

Single-Cell Manifold Preserving Feature Selection (SCMER)

Shaoheng Liang 1, 2. sliang3@mdanderson.org

Vakul Mohanty 1. vmohanty@mdanderson.org

Jinzhuang Dou 1. jdou1@mdanderson.org

Qi Miao 1, 3. qmiao1@mdanderson.org

Yuefan Huang 1, 3. yhuang15@mdanderson.org

Muharrem Müftüoğlu 4. mmuftuoglu@mdanderson.org

Li Ding 5. lding@wustl.edu

Weiyi Peng 6. wpeng2@central.uh.edu

Ken Chen 1,*. kchen3@mdanderson.org

1 Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas, 77030, USA

2 Department of Computer Science, Rice University, Houston, Texas, 77005, USA

3 Department of Biostatistics & Data Science, School of Public Health, The University of Texas Health Science Center at Houston (UTHealth), Houston, Texas, 77030, USA

4 Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, Texas, 77030, USA

5 Department of Medicine, Washington University School of Medicine, St. Louis, MO, 63108

6 Department of Biology and Biochemistry, University of Houston, Houston, Texas, 77024

* Corresponding author

1 ABSTRACT

A key challenge in studying organisms and diseases is to detect rare molecular programs and rare cell populations (RCPs) that drive development, differentiation, and transformation. Molecular features such as genes and proteins defining RCPs are often unknown and difficult to detect from unenriched single-cell data, using conventional dimensionality reduction and clustering-based approaches.

Here, we propose a novel unsupervised approach, named SCMER, which performs UMAP style dimensionality reduction via selecting a compact set of molecular features with definitive meanings.

We applied SCMER in the context of hematopoiesis, lymphogenesis, tumorigenesis, and drug resistance and response. We found that SCMER can identify non-redundant features that sensitively delineate both common cell lineages and rare cellular states ignored by current approaches.

SCMER can be widely used for discovering novel molecular features in a high dimensional dataset, designing targeted, cost-effective assays for clinical applications, and facilitating multi-modality integration.

2 INTRODUCTION

A tissue in a living organism often consists of millions to billions of cells. While the terminally differentiated cells with relatively distinct molecular profiles can be readily distinguished via single-cell RNA sequencing (scRNA-seq) at current sampling depth, many cells involved in development, differentiation, and transformation remain difficult to detect^{1,2}. For example, a fraction of tumor cells in renal cell carcinomas can go through sarcomatoid transformation driven by epithelial to mesenchymal transformation (EMT)^{3,4}; tumor cells in pancreatic ductal adenocarcinomas can transiently express stemness features (e.g., SOX2) at its invasion fronts⁵⁻⁷. These cells can be relatively rare in the sampled

populations, transiently expressing certain molecular features and thereby may not form distinct clusters in high dimensional feature spaces^{8,9}.

To detect characteristic features (e.g., genes, proteins) in a single-cell dataset, studies^{8,10-13} often employ unsupervised clustering followed by one-cluster-vs-all differential expression (DE) analysis. These approaches can detect major cell types governed by lineage features that dominate data variance, but are insensitive to rare but unique features that have relatively small variance and manifest as level gradients within cell-type clusters (a.k.a. cell states)¹⁴. They are also clumsy at detecting features affecting multiple clusters, e.g., transcription factors (TFs) regulating multiple cell types¹⁵, as that involves comparison of an exponentially growing number of cluster combinations. To detect features associated with continuous developmental processes, many studies perform trajectory inference¹⁶ followed by regression analysis to identify correlated features (e.g., Monocle¹⁷). The selection of features depends on trajectories, which could be challenging to infer accurately for complex processes. A detailed comparison was performed by RankCorr¹² across various methods such as statistical tests, logistic regression, MAST¹⁰, scVI¹¹, and COMET¹³.

Most existing approaches regard features as independent variables without exploring their interactions¹⁸. As a result, they tend to identify redundant features (e.g. CD3D, CD3E and CD3G for T cells) that are highly correlated with the inferred clusters or trajectories, but ignore novel, uncorrelated features. Some recent work such as scHOT¹⁸ and SCMarker¹⁹ started to exploit correlational patterns among co- or anti-expressing genes. However, they do not model complex interactions of more than two genes. SCMarker cannot characterize continuous cell states, and scHOT relies on the accuracy of trajectory inference.

To enhance sensitivity in detecting rare features and RCPs, many studies^{20,21} had to slice and dice data spaces in empirical, multifaceted ways⁸ or perform iterative gating²² and re-clustering at variable resolutions, which may lead to biased, irreproducible results.

Increasing the number and variety of molecular features^{23,24} and improving the fidelity of the measurements can help discover RCPs. However, they unavoidably increase the already high cost of experiments. To make assays cost-effective towards clinical applications, it is important to select a compact actionable set of molecular features that unbiasedly represent molecular diversity in high dimensional data. This ability is important for designing and manufacturing customized assays, e.g., 10x targeted gene expression, MissionBio Tapestri²⁵ and NanoString GeoMx²⁶, which perform multi-omics measurements of hundreds of selected DNA, RNA, and proteins.

To address these fundamental challenges, we developed SCMER (Single-Cell Manifold Preserving Feature Selection), which selects an optimal set of features such as genes or proteins from a single-cell dataset. Similar to t-Distributed Stochastic Neighbor Embedding (t-SNE)²⁷ and Manifold Approximation and Projection (UMAP)²⁸, we hypothesize that a manifold defined by pairwise cell similarity scores sufficiently represents the complexity of the data, encoding both global relationship between cell groups and local relationship within cell groups²⁹. By preserving such a manifold while performing feature selection, the most salient features that unbiasedly represent the original molecular diversity will be selected.

SCMER does not require clusters or trajectories, and thereby circumvents the associated biases. It is sensitive to detect diverse features that delineate common and rare cell types, continuously changing cell states, and multicellular programs¹⁵ shared by multiple cell types. It reduces high dimensionality into a compact set of actionable features with definitive biological meanings. This distinguishes SCMER from PCA, t-SNE, UMAP, etc., which result in axes (meta-genes) with complex meanings. SCMER is efficiently implemented in Python using PyTorch³⁰, and supports multicore and GPU acceleration. A one-liner interface is provided for easy use. The open-source implementation is available at <https://github.com/KChen-lab/SCMER>.

3 RESULTS

3.1 THE SCMER APPROACH AND ITS UNIQUE STRENGTH

In a nutshell, SCMER (**Fig. 1a, Methods**) examines a data matrix \mathbf{X} of n cells and D features and calculates a pairwise cell similarity matrix \mathbf{P} (as defined in UMAP²⁸) representing the manifold in \mathbf{X} . It defines a weight vector \mathbf{w} , which transforms \mathbf{X} to $\mathbf{Y} = \mathbf{X}\mathbf{w}$. It then calculates another pairwise cell similarity matrix \mathbf{Q} from \mathbf{Y} and quantifies the level of agreement between \mathbf{P} and \mathbf{Q} using Kullback-Leibler (KL) divergence as defined in t-SNE²⁷. Finally, it uses elastic net to find a sparse and robust solution of \mathbf{w} that minimizes the KL-divergence with Orthant-Wise Limited Memory Quasi-Newton (OWL-QN) algorithm³¹. Features with nonzero weights in \mathbf{w} are deemed chosen. \mathbf{Q} can also be calculated from a different source (e.g., a different technology) instead of \mathbf{X} , which enables a “supervised” and multi-omics mode of SCMER.

A manifold encodes both clusters and continuums of cells. While clusters usually reflect distinct cell types, continuums reflect similar cell types and trajectory of transitioning/differentiating cell states³². SCMER selects optimal features that preserve the manifold and retain inter- and intra-cluster diversity (**Fig. 1b**). It can be applied to discover rich molecular pathways, identify prognostic genes, and design customized DNA/RNA/antibody panels of restricted sizes towards clinical applications.

To elucidate the novel cell populations and features that SCMER uniquely identifies, we generated a synthetic dataset containing a branching trajectory of 4,000 single cells from five major cell types, namely progenitor (Pro), precursor of A and B (PreA and PreB), and mature A and B (A and B) (**Fig. 1c**). Four kinds of features are simulated, those (I) specific to one cell type/cluster, (II) shared by more than one cell type¹⁵, (III) gradually changing over cell states, and (IV) transiently activated (also known as checkpoints³³). We created a total of 180 features including 20 cell-type specific features for each cell type (100 in total), 10 gradually changing features for each branch (20 in total), 5 shared features for

precursors and mature cells (10 in total), 5 transiently expressing features, and 45 random noise features.

Each cell was given a ground-truth pseudo-time in the trajectory, which parameterizes the expected level of a feature. In addition to major cell type labeling, the cells transitioning from precursor to mature are identified as “RCP-A” and “RCP-B”, which overexpress type-IV features. Dispersion was added based on a negative-binomial distribution. In as few as 45 selected features, SCMER recalled all the four types of features. In contrast, the top 45 features determined by a DE analysis are all from type I -- none belonged to type II, III, or IV, while a pseudo-time-based correlation analysis missed type-IV features. As a result, SCMER significantly increased the precision and recall of detecting RCPs, while being comparable to other methods on major cell types (using a k -NN classifier for one cell type at a time; **Table 1** and **Supplementary Note 1**).

Table 1. Precision/recall of detecting RCPs on the synthetic dataset

| Cell types | RCPs | | Major cell types | | | | | |
|-------------|-------------|-------------|------------------|-------------|-------------|-------------|-------------|--|
| | RCP-A | RCP-B | Pro | PreA | PreB | A | B | |
| Abundance | 2.55% | 2.68% | 21.23% | 22.43% | 19.83% | 16.73% | 15.30% | |
| SCMER | 0.82 / 0.68 | 0.87 / 0.67 | 0.97 / 0.96 | 0.95 / 0.96 | 0.94 / 0.94 | 0.95 / 0.96 | 0.94 / 0.93 | |
| DEG | 0.61 / 0.34 | 0.73 / 0.40 | 0.94 / 0.95 | 0.94 / 0.93 | 0.95 / 0.94 | 0.94 / 0.95 | 0.95 / 0.96 | |
| Correlation | 0.48 / 0.36 | 0.43 / 0.28 | 0.91 / 0.96 | 0.76 / 0.67 | 0.76 / 0.67 | 0.88 / 0.95 | 0.88 / 0.92 | |

To comprehensively assess SCMER, we ran it on eight datasets³⁴⁻⁴¹ (including **Supplementary Results**) that involve a variety of biological and technological challenges, such as unresolved borderline cells that blur clustering, continuously changing cell states, multicellular and transient cellular programs. For comparison, we used supervised DE analysis and widely-used unsupervised feature selection methods, including highly expressed genes (HXG), highly variable genes (HVG), SCMarker¹⁹, Monocle¹⁷, and RankCorr¹². SCMER robustly demonstrated the best performance on all the experiments.

3.2 SCMER CHARACTERIZATION OF CELL TYPE AND INTRATUMORAL HETEROGENEITY IN CANCER

Single-cell datasets derived from cancer samples are often highly complex, containing heterogeneous cell types and states in not only tumor cells but also stromal and immune cells. Supervised analysis of cancer data is challenging as cancer cells are highly plastic⁴² and can express novel unknown features, which can heavily confound clustering and trajectory-based analysis. We applied SCMER on a single-cell RNA-seq melanoma dataset containing 4,645 cells from 19 human melanoma samples³⁴. Most cells were annotated as malignant cells, B cells, T cells, macrophages, natural killer (NK) cells, endothelial cells, or cancer associated fibroblasts (CAFs) by the authors based on clustering and DE analysis. However, there were unresolved borderline cells presenting between labeled clusters, which resemble more than one cell types and could be either doublets or RCPs (Fig. 2a). By selecting only 75 genes (Table 2), SCMER clearly preserved the manifold: the resulting UMAP embedding is very similar to the original and the relations among most cell types including the unresolved cells are largely preserved (Fig. 2b and Supplementary Figure 2).

To understand the biological meanings of the selected genes, we compared them with the 11 gene sets described in the original publication that represent important cell types and pathways in the study. The selected genes compactly covered all the 11 gene sets (Table 2).

Table 2. Melanoma features selected by SCMER

| Category | Gene set | All | Highly variable | SCMER | Genes |
|---------------------------------------|-------------------------------------|-----|-----------------|-------|--|
| Resistance / intratumor heterogeneity | AXL | 100 | 52 | 2 | <i>HAPLN3, CD52</i> |
| | MITF | 100 | 28 | 3 | <i>TOB1, TYR, PMEL</i> |
| Cell types | Melanoma | 47 | 19 | 5 | <i>SERPINA3, PRAME, MIA, TYR, PMEL</i> |
| | B cells | 31 | 27 | 2 | <i>IRF8, MS4A1</i> |
| | T cells | 38 | 38 | 7 | <i>TCF7, TIGIT, CD8A, PRDM1, SPOCK2, NKG7, TOX</i> |
| | Macrophages | 92 | 73 | 1 | <i>TYROBP</i> |
| | Endothelial | 95 | 45 | 2 | <i>CDH5, HAPLN3</i> |
| | Cancer associated fibroblasts (CAF) | 88 | 46 | 3 | <i>COLIA2, COLIA1, MFAP4</i> |

| | | | | | |
|---------------|--|-----|-----|---|--|
| Pathways | Exhaustion (consistent across samples) | 28 | 17 | 3 | CD27, TIGIT, CXCL13 |
| | Exhaustion (variable across samples) | 272 | 159 | 7 | <i>HLA-DPA1, HAVCR2, CTLA4, SRGN, IRF8, PDCD1, TOX</i> |
| | Cell cycle | 93 | 54 | 3 | <i>TYMS, AURKB, MCM5</i> |
| None of above | <i>C10orf54, A2M, RGS10, MIR4461, IFITM1, SYTL3, GZMB, DEDD, RGS1, MAGEC2, PTPRC, NME2, CCL4, CCR7, HLA-DQA1, GNLY, OAS2, TXNIP, SELPLG, PAEP, CD69, ELK2AP, GZMH, ANXA1, MTRNR2L2, B2M, PTPRCAP, HLA-DQB1, LSP1, Clorf56, TCL1A, GPR183, CXCR4, RPS4Y1, APOD, KIT, GZMA, VIM, LTB, RPS11, TNFAIP3, CD55, UXSI</i> | | | | |

Interestingly, genes belonging to the known drug resistance AXL program and MITF program were also selected by SCMER. These genes do not preferentially express in a specific cluster (e.g. *PMEL*, *TOB1*, etc. in **Fig. 2d** and **Supplementary Figure 1a,b**).

SCMER also selected novel genes such as *TNFAIP3*, *VIM*, *COLIA1*, and *COLIA2*, which are involved in EMT. The original publication missed *TNFAIP3* and *VIM* and did not mention EMT. *COLIA1* and *COLIA2* were only reported as CAF specific markers. Also found were intra-cluster features for different kinds of immune cells (**Fig. 2e,f** and **Supplementary Figure 1c**) or shared by all cell types (**Supplementary Figure 1d**), which correspond to multicellular programs. *PDCD1* and *TCF7*, which are known to mark different subsets of T cells (exhausted versus memory) were detected, showing anti-correlated (Pearson's $r = -0.24$) expressions within the same cluster (**Supplementary Figure 1c**).

Pathway enrichment analysis⁴³ on the novel genes suggest that these genes are involved in innate immune response (*A2M*, *CD55*, *IFITM1*, *VIM*, *OAS2*, etc.), adaptive immune response (*CD55*, *CCR7*, *GPR183*, *TNFAIP3*, *PTPRC*, etc.), inflammatory abnormality of the skin (*KIT*, *CXCR4*, *TCL1A*, *B2M*, etc.), and other immune pathways (**Supplementary Tables 1 and 2**). These genes thus may be useful in further stratifying malignant and immune cell states (e.g., activation, exhaustion, etc.). Some genes such as *PMEL*, *PDCD1*, and *OAS2* appeared associated with survival outcome in TCGA SKCM patients⁴⁴ (**Fig. 1g,h,i**).

To comprehensively assess the performance of SCMER, we varied the number of selected features and recorded the number of recalled gene sets. SCMER consistently recalled more gene sets than other widely-used methods for any given number of features, demonstrating the best performance (**Fig. 2c**).

We also applied SCMER to a large-scale pan-cancer single-cell transcriptomic study consisting of 198 cell lines and patient samples from 22 cancer types³⁵. SCMER showed the high sensitivity to characterize intra-cluster heterogeneity, identifying recurrent heterogeneous programs shared by a majority of cell lines and by patient tumor samples (**Supplementary Figure 3** and **Supplementary Result 1**).

3.3 SCMER DEFINING CELL SUBTYPES AND STATES IN ILEUM LAMINA PROPRIA IMMUNOCYTES

We further examined SCMER in a complex setting involving many cell subtypes and subtle intra-cluster structure and shared pathways. The dataset contains 39,563 gastrointestinal immune cells collected from inflamed tissues from ten Crohn's disease patients³⁶. As a risk factor to cancer, chronic inflammation involves extensive interaction among various immune cell types such as helper T cells (T_H) and innate lymphoid cells (ILCs), which are regulated by both shared and cell-type specific TFs and cytokines and are difficult to delineate in high dimensional embeddings. The dataset appeared to include 27 cell types and subtypes/states in the original report. Four major cell types, T cells, B cells, phagocytes, and stromal cells each appeared as a cloud in the original embedding (**Fig. 3a**) but can be further dissected into subtypes. For example, T cells were dissected into eight subtypes/states through further clustering.

Circumventing clustering, SCMER selected 250 features from 3,573 highly variable genes (**Supplementary Tables 3 and 4**) with the manifold well preserved. The separability among cell types were comparable with the original embedding, and the manifold of subtypes (RCPs) in each major cell type were well preserved (**Fig. 3b**).

SCMER identified features delineating both clusters and sub-clusters. For example, the well-known lineage features such as *CD79A* (B cells) and *CD7* (T cells) and immune subtype markers such as *FCER2* (naïve B cells) and *ANKRD28* (TRM) were identified (**Fig. 3c,d** and **Supplementary Figure 4a**). Less reported features such as *SEPP1* for M2 macrophages were also among the list (**Supplementary Figure 4b**). The selected features also included genes that encode lysozyme (*LYZ*), complements (*C1QA*, *C1QB*, and *C1QC*), granulysin (*GLNY*), and granzymes (*GZMA*, *GZMB*, *GZMK*, and *GZMH*) (**Fig. 3c and Supplementary Figure 4c**).

NK and ILC1 cells were mixed together in one cluster and can hardly be further dissected based on unsupervised clustering and DE analysis. However, based on the genes selected by SCMER such as *GNLY*, *CCL4*, etc., which displayed dichotomizing levels within the cluster, we were able to further separate NK and ILC1 cells and estimate their abundances (**Supplementary Figure 5**).

SCMER also found TFs that regulate a wide range of cellular activities, including *JUN* and *FOS* (**Fig. 3d**), which are important for immune cell interactions. These features changed gradually among all the cell types, rather than expressing specifically in certain clusters. Other features such as *CD69* (known T cell activation feature) and *ODF2L* (novel T cell subtype feature) also showed gradual change among subtypes instead of clear-cut, on-and-off patterns (**Fig. 3f**). Notably, among our selected features that were not reported in the original publication, *DUSP1*, *DUSP2*, and *DUSP4* (**Fig. 3g and Supplementary Figure 4d**) were from a nucleus-predominant subfamily of dual-specificity phosphatases (DUSPs) family, which affect many cellular processes by regulating MAP kinases. *DUSP1* has been reported as a key regulator⁴⁵ of both innate and adaptive immune responses by inactivating p38 and JNK (c-Jun N-terminal kinase). We found that *DUSP2* and *DUSP4* showed gradients majorly in the T cell cloud, while *DUSP1* was expressed in all the major cell types. These genes are related to a wide range of immune phenotypes including producing inflammatory cytokines and autoimmune responses, which are highly relevant to Crohn's disease.

SCMER again compared favorably to the other methods that selected various numbers of features (**Fig. 3i, Supplementary Result 2**). It was evident that the other methods tended to ignore features associated with intra-cluster heterogeneity and multicellular programs. The novel genes selected by SCMER were also highly enriched in multiple immune pathways⁴³ (humoral immune response, leukocyte migration, complement activation, etc.; **Supplementary Table 5**). Overall, SCMER sensitively preserved different types and levels of heterogeneity in the original data.

3.4 SCMER DISSECTS CONTINUOUS CELL LINEAGE DIFFERENTIATION

Cell differentiation involves many unique patterns of gene expressions, including those gradually changing, shared among cell types, or transiently activating during the process. None of these patterns can be characterized through clustering. Having shown that SCMER identifies both inter- and intra-cluster features, here, we examine if it can identify key differentiation features in hematopoiesis in human bone marrow (BM). A recent study³⁷ sequenced 6,915 BM cells from four healthy donors. The transcriptomes of the cells formed a continuum instead of discrete clusters in the UMAP, reflecting the continuous differentiation process (**Fig. 4a**).

From this dataset, SCMER selected 100 features, which clearly preserved the trajectory of differentiation (**Fig. 4b**). The original publication reported 57 established features belonging to 12 sets (**Table 3**). SCMER recalled all the 12 sets as well as other well-known immune features (e.g. *GNLY* and *CD74*). SCMER also picked up features for less abundant cells that were not reported in the original publication (**Supplementary Table 6**), for example, *CLC* and *PRG2* for granulocyte and macrophage progenitor (GMP) cells (**Supplementary Figure 7a**). Neither of them was prioritized by DE analysis or Monocle¹⁷. Similarly, increasing *GPR183* expression was found in a subgroup of T cells, B cells, and Monocytes (**Fig. 4d**). The function of *GPR183* is not fully known, but emerging evidence shows that it may be a regulator of immune cell migration⁴⁶.

Table 3. Bone Marrow features Recalled by SCMER

| Category | Gene set | All | Highly variable | SCMER | Recalled genes |
|-----------------------------|----------------------|-----|-----------------|-------|---|
| Undifferentiated (HSC/Prog) | Undifferentiated | 7 | 3 | 3 | <i>MEIS1, MSI2, EGR1</i> |
| Myeloid | GMP-Promo | 4 | 4 | 2 | <i>MPO, ELANE</i> |
| | Promo-differentiated | 6 | 4 | 1 | <i>MNDA</i> |
| | Mono | 3 | 3 | 2 | <i>FCN1, C5ARI</i> |
| | cDC | 2 | 1 | 1 | <i>FCER1A</i> |
| | pDC | 6 | 6 | 4 | <i>TCF4, IRF8, JCHAIN, GZMB</i> |
| Erythroid | Ery | 5 | 3 | 2 | <i>HBD, HBB</i> |
| Lymphoid | Pro-B | 7 | 7 | 1 | <i>VPREB1</i> |
| | B | 2 | 2 | 1 | <i>MS4A1</i> |
| | Plasma | 3 | 3 | 2 | <i>MZB1, JCHAIN</i> |
| | T/CLT | 5 | 5 | 2 | <i>CD3D, IL7R</i> |
| | CLT/NK | 7 | 7 | 2 | <i>CCL5, GZMB</i> |
| None of above | | | | | <i>ETSI, HIST1H1C, ATP2B1, HIST1H4C, SOX4, NPR3, HLA-DPA1, GPR183, CTSS, NEAT1, RNASE2, KIAA0087, DUSP1, JUN, EREG, HLF, AREG, CDK6, HSP90B1, FYB1, HBA1, HLA-B, PRTN3, VCAN, HMGN2, CD52, NUSAPI, PRKG2, SLC40A1, TOP2A, HIST1H1B, AHSP, STK17B, S100A6, IGLL1, THBS1, HLA-DRA, CA1, APOOL, HOPX, CALR, KLF6, HMGB2, TXNDC5, PDZD8, FGL2, BCL11A, IQGAPI, PRSS2, HLA-E, CYBB, S100A12, NRIP1, GNLY, SAMSNI, PTPRC, CENPF, PIK3R1, SAMHD1, CENPU, HIST1H2AJ, MT-ND5, PRG2, PSAP, ANXA1, CD74, ASPM, STMN1, CLC, DNTT, S100A10, NAMPT, CSTA, CST3, TXNIP, ZFAS1, SPINK2, SAT1, MKI67</i> |

SCMER identified many genes that displayed gradient along the developmental trajectories, for example, *AHSP* and *CA1* for Erythroid cells, and *VPREB1* for B cells (**Supplementary Figure 7b**). It also identified genes such as *PRTN3* (monocytes) and *PDZD8* (erythroid) that appeared transiently expressed during the developmental process and became dim in terminally differentiated cells (**Fig. 4d**), which were not prioritized by Monocle. Besides, it identified TF genes such as *JUN* and *SOX4*, which play important roles in regulating cell differentiation^{47,48}. We comprehensively evaluated the performance and confirmed that SCMER outperformed the other unsupervised methods in recapitulating molecular diversity (**Fig. 4c**).

3.5 SCMER IDENTIFIES MOLECULAR DRIVERS FROM PERTURBED CELLS

More and more studies using single-cell technologies to investigate heterogeneity of cells in response to a genetic or chemical perturbation⁴⁹. In these experiments, cell state may transition under complex kinetics. To investigate the utility of SCMER in studying cellular responses, we applied it on single-cell data derived from dexamethasone (DEX) treated A549 lung adenocarcinoma cell line³⁸. As reported in the

original publication, the 1,429 cells sampled at 0, 1, and 3 hours after the DEX treatment formed a continuum in the transcriptomic space (**Fig. 5a**), indicating heterogeneous responses of the cell population. After running SCMER on the sci-RNA-seq data, 80 genes were selected, with the manifold and treatment states largely preserved (**Fig. 5b**).

We inferred TF activities based on motif enrichment⁵⁰ in the chromatin accessibility (sci-ATAC-seq) data co-assayed on the same set of cells³⁸ (**Methods**, **Fig. 5c**). Among the top 50 highly variable TFs (**Supplementary Figure 8a**), *NR3C1*, the primary target of DEX³⁸, had the most prominently increasing activity level over treatment time. Other TFs such as *FEV*⁵¹ and the ETS family⁵², also targets of DEX, had decreasing activity levels.

We then correlated the expression levels of the genes selected by SCMER with the activity levels of the top TFs. We found that *FKBP5*, *GALNT18*, *NRCAM*, etc. were positively correlated with *NR3C1*, while *CYP24A1*, *COL5A2*, etc. were negatively correlated (**Supplementary Table 7**, **Supplementary Figure 8**). In particular, *FKBP5*, a factor in the negative feedback loop of glucocorticoid receptor response and regulator of immune processes^{53,54}, had the highest positive correlation ($r = 0.355$) in the whole transcriptome; while *CYP24A1*, which regulates multiple metabolism processes⁵⁵, was the most negative ($r = -0.365$). Cells of high *FKBP5* expression levels came mostly from 1 and 3 hours (**Fig. 5d**), with matched polarized distributions in the RNA and the ATAC embeddings (**Fig. 5g**). Similar patterns were observed between cells of high and those of low *CYP24A1* expression levels (**Fig. 5f,i**).

Interestingly, SCMER also selected a group of genes uncorrelated with prominent TF activities (**Fig. 5j**, **Supplementary Figure 8**). Among them were *MKI67* (e.g., $r = -0.005$ with *NR3C1*) (**Fig. 5e,h**), which encodes proliferation marker protein Ki-67, and other cell-cycle genes such as *CENPF*, *TOP2A*, *RYBP*, *MLH3*, etc. Pathway analysis confirmed that these genes are highly enriched in cell proliferation pathways (**Supplementary Table 8**), indicating that an appreciable fraction of cells continued proliferating despite the treatment. It is not surprising that the levels of these genes were uncorrelated

with chromatin state changes, as it has been shown that cell cycling status has little direct effect on chromatin accessibility⁵⁶. Also among uncorrelated were several cancer cell stemness marker genes⁴³ such as *ACTG1*, *TSC22D1*, and *FNI*, which may indicate that a fraction of cancer cells maintained their stemness during the course of the treatment. These genes would have been missed by a DE analysis supervised by the treatment time.

Taken together, our results demonstrated the superior power of SCMER in discovering features associated with heterogeneous cellular state change in the context of perturbation experiments.

3.6 SCMER ACHIEVES CROSS-MODALITY FEATURE MAPPING

One challenge in applying scRNA-seq for cell-typing is that expression levels of mRNAs can differ substantially from those of homologous proteins, due to post transcriptional modifications⁵⁷. Although performing multi-omics assays may be the ultimate solution, they are currently associated with higher cost and lower throughput. Thus, rather than simply selecting the homologous mRNAs, it is beneficial to identify the set of genes whose expression levels maximally represent cellular diversity at the protein level. This capability can be important for designing targeted, cost-effective assays for preclinical and clinical applications. SCMER is ideally suited for such a purpose, as it allows selecting features in one modality while preserving manifold in another modality.

We ran SCMER on a CITE-seq dataset containing 14,468 bone marrow mononuclear cells (BMNC)³⁹. The protein manifold based on 25 markers was utilized to “supervise” the selection of mRNAs (**Methods**). CITE-seq, which co-assays mRNA and protein markers from the same set of cells, is ideal for obtaining the optimal mapping between mRNAs and proteins (**Fig. 6a,b**).

As shown, the mRNA expression levels of genes homologous to the protein markers, such as *CD4* (a T_h cell marker) and *NCAM1* (CD56, an NK cell marker) offered low power in delineating the corresponding cell types (**Fig. 6d,e**). Some markers, e.g., *CD45RA* (B cells and naïve T cells) and *CD45RO* (memory T

cells) are isoforms of the same gene, *PTPRC*. Consequently, T cell subtypes were less distinguishable in the RNA space than in the protein space (**Fig. 6b**). The differences among CD8 T cell subtypes were even bigger than the differences between CD4 and CD8 T cells.

SCMER selected a set of genes that best preserved the diversity at the protein-level, notably the continuum among naïve CD8 T cells, memory CD8 T cells, and effector CD8 T cells (**Fig. 6c**). It identified genes that are non-homologous to the protein markers but better represent the protein level difference, for example, *GPR183*, *KLRF1*, *CD79B*, and *S100A4* for CD4, CD56, CD45RA, and CD45RO, respectively (**Fig 6d,f**). On the other hand, the SCMER result appeared to better delineate progenitor cells than the protein markers, which demonstrates a strength of integrating complementary modalities.

Similar conclusions were drawn when applying SCMER on another smaller PBMC CITE-seq dataset⁴⁰ with 10 protein markers (**Supplementary Result 3**).

Importantly, the genes selected by SCMER from one donor (14,468 cells) appeared to preserve the cell diversity in another donor (16,204 cells) (**Supplementary Figure 10e-f**), which validated the applicability of SCMER in designing targeted panels for populational level testing.

4 DISCUSSION

SCMER was designed to meet an important need in single-cell molecular data analysis, to sensitively identify non-redundant features that delineate both common cell lineages and rare cellular states ignored by current approaches. It provides an *ab initio* approach for discovering novel genes and features in high dimensional datasets, designing cost-effective assays for potential clinical applications, and assisting multi-modality integration of gene expression, proteins, and other features.

SCMER does not require clusters or trajectories and is not affected by uncertainties in clustering or trajectory inference. It explores alternative explanations via feature selection and reports the most salient

features representing different facets of cells and underlying molecular activities. As a result, on datasets involving hematopoiesis, lymphogenesis, tumorigenesis, and drug resistance and response, SCMER identified features representing major cell types, rare cell populations (RCPs), continuous cell states, and multicellular programs. In our study, it prevailed existing unsupervised methods and often performed better than or comparably to the supervised methods when accurate labeling was possible (**Supplementary Figure 11**). Moreover, SCMER can handle batch effects by treating batches as a stratum, and finding a consensus set of features that preserve the manifold in respective batches (**Methods**). In that manner, it will prioritize genes contributing to biological but not technical variances.

SCMER can run in various supervised modes. It can accept a manifold from a different modality, for example, selecting RNA features under the guidance of a protein manifold (**Supplementary Result 3**). It can fix features preselected by users and find the best “partner” features (**Supplementary Result 4**) or select features from a shortlist (**Supplementary Result 5**). The framework appears effective on cell line and patient data generated by various technologies, including scRNA-seq and mass cytometry⁴¹ (**Supplementary Result 6**). This type of integrative analysis can potentially be extended to other modality combinations such as scRNA with scATAC, or mRNA with miRNA.

SCMER is an efficient method based on the orthant-wise limited memory quasi-Newton (OWL-QN) algorithm³¹. On a dataset with 10,000 cells and 2,000 candidate features, it typically converges in 20 to 40 iterations, which takes 5 to 10 minutes on a desktop computer equipped with a 3.20GHz 6-core Intel Core i7-8700 CPU. GPU acceleration is also supported, and the time consumption is halved with a middle-end NVidia GTX 960M GPU on a laptop computer with a 2.7GHz 4-core Intel Core i7-5700HQ CPU.

Because SCMER detects informative features that represent much wider and more complex biological processes than current methods, we expect it to be of immediate interest in projects producing large numbers of unsorted cells, such as the Human Cell Atlas⁵⁸, the Human BioMolecular Atlas Program (HuBMAP)⁵⁹, the Precancer Atlas⁶⁰ and the Human Tumor Atlas Network⁶¹. It will be beneficial in

various scenarios including biomarker discovery and clinical assay designing. As a first-of-its-kind method designed for manifold preserving feature selection on biomedical data, it can potentially be broadly applied to non-single-cell data, for example, bulk RNA expression²⁹, copy number aberration⁶², and genetic and drug screening data in large cohort studies such as TCGA⁶³, GTEx⁶⁴, Depmap⁶⁵, CTRP⁶⁶, etc.

5 METHODS

5.1 CELL-CELL SIMILARITY

SCMER is inspired by three methods: Stochastic Neighbor-Preserving Feature Selection (SNFS)⁶⁷, t-distributed stochastic neighbor embedding (t-SNE)²⁷ and Uniform Manifold Approximation and Projection (UMAP)²⁸.

t-SNE is one of the most widely used method for data embedding. For a dataset $\mathbf{X} \in \mathbb{R}^{n \times D}$ with n cells and D features, the similarity of a cell i to another cell j is defined as

$$p_{ij} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)}{\sum_{k \neq l} \exp(-\|\mathbf{x}_k - \mathbf{x}_l\|^2/2\sigma^2)},$$

which comprises a cell-cell similarity matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$. σ is a scaling factor. It creates an d -dimensional embedding $\mathbf{Y} \in \mathbb{R}^{n \times d}$. It calculates another cell-cell similarity matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ for \mathbf{Y} , whose entries are

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}.$$

The cost function is defined as the Kullback-Leibler (KL) divergence of \mathbf{P} and \mathbf{Q} , formally

$$C = KL(\mathbf{P} \parallel \mathbf{Q}) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Because UMAP is more sensitivity to both global relationship between cell groups and local relationship within cell groups²⁹, we borrowed a part of the UMAP formulation, i.e.,

$$p_{ij} = \frac{\exp(-(\|\mathbf{x}_i - \mathbf{x}_j\| - \rho_i)/\sigma_i)}{\sum_{k \neq l} \exp(-(\|\mathbf{x}_k - \mathbf{x}_l\| - \rho_i)/\sigma_i)}, q_{ij} = \frac{(1 + (\|\mathbf{y}_i - \mathbf{y}_j\| - \tau_i)/\sigma_i)^{-1}}{\sum_{k \neq l} (1 + (\|\mathbf{y}_k - \mathbf{y}_l\| - \tau_i)/\sigma_i)^{-1}},$$

where $\rho_i = \min\|\mathbf{x}_i - \mathbf{x}_j\|$ and $\tau_i = \min\|\mathbf{y}_i - \mathbf{y}_j\|$. The scaling factor σ_i is chosen such that

$\sum_j \exp(-(\|\mathbf{x}_i - \mathbf{x}_j\| - \rho_i)/\sigma_i) = \log_2 k$, which may be viewed as constructing a soft nearest neighbor graph. We default it to 100 in our experiments. Similar to UMAP, setting it in the range 10 to 1,000 gives very similar results²⁸.

5.2 MARKER SELECTION BY ELASTIC NET

Different from t-SNE and UMAP, instead of allowing \mathbf{Y} to be an arbitrary matrix, we require each column of \mathbf{Y} to be directly taken from a column of \mathbf{X} , i.e., to select a feature. To formally model this procedure.

We use a vector $\mathbf{w} \in \mathbb{R}^D$ to indicate the selection of the features, where 0 means unselected, and set

$$\mathbf{Y} = \mathbf{X}\mathbf{w},$$

which set all unselected features to zero in \mathbf{Y} . In terms of calculating the distances, zeroing out the columns is effectively discarding them. Thus, the definition of \mathbf{Q} is unchanged. Ideally, to select d features, we optimize

$$\min_{\mathbf{w}} C \text{ subject to } \|\mathbf{w}\|_0 = d,$$

where $\|\mathbf{w}\|_0$ is the l_0 -pseudo-norm, i.e., the number of nonzero entries. However, this question is known to be NP-hard, whose determination requires checking all the $\binom{D}{d}$ possibilities. Thus, we fall back to l_1 -norm, the convex approximation of l_0 -pseudo-norm, as in

$$\min_{\mathbf{w}} C + \lambda \|\mathbf{w}\|_1,$$

where l_1 -norm $\|\mathbf{w}\|_1 = \sum_i |w_i|$ and λ is the strength of the regularization. We denote the loss function as

L . The number of chosen features decreases with larger λ . Thus, for a given d , we use a binary search to find a λ . Due to limitations of precision, the specific d may not always be achievable. In that case, we allow for a few more features to be selected, and discard those that are assigned with the lowest weights (Supplementary Note 2). In the result, the features who have nonzero weights in \mathbf{w} are considered selected. The specific weight is not used in downstream analysis (Supplementary Note 2).

The cost, $C = KL(\mathbf{P} \parallel \mathbf{Q})$, is a robust indicator of whether the manifold is successfully retained. A typical range of C is 2.0 – 4.0 when the manifold is reasonably retained. More features (i.e., smaller l_1 -regularization) may be needed if the C is greater than 4.0.

Our model also allows an additional l_2 -regularization (ridge) to form an elastic net model. It may improve the robustness of the panel by slightly increase the redundancy, so that noise or drop-out in one feature has less effects.

5.3 BATCH EFFECT CORRECTION BY STRATIFICATION

Batch effect is a common problem in experiments including multiple samples. For SCMER, the samples are considered a stratum. In specific, a set of \mathbf{P} and \mathbf{Q} can be constructed for each sample, denoted as $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$, while \mathbf{w} is shared by all samples. A cost $C^{(i)}$ can thus be calculated for each sample, and collectively form a new objective $C = \sum_i C^{(i)}$. Thus, SCMER will ignore features that identify different samples and focuses on features that retain cell-cell similarities in all/most samples.

5.4 SUPERVISED MULTI-OMICS MODE

To transfer the manifold in one matrix (\mathbf{X}) to another (\mathbf{X}'), either between different modalities or subsets of features of the same modality, we simply modify the definition of \mathbf{Y} to $\mathbf{Y} = \mathbf{X}'\mathbf{w}$. With all other procedures unchanged, the algorithm is now searching for features in \mathbf{X}' that gives a manifold similar to that of \mathbf{X} . This is also applicable to select features from a shortlist of the original ones.

5.5 USING PRESELECTED FEATURES

In the case that a researcher wants to specify a few features that are known to be useful, we slightly modify the regularization to $\lambda \|\mathbf{V}\mathbf{w}\|_1$, where $\mathbf{V} = \text{diag}(\mathbf{v})$ is a diagonal matrix. If a feature is considered important *a priori*, the corresponding entry in \mathbf{v} is set to 0 to avoid l_1 -regularization. In this “softly-supervised” way, SCMER is more likely to select these features, but may still discard some of them if they are contradicting with the manifold. Thus, in addition, we provide a “hard-supervised” way where a set of features are guaranteed to be kept. Other features are selected to supplement them.

5.6 ORTHANT-WISE LIMITED MEMORY QUASI-NEWTON ALGORITHM

Limited-memory BFGS (L-BFGS) is a widely-used optimization algorithm in the quasi-Newton methods family⁶⁸. It approximates the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm with $O(mD)$ memory, where m can be chosen based on computing resources.

Although L-BFGS usually converge very fast (<20 iterations) for most l_2 -regularized regression problems, it will diverge for l_1 -regularization, whose partial derivative is undefined at $\{\mathbf{w} \mid w_i = 0 \ \exists i\}$:

$$\frac{\partial \|\mathbf{w}\|_1}{\partial w_i} = \frac{\partial \sum_i |w_i|}{\partial w_i} = \frac{\partial |w_i|}{\partial w_i} = \begin{cases} 1 & w_i > 0 \\ \text{undefined} & w_i = 0 \\ -1 & w_i < 0 \end{cases}$$

It should be noted that setting the undefined point to 0 (or any other value) at $w_i = 0$ does not solve the problem as the discontinuity will also break L-BFGS. A modified version of L-BFGS called orthant-wise limited memory quasi-Newton (OWL-QN) algorithm³¹ solves this problem by introducing pseudo-gradients and restrict the optimization to an orthant without discontinuities in the gradient.

L-BFGS optimizer is provided in PyTorch³⁰, in which SCMER is implemented. Based on it, we implemented a special case of OWL-QN algorithm for optimization of the model. Two modifications we made are as follows.

Firstly, we derive the pseudo-gradient, where the pseudo-partial derivative at a discontinuity \mathbf{w}_0 of the loss function $L = C(\mathbf{w}) + \lambda \|\mathbf{w}\|_1$ is defined as

$$\hat{\partial}_i L(\mathbf{w}_0) = \begin{cases} \partial_i^- L(\mathbf{w}_0) & \partial_i^- L(\mathbf{w}_0) > 0 \text{ and } \partial_i^+ L(\mathbf{w}_0) > 0 \\ \partial_i^+ L(\mathbf{w}_0) & \partial_i^- L(\mathbf{w}_0) < 0 \text{ and } \partial_i^+ L(\mathbf{w}_0) < 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{\partial}_i L(\mathbf{w}_0)$ is the pseudo partial derivative and $\partial_i^- L(\mathbf{w}_0)$ is the short hand of

$\lim_{w_i \rightarrow (\mathbf{w}_0)_i^-} \frac{\partial L}{\partial w_i} \Big|_{w_k=(\mathbf{w}_0)_k}$, i.e, the left limit of the partial derivative. Similarly, $\partial_i^+ L(\mathbf{w}_0)$ is the right limit.

Note that the gradient of $C(\mathbf{w})$ is continuous, i.e., $\partial_i^- C(\mathbf{w}_0) = \partial_i^+ C(\mathbf{w}_0) = \partial_i C(\mathbf{w}_0)$. Thus,

$$\hat{\partial}_i L(\mathbf{w}_0) = \begin{cases} \partial_i C(\mathbf{w}_0) - \lambda & \partial_i C(\mathbf{w}_0) - \lambda > 0 \\ \partial_i C(\mathbf{w}_0) + \lambda & \partial_i C(\mathbf{w}_0) + \lambda < 0 \\ 0 & -\lambda \leq \partial_i C(\mathbf{w}_0) \leq \lambda \end{cases}$$

In fact, for L , discontinuities are $\{\mathbf{w} \mid w_i = 0 \ \exists i\}$.

Secondly, we confine the search area in each quasi-Newton optimization step so that it does not cross any discontinuity. Specifically, for our problem where all discontinuities are at 0, when updating \mathbf{w}^t to \mathbf{w}^{t+1} , we reset the value of w_i^{t+1} to 0 if $\text{sign}(w_i^{t+1}) \neq \text{sign}(w_i^t)$. It constrains the optimization to be in the same “orthant” in each iteration.

5.7 DATA PREPROCESSING

For the melanoma data³⁴, which is TPM based, after removing ERCC spike-ins, we processed the data using the standard workflow of SCANPY⁶⁹, including quality control (filtering out genes that are detected in less than 3 cells), normalization (10,000 reads per cell), log transformation, highly variable genes detection (with a loose threshold to filter out noisy genes; not to be confused with the DXG we compared with), and scaling.

For the Ileum Lamina Propria Immunocytes data³⁶, bone marrow data³⁷, and A549 data³⁸, which are UMI based, we used the standard workflow of SCANPY, including quality control (filtering out genes that are

detected in less than 3 cells), normalization (10,000 reads per cell), log transformation, highly variable genes detection, and scaling. We used the stratified approach to suppress batch effect on the Ileum Lamina Propria Immunocytes data.

For protein data in CITE-Seq^{39,40}, we followed the preprocessing of protein data described in the original publication. For mRNA data in CITE-seq, we follow the standard workflow of SCANPY, as described above, except that we did not filter highly variable genes. We preprocessed protein data as mRNA data, without filtering highly variable genes.

5.8 INFERENCE OF TF ACTIVITIES

Because TFs tend to bind at sites with cognate motifs, accessibility at peaks with the motifs reflects their activity. To estimate transcription factor activity from sci-ATAC-seq data, we use chromVAR⁵⁰ package with the default setting. It quantifies accessibility variation across single cells by aggregating accessible regions containing a specific TF motif. The observed accessibility of all peaks containing a TF motif is compared with a background set of peaks normalized for known technical confounders.

5.9 COMPARISON WITH OTHER METHODS

To identify the highly expressed genes (HXG), we used the standard SCANPY⁶⁹ workflow. HXG is defined by the total reads of a gene across all cells. To identify the highly variable genes, we followed the standard scoring method in SCANPY⁶⁹.

SCMarker¹⁹ provides a gene list without ranks. It has two parameters, n and k , which affect the number of resulting features. Based on our observation, n has a minor effect on the result. Thus, we fixed $n = 50$ and tested k from 10 to 1,200 to create feature gene lists of various sizes.

We ran Monocle¹⁷ in unsupervised and supervised manners. For the supervised run, the labels were used directly. The trajectory was inferred using clusters/labels and pseudo-time is calculated. Genes were

ranked by the degree they are explained by functions (which were fitted with cubic splines) of pseudo-time. For the unsupervised run, we clustered the cells and visually confirmed the clusters are concordant with the labels.

We ran RankCorr¹² in both supervised and unsupervised manner. For the supervised run, we used the label from the data directly. For the unsupervised run, we used the Leiden algorithm⁷⁰ for clustering which is the recommended method in SCANPY. Default parameters were used, and the clusters are visually checked that they are reasonable.

For random results, we randomly selected gene sets of given sizes. Reported are mean performance and the critical level of statistically significantly better (or worse) than random as defined by single-sample one-sided z-test at 5% significance level.

6 ADDITIONAL INFORMATION

6.1 ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable in this study.

6.2 CONSENT FOR PUBLICATION

Not applicable in this study.

6.3 AVAILABILITY OF DATA AND MATERIAL

The open source implementation of SCMER available at <https://github.com/KChen-lab/SCMER> under the MIT License. Scripts for reproducing all the results are also included. All original datasets are accessible through the original publications³⁴⁻⁴¹.

6.4 COMPETING INTERESTS

The authors declare that they have no competing interests.

6.5 FUNDING

This project has been made possible in part by the Human Cell Atlas Seed Network Grant (CZF2019-002432 and CZF2019-02425) to KC from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation, grant RP180248 to KC and grant RP200520 to WP from Cancer Prevention & Research Institute of Texas, grant U01CA247760 to KC, grant U24CA211006 to LD, and the Cancer Center Support Grant P30 CA016672 to PP from the National Cancer Institute.

6.6 ACKNOWLEDGEMENTS

The authors would like to thank Hussein Abbas, Yuanxin Wang, Linghua Wang for their comments.

The authors acknowledge the support of the High Performance Computing for research facility at the University of Texas MD Anderson Cancer Center for providing computational resources that have contributed to the research results reported in this paper.

6.7 AUTHORS' CONTRIBUTIONS

SL, MM, WP, LD, and KC conceptualized the project. SL designed the SCMER algorithm and implemented the software. All authors collectively designed the experiments and analyzed the results. All authors drafted the manuscript. All authors have read and approved this paper.

7 REFERENCES

1. Merrell, A. J. & Stanger, B. Z. Adult cell plasticity in vivo: de-differentiation and transdifferentiation are back in style. *Nat. Rev. Mol. Cell Biol.* **17**, 413–425 (2016).
2. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
3. Wang, Z. *et al.* Sarcomatoid Renal Cell Carcinoma Has a Distinct Molecular Pathogenesis, Driver Mutation Profile, and Transcriptional Landscape. *Clin. Cancer Res.* **23**, 6686–6696 (2017).
4. Conant, J. L., Peng, Z., Evans, M. F., Naud, S. & Cooper, K. Sarcomatoid renal cell carcinoma is an example of epithelial–mesenchymal transition. *J. Clin. Pathol.* **64**, 1088–1092 (2011).
5. Lytle, N. K. *et al.* A Multiscale Map of the Stem Cell State in Pancreatic Adenocarcinoma. *Cell* **177**, 572–586.e22 (2019).
6. Sanada, Y. *et al.* Histopathologic Evaluation of Stepwise Progression of Pancreatic Carcinoma with Immunohistochemical Analysis of Gastric Epithelial Transcription Factor SOX2: Comparison of Expression Patterns between Invasive Components and Cancerous or Nonneoplastic Intraductal Components. *Pancreas* **32**, 164–170 (2006).
7. Herreros-Villanueva, M. *et al.* SOX2 promotes dedifferentiation and imparts stem cell-like features to pancreatic cancer cells. *Oncogenesis* **2**, e61–e61 (2013).
8. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
9. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
10. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
11. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

12. Vargo, A. H. S. & Gilbert, A. C. A rank-based marker selection method for high throughput scRNA-seq data. *BMC Bioinformatics* **21**, 477 (2020).
13. Delaney, C. *et al.* Combinatorial prediction of marker panels from single-cell transcriptomic data. *Mol. Syst. Biol.* **15**, e9005 (2019).
14. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
15. Jerby-Arnon, L. & Regev, A. Mapping multicellular programs from single-cell profiles. *bioRxiv* 2020.08.11.245472 (2020) doi:10.1101/2020.08.11.245472.
16. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).
17. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
18. Ghazanfar, S. *et al.* Investigating higher-order interactions in single-cell data with scHOT. *Nat. Methods* **17**, 799–806 (2020).
19. Wang, F., Liang, S., Kumar, T., Navin, N. & Chen, K. SCMarker: Ab initio marker selection for single cell transcriptome profiling. *PLOS Comput. Biol.* **15**, e1007445 (2019).
20. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single cell RNA sequencing. *bioRxiv* 742320 (2020) doi:10.1101/742320.
21. Xiao, Z., Dai, Z. & Locasale, J. W. Metabolic landscape of the tumor microenvironment at single cell resolution. *Nat. Commun.* **10**, 3763 (2019).
22. Liu, B. *et al.* An entropy-based metric for assessing the purity of single cell populations. *Nat. Commun.* **11**, 3155 (2020).
23. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
24. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

25. Miles, L. A. *et al.* Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature* **587**, 477–482 (2020).
26. Amaria, R. N. *et al.* Neoadjuvant immune checkpoint blockade in high-risk resectable melanoma. *Nat. Med.* **24**, 1649–1654 (2018).
27. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
28. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).
29. Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S. & Trapnell, C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.* **11**, 1537 (2020).
30. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
31. Andrew, G. & Gao, J. Scalable training of L^1 -regularized log-linear models. in *Proceedings of the 24th international conference on Machine learning* 33–40 (Association for Computing Machinery, 2007). doi:10.1145/1273496.1273501.
32. Karamitros, D. *et al.* Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nat. Immunol.* **19**, 85–97 (2018).
33. McFaline-Figueroa, J. L. *et al.* A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat. Genet.* **51**, 1389–1398 (2019).
34. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
35. Kinker, G. S. *et al.* Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat. Genet.* **52**, 1208–1218 (2020).
36. Martin, J. C. *et al.* Single-Cell Analysis of Crohn’s Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**, 1493-1508.e20 (2019).

37. van Galen, P. *et al.* Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e24 (2019).
38. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
39. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
40. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
41. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
42. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* **38**, 229-246.e13 (2020).
43. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
44. Anaya, J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Comput. Sci.* **2**, e67 (2016).
45. Huang, C.-Y. & Tan, T.-H. DUSPs, to MAP kinases and beyond. *Cell Biosci.* **2**, 24 (2012).
46. Emgård, J. *et al.* Oxysterol Sensing through the Receptor GPR183 Promotes the Lymphoid-Tissue-Inducing Function of Innate Lymphoid Cells and Colonic Inflammation. *Immunity* **48**, 120-132.e8 (2018).
47. Detavernier, A. *et al.* Monocytes undergo multi-step differentiation in mice during oral infection by *Toxoplasma gondii*. *Commun. Biol.* **2**, 1–14 (2019).
48. Zhang, H., Ye, M., Welner, R. S. & Tenen, D. G. Sox4 Is Required for the Formation and Maintenance of Multipotent Progenitors. *Blood* **124**, 1577–1577 (2014).
49. Dixit, A. *et al.* Perturb-seq: Dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853-1866.e17 (2016).

50. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
51. Pa, N., Lk, W., Ms, S. & Tm, O. Follow-up study of a randomized controlled trial of postnatal dexamethasone therapy in very low birth weight infants: effects on pulmonary outcomes at age 8 to 11 years. *J. Pediatr.* **150**, 345–350 (2007).
52. Srivastava, S. *et al.* ETS Proteins Bind with Glucocorticoid Receptors: Relevance for Treatment of Ewing Sarcoma. *Cell Rep.* **29**, 104-117.e4 (2019).
53. Zannas, A. S., Wiechmann, T., Gassen, N. C. & Binder, E. B. Gene–Stress–Epigenetic Regulation of FKBP5: Clinical and Translational Implications. *Neuropsychopharmacology* **41**, 261–274 (2016).
54. O’Leary, J. C., Zhang, B., Koren, J., Blair, L. & Dickey, C. A. The role of FKBP5 in mood disorders: Action of FKBP5 on steroid hormone receptors leads to questions about its evolutionary importance. *CNS Neurol. Disord. Drug Targets* **12**, 1157–1162 (2013).
55. Tieu, E. W., Tang, E. K. Y. & Tuckey, R. C. Kinetic analysis of human CYP24A1 metabolism of vitamin D via the C24-oxidation pathway. *FEBS J.* **281**, 3280–3296 (2014).
56. Ma, Y., McKay, D. J. & Buttitta, L. Changes in chromatin accessibility ensure robust cell cycle exit in terminally differentiated cells. *PLOS Biol.* **17**, e3000378 (2019).
57. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
58. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, 1–30 (2017).
59. Snyder, M. P. *et al.* The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
60. Spira, A. *et al.* Precancer Atlas to Drive Precision Prevention Trials. *Cancer Res.* **77**, 1510–1541 (2017).
61. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236–249 (2020).

62. Court, C. M. *et al.* Somatic copy number profiling from hepatocellular carcinoma circulating tumor cells. *Npj Precis. Oncol.* **4**, 1–9 (2020).
63. Taskesen, E. *et al.* Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Sci. Rep.* **6**, 24949 (2016).
64. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
65. Tsherniak, A. *et al.* Defining a Cancer Dependency Map. *Cell* **170**, 564-576.e16 (2017).
66. Basu, A. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* **154**, 1151–1161 (2013).
67. Wei, X. & Yu, P. S. Unsupervised Feature Selection by Preserving Stochastic Neighbors. in *Artificial Intelligence and Statistics* 995–1003 (PMLR, 2016).
68. Liu, D. C. & Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **45**, 503–528 (1989).
69. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
70. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

8 FIGURE LEGENDS

Fig. 1: The SCMER approach and its unique strength.

(a) Workflow of SCMER. SCMER selects the features that preserve the manifold from a single-cell omics dataset \mathbf{X} . Features can be selected from either \mathbf{X} or another co-assayed omics \mathbf{X}' . Vector \mathbf{w} indicates the selection. \mathbf{Y} is the dataset after feature selection. \mathbf{P} and \mathbf{Q} are cell-cell similarity matrices for \mathbf{X} and \mathbf{Y} , respectively.

(b) Applications of SCMER. SCMER selects features that preserve the manifold and retain inter- and intra-cluster diversity, and thus can be applied to discover rich molecular pathways, integrate modalities, and design customized DNA/RNA/antibody panels of restricted sizes.

(c) Capabilities of SCMER compared with mainstream label/cluster-based differential expression (DE) analysis methods and correlation-based methods. The hypothetical branching trajectories contains common progenitors (Pro), precursors (PreA, PreB), and mature cells (A, B).

Fig. 2: Results of the data of melanoma patients.

(a) UMAP embedding of the dataset without feature selection. (Macro: macrophages, Endo: endothelial cells, CAF: cancer associated fibroblasts, Unres: unresolved cells; labels are in the same color of dots representing cells.)

(b) UMAP of the dataset using SCMER selected genes.

(c) Recall of gene sets for SCMER, scMarker, Monocle, RankCorr, highly expressed genes (HXG), highly variable genes (HVG), principal component analysis (PCA), and differentially expressed genes (DEG, supervised). X-axis is number of selected genes and Y-axis is number of covered gene sets. A gene set is recalled when at least one gene in the set is selected. Methods recalled more gene sets with fewer genes are of better performance. “Random” shows the expected number of gene sets for randomly

selected markers. The area corresponds to $1.645 \times$ standard deviation on each side. Results above the area is significantly better than random ($p < 0.05$ for one sided z-test).

(d-f) Expression of genes that show intra-cluster gradients.

(g-i) Overall Kaplan-Meier survival curve for selected markers in TCGA SKCM. High and low include patients in above and under 33% percentile, respectively. Each group includes $n = 151$ patients.

Fig. 3: Results of the ileum lamina propria immunocytes data.

- (a)** UMAP embedding of the original dataset (List of abbreviations: **Supplementary Table 14**).
- (b)** UMAP embedding of the same dataset on genes selected by SCMER.
- (c-f)** Examples of expression level of genes selected by SCMER that (c) distinguish major cell types and (d) subtypes, (e) are transcription factors regulating different cell types, and (f) show gradual changes among cell states.
- (g)** Expression level of *DUSP1*. See **Supplementary Figure 4** for *DUSP2* and *DUSP4*.
- (h)** Distribution of expressions in major cell types of genes above.
- (i)** Recall of gene sets for SCMER, scMarker, Monocle, HXG, HVG, PCA, and DEG, similar to **Fig. 2c**.

Fig. 4: Results of the bone marrow hematopoiesis data.

- (a)** UMAP embedding of the original dataset (List of abbreviations: **Supplementary Table 14**).
- (b)** UMAP embedding of the dataset on SCMER selected genes.
- (c)** Recall of gene sets for SCMER, scMarker, Monocle, HXG, HVG, PCA, and DEG, similar to **Fig. 2c**.

(d) Activity of selected markers. Arrows are drawn for a visual reference for the developmental processes.

Fig. 5: Results of the A549 lung cancer cell line data.

(a-c) UMAP embedding of (a) the original sci-RNA-seq dataset, (b) the sci-RNA-seq dataset on SCMER selected markers, and (c) the sci-ATAC-seq peak dataset.

(d-i) Expression of selected genes show in (d-f) RNA space and (g-i) ATAC space. ATAC space only includes co-assayed cells.

(j) Heatmap of expression of selected genes and motif-based activity of highly variable transcription factors (TFs). (Uncor: uncorrelated, Pos: positively correlated, Neg: negatively correlated, with regard to *NR3C1* and *NR3C2*.) *ETV3* and *ETV4* are in the ETS transcription factor family.

Fig. 6: Results of the CITE-seq bone marrow mononuclear cells data.

(a-c) UMAP embedding of original dataset using (a) protein, (b) genes, and (c) SCMER selected genes. T cells and Progenitor cells [HSC, LMPP (lymphoid-primed multipotent progenitors), GMP, and Progenitor of B, Mk (megakaryocyte), RBC, and DC cells], are highlighted for better visual identification. Fully annotated cell types are shown in **Supplementary Figure 10**.

(d-f) Levels of representative (d) proteins, (e) genes, and (f) SCMER selected genes.

Fig. 1

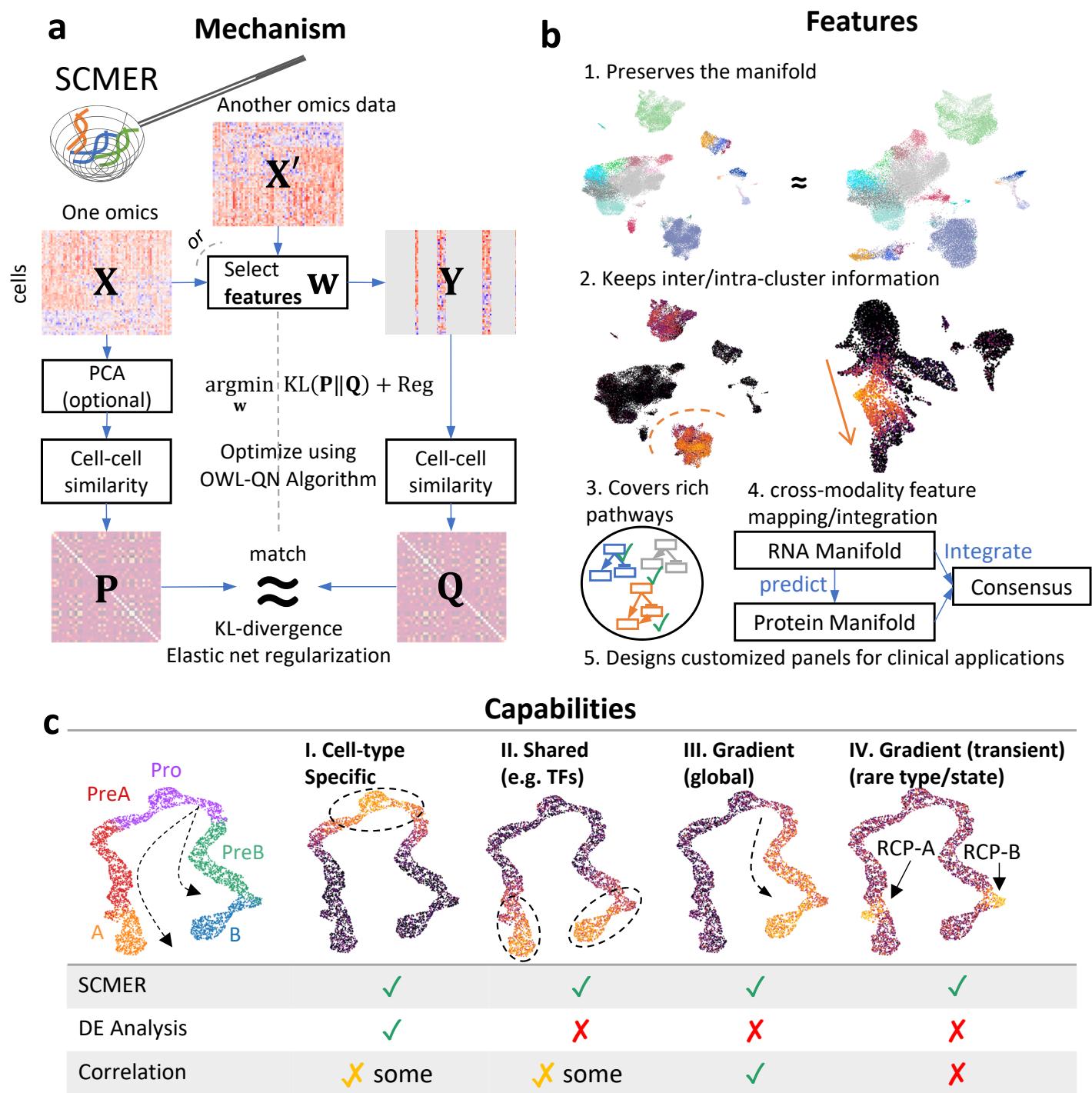
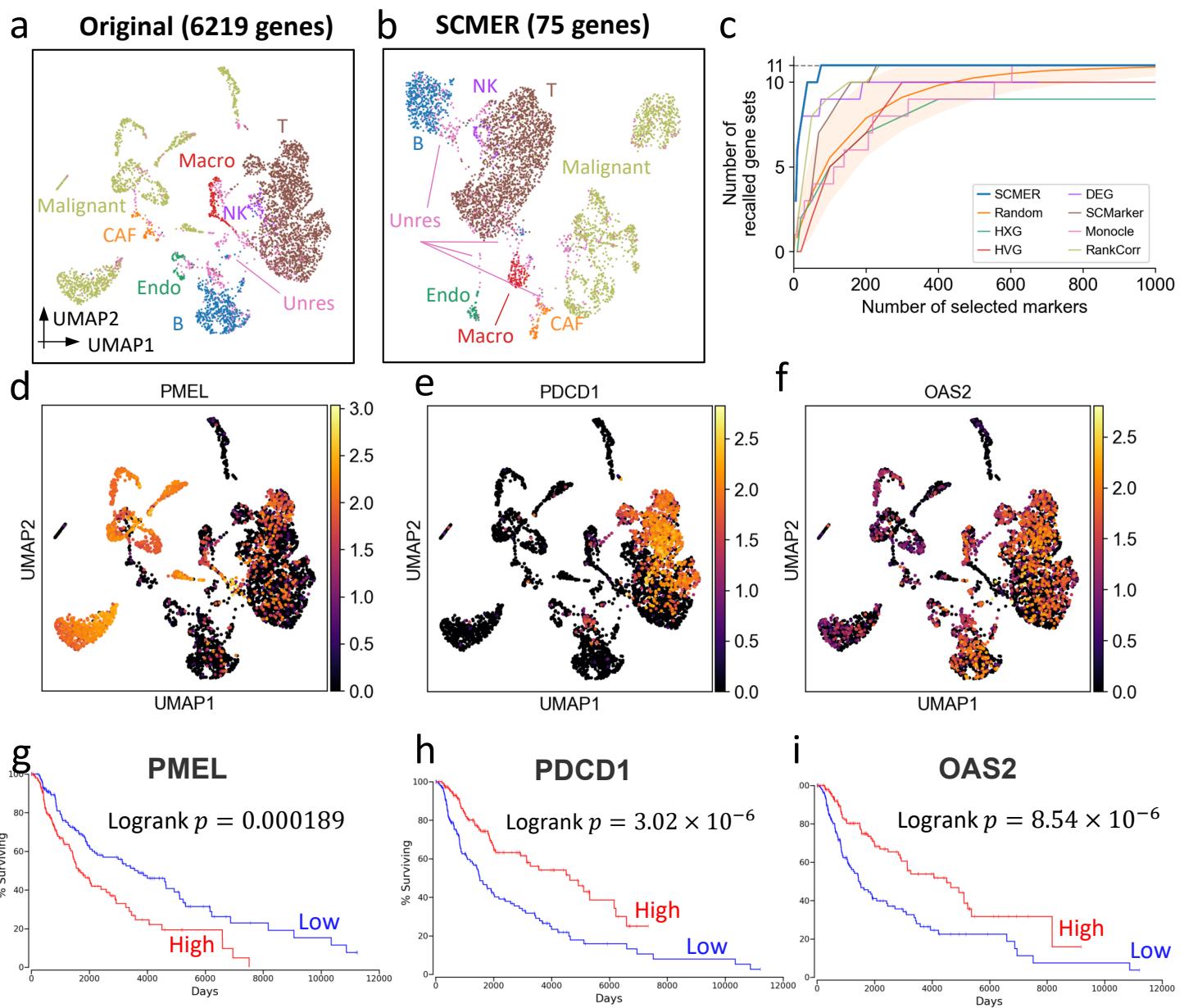


Fig. 2



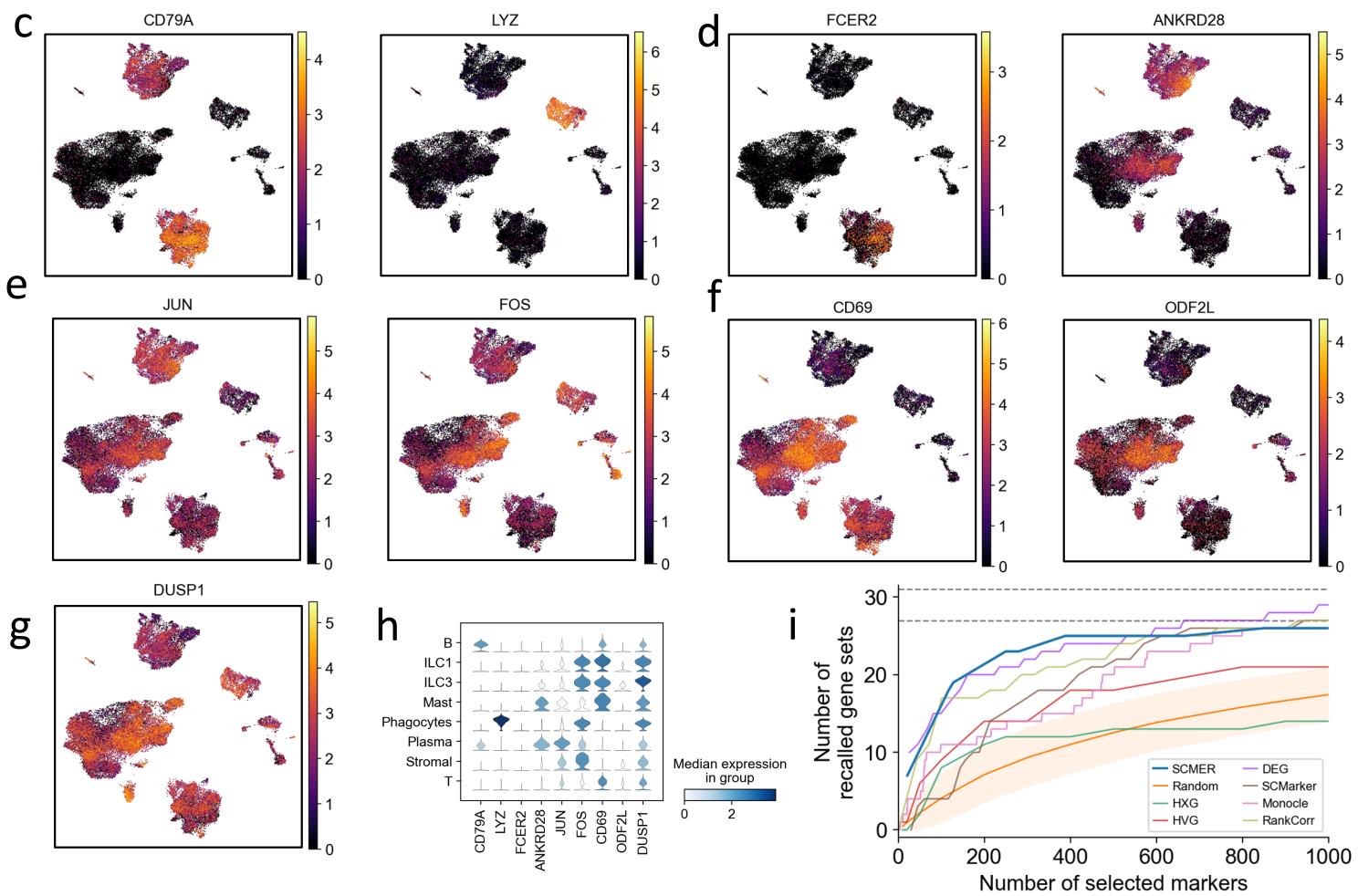
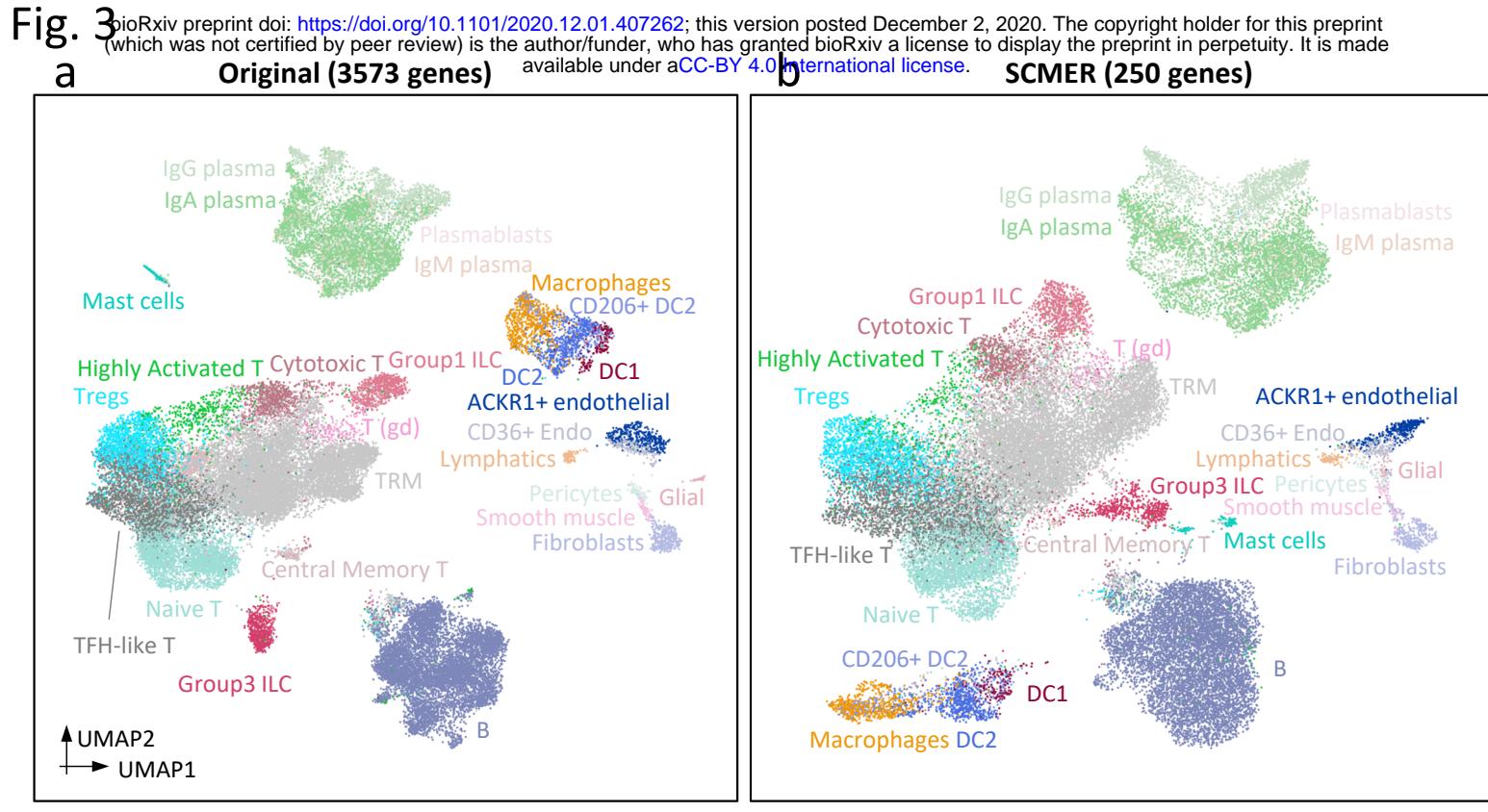
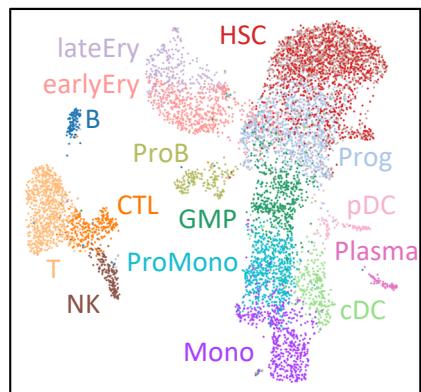
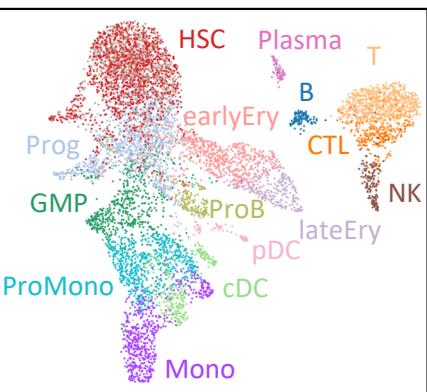


Fig. 4

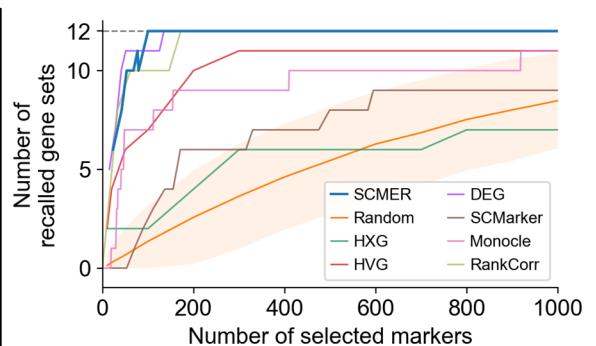
a Original (3352 genes)



b SCMER (100 genes)



c



d

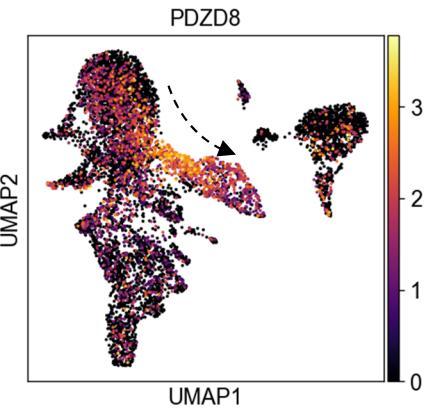
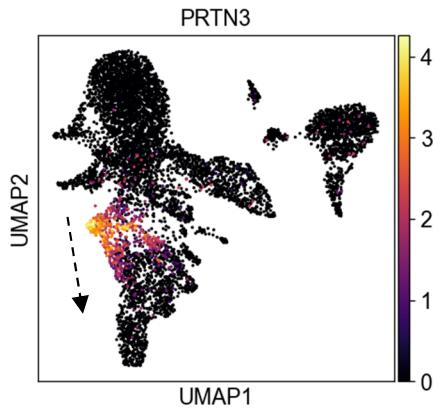
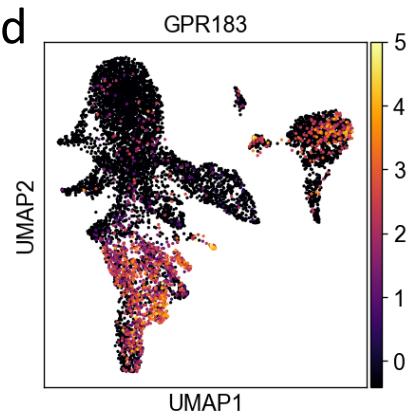


Fig. 5

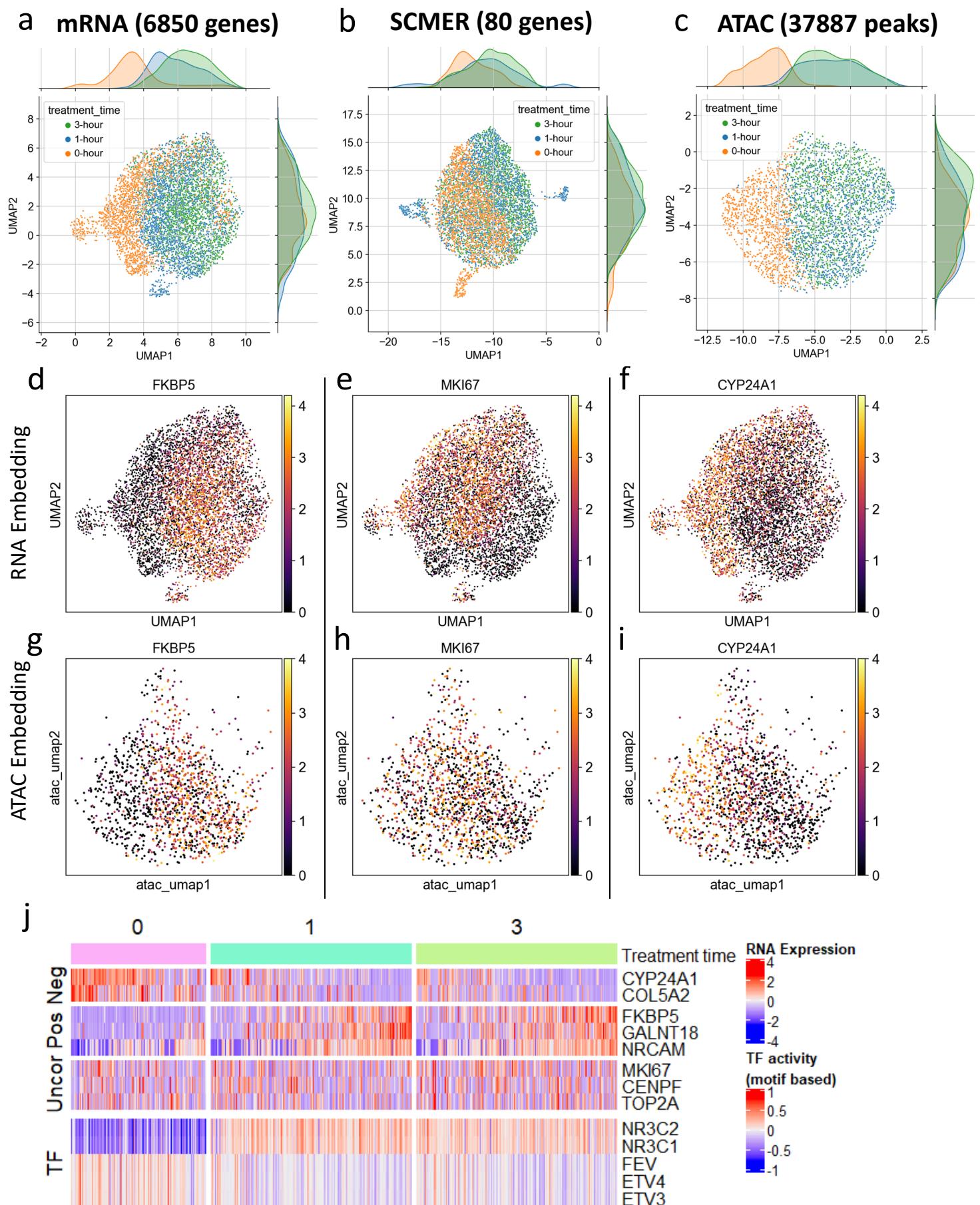


Fig. 6

