that could not be resolved by conventional clustering methods even at high resolution (S5A Fig). Notably, GiniClust3 [35] also did not detect these two rare clusters (S5B Fig). Doublet detection using Scrublet [36] indicated that none of the cells in R1 and R2 were predicted as doublets (S5D Fig). While these cells were closely positioned within larger populations in the standard UMAP embedding, they became clearly separated from surrounding clusters in the rpcUMAP space (Fig 4A), suggesting the potential presence of subtle but consistent transcriptional differences that are revealed by scMagnifier. To characterize these clusters, we first grouped the remaining cells according to their original cell-type annotations. We then performed differential expression analysis for R1, R2 and each annotation-defined cluster, obtaining cluster-specific differentially expressed gene (DEG) sets for all groups. These cluster-specific DEG sets were subsequently used to quantify transcriptional similarity between R1, R2 and annotated cell populations using the Jaccard coefficient (**See "Jaccard coefficient calculation" section of the Methods**). R1 showed similarity with germinal centers (GC) B cells in the dark zone (DZ) (Jaccard = 0.33) and lower similarity with mucosa-associated lymphoid tissue (MALT) B cells (Jaccard = 0.12), while R2 exhibited a similarity with GC B cells in the DZ (Jaccard = 0.41) (S5C Fig and Table Q in S1 Text). Consistent with these results, dot plots of DEG expression (Fig 4B) revealed that R1 shares partial expression patterns with both MALT B cells and GC B cells in the DZ, while also displaying distinct expression distributions. In contrast, R2 more closely resembled GC B cells in the DZ but still exhibited differences. Notably, cells in R1 and R2 were originally annotated as MALT B cells and GC B cells in the DZ (Fig 4A). Together, these results suggest that R1 likely represents a finer-grained subpopulation within the MALT B cell compartment, and R2 likely represents a finer-grained subpopulation within GC B cells in the DZ.

To further examine the transcriptional differences among these populations, we next visualized the significantly upregulated genes for R1, R2, MALT B cells and GC B cells in the DZ using a heatmap (Fig 4C). R1 exhibits significantly higher expression of *CCND2* and *UBE2S* relative to the other cell clusters. It has been reported that *CCND2* is implicated in promoting B cell cell-cycle entry and has been linked to proliferative B cell states [37], while *UBE2S* contributes to mitotic progression and supports proliferation in multiple cellular systems [38]. Together, these findings suggest that R1 represent a rare subpopulation of MALT B cells associated with proliferation and activation. By contrast, R2 shows higher expression of *EBI3* and *TLR10* relative to the other cell clusters. *EBI3* has been reported in activated B cells and can modulate B cell differentiation and immune responses [39], and *TLR10* has been associated with activation and mucosal immune regulation [40]. These observations suggest that R2 represents a rare subpopulation of B cells associated with activation and immune regulation, likely resembling a MALT-like subcluster rather than a simple proliferative GC-DZ program.

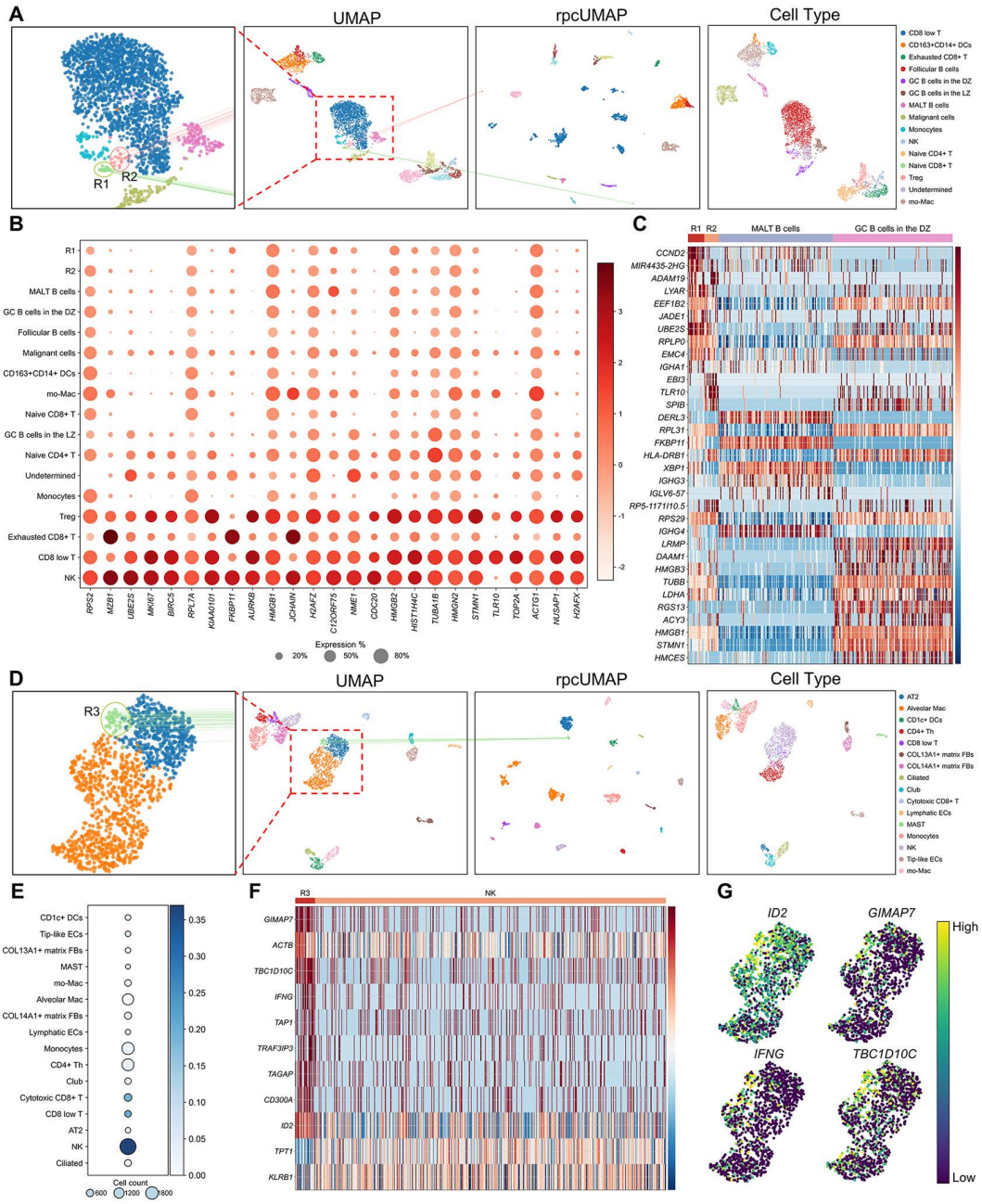

Fig 4. scMagnifier detects rare immune cell populations in lung adenocarcinoma datasets. (A) UMAP and rpcUMAP visualizations of scMagnifier results on the EBUS_10 dataset, highlighting the R1 and R2 clusters. The connecting lines denote the positional correspondence of R1 and R2 between the UMAP and rpcUMAP embedding spaces. (B) Dot plots showing the expression profiles of the top 15 DEGs individually selected from R1 and R2, across R1, R2 and the remaining cell-type-annotated cells. (C) Heatmap of gene expression for the top 10 upregulated genes selected from each of the four clusters (R1, R2, MALT B cells and GC B cells in the DZ). (D) UMAP and rpcUMAP visualizations of scMagnifier results on the LUNG_N30 dataset, highlighting the R3 clusters. (E) Bubble plot showing Jaccard coefficient values between R1 and clusters of the remaining cells grouped by cell-type annotations. (F) Heatmap of gene expression for the top 10 upregulated genes selected from each of the two clusters (R3 and NK). (G) UMAP heatmaps showing the expression of *ID2*, *IFNG*, *GIMAP7* and *TBC1D10C* in R3 and NK cells.

https://doi.org/10.1371/journal.pcbi.1014167.g004

Similarly, we next applied scMagnifier to the LUNG_N30 dataset and identified a small cluster, R3 (36 cells, 1.25%), that was not detected by conventional high-resolution clustering and Ginclust3 (S5A and S5B Fig). In the standard UMAP embedding, R3 appeared connected to neighboring cells, whereas in rpcUMAP it formed a clearly separated cluster (Fig 4D). Jaccard similarity between R3 DEGs and annotated clusters reveals highest overlap with natural killer (NK) cells (Jaccard = 0.37) (Fig 4E and Table R in S1 Text), and R3 cells were originally annotated as NK (Fig 4D), supporting the hypothesis that R3 is a rare NK subpopulation. To investigate this further, we compared significantly upregulated genes for R3 and NK cells (Fig 4F). R3 exhibits significantly higher expression of *ID2*, *IFNG*, *GIMAP7* and *TBC1D10C* relative to NK cell clusters, which is also clearly reflected in the UMAP visualization (Fig 4G). It has been reported that *ID2* is a key transcription factor regulating NK cell development and maturation [41], while *IFNG* is a hallmark effector cytokine of NK cells [42]. *GIMAP7* and *TBC1D10C* are involved in lymphocyte activation and regulation [43,44]. Together, these findings suggest that R3 represents a rare subpopulation of NK cells associated with a distinct activation or maturation state, potentially corresponding to the $CD56^{bright}$ NK cells, which are known to express higher levels of *IFNG*, and are often linked to early stages of activation and immune responses [45].

In summary, scMagnifier successfully identified rare subpopulations in the EBUS_10 and LUNG_N30 datasets, highlighting its potential to uncover biologically meaningful yet low-abundance cell populations that are often overlooked by conventional clustering methods.

Integration of scMagnifier with STAGATE reveals tumor cell subtypes and spatial organization in ovarian cancer

The identification of tumor cell subtypes and their spatial organization plays a critical role in understanding cancer progression and therapeutic response [46]. STAGATE enables the accurate identification of spatial domains by learning low-dimensional latent embeddings [47], while scMagnifier provides fine-grained types of cells. Hence, we integrated scMagnifier with STAGATE (**See "Integration of scMagnifier with STAGATE for Spatial Transcriptomics Analysis" section of the Methods**), aiming to identify biologically meaningful tumor subtypes and illustrate their spatial domains. We used the spatial transcriptomics dataset of epithelial ovarian cancer from the SPATCH website [48], which incorporates both detailed cell annotations and spatial layer annotations. For our analysis, we selected a subset of cells located at the center of spatial coordinates comprising 48,793 cells (Fig 5B).

Epithelial ovarian cancer, a malignancy originating from epithelial cells, is characterized by significant cellular heterogeneity and complex tumor microenvironment [49]. Based on cell annotations and spatial layer annotations provided by the dataset (Fig 5A), we determined the approximate location of the tumor's core regions (The overlapping region of Epithelial cell clusters in the cell type annotation map and Layer 3 regions in the spatial layer annotation map). Subsequently, using the final clustering results from scMagnifier combined with STAGATE, we identified five distinct subclusters of tumor cells (Cluster 0, Cluster 2, Cluster 4, Cluster 7 and Cluster 9), which correspond to potential tumor core regions (Fig 5A and 5B). Using differential gene expression analysis and functional enrichment analysis, we confirmed that the five tumor subclusters exhibit distinct molecular and functional differences (Fig 5C and 5D).

Notably, we observed that the spatial localization of Cluster 2 closely overlapped with the deeply stained regions in H&E histology (Fig 5E). Differential gene expression analysis revealed that Cluster 2 displayed high expression of *IGF2* (Fig 5C). It has been reported *IGF2* is involved in tumor growth and invasion mechanisms [50]. Consistent with this, functional enrichment analysis identified pathways related to negative regulation of apoptosis (Adjusted p-value = 2.0×10^{-7}) and extracellular structure organization (Adjusted p-value = 3.9×10^{-8}) in Cluster 2 (Fig 5D), indicating that these cells may actively evade cell death and promote metastatic potential. This functional profile aligns well with *IGF2*-mediated proliferative and anti-apoptotic signaling, a characteristic feature of aggressive epithelial tumor subpopulations [50]. Importantly, deep staining in H&E images generally indicates high cellular density and malignancy, which are hallmarks of invasive tumor subpopulations [51], and further supports that Cluster 2 corresponds to a highly aggressive tumor subpopulation. Collectively, this spatial correlation suggests that, by integrating scMagnifier with STAGATE, we were able to identify

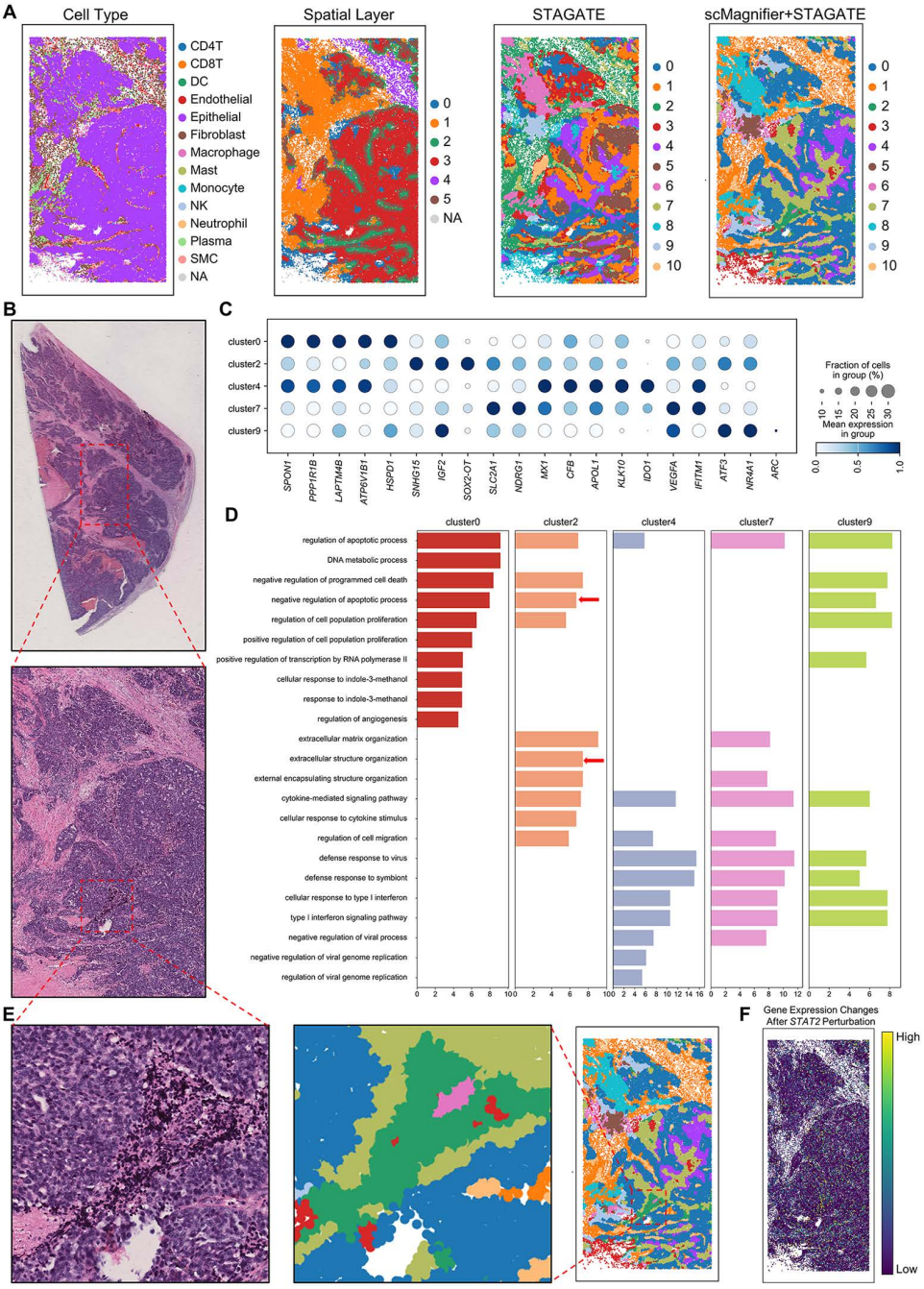

Fig 5. The integration of scMagnifier and STAGATE identifies tumor cell subtypes and their spatial organization in ovarian cancer. (A) Spatial visualization of the ovarian cancer dataset. Left two panels: cell-type annotations and spatial layer annotations provided with the dataset. Right two panels: clustering results from STAGATE alone (Leiden, resolution = 0.3) and from scMagnifier combined with STAGATE. (B) H&E-stained image highlighting the specific regions analyzed in the dataset used for our experiments. (C) Bubble plot of the top 5 differentially expressed genes for each of the five tumor subclusters identified by scMagnifier+STAGATE. (D) GO enrichment analysis of the five tumor subclusters, showing the top 10 significantly enriched GO terms for each subcluster. (E) Spatial correspondence between histology and scMagnifier+STAGATE clustering. Left panel: dark regions highlighted in the H&E-stained image. Right panel: spatial map showing the location of cluster 2 identified by scMagnifier+STAGATE. (F) Spatial heatmap showing the sum of absolute expression changes for *STAT2* and its target genes across cells after *STAT2* perturbation.

https://doi.org/10.1371/journal.pcbi.1014167.g005

high-invasion tumor regions without relying on traditional histological images, a capability that was not achievable with STAGATE alone.

To further investigate why this specific region was detected, we analyzed the impact of TF genes perturbation within the scMagnifier framework. After perturbing *STAT2*, we calculated the changes in gene expression levels of *STAT2* and its target genes, and then visualized the resulting changes on the spatial map (Fig 5F). Notably, the region corresponding to cluster 2 became significantly more prominent compared to its surrounding areas, suggesting that this cluster was more strongly affected by *STAT2* perturbation. This increased prominence allowed for the successful identification of this region after *STAT2* perturbation (S6 Fig), indicating that perturbation amplified the underlying biological differences within this cluster. In summary, the integration of scMagnifier with STAGATE enables the revelation of tumor cell subtypes and spatial organization.

Discussion

Accurately delineating cell subpopulations and their boundaries remains challenging, particularly when transcriptional differences are subtle or obscured by noise, batch effects. In this study, we presented scMagnifier, a regulatory perturbation-driven consensus clustering framework designed to resolve fine-grained cell subtypes. scMagnifier first amplifies subtle regulatory differences through systematic perturbation of TF genes to reveal latent subclusters, and then integrates clustering outcomes across multiple perturbations via consensus clustering to obtain stable subtype assignments and well-defined boundaries. We demonstrated practical utility of scMagnifier in several applications. We found scMagnifier clearly reveals hidden heterogeneity within MAIT/Th1-Th17 populations. We additionally substantiated the ability of scMagnifier to detect rare cell populations. Finally, by integrating scMagnifier and STAGATE, we identified five subtypes of ovarian cancer and detected invasive regions without incorporating pathological images.

Recent studies have demonstrated that GRN perturbation models are capable of uncovering latent biological signals [7–9]. Building on this framework, we showed that perturbing TF genes at the gene expression level amplifies the underlying biological differences, thereby enabling the identification of subclusters (Figs 2E and 5F). It is precisely this perturbation strategy that supports the key advantage of scMagnifier. However, our perturbations do not explicitly model true post-perturbation cellular states, as they rely on GRN-based signal propagation at the transcriptomic level rather than learning distributional shifts induced by real perturbations. To address this limitation and enhance biological fidelity, future extensions could integrate scMagnifier with optimal transport-based perturbation models such as CellOT [52], which learn mappings between control and perturbed single-cell distributions, thereby improving the biological realism of perturbation effects while retaining scMagnifier's ability to amplify regulatory differences for fine-grained subcluster discovery.

scMagnifier effectively identifies heterogeneous cell populations, which form clearly separated clusters from the remaining cells in the rpcUMAP embedding (Figs 3A, 4A and 4D). This clear separation stems from two synergistic factors: first, the perturbation of TF genes amplifies the underlying regulation-informed transcriptional differences between cell subpopulations; second, unlike standard UMAP, rpcUMAP uniquely incorporates perturbation-derived intercellular distance information. By integrating this perturbation-based metric, rpcUMAP further exaggerates the dissimilarities between distinct clusters, thereby achieving the enhanced separation observed in the embedding. However, many cell states are defined by features beyond transcript abundance (such as protein markers, chromatin accessibility) and therefore may remain ambiguous in transcription-only analyses. Extending scMagnifier to integrate multimodal measurements via modality-aware distance fusion or multimodal GRN inference would increase sensitivity and specificity for cell type detection.

Methods

Data preprocessing

All datasets were prepared in the AnnData format and processed using Scanpy [14]. First, lowly expressed genes were filtered out by retaining only genes with at least one count across all cells. Expression values were then normalized per cell

to a total UMI count of 10,000. Highly variable genes (HVGs) were identified on the normalized non-log-transformed expression matrix, and the top 2,000 HVGs were retained. After subsetting the dataset to these HVGs, expression values were renormalized per cell to a total UMI count of 10,000 to account for the effects of gene filtering. Importantly, the renormalized non-log-transformed GEM was retained and used for downstream cluster-specific GRN construction and TF gene perturbation analyses. Subsequently, the expression values were log-transformed and scaled to a maximum value of 10. Principal component analysis was performed with 20 components, followed by the construction of a KNN graph (k = 10) and UMAP for dimensionality reduction. Finally, initial cell clustering results were obtained using the Leiden [19] or Louvain [20] algorithm.

Cluster-specific GRN construction

Cluster-specific GRNs were constructed following the standard CellOracle framework [7]. As prior regulatory information, we employed the human promoter-based GRN provided by CellOracle as our basic GRN, which defines potential TF gene-target gene relationships. Cluster-specific GRNs were constructed based on the basic GRN and initial clustering results, using the renormalized but non-log-transformed GEM stored during data preprocessing. Firstly, dimensionality reduction was performed using PCA (oracle.perform_PCA() function), followed by KNN-based imputation (oracle.knn_imputation() function) to alleviate data sparsity. Regulatory relationships between TF genes and their putative target genes were modeled using regression-based approaches (oracle.get_links() function), and cluster-specific regulatory links were further identified and filtered based on statistical significance and network scores (links.filter_links() function). Finally, quantitative GRN models for perturbation simulation were fitted (oracle.fit_GRN_for_simulation() function). All parameters were set according to the default or recommended settings in CellOracle.

TF genes perturbation on the level of gene expression

TF gene perturbation was performed on the renormalized but non-log-transformed GEM retained during data preprocessing ($\mathbf{X} \in R^{N \times G}$). This design ensured that perturbations were applied in the original linear expression space, thereby being consistent with the linear regulatory assumptions underlying the GRN models, which were inferred from the same renormalized non-log-transformed matrix.

Candidate TF genes for perturbation were selected by intersecting the TF genes from the basic GRN with the top 2,000 HVGs. Perturbations were applied in a cluster-wise manner: let the dataset be partitioned into C clusters according to the initial cluster results and let I_c denote the set of cell indices belonging to cluster c. For a selected perturbed TF gene g_0 and perturbation fold change $\mu (\mu \in \{-0.1, 0.1\})$, the initial, cluster-specific perturbation matrix $\Delta\mathbf{X}^{(0,c)} \in R^{N \times G}$ was defined by

$$\Delta X_{i,g}^{(0,c)} = \begin{cases} \mu \cdot X_{i,g_0}, & i \in I_c \text{and} g = g_0, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $X_{i,g}$ denotes the original expression of gene g in cell i. We applied a uniform ±10% fold-change for all candidate TFs. Sensitivity analysis (S12 Fig) indicates that this default value avoids introducing excessive noise while maintaining clustering stability.

For each cluster c, let $C^{(c)} \in R^{G \times G}$ denote the cluster-specific regulatory coefficient matrix. Propagation of the initial perturbation through the cluster-specific GRN was performed iteratively to capture multi-step regulatory effects. For iteration $k = 1, \ldots, n$ (we use $n = 3$ iterations), we computed

$$\Delta\mathbf{X}^{(k,c)} = \Delta\mathbf{X}^{(k-1,c)}\mathbf{C}^{(c)}. \tag{2}$$

To preserve the intended perturbation on the TF itself across iterations, the perturbed-TF column was restored to its initial values before proceeding to the next iteration.

After n iterations, the propagated perturbation for cluster c was $\Delta\mathbf{X}^{(n,c)}$. The overall propagated perturbation affecting all cells was obtained by summing cluster-wise contributions:

$$\Delta\mathbf{X}_{\text{prop}} = \sum_{c=1}^{C} \Delta\mathbf{X}^{(n,c)}. \tag{3}$$

The post-perturbation GEM was then formed by adding the propagated perturbation back to the original expression matrix:

$$\mathbf{X}_{\text{perturbed}} = \mathbf{X} + \Delta\mathbf{X}_{\text{prop}}, \tag{4}$$

and negative values were clipped to zero to maintain non-negativity. Finally, $\mathbf{X}_{\text{perturbed}}$ was processed through the downstream analytical pipeline consistent with that used in our data preprocessing to produce the post-perturbation clustering.

This procedure was repeated independently for each perturbation candidate TF gene producing a collection of clustering outcomes that together formed the perturbation-driven ensemble.

Consensus clustering

To integrate perturbation-induced clustering results and generate a consensus, we first transformed the clustering assignments into a one-hot matrix. For each clustering result, we assigned a "1" to the cluster a cell belongs to and "0" to other clusters, producing a sparse matrix for each clustering result. These one-hot matrices were then concatenated to form a high-dimensional matrix where each row represents a cell, and each column corresponds to a particular cluster assignment across all perturbation experiments. Next, we computed the pairwise cosine distance between cells based on this one-hot matrix (S9 Fig). The cosine distance between two cells i and j, was calculated as

$$D_{\text{onehot}}(i,j) = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|\|\mathbf{v}_j\|}, \tag{5}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the one-hot vectors for cells i and j.

In parallel, we calculated the Euclidean distance between cells in the embedding space (PCA-reduced space in single-batch scRNA-seq datasets) derived from the original GEM. The Euclidean distance between cells i and j in the k-dimensional ($k = 20$) embedding space was given by

$$D_{\text{emb}}(i,j) = \sqrt{\sum_{k=1}^{20} (X_{ik} - X_{jk})^2}, \tag{6}$$

where X_{ik} and X_{jk} are the k-th embedding values of cells i and j.

Both the one-hot distance matrix and embedding distance matrix were min-max normalized to ensure their scales were consistent by

$$\overline{D}(i,j) = \frac{D(i,j) - \min(\mathbf{D})}{\max(\mathbf{D}) - \min(\mathbf{D})}. \tag{7}$$

Then we computed a weighted sum of the normalized one-hot and embedding distances by

$$D_{\text{combined}}(i,j) = \alpha\overline{D}_{\text{emb}}(i,j) + (1-\alpha)\overline{D}_{\text{onehot}}(i,j), \tag{8}$$

where α is set as 0.8 by default (S7 and S8 Figs).

We next computed rpcUMAP with the combined distance matrix as the precomputed distance metric (via umap.umap_.UMAP() function) and set the number of nearest neighbors to 10. We further constructed a KNN graph (k = 10) from the combined distance matrix and applied the clustering algorithm (consistent with the clustering approach implemented in our data processing), with the resolution parameter specified as 1.5.

Cluster merging

We performed cluster merging to produce the final stable consensus clusters. This process included two stages: centroid-based merging and merging of small clusters.

In the first stage, we calculated the centroids of each cluster. The centroid of a cluster was computed from the HVG expression matrix, which corresponded to the final output of our preprocessing pipeline, where gene expression values had already been normalized, log-transformed and scaled. Specifically, the centroid of a cluster l was defined as the arithmetic mean of the HVG expression vectors of all cells assigned to that cluster:

$$\mathbf{c}_l = \frac{1}{|S_l|}\sum_{i \in S_l} \mathbf{X}_i, \tag{9}$$

where S_l represents the set of cells assigned to cluster l, $|S_l|$ is the number of cells in cluster l and $\mathbf{X}_i$ denotes the HVG expression vector of cell i.

The centroids were then used to calculate the pairwise Euclidean distances between them:

$$d_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|_2 = \sqrt{\sum_{k=1}^{n}(c_{ik} - c_{jk})^2}, \tag{10}$$

where c_{ik} and c_{jk} represent the expression values of gene k in the centroids $\mathbf{c}_i$ and $\mathbf{c}_j$ respectively, and n is the number of HVGs. Then we computed the median of the minimum nearest-neighbor distances for each cluster, and multiplied this value by a scaling factor (set as 0.75 by default, S10 Fig) to obtain the final distance threshold for merging similar clusters:

$$d_{\text{threshold}} = \text{median}(\min_{j \neq i} d_{ij}) \times \text{scaling factor}. \tag{11}$$

If the centroid distance between two clusters was smaller than this threshold, the clusters were merged.

In the second stage, we merged small clusters. Small clusters were defined as those containing fewer cells than a minimum threshold, set as a fraction of the total number of cells (default: 1% of total cells). For each small cluster, the nearest non-small cluster was identified based on the centroid distance, and the small cluster was merged with this centroid-nearest non-small cluster. This step was originally intended to merge boundary-associated cells misidentified as small clusters. However, if the algorithm is applied to the task of identifying rare cell populations, the threshold can be reduced (e.g., set 0.1% of total cells) to retain low-abundance cell populations. Notably, adjusting this threshold also enables flexible control over the final number of clusters obtained after merging.

Extension of scMagnifier for multi-batch datasets

scMagnifier can be adapted to handle multi-batch datasets by integrating it with batch effect correction methods that operate in the dimensionality reduction space, such as Harmony [15], Scanorama [16] or scVI [17]. Specifically, during the algorithm workflow, we make two key adjustments. First, during data preprocessing and TF genes perturbation, we

replace PCA-based dimensionality reduction with a batch effect correction method. Second, in the consensus clustering step, we use the embedding obtained from the batch effect correction method to compute cell distances in the reduced space, instead of using the PCA matrix. All other steps remain unchanged.

Comparisons between rpcUMAP and conventional UMAP

The traditional UMAP embeddings were computed directly from the preprocessed gene expression matrix using Scanpy. Specifically, the input matrix consisted of normalized expression values for highly variable genes, followed by PCA to reduce dimensionality. The UMAP algorithm then constructs a k-nearest neighbor graph based on Euclidean distances in PCA space and optimizes a low-dimensional embedding that preserves local neighborhood relationships. This standard UMAP captures the overall transcriptional variation driven solely by the biological information contained in the gene expression matrix. In contrast, rpcUMAP not only incorporates this gene-expression-driven information, but also integrates perturbation-informed cell-cell relationships. As a result, rpcUMAP provides a more informative visualization of how perturbations affect the cellular landscape, leading to clearer separation and interpretation of cell states compared with traditional UMAP.

Benchmarking setup and parameters

In our benchmarks, we ensured that the number of output clusters was kept consistent with the annotated cell-type count. Specifically, during the data preprocessing step, we adjusted the clustering resolution to match the number of clusters in the cell annotations. This resolution was maintained throughout the TF genes perturbation process. Finally, during the cluster merging stage, we controlled the final number of clusters by adjusting the minimum threshold to ensure it aligned with the annotated cell-type count. All other parameters across algorithms were kept at their default settings.

Ablation experiments

To evaluate the contribution of individual components of scMagnifier, we performed ablation experiments comparing four configurations: (i) full scMagnifier, (ii) without cluster merging, (iii) without GEM-derived distances and (iv) without perturbation-derived distances. These experiments were conducted on three datasets: LN_04, EBUS_10 and LUNG_N09. For each configuration, we ran the complete pipeline and computed clustering metrics including ARI, NMI and silhouette score.

During the comparison, the number of output clusters was controlled to match the annotated cell-type count. The ablation of GEM-derived distances was implemented by setting the weight parameter α in the consensus clustering step to 0, effectively removing contribution from GEM-based similarities. Conversely, the ablation of perturbation-derived distances was performed by setting $\alpha = 1$, effectively ignoring perturbation-driven distances. The removal of cluster merging simply skipped the cluster merging stage after consensus clustering.

Identifying differentially expressed genes

We used the Wilcoxon test implemented in SCANPY [14] to identify DEGs. For general analyses, DEGs were selected based on the ranking scores generated by the test combined with an adjusted p-value (Benjamin-Hochberg-corrected FDR) < 0.05. When analyzing whether rare cells constitute distinct subtypes, we selected strictly upregulated differentially expressed genes (DEGs) using dual thresholds: adjusted p-value (Benjamin-Hochberg-corrected FDR) < 0.05 and log fold change (logFC) > 0.25 (Fig 4C and 4F).

Pathway enrichment analysis

We first used the Wilcoxon test implemented in SCANPY [14] to identify DEGs. For each cluster, DEGs were sorted by adjusted p-value (ascending) and logFC (descending), and the top 200 DEGs were selected for enrichment analysis.

Subsequently over-representation analysis (ORA) was conducted using the enrichr function in the GSEApy [53] package for two functional databases: GO Biological Process 2021 and KEGG 2021 (Human), with a significance threshold of adjusted p-value < 0.05.

TF gene importance score calculation

To quantify how strongly each perturbed TF influences a given consensus cluster, we computed a per-gene, per-cluster importance score. Let G be the set of candidate TF genes and let I be the set of cells. For a perturbation of gene $g \in G$, we denoted post-perturbation GEM by $\mathbf{X}^{(g)} \in R^{|I|\times P}$ (P is the number of genes) and the original (renormalized but non-log-transformed) GEM by $\mathbf{X}$.

We first calculated a binary change indicator to quantify perturbation-induced cluster assignment shifts for each cell. For each perturbation gene g, we first established a label mapping between original and perturbed clusters to account for potential relabeling of identical clusters. Let L_i be the original label of cell i and $P_i^{(g)}$ the label of cell i after perturbing gene g. We defined the overlap matrix $\mathbf{M}^{(g)}$:

$$M^{(g)}(a,b) = \left|\left\{i : L_i = a \text{ and } P_i^{(g)} = b\right\}\right|. \tag{12}$$

We obtained a mapping $m^{(g)}$ from each original cluster a to a perturbed cluster by choosing the perturbed label with maximal overlap:

$$m^{(g)}(a) = \operatorname{argmax}_b M^{(g)}(a,b). \tag{13}$$

Using this mapping, we defined the per-cell binary change indicator

$$B_{g,i} = I\left(P_i^{(g)} \neq m^{(g)}(L_i)\right), \tag{14}$$

where $I(\cdot)$ is the indicator function. The binary matrix $\mathbf{B} \in \{0,1\}^{|G|\times|I|}$ collected $B_{g,i}$.

We then computed a continuous change metric to quantify the magnitude of global expression profile alterations for each cell under perturbation g. Let μ_j and σ_j be the mean and standard deviation (across cells) of $\log(1 + X_{:,j})$ for gene j, computed from the original GEM $\mathbf{X}$. We defined

$$z_{i,j} = \frac{\log(1+X_{i,j}) - \mu_j}{\sigma_j},\ z_{i,j}^{(g)} = \frac{\log(1+X_{i,j}^{(g)}) - \mu_j}{\sigma_j}. \tag{15}$$

The continuous change for cell i under perturbation g was

$$C_{g,i} = \|z_{i,\cdot}^{(g)} - z_{i,\cdot}\|_2 = \sqrt{\sum_j (z_{i,j}^{(g)} - z_{i,j})^2}. \tag{16}$$

The continuous matrix $\mathbf{C} \in R^{|G|\times|I|}$ collected $C_{g,i}$.

To combine the binary and continuous signals on a common scale, each row $\mathbf{C}_{g,\cdot}$ was min-max normalized across cells:

$$\overline{C}_{g,i} = \frac{C_{g,i} - \min_i C_{g,i}}{\max_i C_{g,i} - \min_i C_{g,i}}. \tag{17}$$

The two signals were then combined to yield a per-gene per-cell combined perturbation score

$$S_{g,i} = B_{g,i} + \overline{C}_{g,i}. \tag{18}$$

Given consensus cluster k with cell set I_k, the importance of TF gene g for cluster k is the mean of perturbation scores across cells in that cluster:

$$\text{score}_{g,k} = \frac{1}{|I_k|} \sum_{i \in I_k} S_{g,i}. \tag{19}$$

Jaccard coefficient calculation

To assess the transcriptional similarity between the rare cell clusters and biologically annotated cell populations, we quantified the overlap of their cluster-specific DEG sets using the Jaccard similarity coefficient. For each cluster, cluster-specific DEGs were first identified and the top 50 DEGs were selected to form distinct gene sets for each cluster. The Jaccard coefficient between the gene set of a rare cell cluster G_1 and the gene set of an annotated cell population G_2 was calculated using the following formula:

$$J = \frac{|G_1 \bigcap G_2|}{|G_1 \bigcup G_2|} \tag{20}$$

Integration of scMagnifier with STAGATE for spatial transcriptomics analysis

To integrate scMagnifier with STAGATE [47] for spatial transcriptomics analysis, we followed the same extension strategy as described for applying scMagnifier to multi-batch datasets. Specifically, STAGATE was used to generate a low-dimensional embedding that captures spatial context, and this embedding replaced the PCA space in the scMagnifier workflow. During data preprocessing and TF gene perturbation, the STAGATE-derived latent embedding was used for downstream analyses instead of PCA. In the consensus clustering step, cell-cell distances in the reduced space were computed based on the STAGATE embedding, while all other steps of the scMagnifier algorithm remained unchanged.

In our dataset, Spatial neighbor graphs in STAGATE were constructed using a KNN-based strategy ($k = 10$). STAGATE was trained for 300 epochs to obtain stable spatial embeddings. Downstream clustering was performed using the Leiden algorithm with a resolution of 0.3.

Supporting information

S1 Fig. Schematic diagram of cluster merging step in scMagnifier.
(TIF)

S2 Fig. Benchmarking of scMagnifier. (A) Bar plots illustrating the performance of nine algorithms on four single-batch datasets, evaluated using the silhouette score. (B) Bar plots illustrating the silhouette score for seven batch-correction methods applied alone or in combination with scMagnifier, evaluated on two multi-batch datasets. (C) Cell type mixing entropy of embeddings derived from scMagnifier, PCA, and scVI, evaluated on four single-batch datasets. Entropy of cell type mixing was computed based on neighbors in the corresponding latent space for each method: PCA-based neighbors for PCA, scVI-based neighbors for scVI, and mixed-distance neighbors for scMagnifier. (D) Bar plots depicting batch mixing entropy for seven batch-correction approaches applied alone or integrated with scMagnifier, assessed on two multi-batch datasets.
(TIF)

S3 Fig. UMAP and rpcUMAP visualization comparison of clustering results before and after cluster merging in six datasets (LN_04, EBUS_10, LUNG_N09, LUNG_N34, Pancreas and BMMC), where the merging operation was performed by adjusting the minimum threshold to align the number of clusters with the number of cell types (See "Cluster merging" section of the Methods).
(TIF)

S4 Fig. scMagnifier uncovers hidden heterogeneity among MAIT/Th1-Th17 cell populations (A) UMAP visualization of clustering results via the Leiden clustering algorithm with increasing resolution parameters (resolution = 0.7, 1.0, 1.3), highlighting MAIT cells and a Th1/Th17-MAIT mixed population were grouped into a single cluster. (B) Bar plot of the top 10 TF genes ranked by importance scores in cluster 2.
(TIF)

S5 Fig. scMagnifier detects rare cell populations in lung adenocarcinoma datasets (A) UMAP visualization of clustering results obtained via standard clustering method at high resolution (resolution = 1.5) in EBUS_10 and LUNG_N30 datasets, where rare cell populations fail to be identified as individual clusters. (B) UMAP visualization of rare cell identification results by GiniClust3. (C) Bubble plot showing Jaccard coefficient values between R1/R2 and clusters of the remaining cells grouped by cell-type annotations. (D) UMAP visualization of doublet detection results in the EBUS_10 dataset.
(TIF)

S6 Fig. Spatial visualization of clustering results after STAT2 gene perturbation using STAGATE (resolution = 0.3).
(TIF)

S7 Fig. Comparison of rpcUMAP visualization plots under different α parameters (0.0, 0.2, 0.4, 0.6, 0.8, 1.0) across three datasets (LN_04, EBUS_10 and LUNG_N09). As the α parameter decreases gradually, more small clusters emerge in rpcUMAP visualizations, resulting in an excessively fragmented clustering profile. Thus, we selected α = 0.8 as the default setting, which not only enables the separation of truly distinct clusters by incorporating perturbation-derived information but also avoids over-separation and the generation of numerous boundary small clusters that interfere with result interpretation.
(TIF)

S8 Fig. Comparison of rpcUMAP visualization plots under different α parameters (0.0, 0.2, 0.4, 0.6, 0.8, 1.0) across three datasets (LUNG_N34, Pancreas and BMMC).
(TIF)

S9 Fig. Comparison of rpcUMAP visualization plots in four datasets (LN_04, EBUS_10, LUNG_N09, LUNG_N34) using Euclidean distance or cosine distance to calculate intercellular distances of one-hot matrices. The visualization results derived from cosine distance are less fragmented than those from Euclidean distance, thus cosine distance was selected as the default setting.
(TIF)

S10 Fig. UMAP visualization of cluster merging results under different scaling factors (0.75, 1.0) in two datasets (EBUS_10, LUNG_N34). When the scaling factor = 1.0, clusters are over-merged, with different cell type clusters combined into one. Thus, 0.75 was selected as the default value to balance merging efficacy and cluster specificity.
(TIF)

S11 Fig. Performance Evaluation of scMagnifier Core Modules via Ablation Experiments.
(TIF)

S12 Fig. Sensitivity analysis of scMagnifier to perturbation fold-change using ARI, NMI and Silhouette Score.
(TIF)

S13 Fig. scMagnifier cluster stability across a range of resolutions. Plateaus indicate stable cluster numbers, with corresponding UMAPs confirming biologically meaningful structure. The resolution parameter in the consensus clustering step was varied from 0.2 to 2.0 with increments of 0.2 (all other parameters were kept at their default values), and the number of detected clusters was plotted as a function of resolution.
(TIF)

S1 Text. Supplementary data materials.
(PDF)

Author contributions

Conceptualization: Zhenhui He, Kangning Dong.

Data curation: Zhenhui He.

Formal analysis: Zhenhui He.

Funding acquisition: Kangning Dong.

Investigation: Zhenhui He.

Methodology: Zhenhui He, Kangning Dong.

Project administration: Kangning Dong.

Resources: Zhenhui He, Kangning Dong.

Software: Zhenhui He.

Supervision: Kangning Dong.

Validation: Zhenhui He.

Visualization: Zhenhui He.

Writing – original draft: Zhenhui He.

Writing – review & editing: Zhenhui He, Kangning Dong.

References

1. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015;58(4):610–20. https://doi.org/10.1016/j.molcel.2015.04.005 PMID: 26000846
2. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–82. https://doi.org/10.1038/s41576-018-0088-9 PMID: 30617341
3. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746. https://doi.org/10.15252/msb.20188746 PMID: 31217225
4. Imoto Y, Nakamura T, Escolar EG, Yoshiwaki M, Kojima Y, Yabuta Y, et al. Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. Life Sci Alliance. 2022;5(12):e202201591. https://doi.org/10.26508/lsa.202201591 PMID: 35944930
5. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017;14(11):1083–6. https://doi.org/10.1038/nmeth.4463 PMID: 28991892
6. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167(7):1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038 PMID: 27984732
7. Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA. Dissecting cell identity via network inference and in silico gene perturbation. Nature. 2023;614(7949):742–51. https://doi.org/10.1038/s41586-022-05688-9 PMID: 36755098
8. Li C, Shao X, Zhang S, Wang Y, Jin K, Yang P, et al. scRank infers drug-responsive cell types from untreated scRNA-seq data using a target-perturbed gene regulatory network. Cell Rep Med. 2024;5(6):101568. https://doi.org/10.1016/j.xcrm.2024.101568 PMID: 38754419

9. Osorio D, Zhong Y, Li G, Xu Q, Yang Y, Tian Y, et al. scTenifoldKnk: An efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. Patterns (N Y). 2022;3(3):100434. https://doi.org/10.1016/j.patter.2022.100434 PMID: 35510185
10. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6. https://doi.org/10.1038/nmeth.4236 PMID: 28346451
11. Cui Y, Zhang S, Liang Y, Wang X, Ferraro TN, Chen Y. Consensus clustering of single-cell RNA-seq data by enhancing network affinity. Brief Bioinform. 2021;22(6):bbab236. https://doi.org/10.1093/bib/bbab236 PMID: 34160582
12. Geddes TA, Kim T, Nan L, Burchfield JG, Yang JYH, Tao D, et al. Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis. BMC Bioinformatics. 2019;20(Suppl 19):660. https://doi.org/10.1186/s12859-019-3179-5 PMID: 31870278
13. Kim H, Park I, Park J-E, Kim JK, Seo M, Kim JK. scICE: enhancing clustering reliability and efficiency of scRNA-seq data with multi-cluster label consistency evaluation. Nat Commun. 2025;16(1):6031. https://doi.org/10.1038/s41467-025-60702-8 PMID: 40603842
14. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0 PMID: 29409532
15. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289–96. https://doi.org/10.1038/s41592-019-0619-0 PMID: 31740819
16. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019;37(6):685–91. https://doi.org/10.1038/s41587-019-0113-3 PMID: 31061482
17. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8. https://doi.org/10.1038/s41592-018-0229-2 PMID: 30504886
18. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11(1):2285. https://doi.org/10.1038/s41467-020-16164-1 PMID: 32385277
19. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233. https://doi.org/10.1038/s41598-019-41695-z PMID: 30914743
20. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008
21. Quah FX, Hemberg M. SC3s: efficient scaling of single cell consensus clustering to millions of cells. BMC Bioinformatics. 2022;23(1):536. https://doi.org/10.1186/s12859-022-05085-z PMID: 36503522
22. Schubert E, Sander J, Ester M, Kriegel HP, Xu X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Trans Database Syst. 2017;42:1–21.
23. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. WIREs Data Min & Knowl. 2011;2(1):86–97. https://doi.org/10.1002/widm.53
24. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst. 2016;3(4):346-360.e4. https://doi.org/10.1016/j.cels.2016.08.011 PMID: 27667365
25. Luecken MD, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2). 2021.
26. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36(3):964–5. https://doi.org/10.1093/bioinformatics/btz625 PMID: 31400197
27. Ippolito GC, Dekker JD, Wang Y-H, Lee B-K, Shaffer AL 3rd, Lin J, et al. Dendritic cell fate is determined by BCL11A. Proc Natl Acad Sci U S A. 2014;111(11):E998-1006. https://doi.org/10.1073/pnas.1319228111 PMID: 24591644
28. Wu L, Xue Z, Jin S, Zhang J, Guo Y, Bai Y, et al. huARdb: human Antigen Receptor database for interactive clonotype-transcriptome analysis at the single-cell level. Nucleic Acids Res. 2022;50(D1):D1244–54. https://doi.org/10.1093/nar/gkab857 PMID: 34606616
29. Wang H, Souter MNT, de Lima Moreira M, Li S, Zhou Y, Nelson AG, et al. MAIT cell plasticity enables functional adaptation that drives antibacterial immune protection. Sci Immunol. 2024;9(102):eadp9841. https://doi.org/10.1126/sciimmunol.adp9841 PMID: 39642244
30. Cao H, Diao J, Liu H, Liu S, Liu J, Yuan J, et al. The Pathogenicity and Synergistic Action of Th1 and Th17 Cells in Inflammatory Bowel Diseases. Inflamm Bowel Dis. 2023;29(5):818–29. https://doi.org/10.1093/ibd/izac199 PMID: 36166586
31. Villarino AV, Gallo E, Abbas AK. STAT1-activating cytokines limit Th17 responses through both T-bet-dependent and -independent mechanisms. J Immunol. 2010;185(11):6461–71. https://doi.org/10.4049/jimmunol.1001343 PMID: 20974984
32. Garner LC, Amini A, FitzPatrick MEB, Lett MJ, Hess GF, Filipowicz Sinnreich M, et al. Single-cell analysis of human MAIT cell transcriptional, functional and clonal diversity. Nat Immunol. 2023;24(9):1565–78. https://doi.org/10.1038/s41590-023-01575-1 PMID: 37580605
33. Jindal A, Gupta P, Jayadeva, Sengupta D. Discovery of rare cells from voluminous single cell expression data. Nat Commun. 2018;9(1):4719. https://doi.org/10.1038/s41467-018-07234-6 PMID: 30413715
34. Torre E, Dueck H, Shaffer S, Gospocic J, Gupte R, Bonasio R, et al. Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. Cell Syst. 2018;6(2):171-179.e5. https://doi.org/10.1016/j.cels.2018.01.014 PMID: 29454938
35. Dong R, Yuan G-C. GiniClust3: a fast and memory-efficient tool for rare cell type identification. BMC Bioinformatics. 2020;21(1):158. https://doi.org/10.1186/s12859-020-3482-1 PMID: 32334526

36. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. 2019;8(4):281-291.e9. https://doi.org/10.1016/j.cels.2018.11.005 PMID: 30954476
37. Pae J, Ersching J, Castro TBR, Schips M, Mesin L, Allon SJ, et al. Cyclin D3 drives inertial cell cycling in dark zone germinal center B cells. J Exp Med. 2021;218(4):e20201699. https://doi.org/10.1084/jem.20201699 PMID: 33332554
38. Liu Z, Xu L. UBE2S promotes the proliferation and survival of human lung adenocarcinoma cells. BMB Rep. 2018;51(12):642–7. https://doi.org/10.5483/BMBRep.2018.51.12.138 PMID: 30545437
39. Ma N, Fang Y, Xu R, Zhai B, Hou C, Wang X, et al. Ebi3 promotes T- and B-cell division and differentiation via STAT3. Mol Immunol. 2019;107:61–70. https://doi.org/10.1016/j.molimm.2019.01.009 PMID: 30660991
40. Su S-B, Tao L, Deng Z-P, Chen W, Qin S-Y, Jiang H-X. TLR10: Insights, controversies and potential utility as a therapeutic target. Scand J Immunol. 2021;93(4):e12988. https://doi.org/10.1111/sji.12988 PMID: 33047375
41. Li Z-Y, Morman RE, Hegermiller E, Sun M, Bartom ET, Maienschein-Cline M, et al. The transcriptional repressor ID2 supports natural killer cell maturation by controlling TCF1 amplitude. J Exp Med. 2021;218(6):e20202032. https://doi.org/10.1084/jem.20202032 PMID: 33857289
42. Abel AM, Yang C, Thakar MS, Malarkannan S. Natural Killer Cells: Development, Maturation, and Clinical Utilization. Front Immunol. 2018;9:1869. https://doi.org/10.3389/fimmu.2018.01869 PMID: 30150991
43. Limoges M-A, Cloutier M, Nandi M, Ilangumaran S, Ramanathan S. The GIMAP Family Proteins: An Incomplete Puzzle. Front Immunol. 2021;12:679739. https://doi.org/10.3389/fimmu.2021.679739 PMID: 34135906
44. Villagomez FR, Diaz-Valencia JD, Ovalle-García E, Antillón A, Ortega-Blake I, Romero-Ramírez H, et al. TBC1D10C is a cytoskeletal functional linker that modulates cell spreading and phagocytosis in macrophages. Sci Rep. 2021;11(1):20946. https://doi.org/10.1038/s41598-021-00450-z PMID: 34686741
45. Poli A, Michel T, Thérésine M, Andrès E, Hentges F, Zimmer J. CD56bright natural killer (NK) cells: an important NK cell subset. Immunology. 2009;126(4):458–65. https://doi.org/10.1111/j.1365-2567.2008.03027.x PMID: 19278419
46. Arora R, Cao C, Kumar M, Sinha S, Chanda A, McNeil R, et al. Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. Nat Commun. 2023;14(1):5029. https://doi.org/10.1038/s41467-023-40271-4 PMID: 37596273
47. Dong K, Zhang S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. Nat Commun. 2022;13(1):1739. https://doi.org/10.1038/s41467-022-29439-6 PMID: 35365632
48. Ren P, Zhang R, Wang Y, Zhang P, Luo C, Wang S, et al. Systematic benchmarking of high-throughput subcellular spatial transcriptomics platforms across human tumors. Nat Commun. 2025;16(1):9232. https://doi.org/10.1038/s41467-025-64292-3 PMID: 41107232
49. Yang Y, Yang Y, Yang J, Zhao X, Wei X. Tumor Microenvironment in Ovarian Cancer: Function and Therapeutic Strategy. Front Cell Dev Biol. 2020;8:758. https://doi.org/10.3389/fcell.2020.00758 PMID: 32850861
50. Dong Y, Li J, Han F, Chen H, Zhao X, Qin Q, et al. High IGF2 expression is associated with poor clinical outcome in human ovarian cancer. Oncol Rep. 2015;34(2):936–42. https://doi.org/10.3892/or.2015.4048 PMID: 26063585
51. Malhotra S, Kazlouskaya V, Andres C, Gui J, Elston D. Diagnostic cellular abnormalities in neoplastic and non-neoplastic lesions of the epidermis: a morphological and statistical study. J Cutan Pathol. 2013;40(4):371–8. https://doi.org/10.1111/cup.12090 PMID: 23398548
52. Bunne C, Stark SG, Gut G, Del Castillo JS, Levesque M, Lehmann K-V, et al. Learning single-cell perturbation responses using neural optimal transport. Nat Methods. 2023;20(11):1759–68. https://doi.org/10.1038/s41592-023-01969-x PMID: 37770709
53. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023;39(1):btac757. https://doi.org/10.1093/bioinformatics/btac757 PMID: 36426870