Computational and Systems Biology

Pseudo-grading of tumor subpopulations from single-cell transcriptomic data using Phenotype Algebra

Reviewed Preprint
v1 • August 20, 2024
Not revised

Namrata Bhattacharya, Anja Rockstroh, Sanket Suhas Deshpande, Sam Koshy Thomas, Anunay Yadav, Chitrita Goswami, Smriti Chawla, Pierre Solomon, Cynthia Fourgeux, Gaurav Ahuja, Brett G Hollier, Himanshu Kumar, Antoine Roquilly, Jeremie Poschmann, Melanie Lehman, Colleen C Nelson, Debarka Sengupta

Australian Prostate Cancer Research Centre-Queensland, Faculty of Health, School of Biomedical Sciences, Centre for Genomics and Personalised Health, Queensland University of Technology, Brisbane, Queensland-4000, Australia • Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India • Translational Research Institute, Princess Alexandra Hospital, Woolloongabba, Queensland-4102, Australia • Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India • School of Mathematical Sciences, The University of Adelaide, North Terrace, Adelaide, SA-5005, Australia • Center for Computational Biomedicine, Harvard Medical School, Boston, MA-02115, USA • Nantes Université, CHU Nantes, INSERM, Center for Research in Transplantation and Translational Immunology, UMR, 1064, Nantes, France • Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India • Laboratory of Immunology and Infectious Disease Biology, Department of Biological Sciences, Indian Institute of Science Education and Research (IISER), Bhopal, India • Vancouver Prostate Centre, Department of Urologic Sciences, University of British Columbia, Vancouver, Canada

Abstract

Single-cell RNA-sequencing (scRNA-seq) coupled with robust computational analysis facilitates the characterization of phenotypic heterogeneity within tumors. Current scRNA-seq analysis pipelines are capable of identifying a myriad of malignant and non-malignant cell subtypes from single-cell profiling of tumors. However, given the extent of intra-tumoral heterogeneity, it is challenging to assess the risk associated with individual malignant cell subpopulations, primarily due to the complexity of the cancer phenotype space and the lack of clinical annotations associated with tumor scRNA-seq studies. To this end, we introduce SCellBOW, a scRNA-seq analysis framework inspired by document embedding techniques from the domain of Natural Language Processing (NLP). SCellBOW is a novel computational approach that facilitates effective identification and high-quality visualization of single-cell subpopulations. We compared SCellBOW with existing best practice methods for its ability to precisely represent phenotypically divergent cell types across multiple scRNA-seq datasets, including our in-house generated human splenocyte and matched peripheral blood mononuclear cell (PBMC) dataset. For malignant cells, SCellBOW estimates the relative risk associated with each cluster and stratifies them based on their aggressiveness. This is achieved by simulating how the presence or absence of a specific malignant cell subpopulation influences disease prognosis. Using SCellBOW, we identified a hitherto unknown and pervasive AR−/NE_{low} (androgen-receptor-negative, neuroendocrine-low)

malignant subpopulation in metastatic prostate cancer with conspicuously high aggressiveness. Overall, the risk-stratification capabilities of SCellBOW hold promise for formulating tailored therapeutic interventions by identifying clinically relevant tumor subpopulations and their impact on prognosis.

eLife assessment

This **valuable** study introduces SCellBOW, a novel tool leveraging natural language processing techniques to enhance cell clustering and infer survival risks from single-cell RNA sequencing data. The methodology and results are **convincing**, demonstrating superior clustering performance and the ability to assign risk scores to cancer cell clusters across multiple datasets. SCellBOW's unique approach promises significant advancements in understanding cancer cell heterogeneity and identifying aggressive cancer cell subgroups.

https://doi.org/10.7554/eLife.98469.1.sa3

Introduction

Intra-and inter-tumoral heterogeneity are pervasive in cancer and manifest as a constellation of molecular alterations in tumor tissues. The late-stage clonal proliferation, partial selective sweeps, and spatial segregation within the tumor mass collectively orchestrate lineage plasticity and metastasis[1]. In collaboration with non-malignant cell types in the tumor stroma, malignant cells with distinct genetic and phenotypic properties create complex and dynamic ecosystems, rendering the tumors recalcitrant to therapies[2]. Thus, the phenotypic characterization of malignant cell subpopulations (subclones) is critical to understanding the underlying mechanisms of resistive behavior. The widespread adoption of single-cell RNA-sequencing (scRNA-seq) has enabled the profiling of individual cells, thereby obtaining a high-resolution snapshot of their unique molecular landscapes[3,4]. A precise understanding of cell-to-cell functional variability captured by scRNA-seq profiles is crucial in this context. These molecular profiles assist in robust deconvolution of the oncogenic processes instigated by various selection pressures exerted by anticancer agents. They also facilitate understanding the cross-talks between malignant and non-malignant cell types within the tumor microenvironment[5]. To effectively analyze tumor scRNA-seq data, various specialized techniques have been developed. These techniques assist in proactive investigation of complex and elusive cell populations[6,7], regulatory gene interactions[8], neoplastic cell lineage trajectories[9], and expression-based inference of copy number variations[10]. While these computational techniques have been successful in gaining novel biological insights, their adoption in clinical settings is still elusive.

Over the past years, leading consortia such as The Cancer Genome Atlas (TCGA)[11] and other large-scale independent studies have established reproducible molecular subtypes of cancers with divergent prognoses. For instance, metastatic prostate cancer has typically been categorized based on the androgen receptor (AR) activity or neuroendocrine (NE) program: the less aggressive AR+/NE– (AR prostate cancer, ARPC) and the highly aggressive AR–/NE+ (NE prostate cancer, NEPC)[12]. More recent studies have identified additional phenotypes, such as the AR–/NE– (double-negative prostate cancer, DNPC) and AR+/NE+ (amphicrine prostate cancer, AMPC)[13,14]. These findings underscore the importance of considering tumor heterogeneity while dictating the differential treatment regimes to improve patient outcomes[15]. These studies are predominantly based on bulk omics assays, which precludes the detectability of fine-grained

molecular subtypes of clinical relevance. To address this, there is an urgent need to develop novel analytical approaches that are capable of exploiting single-cell omics profiles for risk attribution to malignant cell subtypes.

A growing number of NLP-based methods have recently been gaining popularity for their ability to predict patients' survival based on their clinical records and for facilitating efficient analyses of high-dimensional scRNA-seq data[16–19]. For instance, Nunez *et al.*[16] employ language models to predict the survival outcomes of breast cancer patients based on their initial oncologist consultation documents. Similarly, in the domain of scRNA-seq analysis, scETM[18] uses topic models to infer biologically relevant cellular states from single-cell gene expression data. Recently, scBERT[19], an adaptation of attention-based transformer architecture, has been developed for cell type annotation in single-cell. Motivated by the objective to efficiently associate survival risk directly with cellular subpopulations extracted from scRNA-seq data, we introduce SCellBOW (single-cell bag-of-words), a Doc2vec[20] inspired transfer learning framework for single-cell representation learning, clustering, visualization, and relative risk stratification of malignant cell types within a tumor. SCellBOW intuitively treats cells as documents and genes as words. SCellBOW learned latent representations capture the semantic meanings of cells based on their gene expression levels. Due to this, cell type or condition-specific expression patterns get adequately captured in cell embeddings. We demonstrate the utility of SCellBOW in clustering cells based on their semantic similarities across multiple scRNA-seq datasets, spanning from normal prostate to pancreas and peripheral blood mononuclear cells (PBMC). As an extended validation, we apply SCellBOW to an in-house scRNA-seq dataset comprising human PBMC and matched splenocytes. We observed that SCellBOW outperforms existing best-practice single-cell clustering methods in its ability to precisely represent phenotypically divergent cell types.

Beyond robust identification of cellular clusters, the latent representation of single-cells provided by SCellBOW captures the'semantics' associated with cellular phenotypes. In line with *word algebra* supported by NLP models, which explores word analogies based on semantic closeness (e.g., adding a word vector associated *royal* to *man* brings it closer to *king*), we conjectured that cellular embeddings can reveal biologically intuitive outcomes through algebraic operations. Thus, we aimed to replicate this feature in the single-cell phenotype space to introduce *phenotype algebra* and apply the same in attributing prognostic value to cancer cell subpopulations identified from scRNA-seq studies. With the help of *phenotype algebra,* it is possible to simulate the exclusion of the phenotype associated with a specific malignant cell subpopulation and relate the same to the disease outcome. Selectively removing a specific subtype from the whole tumor allows us to categorize the aggressiveness of individual subtypes under negative selection pressure while preserving the impact of the residual tumor. This information can ultimately aid therapeutic decision-making and improve patient outcomes. As a proof of concept, we validate the phenotype *algebra* component of SCellBOW for pseudo-grading malignant cell subtypes across three independent cancer types: glioblastoma multiforme, breast cancer, and metastatic prostate cancer.

Finally, we demonstrate the utility of SCellBOW to uncover and define novel subpopulations in a cancer type based on biological features relating to disease aggressiveness and survival probability. Metastatic prostate cancer encompasses a range of malignant cell subpopulations, and characterizing novel or more fine-grained subtypes with clinical implications on patient prognosis is an active field of research[21,22]. To explore this, we applied our *phenotype algebra* on the SCellBOW clusters as opposed to predefined subtypes. Through our analysis using SCellBOW, we identify a dedifferentiated metastatic prostate cancer subpopulation. This subpopulation is nearly ubiquitous across numerous metastatic prostate cancer samples and demonstrates greater aggressiveness compared to any known molecular subtypes. To the best of our knowledge, SCellBOW pioneered the application of a NLP-based model to attribute survival risk to malignant cell subpopulations from patient tumors. SCellBOW holds promise for tailored therapeutic interventions by identifying clinically relevant subpopulations and their impact on prognosis.

Results

Overview of SCellBOW

Correlating genomic readouts of tumors with clinical parameters has helped us associate molecular signatures with disease prognosis[23]. Given the prominence of phenotypic heterogeneity in tumors, it is important to understand the connection between molecular signatures of clonal populations and disease aggressiveness. Existing methods map tumor samples to a handful of well-characterized molecular subtypes with known survival patterns primarily obtained from bulk RNA-seq studies. However, bulk RNA-seq measures the average level of gene expression distributed across millions of cells in a tissue sample, thereby obscuring the intra-tumoral heterogeneity[24]. This limitation prompted us to turn to tumor scRNA-seq. Tumor scRNA-seq studies have successfully revealed the extent of gene expression variance across individual malignant cells, contributing to an in-depth understanding of the mechanisms driving cancer progression[25]. However, to date, most studies use marker genes identified from bulk RNA-seq studies to characterize malignant cell clusters identified from scRNA-seq data. This approach is inadequate since every tumor is unique when looked through the lenses of single-cell expression profiles[26]. Bulk markers alone cannot adequately capture the intricate heterogeneity present within tumors, as different tumors may exhibit distinct phenotypes with varying degrees of aggressiveness at the single-cell level. While existing methods effectively reveal the subpopulations, they are insufficient in associating malignant risk with specific cellular subpopulations identified from scRNA-seq data. The proposed approach, SCellBOW, can effectively capture the heterogeneity and risk associated with each phenotype, enabling the identification and assessment of malignant cell subtypes in tumors directly from scRNA-seq gene expression profiles, thereby eliminating the need for marker genes (**Fig. 1**). Below, we highlight key constructions and benefits of this new approach.

SCellBOW adapts the popular document-embedding model Doc2vec for single-cell latent representation learning, which can be used for downstream analysis. SCellBOW represents cells as documents and gene names as words. Gene expression levels are encoded by'word frequencies', i.e., variation in gene expression values is captured by introducing gene names in proportion to their intensities in cells. SCellBOW encodes each cell into low-dimensional embeddings. SCellBOW extracts the gene expression patterns of individual cells from a relatively large unlabelled source scRNA-seq dataset by pre-training a shallow source network. The neuronal weights estimated during pre-training are transferred to a relatively smaller unlabelled scRNA-seq dataset, where they are refined based on the gene expression patterns in the target dataset. SCellBOW leverages the variability in gene expression values to subsequently cluster cells according to their semantic similarity.

Building upon SCellBOW's capacity to preserve the semantic meaning of individual cells, we attempted to establish meaningful associations among cellular phenotypes. SCellBOW offers a remarkable feature for executing algebraic operations such as'+' and'–' on single-cells in the latent space while preserving the biological meanings. This feature catalyzes the simulation of the residual phenotype of tumors, following positive and negative selection of specific malignant cell subtypes in a tumor. This could potentially be used to identify the contribution of that phenotype to patient survival. For example, the'–' operation can be used to predict the likelihood of survival by eliminating the impact of a specific aggressive cellular phenotype from the whole tumor. We empirically show the retention of such semantic relationships in the context of cellular phenotypes of cancer cells using *phenotype algebra*. SCellBOW uses algebraic operations to compare and analyze the importance of cellular phenotypes working independently or in combination toward the aggressiveness of the tumor in a way that is not possible using traditional methods. *Phenotype algebra* can be performed either on the pre-defined cancer subtypes or

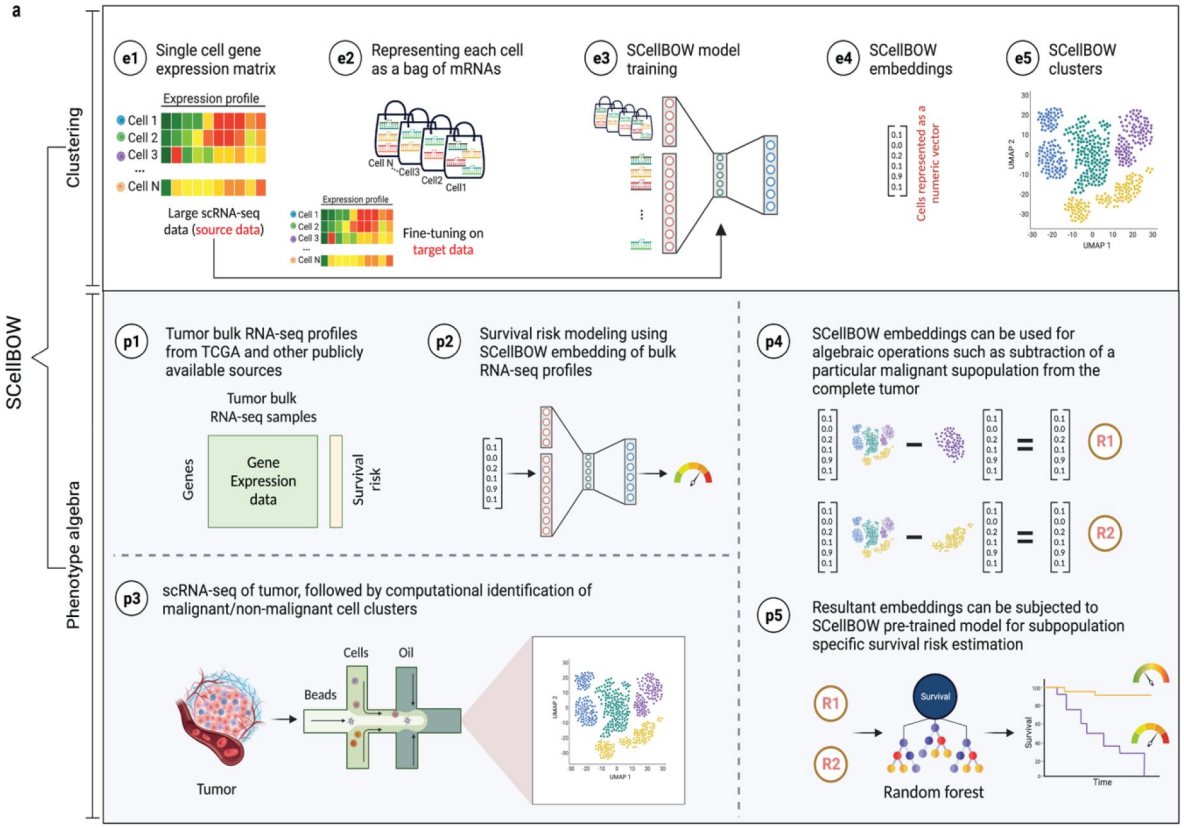

Fig. 1.

SCellBOW workflow.

a, Schematic overview of SCellBOW workflow for identifying subclones and assessing subclonal tumor aggressiveness. For SCellBOW clustering, firstly, a corpus was created from the gene expression matrix, where cells were analogous to documents and genes to words. Next, the pre-trained model was retrained with the vocabulary of the target dataset. Then, clustering was performed on embeddings generated from the neural network. For SCellBOW *phenotype algebra*, vectors were created for reference (*total tumor*) and queries. Then, the query vector was subtracted from the reference vector to calculate the predicted risk score using a bootstrapped random survival forest. Finally, survival probability was evaluated and phenotypes were stratified by the median predicted risk score.

SCellBOW clusters. Drawing from the overall performance of SCellBOW in accurately clustering and ranking cellular phenotypes by aggressiveness, we further analyzed multi-patient scRNA-seq data of metastatic prostate cancer and characterized an unknown, de-differentiated AR−/NE_{low}subpopulation of malignant cells.

SCellBOW robustly dissects tissue heterogeneity

Malignant cells are far more heterogeneous compared to associated normal cells. Clustering is often the first step toward recognizing subclonal lineages in a tumor sample. Clustering single-cells based on their molecular profiles can potentially identify rare cell populations with distinct phenotypes and clinical outcomes. To evaluate the strength of the clustering ability of SCellBOW, we benchmarked our method against five existing scRNA-seq clustering tools[27–31] (**Supplementary Table 1**). Among these packages, Seurat[27] and Scanpy[28] are the most popular, and both employ graph-based clustering techniques. DESC[29] is a deep neural network-based scRNA-seq clustering package, whereas ItClust[30] and scETM[31] are transfer learning methods. All the packages are resolution-dependent except for ItClust. ItClust automatically selects the resolution with the highest silhouette score. We used a number of scRNA-seq datasets to evaluate the tools cited above (**Supplementary Table 2**). For the objective evaluation of performance, we used an adjusted Rand index (ARI) and normalized mutual information (NMI) to compare clusters with known cell annotations (**Supplementary Table 3)**. For all methods, except for ItClust, we computed overall ARI and NMI for different values of resolution ranging between 0.2 to 2.0. We computed the cell type silhouette index (SI) based on known annotations to measure the signal-to-noise ratio of low-dimensional single-cell embeddings.

We constructed three use cases leveraging publicly available scRNA-seq datasets. Each instance constitutes a pair of single-cell expression datasets, of which the source data is used for self-supervised model training and the target data for model fine-tuning and analysis of the clustering outcomes. In all cases, the target data has associated cell type annotations derived from fluorescence-activated cell sorting (FACS) enriched pure cell subpopulations. The first use case consists of non-cancerous cells from prostate cancer patients[32] (120,300 cells) as the source data and cells from healthy prostate tissues[33] (28,606 cells) as the target data. This use case was designed to assess the resiliency of SCellBOW to the presence of disease covariates in a large scRNA-seq dataset. The second use case comprises a large PBMC dataset[34] (68,579 cells) as the source data, whereas the target data was sourced from a relatively small FACS-annotated PBMC dataset (2,700 cells) from the same study. The third use case, as source data, comprises pancreatic cells from three independent studies processed with different single-cell profiling technologies (inDrop[35], CEL-Seq2[36], SMARTer[37]). The target data used in this case was from an independent study processed with a different technology, Smart-Seq2[38] (**Fig. 2**, **Supplementary Fig. 2**). For the normal prostate, PBMC, and pancreas datasets, SCellBOW produced the highest ARI scores across most notches of the resolution spectrum (0.2 to 2.0) (**Fig. 2g-i**). We observed a similar trend in the case of NMI (**Fig. 2j**). SCellBOW exhibited the highest NMI compared to other tools for the normal prostate, PBMC, and pancreas datasets. For further deterministic evaluation of the different methods across the datasets, we set 1.0 as the default resolution for calculating cell type SI (**Fig 2k**). In the PBMC and pancreas datasets, SCellBOW yielded the highest cell type SI for both datasets. SCellBOW and Seurat were comparable in performance for the pancreas dataset, outperforming other tools. We observed poor performance by DESC and ItClust across all the datasets in terms of cell type SI. In terms of cell type SI, Seurat and scETM showed improved results in the normal prostate dataset. However, SCellBOW outperformed both Seurat and scETM in terms of overall ARI and NMI for the same dataset, indicating higher clustering accuracy. To further evaluate the cluster quality in the normal prostate dataset, we compared the known cell types to the predicted clusters. Known cell types such as basal epithelial, luminal epithelial, and smooth muscle cells were grouped into homogeneous clusters by SCellBOW (**Supplementary Fig. 2**). We observed that the majority of fibroblasts and endothelial cells were mapped by SCellBOW to single clusters, unlike Seurat and scETM. SCellBOW retained hillock cells in close proximity to both basal epithelial and club cells, in

contrast to Seurat, which only includes basal epithelial cells. We compared SCellBOW with two additional methods published recently – scBERT[19] and scPhere[39]. Among these, scBERT is a transfer learning-based method built upon BERT[40], scPhere is a deep generative model designed for embedding single cells into low-dimensional hyperspherical or hyperbolic spaces. While scBERT offers competitive performance, scPhere consistently falls behind the other methods (**Supplementary Fig. 3 and Supplementary Note 1**).

Our analyses on the public datasets confirm the robustness of SCellBOW compared to the prominent single-cell analysis tools, including the prominent transfer learning methods. To this end, we applied SCellBOW to investigate a more challenging task of analyzing an in-house scRNA-seq data comprising splenocytes and matched PBMCs from two healthy and two brain-dead donors. Given multiple covariates, such as the origin of the cells and the physiological states of the donors, analysis of this scRNA-seq data presents a challenging use case. We used the established high-throughput scRNA-seq platform CITE-seq[41] to pool eight samples into a single experiment. After post-sequencing quality control, we were left with 4,819 cells. We annotated the cells using Azimuth[42], with occasional manual interventions (**Supplementary Note 2, Supplementary** Fig. 4d-f). We quantitatively evaluated SCellBOW and the rest of the benchmarking methods by measuring ARI, NMI, and cell type SI (**Fig. 3e**, **Supplementary** Fig. 3d-f**)**. While most methods did reasonably well, SCellBOW offered an edge. We observed the best results in SCellBOW in terms of ARI, NMI, and cell type SI compared to other tools (**Supplementary Table 3**). SCellBOW yielded clusters largely coherent with the independently done cell type annotation using Azimuth (**Fig 3**, **Supplementary** Fig. 5). Further, most clusters harbor cells from all the donors, indicating that the sample pooling strategy was effective in reducing batch effects. B cells, T cells, and NK cells map to SCellBOW clusters where the respective cell types are the majority. Most CD4+ T cells map to CL0 and CL9 (here, CL is used as an abbreviation for cluster) (**Fig 3f**). CL0 is shared between CD8+ T cells, CD4+ T cells, and Treg cells, which originate from the same lineage. CL4 majorly consists of NK cells with a small fraction of CD8+ T cells, which is not unduly deviant from the PBMC lineage tree. As a control, we performed similar analyses of the Scanpy clusters (**Fig 3g**). While Scanpy performed reasonably well, misalignments could be spotted. For example, CD4+ and CD8+ T cells were split across many clusters with mixed cell type mappings. SCellBOW maps CD14 monocytes to a single cluster, whereas Scanpy distributes CD14 monocytes across two clusters (CL1 and CL8), wherein CL8 is equally shared with conventional Dendritic Cells (cDC).'Eryth' annotated cells are indignantly mapped to different clusters by both SCellBOW and Scanpy. This could be due to the Azimuth's reliance on high mitochondrial gene expression levels for annotating erythroid cells[42] **(Supplementary Note 3)**. However, SCellBOW CL6 is dominated by Erythroid cells with marginal interference from other cell types. In the case of Scanpy, no such cluster could be detected. Discerning phenotypic heterogeneity from the expression profiles of seemingly similar cells is a challenging task. The performance of SCellBOW, with near-ground-truth cell type annotation, in such a scenario confirmed its ability to adequately decipher the underlying cellular heterogeneity and provide robust cell type clustering.

SCellBOW facilitates survival-risk attribution of tumor subpopulations

Every cancer features unique genotypic as well as phenotypic diversity, impeding the personalized management of the disease. Patient survival is arguably the most important clinical parameter to assess the utility of clinical interventions. The widespread adoption of genomics in cancer care has allowed correlating molecular portraits of tumors with patient survival in all major cancers. This, in turn, has broadened our understanding of the aggressiveness and impact of diverse molecular subtypes associated with a particular cancer type. Single-cell expression profiling of cancers allows the detection and characterization of tumor-specific malignant cell subtypes. This presents an opportunity to gauge the aggressiveness of each cancer cell subtype in a tumor. A few studies suggested that there is an association between the tumorigenicity of stromal cells in tumor

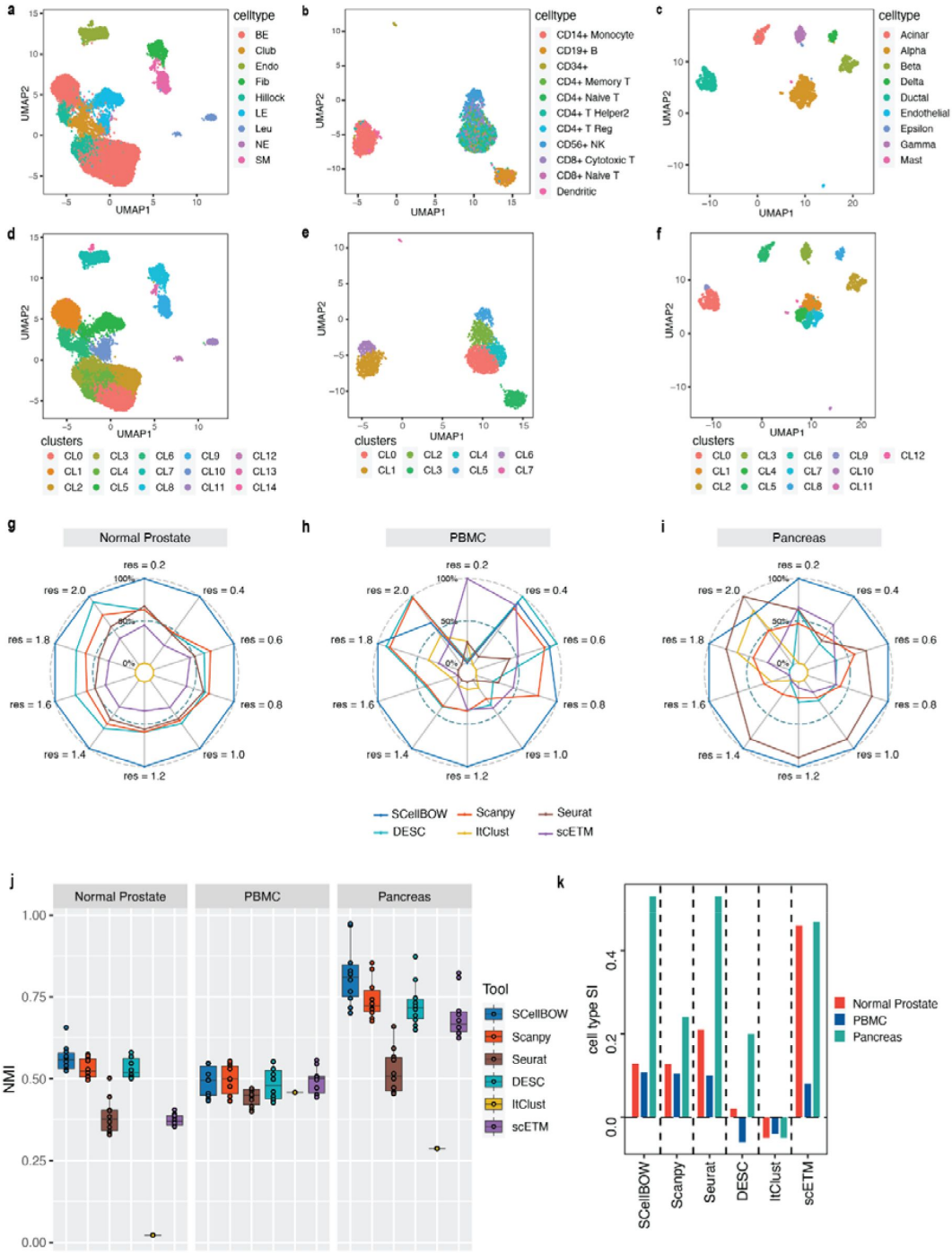

Fig. 2.

Evaluation of single-cell representations using SCellBOW.

a-c, UMAP plots for the normal prostate (a), PBMC (b), and pancreas (c) datasets. The coordinates are colored by cell types.
d-f, UMAP plots for normal prostate (d), PBMC (e), and pancreas (f) datasets, where the coordinates are colored by SCellBOW clusters. CL is used as an abbreviation for cluster.
g-i, Radial plot for the percentage of contribution of different methods towards ARI for various resolutions ranging from 0.2 to 2.0. ItClust is a resolution-independent method; thus, the ARI is kept constant across all the resolutions.
j, Box plot for the NMI of different methods across different resolutions ranging from 0.2 to 2.0 in steps of 0.2.
k, Bar plot for the cell type silhouette index (SI) for different methods. The default resolution was set to 1.0.

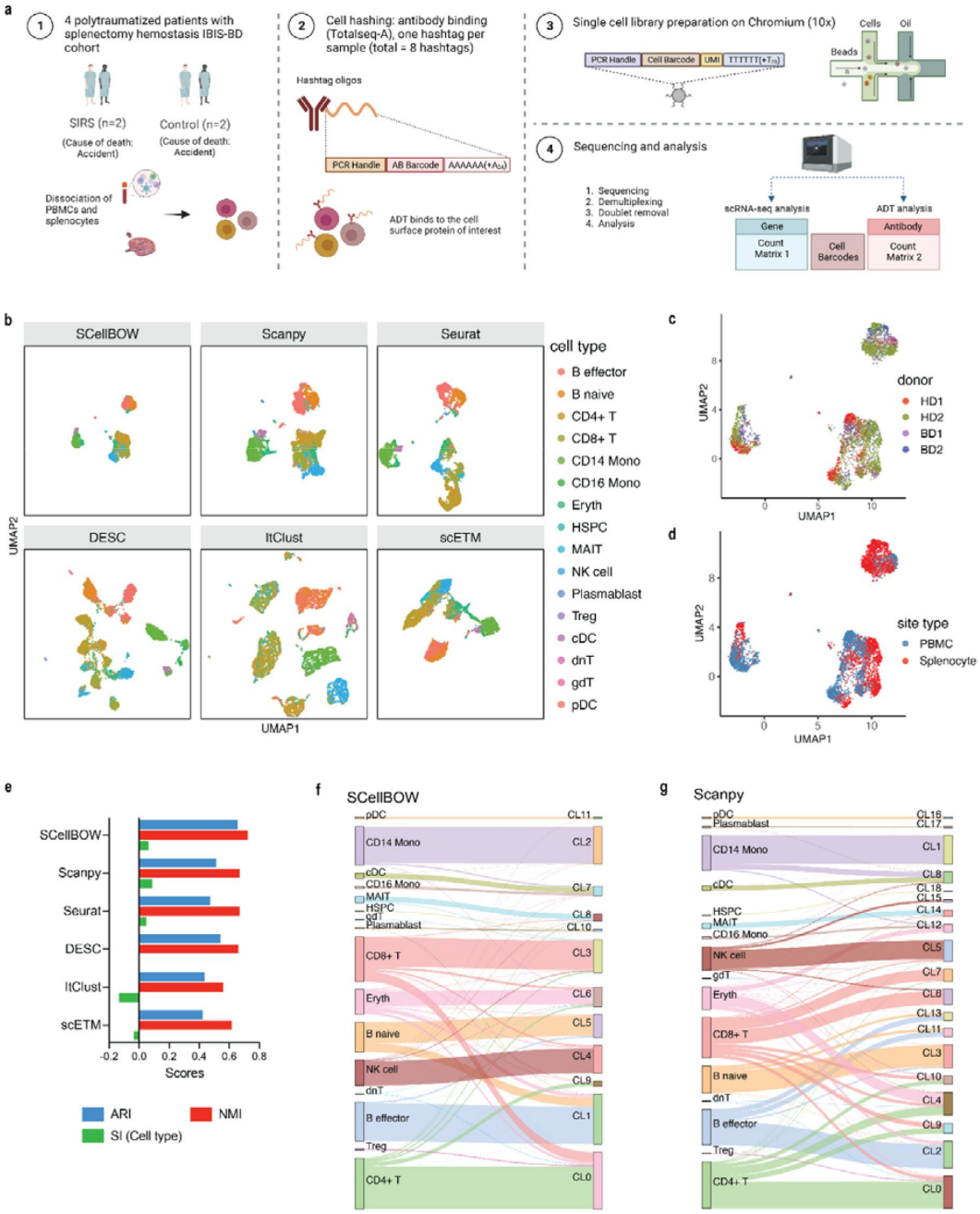

Fig. 3.

Evaluation of in-house splenocytes and matched PBMCs.

a, An experiment schematic diagram highlighting the sites of the organs for tissue collection and sample processing. In this matched PBMC-splenocyte CITE-seq experiment, PBMCs and splenocytes were collected, followed by high-throughput sequencing and downstream analyses.

b, The UMAP plots for the embedding of SCellBOW compared to different benchmarking methods. The coordinates of all the plots are colored by cell type annotation results using Azimuth. c-d, UMAP plots for SCellBOW embedding colored by donors (c) and cell types (d).

e-f, Alluvial plots for Azimuth cell types mapped to SCellBOW clusters (e) and Scanpy clusters

(f). The resolution of SCellBOW was set to 1.0. CL is used as an abbreviation for cluster. g, Bar plot for ARI, NMI, cell type SI at resolution 1.0.

microenvironments and patient survival[43]. This has sparked debate about the utility of gene expression in profiling single tumor cells. SCellBOW estimates the relative aggressiveness of different malignant cell subtypes using survival information as a surrogate. SCellBOW helps achieve this by enabling algebraic operations (subtraction or addition between two expression vectors or phenotypes) in the embedding space. We refer to this as *phenotype algebra*. We utilized this ability to find an association between the embedding vectors, representing *total tumor – a specific malignant cell cluster* with tumor aggressiveness. This immediately opens up a way to infer the level of aggressiveness of a particular cluster of cancer cells obtained through single-cell clustering. Subtracting an aggressive phenotype from the whole tumor (average of SCellBOW embeddings across all malignant cells in a tumor) would better the odds of survival relative to dropping a subtype under negative selection pressure.

As proof of concept, we first validated our approach on glioblastoma multiforme (GBM), which has been studied widely employing single-cell technologies. GBM has three well-characterized malignant subtypes: proneural (PRO), classical (CLA), and mesenchymal (MES)[44,45]. We obtained known markers of PRO, CLA, and MES to annotate 4,508 malignant GBM cells obtained from a single patient reported by Couturier and colleagues[46] and used it as our target data (**Fig. 4b**, **Supplementary Methods 2.1**). As the source data, we used the GBM scRNA-seq data from Neftel *et al.*[47]. The survival data consisted of 613 bulk GBM samples with paired survival information from the TCGA consortium. We constructed the following query vectors for the survival prediction task: *total tumor*, i.e., average of embeddings of all malignant cells, *total tumor* – (MES+CLA), *total tumor* – MES/CLA/PRO (individually). We conjectured that for the most aggressive malignant cell subtype, *total tumor – subtype specific pseudo-bulk*, would yield the biggest drop in survival risk relative to the *total tumor*. Survival risk predictions associated with *total tumor* – MES/CLA/PRO thus obtained reaffirmed the clinically known aggressiveness order, i.e., CLA >-MES >-PRO, where CLA succeeds the rest of the subtypes in aggressiveness[48] (**Fig. 4c, d**). More complex queries can be formulated, such as *total tumor* – (MES+CLA), which indicates that the tumor does not comprise the two most aggressive phenotypes, CLA and MES, and instead consists only of PRO cells. Hypothetically, this represents the most favorable scenario, as testified by the *phenotype algebra*[45,49].

We performed a similar benchmarking on well-established PAM50-based[50] breast cancer (BRCA) subtypes: luminal A (LUMA), luminal B (LUMB), HER2-enriched (HER2), basal-like (BASAL), and normal-like (NORMAL) (**Fig. 4a**). We used 24,271 cancer cells from Wu *et al.*[51] as the source data, 545 single-cell samples from Zhou *et al.*[52,53] as the target data. We used 1,079 bulk BRCA samples with paired survival information from TCGA as the survival data. SCellBOW-predicted survival risks for the different subtypes were generally in agreement with the clinical grading of the PAM50 subtypes[54,55]. Exclusion of LUMA from *total tumor* yielded the highest risk score, indicating that LUMA has the best prognosis, followed by HER2 and LUMB, whereas BASAL and NORMAL were assigned worse prognosis[56–58] (**Fig. 4e, f**). We observed an interesting misalignment from the general perception about the relative aggressiveness of the NORMAL subtype-removal of this subtype from *total tumor* indicated the highest improvement in prognosis. The NORMAL subtype is a poorly characterized and rather heterogeneous category. Recent evidence suggests that these tumors potentially represent an aggressive molecular subtype and are often associated with highly aggressive claudin-low tumors[54,59]. These benchmarking studies on the well-characterized cancer subtypes of GBM and BRCA affirm SCellBOW's capability to preserve the desirable characteristics in the resulting phenotypes obtained as outputs of algebraic expressions involving other independent phenotypes and operators such as +/–. To further validate the efficiency of SCellBOW in learning single-cell latent representation, we have illustrated the effectiveness of fixed-length embeddings obtained from SCellBOW for *phenotype algebra* compared to other methods-scETM, scBERT, and scPhere (**Supplementary Fig. 6 and Supplementary Note 4**). Our results showed that SCellBOW learned latent representation of

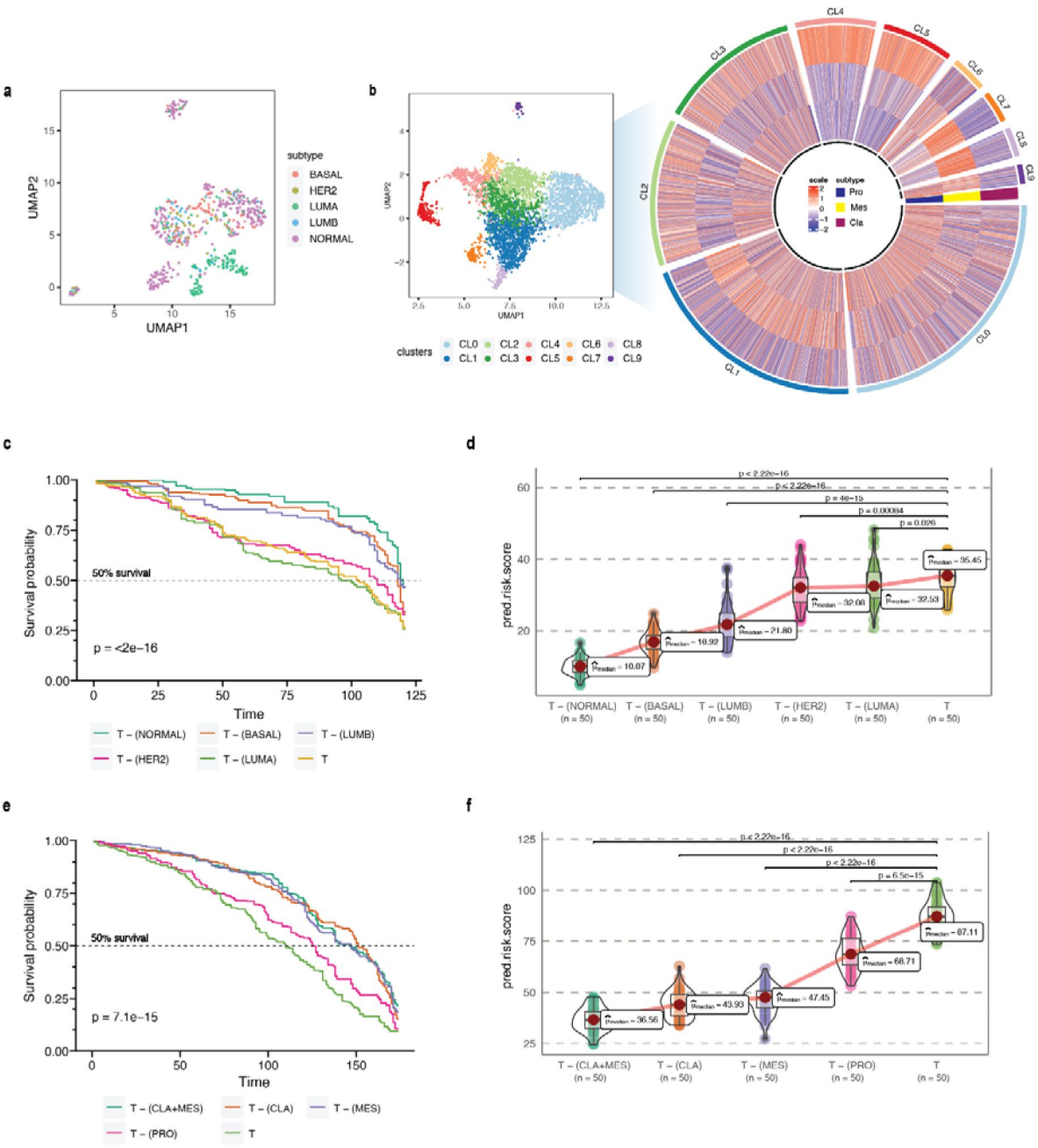

Fig. 4.

Subclonal survival risk inference.

a, UMAP plot for the embedding of BRCA target dataset colored by PAM50 molecular subtype.
b, Heatmap for GSVA score for three molecular subtypes of GBM: CLA, MES, and PRO, grouped by SCellBOW clusters at resolution 1.0.
c, Survival plot for BRCA molecular subtypes based on *phenotype algebra*. The *total tumor* is denoted by *T*.
d, Violin plot for predicted risk scores for BRCA molecular subtypes.
e, Survival plot for GBM molecular subtypes based on *phenotype algebra*. f, Violin plot for predicted risk scores for GBM molecular subtypes.

single-cells accurately captures the'semantics' associated with cellular phenotypes and allows algebraic operations such as'+' and'–'. Whereas scETM, scBERT, and scPhere fail to stratify cancer clones in terms of their aggressiveness and contribution to disease prognosis.

SCellBOW enables pseudo-grading of metastatic prostate cancer cells

In prostate cancer, the processes of transdifferentiation and dedifferentiation are vital in metastasis and treatment resistance[60] (**Fig. 5**). Prostate cancer originates from secretory prostate epithelial cells, where AR, a transcription factor regulated by androgen, plays a key role in driving the differentiation[61,62]. Androgen-targeted therapies (ATTs) constitute the primary treatment options for metastatic prostate cancer, and they are most effective in well-differentiated prostate cancer cells with high AR activity[63]. After prolonged treatment with ATTs, the cancers eventually progress towards metastatic castration-resistant prostate cancer (mCRPC), which is highly recalcitrant to therapy[64]. In response to more potent ATTs, the prostate cancer cells adapt to escape reliance on AR with low AR activity[65]. The loss of differentiation pressure results in altered states of lineage plasticity in prostate tumors[66,67]. The most well-defined form of treatment-induced plasticity is neuroendocrine transformation. NEPC is highly aggressive that often manifests with visceral metastases, and currently lacks effective therapeutic options[66,68]. Recent studies have pointed toward the existence of additional prostate cancer phenotypes that emerge through lineage plasticity and metastasis of malignant cells[14]. This includes malignant phenotypes such as low AR signaling and DNPC, which lack AR activity and NE features. These additional phenotypes, resulting from the mechanisms of resistance to AR inhibition, can likewise be characterized by distinct gene expression patterns. Presumably, these phenotypes represent an intermediate or transitory state of the progression trajectory from high AR activity to neuroendocrine transdifferentiation[69].

Here, we performed a pooled analysis of scRNA-seq target data consisting of 836 malignant cells derived from tumors collected from 11 tumors (mCRPC)[70]. We initially classified cells into three categories-AR activity high (AR+/NE–, ARAH), AR activity low (AR_{low}/NE–, ARAL), and neuroendocrine (AR–/NE+, NE). The classification was based on the known molecular signatures associated with ARPC and NEPC genes[69] (**Fig. 5b**, **Supplementary Methods 2.2**). We used the pre-trained model built on the Karthaus *et al.*[32] dataset for transfer learning. We used 81 advanced metastatic prostate cancer patient samples with paired survival information from Abida *et al.*[23] as the survival data. We subsequently assessed the relative aggressiveness of these high-level categories using *phenotype algebra* (**Fig. 5e, f**). We observed that subtracting the latent signature associated with the NE subtype from the tumor led to the largest drop in the predicted risk score, aligning with the anticipated order of survival[71]. In contrast, removing the ARAH subtype from the *total tumor* had a minimal impact on the predicted risk score. Furthermore, subtracting the ARAL subtype resulted in a predicted risk score between ARAH and NE, which supports the hypothesis that ARAL represents an intermediate state between ARAH and NE[69]. Upon further examination of the tumor prognosis by removing the subtypes in a combined state (ARAH+ARAL+NE), the *total tumor* had the highest positive improvement. The survival probability graph also followed the same order.

Grouping prostate cancer cells into three high-level categories is an oversimplified view of the actual heterogeneity of advanced prostate cancer biology. Herein, SCellBOW clusters could be utilized to discover novel subpopulations based on the single-cell expression profiles that result from therapy-induced lineage plasticity. Subsequently, *phenotype algebra* can assign a relative rank to these clusters under a negative selection pressure based on their aggressiveness. We utilized this concept to cluster the malignant cells from He *et al.* using SCellBOW, resulting in eight clusters, and then we predicted the relative risk for each cluster (**Fig. 6a, b**). This approach

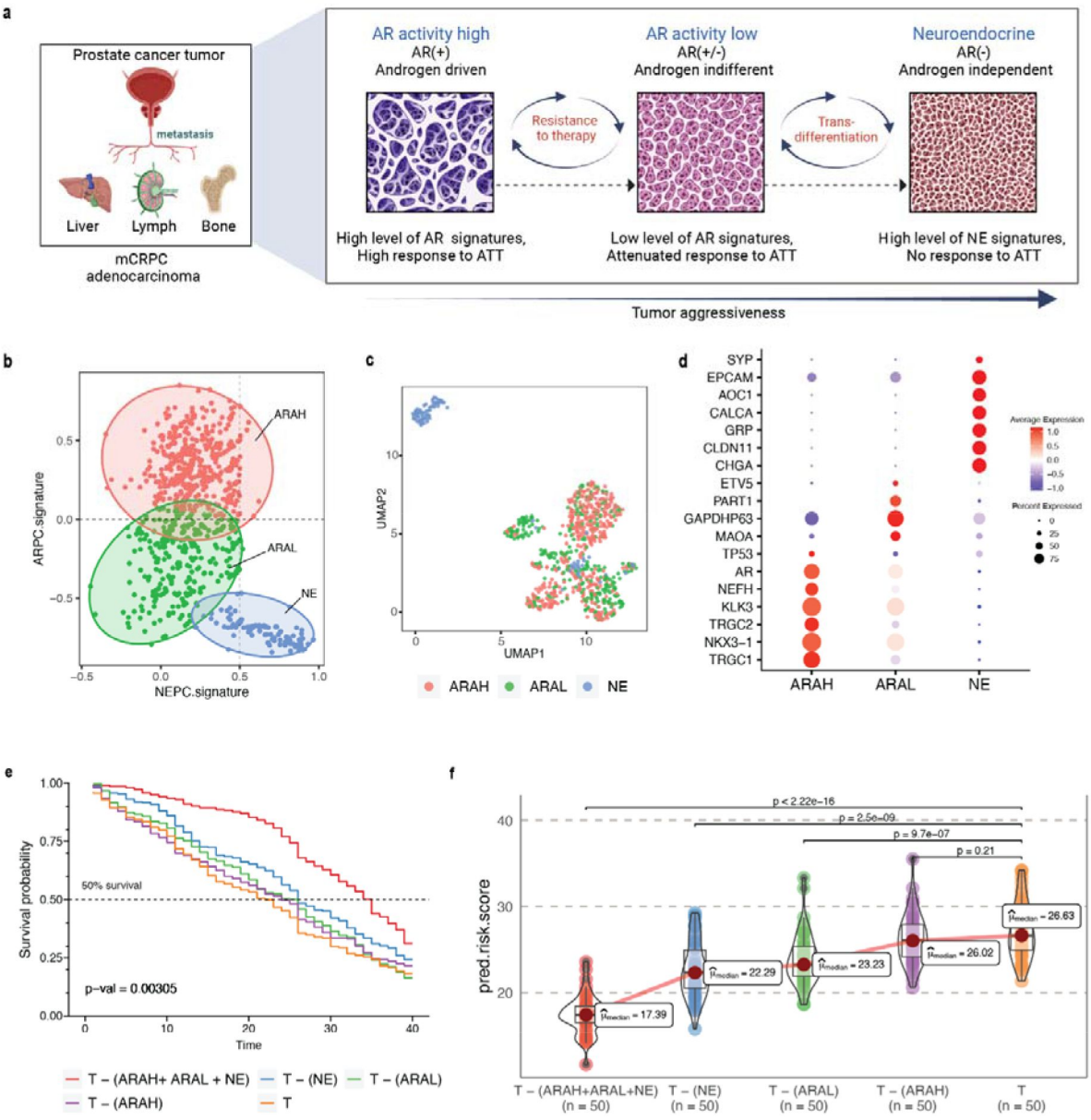

Fig. 5.

Phenotype algebra of metastatic prostate cancer data based on AR-and NE-activity.

a, Schematic of the transdifferentiation states underlying lineage plasticity that occurs during mCRPC progression from an ARPC to NEPC.

b, Scatter plot of GSVA scores of ARPC and NEPC gene sets, K-means clustering was used to allocate cells into the three high-level ARAH, ARAL, and NEPC categories.

c, UMAP plot for projection of SCellBOW embedding colored by ARAH, ARAL, and NEPC.

d, Heatmap showing the top differentially expressed genes (y-axis) between each high-level category (x-axis) and all other cells, tested with a Wilcoxon rank-sum test.

e, Survival plot for mCRPC phenotypes based on *phenotype algebra*. The t*otal tumor* is denoted by *T*.

f, Violin plot for predicted risk scores for mCRPC phenotypes-ARAH, ARAL, and NEPC.

enables a novel and more refined understanding of lineage plasticity states and characteristics that determine aggressiveness during prostate cancer progression. This goes beyond the conventional categorization into ARAH, ARAL, and NE.

To further elucidate these altered cellular programs, we performed a gene set variation analysis (GSVA)[72] based on the AR-and NE-activity (**Fig. 6e**, **Supplementary Methods 3, and Supplementary Data 1**). Our result showed that CL4 is characterized by the highest expression of NE-associated genes and the absence of AR-regulated genes, indicating the conventional NEPC subtype. Despite CL4 having the strongest NEPC signatures, eliminating CL2 from the tumor conferred an even higher aggressiveness level. Overall, we observed that, unlike other clusters, CL2 is composed of cells from the majority of drug-treated patients and multiple metastatic sites (**Fig. 6d**). This highlights that the clustering is not confounded by the individuals or the tissue origin, as often observed during integrative analysis of tumor scRNA-seq data. CL2 instead features a unique gene expression profile common to these cells. Moreover, CL2 has a mixed signature entailing ARAH, ARAL, and NE, indicating the emergence of a more transdifferentiated subtype as a consequence of therapy-induced lineage plasticity (**Fig. 6c**). Even though CL2 shows NE signature, it is distinguished by the gene expression signature induced by the inactivation of the androgen signaling pathway due to ATTs. As a consequence, cells manifesting this novel signature are grouped into a single cluster. Among all clusters, CL1 and CL3 resemble the traditional ARAH subtype. According to our *phenotype algebra* model, excluding CL6 and CL7 from the *total tumor* yielded the highest risk score. Brady *et al.*[13] have broadly partitioned metastatic prostate cancer into six phenotypic categories using digital spatial profiling (DSP) transcript and protein abundance data in spatially defined metastasis regions. To categorize the SCellBOW clusters into these broad phenotypes, we performed Pearson's correlation test between averaged DSP expression measurements of the six phenotypes and averaged scRNA-seq expression of the SCellBOW clusters (**Fig. 6f**). The results revealed that CL2 has the highest correlation with the phenotype defined by lack of expression of AR signature genes and low or heterogeneous expression of NE-associated genes (AR–/NE_{low}). Similarly, as expected, CL4 showed the highest correlation with the NEPC phenotype defined by positive expression of NE-associated genes without AR activity (AR–/NE+). Meanwhile, CL6 exhibits a closer resemblance to a DNPC phenotype (AR–/NE–).

To gain a deeper understanding of these modified cellular processes in CL2 compared to other clusters, we conducted cluster-wise functional GSVA based on the hallmarks of cancer (**Fig 6g**). We observed that CL2 exhibited the least prostate cancer signature, indicating that this cluster has deviated from prototypical prostate cancer behavior. It has rather dedifferentiated into a more aggressive phenotype as a consequence of therapy-induced lineage plasticity[73]. CL2 exhibited the highest enrichment of genes related to cancer stemness compared to other clusters. CL2 showed pronounced repression of epithelial genes that are downregulated during the epithelial-to-mesenchymal transition (EMT). Furthermore, there is a lack of expression of genes that are upregulated during the reversion of mesenchymal to epithelial phenotype (MET). Thus, the cells in CL2 are undergoing a process of dedifferentiation from being epithelial cells and activating the mesenchymal gene networks. Existing studies have reported that adaptive resistance is positively correlated with the acquisition of mesenchymal traits in cancer[74,75]. CL2 was enriched with signatures associated with metastasis and drug resistance. Specifically, CL2 showed downregulation of the genes involved in drug metabolism and cellular response while upregulation of the genes associated with drug resistance. In cancer cells, the acquisition of stemness-like as well as cell dedifferentiation (mesenchymal) traits can facilitate the formation of metastases and lead to the development of drug resistance[76]. Further analysis of the metastatic potential of CL2 indicated that the cluster is enriched with genes associated with vasculature and angiogenesis. Tumors induce angiogenesis in the veins and capillaries of the host tissue to become vascularized, which is crucial for their growth and metastasis[77]. Thus, based on our findings, we anticipate that CL2 corresponds to a highly aggressive and dedifferentiated subclone of mCRPC within the lineage plasticity continuum, correlating to poor patient survival and positive

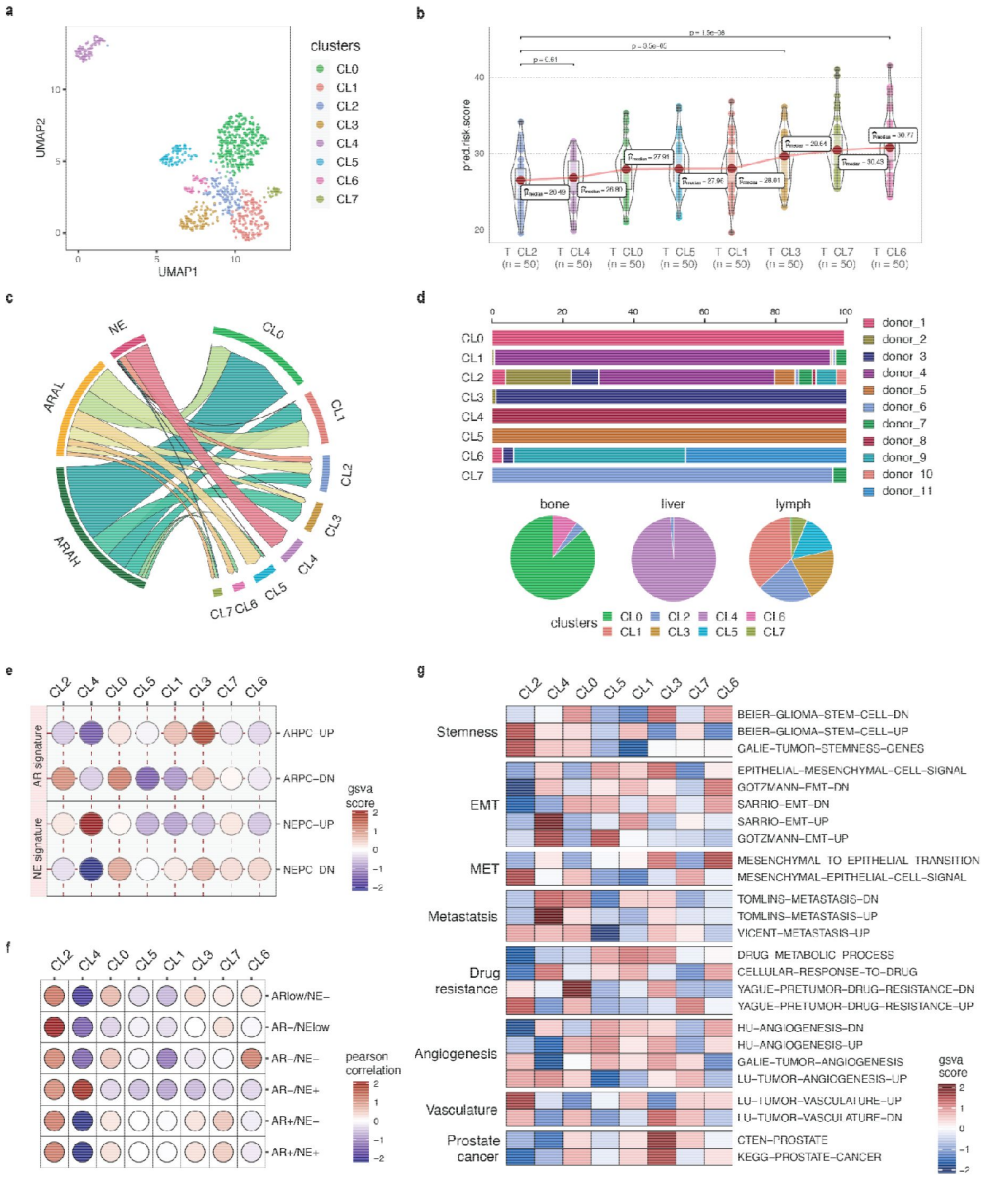

Fig. 6.

Phenotype algebra of metastatic prostate cancer data based on SCellBOW clusters.

a, UMAP plot for projection of embeddings with coloring based on the SCellBOW clusters at resolution 0.8. CL is used as an abbreviation for cluster.

b, Violin plot of *phenotype algebra*-based cluster-wise risk scores for SCellBOW clusters based on *phenotype algebra*-based predictions.

c, Illustration of the distribution of cells from the three high-level groups-ARAH, ARAL, and NEPC across the SCellBOW clusters.

d, Patient and organ site distribution across the SCellBOW clusters.

e, Bubble plot of row-scaled GSVA scores for custom curated gene sets containing activated and repressed AR-and NE-signatures.

f, Correlation plot of six phenotypic categories based on DSP gene expression correlated with the SCellBOW clusters based on scRNA-seq gene expression. The six phenotypic categories are defined by Brady *et al.* based on the activity of AR and NE programs.

g, Top gene sets correlated with SCellBOW clusters. Signatures were collected from the C2 ''curated'', C5 ''Gene Ontology'', and H ''hallmark'' gene sets from mSigDB[94]. Ranking by row scaled GSVA scores of one cluster against all others.

metastatic status. Our cluster-wise *phenotype algebra* results imply that the traits of the androgen and neuroendocrine signaling axes are not the exclusive defining features of the predicted risk ranking and that other yet under-explored biological programs play additional important roles.

Overall discussion

In this work, we introduce SCellBOW, a scRNA-seq analysis framework inspired by NLP-based transfer learning approach that can be utilized to decipher molecular heterogeneity from scRNA-seq profiles and infer survival risks associated with the individual malignant subpopulations in tumors. In this novel approach, cells are treated as a bag-of-molecules, similar to representing a document as a "bag" of words (BOW), representing molecules in proportion to their abundance within each cell. SCellBOW uses a document embedding architecture to preserve the cellular similarities observed in the gene expression space. SCellBOW-learned neuronal weights are transferable. The model can use the knowledge acquired from a source dataset to warm start the learning process on a target dataset. Although Doc2vec is not commonly used for transfer learning, our experiments, which involve repurposing a pre-trained, modified Doc2vec model for scRNA-seq analysis, establish promising use cases for transfer learning (**Supplementary Fig. 7 and Supplementary Note 5**).

SCellBOW accomplishes two major tasks: a. single-cell representation learning under a transfer learning framework facilitating high-quality cell clustering and visualization of scRNA-seq profiles. b. *phenotype algebra*, enabling the attribution of survival risk to malignant cell clusters obtained from tumor scRNA-seq data. For both tasks, we compared SCellBOW with several state-of-the-art single-cell analysis tools, including ones based on complex language modeling architectures (e.g., scETM uses a topic model, scBERT uses a transformer-inspired model). SCellBOW exhibited consistency in characterizing cellular heterogeneity through clustering and survival risk (aggressiveness) stratification of malignant cell subtypes.

With the SCellBOW, we observe considerable improvement in the quality of clustering and *phenotype algebra* results obtained from scRNA-seq target datasets. The existing single-cell transfer learning tools, such as scETM, ItClust, and scBERT, are pre-trained on large amounts of pre-annotated single-cell data to achieve good performance. However, using source data annotations as a reference limits the identification of new cell types in the target data. Moreover, the degree of imbalance in the cell type distribution substantially influences the performance of these tools[78]. It is important to realize that exploratory studies involving single-cell expression profiling require unsupervised data analysis. SCellBOW addresses these challenges by allowing self-supervised pre-training on gene expression datasets, learning the general syntax and semantic patterns from the unlabelled scRNA-seq dataset. Additionally, it is less susceptible to overfitting[79], especially with small training datasets, owing to the shallow neural network used in SCellBOW, which enables faster convergence compared to other studied transfer learning methods (**Supplementary Table 5**). Notably, SCellBOW, unlike ItClust and scETM, does not limit model training to genes intersecting between the source and the target datasets and is independent of the source data cell type annotation. Moreover, while methods such as scETM, ItClust, and scBERT utilize numerical gene expression values directly as an input to their network architectures, SCellBOW, for the first time, encodes gene expression profiles into documents in their native format.

We compared the clustering capability of SCellBOW to popular single-cell analysis techniques as well as some recently proposed transfer learning approaches. An array of scRNA-seq datasets was used for the same, representing different sizes, cell types, batches, and diseases. As evident from the comparative analyses, SCellBOW exhibits robustness and consistency across all data and metrics. We also reported tangible benefits of transfer learning from apt source data. For example, transfer learning with a large number of tumor-adjacent normal cells from the prostate as the source and healthy prostate cells as the target adequately portrayed the cell type diversity therein.

SCellBOW reliably clustered pancreatic islet-specific cells that were processed using varied single-cell technologies. The prevalence of single-cell expression profiling and the heterogeneity of PBMCs make it one of the best-studied tissue systems in humans[80]. We isolated matched PBMCs and splenocytes from two healthy donors and two brain-dead donors (the cause of death is subarachnoid hemorrhage on aneurysm rupture). The utilization of this data presents a more intricate problem, where the cells originate from diverse biological and technical replicates, conditions, individuals, and organs. SCellBOW-based analysis of PBMCs with near-ground-truth cell type annotation confirmed its ability to adequately decipher underlying cellular heterogeneity. Our results allude to a visible improvement in scRNA-seq analysis outcomes, even with small sample sizes and multiple covariates, when contrasted with other best-practice single-cell clustering methods.

Beyond robust identification of cell type clusters, SCellBOW uses algebraic operations to analyze the cellular phenotypes that could potentially identify their contribution to patient survival outcomes. We have leveraged the power of word algebra through document embeddings to perform pseudo-grading of cancer subtypes based on their aggressiveness. This involves comparing the likelihood of a subtype being eliminated from the entire tumor under negative selection pressure. In addition to overall survival probability, SCellBOW assigns a risk score to discern the differences between equally aggressive subclones that may be hard to decipher from their survivability profiles. The *phenotype algebra* module exhibits resilience in describing the risk associated with malignant cell subpopulations arising from various types of cancer, which otherwise cannot be accomplished using gene expression data (**Supplementary Fig. 8**h-j). We demonstrated a potential use case in estimating subclonal survival risk in three cancer types: GBM, BRCA, and mCRPC. Several examples were shown where simple algebraic operators such as'+' and'–' could derive clinically intuitive outcomes. For example, subtracting the CLA phenotype from the whole tumor resulted in an improvement in the survival risk in a GBM patient compared to MES and PRO subtypes. Upon further probing of the tumor prognosis by removal of the subtypes in combinations from the tumor, specifically CLA and MES subtypes, which are known to be the most aggressive, we observed the highest improvement. To summarize, SCellBOW allows simulations involving multiple phenotypes. Our subsequent investigation focused on a BRCA dataset that harbors a more complex subtype structure. We observed a deviance in our results from the general perception about the relative aggressiveness of the NORMAL subtype in breast cancer. Notably, removing the NORMAL subtype from the *total tumor* was associated with the highest improvement in prognosis. This contradicts the common assumption that a NORMAL subtype is an artifact resulting from a high proportion of normal cells in the tumor specimen[81]. Despite indications that these tumors often do not respond to neoadjuvant chemotherapy[58], the clinical significance of the NORMAL subtype remains uncertain due to a limited number of studies. The NORMAL breast cancer cells are poorly characterized and heterogeneous in nature. As per the recent classification of the breast cancer subtypes, the NORMAL subtype has been identified to be a potentially aggressive molecular subtype, referred to as claudin-low tumors[59]. This evidence suggests that the NORMAL subtype of breast cancer is potentially an aggressive molecular subtype, which is consistent with the prediction made by SCellBOW.

In advanced metastatic prostate cancer, molecular subtypes are still poorly defined, and characterizing novel or more fine-grained subtypes with clinical implications on patient prognosis is an active field of research. As a proof-of-concept, we executed SCellBOW on the three mCRPC subtypes, namely ARAH, ARAH, and NE. We observed that eliminating the subtypes individually and in combinations (ARAH+ARAL+NE) exhibited a clinically intuitive change in prognosis using SCellBOW. To date, mCPRC classification has largely been confined to the gradients of AR and NE activities. There is considerable scope for embracing more fine-grained subtypes to better explain clonal selection and epithelial plasticity in drug resistance in wider use cases. Convinced by the overall performance of SCellBOW, we applied *phenotype algebra* on the SCellBOW clusters obtained from the He et al. mCRPC dataset[70]. We observed a misalignment from the general perception that the neuroendocrine subtype features the worst prognosis. Our results pointed

towards the existence of a more aggressive and dedifferentiated subclone in the lineage plasticity continuum. This subpopulation of de-differentiated cells in mCRPC, AR−/NE_{low}, exhibit greater aggressiveness and have distinct gene expression patterns that do not match those of any previously identified molecular subtypes. This novel subclone shares gene expression signatures with both androgen-repressed and neuroendocrine-activated genes. GSVA analysis of this hitherto unknown phenotype offered insights into its putative functional characteristics, which can be broadly defined by stemness, EMT, MET, and drug resistance. To summarize, SCellBOW clustering combined with *phenotype algebra* can provide novel insights into factettes of cancer biology. This can be used to delineate features of lineage plasticity and aid in better classification of molecular phenotypes with clinical significance. Once the subpopulations are identified, delving into the intricate mechanisms governing their resistant behavior can provide invaluable insights into for designing new drugs. Further, such a tool can empower medical oncologists and oncology researchers to develop more personalized and systemic treatment regimes for cancer patients.

Conclusion

SCellBOW is a scRNA-seq analysis tool that can be utilized to decipher molecular heterogeneity from scRNA-seq profiles and infer survival risks associated with the individual malignant subpopulations in tumors. Our main contributions are as follows: a. We proposed SCellBOW that learns a distributed representation (fixed-length embedding per cell) of single-cells. Despite being a computationally inexpensive and simple architecture, SCellBOW outperforms transformer-based approaches such as scBERT on both single-cell clustering and cancer cell risk stratification tasks. b. We also introduced the malignant cell “pseudo-grading” task as a computationally tractable problem. The pseudo-grading module of SCellBOW performs joint representation learning of externally sourced tumor bulk RNA sequencing data (e.g., TCGA) and tumor scRNA-seq data of interest. Finally, a survival prediction model trained on bulk expression profiles is used for survival risk stratification for each of the malignant cell clusters derived from the tumor scRNA-seq data. We compared SCellBOW with existing best practice methods for its ability to precisely represent phenotypically divergent cell types across multiple scRNA-seq datasets, including an in-house generated human splenocyte and matched PBMC dataset. SCellBOW has proven effective in characterizing poorly defined metastatic prostate cancer. We identified a subpopulation of dedifferentiated cells in mCRPC, AR−/NElow, that exhibit greater aggressiveness and have distinct gene expression patterns that do not match those of any previously identified molecular subtypes. We could trace this back in a large-scale spatial omics atlas of 141 well-characterized metastatic prostate cancer samples at the spot resolution. In the case of breast cancer, our results indicated that the normal-like subtype, which was previously considered an artifact of high normal cell content in the tumor sample, may be among the most aggressive ones, which is concordant with recent reports. In conclusion, the robust clustering and risk-stratification capabilities of SCellBOW hold promise for tailored therapeutic interventions by identifying clinically relevant subpopulations and their impact on prognosis.

Material and methods

Overview of SCellBOW components and functionalities

SCellBOW has two main applications: **a. *Cell embedding and clustering.*** In this work, we demonstrat single-cell representation learning using a source and target scRNA-seq data. The source data is typically any large scRNA-seq data capable of priming a neural architecture for improved representation learning on the target data of interest. Note that in this case, cells in the source or the target data need not be annotated. This transfer learning capability is meant to obtain improved fixed-length embeddings for cells in the target data by leveraging the

transcriptomic patterns from existing datasets. **b. *Malignant cluster specific survival risk attribution.*** This feature is only applicable to malignant cell clusters. SCellBOW's *phenotype algebra module* predicts the aggressiveness associated with each malignant cell cluster. SCellBOW computes fixed-length embeddings jointly for scRNA-seq target dataset (from the tumor under study) and tumor bulk RNA-seq profiles with patient-survival data (e.g., TCGA). Notably, embeddings of transcriptomes from scRNA-seq target data and bulk RNA-seq survival data have been determined concurrently by recalibrating the pre-trained model. A survival prediction model trained on the bulk RNA-seq embeddings and patient-survival data is then tested on the embeddings of the target data to make predictions about the degree of aggressiveness exhibited by the tumor variants.

The steps involved in the above two tasks are depicted in **Figure 1** (**Fig. 1**). The granular details involved in each of these steps can be found in the below subsections.

Data preparation for clustering

SCellBOW follows the same preprocessing procedure for the source and target data. SCellBOW first filters the cells with less than 200 genes expressed and genes that are expressed in less than 20 cells. The thresholds may vary depending on the dimension of the dataset (**Supplementary Table 2**). After eliminating low-quality cells and genes, SCellBOW log-normalizes the gene expression data matrix. In the first step, CPM normalization is performed, where the expression of each gene in each cell is divided by the total gene expression of the cell, multiplied by 10,000. In the second step, the normalized expression matrix is natural-log transformed after adding one as a pseudo count. Highly variable genes are selected using the highly_variable_genes() function from the Scanpy package. SCellBOW performs a z-score scaling on the log-normalized expression matrix with the selected highly variable genes. SCellBOW can handle data in different formats, including UMI count, FPKM, and TPM. The UMI count data follow the same preprocessing procedure as above. SCellBOW skips the normalization step for TPM and FPKM since their lengths have been normalized.

Creating a corpus from source and target data

The gene expression data matrix is analogous to the term-frequency matrix, which represents the frequency of different words in a set of documents. In the context of genomic data, a similar concept can be used to represent the expression level of a gene (words) in a set of cells (documents). Let $E \in R^{G \times C}$ be the gene expression matrix obtained from a scRNA-seq experiment, where each value $E_{g,c}$ of the matrix indicates the expression value of a gene $g \in G$ in a cell $c \in C$ obtained after Scanpy preprocessing. SCellBOW generates embeddings by taking two input datasets: a source and a target data matrix. The source dataset contains the initial weights of the neural network model, while the target data contains the cells that require clustering or malignancy potential ranking. Before generating the embeddings, SCellBOW performs feature scaling on the data matrix, rescaling each feature to a range of [0, 10] as follows.

$$E'_{g,c} = int\left(\frac{E_{g,c} - min}{max - min}\right), \quad (1)$$

where scaling is applied to each entry with max = 10 and min = 0. This establishes the equivalence between term-frequency and gene expression matrix. Here, we consider that a specific gene is a word, and the expression of the gene in a cell corresponds to the number of copies of the word in a document. To scale the data matrix, we have used the MinMaxScaler() function from the scikit-learn package[82]. To create the corpus, SCellBOW duplicates the name of the expressed gene in a cell as many times as the gene is expressed. SCellBOW shuffles the genes in each cell to ensures a uniform distribution of genes across the cell, removing any positional bias within the dataset[19]. We verified that this randomization does not change the outcomes dramatically (**Supplementary Fig. 9**). The resulting gene names are treated as the words in the document. To build the vocabulary from a sequence of cells, a tag number is associated with each cell using the

TaggedDocument() function in the Gensim package[83]. For each gene in the documents, a token is assigned using a module called tokenize with a word_tokenize() function in NLTK package[84] that splits gene names into tokens.

The SCellBOW network

SCellBOW produces a low-dimensional fixed-length embedding of the single-cell transcriptome using transfer learning. The data matrix is transformed from $E' \in Z^{G \times C}$ feature space to $E'' \in R^{d \times C}$ latent feature space of d dimensions, with $d \ll G$. To generate the embeddings, SCellBOW trains a Doc2vec distributed memory model of paragraph vectors (PV-DM) model. The PV-DM model is similar to bag-of-words models in Word2vec[85]. The training corpus in Doc2vec contains a set of documents, each containing a sequence of words $W = \{w_1, w_2, .., w_T\}$ that forms a vocabulary V. The words within each document are treated as shared among all documents. The training objective of the model is to maximize the probability of predicting the target word w_t, given the context words that occur within a fixed-size window of size n around w_t in the whole corpus as follows.

$$maximize\left(\frac{1}{T}\sum_{t=1}^{T} \quad log\ log\ \ Pr(w_t \mid w_{t-n}, .., w_{t+n})\right). \tag{2}$$

The probability can be modeled using the hierarchical softmax function as follows.

$$Pr(w_t \mid w_{t-n}, .., w_{t+n}) = \frac{e^{y w_t}}{\sum_i \quad e^{y_i}}, \tag{3}$$

where each of y_i computes the log probability of the word w_t normalized by the sum of the log probabilities of all words in V. This is achieved by adjusting the weights of the hidden layer of a neural network. To build the Doc2vec model, we used the doc2vec() function in the Gensim library. The initial learning rate was set to 0.025, and the window size to 5. We chose the PV-DM training algorithm and set the embedding vector size to 300 as the default parameter. The choice of embedding vector size may be adjusted according to the size and dimensionality of the dataset.

Fine-tuning using transfer learning

At first, SCellBOW is pre-trained with a source data matrix $E'_{src} \in Z^{G_{src} XC_{src}}$ with G_{src} genes and C_{src} cells. The source data matrix sets the initial weights of the neural network model. During transfer learning, SCellBOW fine-tunes the weights learned by the pre-trained model from E'_{src} using a target dataset $E'_{trg} \in Z^{G_{trg} XC_{trg}}$ with G_{trg} genes and C_{trg} cells. This facilitates faster convergence of the neural network compared to starting from randomly initialized weights. The output layer of the network is a fixed-length low-dimension embedding (a vector representation) for each cell $c \in C_{trg}$. To infer the latent structure of E_{trg} single-cell corpus, we used the infer_vector() function in the Gensim package to produce the dimension-reduced vectors for each cell in C_{trg}. This step ensures the network can map the target data into a low dimensional embedding space R^d; i.e., $E' \rightarrow E''$,where $E'' \in R^{d \times C}$. The resulting embeddings can be used for various downstream analyses of the single-cell data, such as clustering, visualization of cell types, and *phenotype algebra*.

Visualization of SCellBOW clusters

SCellBOW maps the cells to low-dimensional vectors in such a way that two cells with similar gene expression patterns will have the least cosine distance between their inferred vectors.

$$similarity(a, b) = \frac{a \,.b}{\|a\| \times \|b\|}. \tag{6}$$

After generating low-dimensional embeddings for the cells, SCellBOW identifies the groups of cells with similar gene expression patterns. To determine the clusters, SCellBOW uses the Leiden algorithm[86] on the embedding matrix of the target dataset. We used the leiden() function in the

Scanpy package with a default resolution of 1.0. The resolution might vary in the reported results depending on the dimension of the target dataset. To visualize the clusters within a two-dimensional space, SCellBOW uses the umap() function from the Scanpy library.

Data preparation for phenotype algebra

To perform phenotype algebra, three independent datasets are required. Two of the three datasets are scRNA-seq datasets used for transfer learning (source and target data). SCellBOW preprocesses the source data matrix using the standard preprocessing steps. SCellBOW uses an additional bulk RNA-seq gene expression matrix (referred to as survival data). The samples are paired with survival information (e.g., vital status, days to follow-up, days to death). In the survival data, samples without follow-up time or survival status and samples with clinical information but no corresponding RNA-seq data were excluded. SCellBOW accounts for unequal cell distribution across different classes in the target data matrix. SCellBOW up samples the imbalanced target dataset by generating synthetic samples from the minority class. The value of the synthetic minority class sample is determined by interpolating between its neighboring cells from the same class. We used SMOTE[87] from the imblearn python library[88]. This confirms that the cell type proportions do not confound the output of phenotype algebra.

Generating vectors for phenotype algebra

SCellBOW constructs different pseudo-bulk vectors from the target dataset. The vectors are constructed either based on the user-defined subtypes or SCellBOW clusters. These vectors are created by calculating the mean gene expression of cells belonging to each phenotype. SCellBOW first constructs a reference pseudo-bulk vector by taking the average of all malignant cells (*total tumor*). This accounts for the tumor in its entirety and acts as the reference for the relative ranking of the different groups. Similarly, for each group in the dataset, SCellBOW then constructs a query pseudo-bulk vector by taking the average gene expression of the cells in that individual group. SCellBOW can also perform addition or subtraction operations on the query vectors of two or more cellular phenotypes to evaluate their risk in a combined state such as T-(a + b), where T is the embedding for *total tumor*, a and b are the embeddings for query vectors of two distinct cellular phenotypes. Following this, SCellBOW concatenates the bulk RNA-seq data matrix and the reference and query vectors based on common genes. We used the concatenate() function in the AnnData python package[89]. The resulting combined dataset is then passed to the pre-trained model, which maps it to a lower-dimensional embedding space.

Survival risk attribution

SCellBOW infers the survival probability and predicted risk score for the user-defined tumor subtypes and SCellBOW clusters. The survival risk of the different groups is predicted by fitting a random survival forest (RSF)[90] machine learning model with SCellBOW embeddings. At first, the RSF is trained with the survival information combined with bulk RNA-seq embeddings obtained from transfer learning. SCellBOW subtracts each query vector from the reference vector and fits the difference of these vectors as input to the RSF model to infer the survival probability $S(t)$. The survival probability computes the probability of occurrence of an event beyond a given time point t as follows.

$$[T < t] = \int_{-\infty}^{t} f(x)dx, \tag{4}$$

$$S(t) = 1 - Pr\,Pr\,[T < t] = Pr\,Pr\,[T > t]\,, \tag{5}$$

where T denotes the waiting time until the event occurs and f(x) is the probability density function for the occurrence of an event. SCellBOW computes the survival probability using the predict_survival_function() from the scikit-survival RandomSurvivalForest python package[91] with n_estimators = 1000.

In addition to survival probability, SCellBOW can estimate the relative aggressiveness of different malignant phenotypes by assigning a risk score for distinct groups. To infer the predicted risk score, SCellBOW first trains 50 bootstrapped RSF models using 80% of the training set for each iteration. The training data is sub-sampled using different seeds for every iteration. We used the predict() function from the scikit-survival package to compute the risk score of each of the input vectors. SCellBOW derives the median of the predicted risk score for each group from the 50 bootstrapped models. The *phenotype algebra* model assigns groups with shorter survival times a lower rank by considering all possible pairs of groups in the data. The groups with a lower predicted risk score after removal from the reference pseudo-bulk are considered more aggressive, as they are associated with shorter survival times.

Description of datasets for model evaluation

To evaluate the performance of SCellBOW, we used fifteen publicly available scRNA-seq datasets and an in-house scRNA-seq dataset spanning different cell types, sizes, and diseases (**Supplementary Table 2)**. Four use cases were constructed to benchmark the clustering efficiency, each involving a pair of single-cell expression datasets. In the first use case, we used ~120,300 non-cancerous human prostate cells from Karthaus *et al.*[32] and ~28,600 healthy prostate cells from Henry *et al.*[33]. The second use case consisted of two PBMC datasets from Zheng *et al.*[34] containing approximately 68,000 cells and 2,700 cells. For the third use case, we combined three independent batches of pancreatic islet cells from Baron *et al.*[35], Muraro *et al.*[36], and Wang *et al.*[37], with a total of 11,181 cells as the source data and 2,068 cells from Segerstolpe *et al.*[38] as the target data. In the fourth use case, we used the Zheng *et al.* data containing approximately 68,000 cells as the source data and the in-house dataset isolated from matched spleen and PBMC samples from multiple patients as the target data.

To validate the accuracy of *phenotype algebra*, we constructed three use cases from publicly available tumor datasets comprising malignant cells from GBM, BRCA, and mCRPC patient tumors. Each instance involved a pair of single-cell expression datasets and a bulk RNA-seq dataset paired with clinical information. For ease of reference, we stick to the following nomenclature – a) the source data comprises the scRNA-seq dataset used for model pre-training, b) the target data comprises malignant cells under investigation, and c) the survival data comprises tumor bulk RNA-seq samples. In the case of GBM, we used two scRNA-seq datasets comprising approximately 12,074 cells and 4,508 cells obtained from Neftel *et al.*[47] and Couturier *et al.*[46], respectively. In the case of BRCA, we retrieved triple-negative breast cancer scRNA-seq datasets from Wu *et al.*[51] and Zhou *et al.*[53] with approximately 24,271 cells and 545 malignant cells, respectively. In both the use cases, the bulk RNA-seq was obtained from TCGA (*https://portal.gdc.cancer.gov*). In the third use case, we acquired author-annotated malignant cells from He *et al.*[70]. A total of 836 malignant cells were derived from 11 patients and three metastasis sites: bone, lymph node, and liver. We retrieved 81 mCRPC bulk RNA-seq samples with paired survival from Abida *et al.*[23]. We obtained the gene expression of 141 regions of interest determined by digital spatial gene expression profiling of mCRPC from Brady *et al.*[13]. We used this data to correlate the He *et al.* scRNA-seq gene expression with the DSP gene expression of the six phenotypic categories based on the AR-and NE-activity: AR+/NE−, AR_{low}/NE−, AR−/NE−, AR−/NE_{low}, AR+/NE+, and AR−/NE+. Details of the datasets are described in **Supplementary Methods (Supplementary Methods 1)**.

Isolation of matched PBMC and splenocyte samples

In this study, we isolated PBMC and splenocyte from four patients (two healthy donors, HD1 18 years old male and HD2 61 years old female; two brain-dead donors, BD1 57 years old female and BD2 58 years old female). PBMCs were isolated from blood after Ficoll gradient selection. Spleen tissue was mechanically dissociated and then digested with Collagenase and DNAseq. Splenocytes were finally isolated after Ficoll gradient selection. Splenocytes and PBMCs were stored in DMSO

with 10% FBS in liquid nitrogen at the Biological Resource Centre for Biobanking (CHU Nantes, Hotel Dieu, Centre de Ressources Biologiques). This biocollection was authorized in May 2013 by the French Agence de la Biomedecine (PFS13-009).

Library preparation

For the in-house dataset, we performed CITE-seq using Hashtag Oligos (HTO) to pool samples into a single 10X Genomics channel for scRNA-seq[92]. The HTO binding was performed following the specified protocol (Total-seq B). After thawing and cell washing, 1M cells were centrifuged and resuspended in 100uL PSE buffer (PBS/FBS/EDTA). Cells were incubated with 10ul human FcR blocking reagent for 10 min at 4°C. 1 uL of a Hashtag oligonucleotide (HTO) antibody (Biolegend) was then added to each sample and incubated at 4°C for 30 min. Cells were then washed in 1mL PSE, centrifuged at 500g for 5 min, and resuspended in 200uL PSE. Cells were stained with 2uL DAPI, 60uM filtered, and viable cells were sorted (ARIA). Cells were checked for counting and viability, then pooled and counted again. Cell viability was set to < 95%. Cells of HD1 spleen were lost during the washing step and thus not used during further downstream processing. Cells were then loaded on one channel of Chromium Next Controller with a 3' single-cell Next v3 kit. We followed protocol CG000185, Rev C, until the library generation stage. For the HTO library, we followed the protocol of the 10X genomics 3' feature barcode kit (PN-1000079) to generate HTO libraries.

Next-generation sequencing and post-sequencing quality control

We sequenced 310pM of pooled libraries on a NOVAseq6000 instrument with an S1(v1) flow-cell. The program was run as follows: Read1 29 cycles / 8 cycles (i7) /0 (i5)/Read2 93 cycles (Standard module, paired-end, two lanes). The FASTQ files were demultiplexed with CellRanger v3.0.1 (10X Genomics) and aligned on the GRCh38 human reference genome. We recovered a total of 6,296 cells with CellRanger, and their gene expression matrices were loaded on R.

We performed the downstream analysis of the in-house matched PBMC-splenocyte dataset in R using Seurat 4.0. For RNA and HTO quantification, we selected cell barcodes detected by both RNA and HTO. We demultiplexed the cells based on their HTO enrichment using the HTODemux() function in the Seurat R package with default parameters[27,93]. We subsequently eliminated doublet HTOs (maximum HTO countlJ>lJ1) and negative HTOs. Singlets were used for further analysis, leaving 4,819 cells and 33,538 genes. We annotated each of the cell barcodes using HTO classification as PBMC and Spleen based on the origin of cells and HD1, HD2, BD1, and BD2 based on the patients (**Supplementary Fig. 4**). We performed automatic cell annotation using the Seurat-based Azimuth using human PBMC as the reference atlas[42] (**Supplementary Note. 2)**. We defined six major cell populations: B cells, CD4 T cells, CD8 T cells, natural killer (NK) cells, monocytes, and dendritic cells. We then manually verified the annotation based on the RNA expression of known marker genes (**Supplementary Fig. 4**d **and Supplementary Table 4**).

Availability of data and materials

The in-house matched PBMC-splenocyte scRNA-seq expression data generated in this study is available in the GEO with accession number GSE221007. Details of the public datasets analyzed in this paper are described in **Supplementary Table 2**.

All source codes are available at GitHub *https://github.com/cellsemantics/SCellBOW*.

Acknowledgements

DS acknowledges the support of the ihub-Anubhuti-iiitd Foundation set up under the NM-ICPS scheme of the DST. JP, AR, PS, and SF thank the biological resource centre for biobanking (CHU Nantes, Hôtel Dieu, Centre de Ressources Biologiques (CRB), Nantes, F-44093, France (BRIF: BB-0033-00040)) and the Genomics Core Facility GenoA, member of Biogenouest and France Genomique, and to the Bioinformatics Core Facility BiRD, member of Biogenouest and Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013) for the use of their resources and their technical support. Parts of the schematics were created with BioRender.com.

Author contributions

DS conceived the study with CCN. JP, Antoine Roquilly, PS, and CF contributed to the experimental design of the matched PBMC-splenocyte CITE-seq study. NB developed the SCellBOW python package and conducted data analyses with assistance from Anja Rockstroh and SSD. SSD and SKT assisted NB in model training. DS supervised the algorithm development. CCN supervised cancer-related analyses and interpretation. Anja Rockstroh supervised and performed specific data analysis and data interpretation tasks. Anja Rockstroh, HK, JP, GA, ML, and BGH assisted in the biological interpretation of results. DS, CCN, and ML supervised the entire study. All authors substantially contributed to manuscript writing and reviewing.

Conflict of interest

DS and GA are stockholders at CareOnco BioTech. Pvt. Ltd. The remaining authors declare no competing interests.

References

1. Dentro S. C., et al. (2021) **Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes** *Cell* **184**:2239–2254

2. Lawson D. A., Kessenbrock K., Davis R. T., Pervolarakis N., Werb Z (2018) **Tumour heterogeneity and metastasis at single-cell resolution** *Nat. Cell Biol* **20**:1349–1360

3. Bhattacharya N., Nelson C. C., Ahuja G., Sengupta D (2021) **Big data analytics in singlelJcell transcriptomics: Five grand opportunities** *Wiley Interdiscip. Rev. Data Min. Knowl. Discov* **11**

4. Kanev K., et al. (2021) **Tailoring the resolution of single-cell RNA sequencing for primary cytotoxic T cells** *Nat. Commun* **12**

5. Dagogo-Jack I., Shaw A. T (2018) **Tumour heterogeneity and resistance to cancer therapies** *Nat. Rev. Clin. Oncol* **15**:81–94

6. Pang L., et al. (2019) **Discovering Rare Genes Contributing to Cancer Stemness and Invasive Potential by GBM Single-Cell Transcriptional Analysis** *Cancers* **11**

7. Poonia S., et al. (2023) **Marker-free characterization of full-length transcriptomes of single live circulating tumor cells** *Genome Res* **33**:80–95

8. Chapman A. R., et al. (2022) **Correlated gene modules uncovered by high-precision single-cell transcriptomics** *Proc. Natl. Acad. Sci. U. S. A* **119**

9. Simeonov K. P., et al. (2021) **Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states** *Cancer Cell* **39**:1150–1162

10. Tickle T., Tirosh I., Georgescu C., Brown M., Haas B. (2019) **inferCNV of the Trinity CTAT Project. Klarman Cell Observatory, Broad Institute of MIT and Harvard**

11. Weinstein J. N., et al. (2013) **The Cancer Genome Atlas Pan-Cancer analysis project** *Nat. Genet* **45**:1113–1120

12. Beltran H., et al. (2016) **Divergent clonal evolution of castration-resistant neuroendocrine prostate cancer** *Nat. Med* **22**:298–305

13. Brady L., et al. (2021) **Inter-and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling** *Nat. Commun* **12**

14. Han H., et al. (2022) **Mesenchymal and stem-like prostate cancer linked to therapy-induced lineage plasticity and metastasis** *Cell Rep* **39**

15. Chawla S., et al. (2022) **Gene expression based inference of cancer drug sensitivity** *Nat. Commun* **13**

16. Nunez J.-J., Leung B., Ho C., Bates A. T., Ng R. T (2023) **Predicting the Survival of Patients With Cancer From Their Initial Oncology Consultation Document Using Natural Language Processing** *JAMA Netw Open* **6**

17. Kim S., et al. (2021) **Deep-Learning-Based Natural Language Processing of Serial Free-Text Radiological Reports for Predicting Rectal Cancer Patient Survival** *Front. Oncol* **11**

18. Zhao Y., Cai H., Zhang Z., Tang J., Li Y (2021) **Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data** *Nat. Commun* **12**

19. Yang F., et al. (2022) **scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data** *Nature Machine Intelligence* **4**:852–866

20. Le Q. V., Mikolov T (2014) **Distributed Representations of Sentences and Documents** *arXiv [cs.CL*

21. Kfoury Y., et al. (2021) **Human prostate cancer bone metastases have an actionable immunosuppressive microenvironment** *Cancer Cell* **39**:1464–1478

22. Feng E., et al. (2022) **Intrinsic Molecular Subtypes of Metastatic Castration-Resistant Prostate Cancer** *Clin. Cancer Res* **28**:5396–5404

23. Abida W., et al. (2019) **Genomic correlates of clinical outcome in advanced prostate cancer** *Proc. Natl. Acad. Sci. U. S. A* **116**:11428–11436

24. Zhu S., Qing T., Zheng Y., Jin L., Shi L (2017) **Advances in single-cell RNA sequencing and its applications in cancer research** *Oncotarget* **8**:53763–53779

25. Baslan T., Hicks J (2017) **Unravelling biology and shifting paradigms in cancer with single-cell sequencing** *Nat. Rev. Cancer* **17**:557–569

26. Zhang Y., et al. (2021) **Precision treatment exploration of breast cancer based on heterogeneity analysis of lncRNAs at the single-cell level** *BMC Cancer* **21**

27. Butler A., Hoffman P., Smibert P., Papalexi E., Satija R (2018) **Integrating single-cell transcriptomic data across different conditions, technologies, and species** *Nat. Biotechnol* **36**:411–420

28. Wolf F. A., Angerer P., Theis F. J (2018) **SCANPY: large-scale single-cell gene expression data analysis** *Genome Biol* **19**

29. Li X., et al. (2020) **Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis** *Nat. Commun* **11**

30. Hu J., et al. (2020) **Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis** *Nat Mach Intell* **2**:607–618

31. Stein-O'Brien G. L., et al. (2021) **Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species** *Cell Syst* **12**

32. Karthaus W. R., et al. (2020) **Regenerative potential of prostate luminal cells revealed by single-cell analysis** *Science* **368**:497–505

33. Henry G. H., et al. (2018) **A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra** *Cell Rep* **25**:3530–3542

34. Zheng G. X. Y., et al. (2017) **Massively parallel digital transcriptional profiling of single cells** *Nat. Commun* **8**:1–12

35. Baron M., et al. (2016) **A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter-and Intra-cell Population Structure** *Cell systems* **3**

36. Muraro M. J., et al. (2016) **A Single-Cell Transcriptome Atlas of the Human Pancreas** *Cell Syst* **3**:385–394

37. Wang Y. J., et al. (2016) **Single-Cell Transcriptomics of the Human Endocrine Pancreas** *Diabetes* **65**:3028–3038

38. Segerstolpe Å., et al. (2016) **Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes** *Cell Metab* **24**

39. Ding J., Regev A (2021) **Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces** *Nat. Commun* **12**

40. Devlin J., Chang M.-W., Lee K., Toutanova K (2018) **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding** *arXiv [cs.CL*

41. Stoeckius M., et al. (2017) **Simultaneous epitope and transcriptome measurement in single cells** *Nat. Methods* **14**:865–868

42. Hao Y., et al. (2021) **Integrated analysis of multimodal single-cell data** *Cell* **184**:3573–3587

43. Dwivedi B., Bhasin M (2021) **Survival Genie: A Web Portal for Single-Cell Data, Gene-Ratio, and Cell Composition-Based Survival Analyses** *Blood* **138**

44. Tang Y., et al. (2021) **Identification of five important genes to predict glioblastoma subtypes** *Neurooncol Adv* **3**

45. Behnan J., Finocchiaro G., Hanna G (2019) **The landscape of the mesenchymal signature in brain tumours** *Brain* **142**:847–866

46. Couturier C. P., et al. (2020) **Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy** *Nat. Commun* **11**

47. Neftel C., et al. (2019) **An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma** *Cell* **178**:835–849

48. Lin N., et al. (2014) **Prevalence and clinicopathologic characteristics of the molecular subtypes in malignant glioma: a multi-institutional analysis of 941 cases** *PLoS One* **9**

49. Patel A. P., et al. (2014) **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma** *Science* **344**:1396–1401

50. Parker J. S., et al. (2009) **Supervised risk predictor of breast cancer based on intrinsic subtypes** *J. Clin. Oncol* **27**:1160–1167

51. Wu S. Z., et al. (2020) **Stromal cell diversity associated with immune evasion in human triplel negative breast cancer** *The EMBO Journal* **39** https://doi.org/10.15252/embj.2019104063

52. Karaayvaz M., et al. (2018) **Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq** *Nat. Commun* **9**

53. Zhou S., et al. (2021) **Single-cell RNA-seq dissects the intratumoral heterogeneity of triple-negative breast cancer based on gene regulatory networks** *Mol. Ther. Nucleic Acids* **23**:682–690

54. Mathews J. C., et al. (2019) **Robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes** *NPJ Breast Cancer* **5**

55. Weigelt B., et al. (2010) **Breast cancer molecular profiling with single sample predictors: a retrospective analysis** *Lancet Oncol* **11**:339–349

56. Hennigs A., et al. (2016) **Prognosis of breast cancer molecular subtypes in routine clinical care: A large prospective cohort study** *BMC Cancer* **16**:1–9

57. Ahn H. J., Jung S. J., Kim T. H., Oh M. K., Yoon H.-K (2015) **Differences in Clinical Outcomes between Luminal A and B Type Breast Cancers according to the St. Gallen Consensus 2013** *Journal of Breast Cancer* **18** https://doi.org/10.4048/jbc.2015.18.2.149

58. Yersal O. (2014) **Biological subtypes of breast cancer: Prognostic and therapeutic implications** *World Journal of Clinical Oncology* **5** https://doi.org/10.5306/wjco.v5.i3.412

59. Liu Z., Zhang X.-S., Zhang S (2014) **Breast tumor subgroups reveal diverse clinical prognostic power** *Sci. Rep* **4**

60. Gupta P. B., Pastushenko I., Skibinski A., Blanpain C., Kuperwasser C (2019) **Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance** *Cell Stem Cell* **24**:65–78

61. Formaggio N., Rubin M. A., Theurillat J.-P (2021) **Loss and revival of androgen receptor signaling in advanced prostate cancer** *Oncogene* **40**:1205–1216

62. Stelloo S., et al. (2015) **Androgen receptor profiling predicts prostate cancer outcome** *EMBO Mol. Med* **7**:1450–1464

63. Lonergan P. E., Tindall D. J (2011) **Androgen receptor signaling in prostate cancer development and progression** *J. Carcinog* **10**

64. Einstein D. J., et al. (2021) **Metastatic Castration-Resistant Prostate Cancer Remains Dependent on Oncogenic Drivers Found in Primary Tumors** *JCO Precis Oncol* **5**

65. Augello M. A., Den R. B., Knudsen K. E. (2014) **AR function in promoting metastatic prostate cancer** *Cancer and Metastasis Reviews* **33**:399–411 https://doi.org/10.1007/s10555-013-9471-3

66. Antonarakis E. S (2019) **Targeting lineage plasticity in prostate cancer** *The Lancet Oncology* **20**:1338–1340 https://doi.org/10.1016/s1470-2045(19)30497-8

67. Beltran H., et al. (2019) **The Role of Lineage Plasticity in Prostate Cancer Therapy Resistance** *Clin. Cancer Res* **25**:6916–6924

68. Yamada Y., Beltran H. (2021) **Current Oncology Reports** https://doi.org/10.1007/s11912-020-01003-9

69. Labrecque M. P., et al. (2019) **Molecular profiling stratifies diverse phenotypes of treatment-refractory metastatic castration-resistant prostate cancer** *J. Clin. Invest* **129**:4492–4505

70. He M. X., et al. (2021) **Transcriptional mediators of treatment resistance in lethal prostate cancer** *Nat. Med* **27**:426–433

71. Wang H. T., et al. (2014) **Neuroendocrine Prostate Cancer (NEPC) Progressing From Conventional Prostatic Adenocarcinoma: Factors Associated With Time to Development of NEPC and Survival From NEPC Diagnosis—A Systematic Review and Pooled Analysis** *Journal of Clinical Oncology* **32**:3383–3390 https://doi.org/10.1200/jco.2013.54.3553

72. Hänzelmann S., Castelo R., Guinney J (2013) **GSVA: gene set variation analysis for microarray and RNA-seq data** *BMC Bioinformatics* **14**

73. Merkens L., et al. (2022) **Aggressive variants of prostate cancer: underlying mechanisms of neuroendocrine transdifferentiation** *J. Exp. Clin. Cancer Res* **41**

74. Hangauer M. J., et al. (2017) **Drug-tolerant persister cancer cells are vulnerable to GPX4 inhibition** *Nature* **551**:247–250

75. Catapano J., et al. (2022) **Acquired drug resistance interferes with the susceptibility of prostate cancer cells to metabolic stress** *Cell. Mol. Biol. Lett* **27**

76. Castellón E. A., Indo S., Contreras H. R. (2022) **Cancer Stemness/Epithelial–Mesenchymal Transition Axis Influences Metastasis and Castration Resistance in Prostate Cancer: Potential Therapeutic Target** *Int. J. Mol. Sci* **23**

77. Lugano R., Ramachandran M., Dimberg A (2020) **Tumor angiogenesis: causes, consequences, challenges and opportunities** *Cell. Mol. Life Sci* **77**:1745–1770

78. Khan S. A., et al. (2023) **Reusability report: Learning the transcriptional grammar in single-cell RNA-sequencing data using transformers** *Nature Machine Intelligence* **5**:1437–1446

79. Wu X., Yang F., Zhou T., Lin X. (2021) **Rethinking the Impacts of Overfitting and Feature Quality on Small-scale Video Classification** *Proceedings of the 29th ACM International Conference on Multimedia* :4760–4764

80. Phongpreecha T., et al. (2020) **Single-cell peripheral immunoprofiling of Alzheimer's and Parkinson's diseases** *Sci Adv* **6**

81. Strehl J. D., Wachter D. L., Fasching P. A., Beckmann M. W., Hartmann A (2011) **Invasive Breast Cancer: Recognition of Molecular Subtypes** *Breast Care* **6**:258–264

82. Pedregosa F., et al. (2011) **Scikit-learn: Machine learning in Python** *Journal of machine Learning research* **12**:2825–2830

83. Rehurek R., Sojka P. (2011) **Gensim--python framework for vector space modelling**

84. Bird S., Klein E., Loper E (2009) **Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit**

85. Mikolov T., Chen K., Corrado G., Dean J (2013) **Efficient Estimation of Word Representations in Vector Space** *arXiv [cs.CL]*

86. Traag V. A., Waltman L., van Eck N. J (2019) **From Louvain to Leiden: guaranteeing well-connected communities** *Sci. Rep* **9**

87. Huang P. J. (2015) **Classification of Imbalanced Data Using Synthetic Over-sampling Techniques**

88. Lemaitre G., Nogueira F., Aridas C. K (2016) **Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning** *arXiv [cs.LG*

89. Virshup I., Rybakov S., Theis F. J., Angerer P., Alexander Wolf F. (2021) **anndata: Annotated data** *bioRxiv* https://doi.org/10.1101/2021.12.16.473007

90. Ishwaran H., Kogalur U. B., Blackstone E. H., Lauer M. S. (2008) **Random survival forests** *The Annals of Applied Statistics* **2** https://doi.org/10.1214/08-aoas169

91. Pölsterl S. (2020) **scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn** *J. Mach. Learn. Res* **21**:1–6

92. Abidi A., et al. (2020) **Characterization of Rat ILCs Reveals ILC2 as the Dominant Intestinal Subset** *Front. Immunol* **11**

93. Stoeckius M., et al. (2018) **Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics** *Genome Biol* **19**

94. Liberzon A., et al. (2015) **The Molecular Signatures Database (MSigDB) hallmark gene set collection** *Cell Syst* **1**:417–425

Editors

Reviewing Editor
Seunghee Hong
Yonsei University, Seoul, Korea, the Republic of

Senior Editor
Murim Choi
Seoul National University, Seoul, Korea, the Republic of

Reviewer #1 (Public Review):

Summary:

This review evaluates the SCellBOW framework, which applies phenotype algebra to obtain vectors from cancer subclusters or user-defined subclusters.

Strengths:

SCellBOW employs an innovative application of NLP-inspired techniques to analyze scRNA-seq data, facilitating the identification and visualization of phenotypically divergent cell subpopulations.

The framework demonstrates robustness in accurately representing various cell types across multiple datasets, highlighting its versatility and utility in different biological contexts.

By simulating the impact of specific malignant subpopulations on disease prognosis, SCellBOW provides valuable insights into the relative risk and aggressiveness of cancer subpopulations, which is crucial for personalized therapeutic strategies.

The identification of a previously unknown and aggressive AR−/NElow subpopulation in metastatic prostate cancer underscores the potential of SCellBOW in uncovering clinically significant findings.

Weaknesses:

The reliance on bulk RNA-seq data as a reference raises concerns about potentially misleading results due to the presence of RNA expression from immune cells in the TME. It is unclear if SCellBOW adequately addresses this issue, which could affect the accuracy of the cancer subcluster vectors.

The method of extracting vectors in phenotype algebra appears to be a straightforward subtraction operation. This simplicity might limit its efficiency in excluding associations with phenotypes from specific subpopulations, potentially leading to inaccurate interpretations of the data.

The review would benefit from additional validation studies to assess the effectiveness of SCellBOW in distinguishing between cancerous and non-cancerous signals, particularly in heterogeneous tumor environments.

Further clarification on how SCellBOW handles mixed-cell populations within bulk RNA-seq data would strengthen the evaluation of its applicability and reliability in diverse research settings.

https://doi.org/10.7554/eLife.98469.1.sa2

Reviewer #2 (Public Review):

Summary:

The authors developed a novel tool, SCellBOW, to perform cell clustering and infer survival risks on individual cancer cell clusters from the single-cell RNA seq dataset. The key ideas/techniques used in the tool include transfer learning, bag of words (BOW), and phenotype algebra which is similar to word algebra from natural language processing (NLP). Comparisons with existing methods demonstrated that SCellBOW provides superior clustering results and exhibits robust performance across a wide range of datasets. Importantly, a distinguishing feature of SCellBOW compared to other tools is its ability to assign risk scores to specific cancer cell clusters. Using SCellBOW, the authors identified a new group of prostate cancer cells characterized by a highly aggressive and dedifferentiated phenotype.

Strengths:

The application of natural language processing (NLP) to single-cell RNA sequencing (scRNA-seq) datasets is both smart and insightful. Encoding gene expression levels as word frequencies is a creative way to apply text analysis techniques to biological data. When combined with transfer learning, this approach enhances our ability to describe the heterogeneity of different cells, offering a novel method for understanding the biological behavior of individual cells and surpassing the capabilities of existing cell clustering methods. Moreover, the ability of the package to predict risk, particularly within cancer datasets, significantly expands the potential applications.

Weaknesses:

Given the promising nature of this tool, it would be beneficial for the authors to test the risk-stratification functionality on other types of tumors with high heterogeneity, such as liver and

pancreatic cancers, which currently lack clinically relevant and well-recognized stratification methods. Additionally, it would be worthwhile to investigate how the tool could be applied to spatial transcriptomics by analyzing cell embeddings from different layers within these tissues.

https://doi.org/10.7554/eLife.98469.1.sa1

Author response:

> ***Reviewer #1:***
>
> *This review evaluates the SCellBOW framework, which applies phenotype algebra to obtain vectors from cancer subclusters or user-defined subclusters.*
>
> *Strengths:*
>
> *SCellBOW employs an innovative application of NLP-inspired techniques to analyze scRNA-seq data, facilitating the identification and visualization of phenotypically divergent cell subpopulations. The framework demonstrates robustness in accurately representing various cell types across multiple datasets, highlighting its versatility and utility in different biological contexts. By simulating the impact of specific malignant subpopulations on disease prognosis, SCellBOW provides valuable insights into the relative risk and aggressiveness of cancer subpopulations, which is crucial for personalized therapeutic strategies. The identification of a previously unknown and aggressive AR−/NElow subpopulation in metastatic prostate cancer underscores the potential of SCellBOW in uncovering clinically significant findings.*
>
> *Major concerns:*
>
> *The reliance on bulk RNA-seq data as a reference raises concerns about potentially misleading results due to the presence of RNA expression from immune cells in the TME. It is unclear if SCellBOW adequately addresses this issue, which could affect the accuracy of the cancer subcluster vectors.*

To address the concern about potentially misleading results due to the TME when using bulk RNA-seq data as a reference:

a. We account for systematic biases between the single-cell and bulk transcriptomics readouts by creating pseudo-bulk profiles for single-cell clusters, enabling more accurate comparisons.

b. We encode expressions into word vectors and co-embed them together. By doing this, we mitigate any possibility of systematic differences in the embedding.

c. It is imperative that we subject both single-cell and bulk data through the same treatments because otherwise, it will be difficult to perform algebraic operations on them.

d. We rely on tumor bulk transcriptomics data from TCGA due to its high sample size and patient meta-data such as information pertaining to patient survival.

We will discuss this in the revised manuscript.

> *The method of extracting vectors in phenotype algebra appears to be a straightforward subtraction operation. This simplicity might limit its efficiency in excluding associations with phenotypes from specific subpopulations, potentially leading to inaccurate interpretations of the data.*

Vector algebra operations are not done in the gene expression space (i.e., gene expression vectors associated with tumor samples), rather we process the single cell and bulk expression profiles through multiple steps (pseudo-bulk vector generation for single cell clusters, mapping gene expression values to word frequencies as better understood by the Doc2vec neural networks etc.) to ensure their embeddings are consistent and capture intricate phenotypic information. We have demonstrated this through rigorous validation of the clusters yielded on various types of healthy and diseased samples. Furthermore, we have demonstrated the consistency of the vector algebra operations on known cancer subtypes in breast cancer, glioblastoma, and prostate cancer.

We will discuss this in the revised manuscript.

> *The review would benefit from additional validation studies to assess the effectiveness of SCellBOW in distinguishing between cancerous and non-cancerous signals, particularly in heterogeneous tumor environments.*

In our study, we are primarily interested in signals from malignant cells. However, we may consider scRNA-seq data with stromal cells and test whether SCellBOW can identify the influence of different stromal cell types on cancer aggressiveness.

> *Further clarification on how SCellBOW handles mixed-cell populations within bulk RNA-seq data would strengthen the evaluation of its applicability and reliability in diverse research settings.*

We will elaborate on our discussion in the Result as well as Discussion sections.

> ***Reviewer #2:***
>
> *The authors developed a novel tool, SCellBOW, to perform cell clustering and infer survival risks on individual cancer cell clusters from the single-cell RNA seq dataset. The key ideas/techniques used in the tool include transfer learning, bag of words (BOW), and phenotype algebra which is similar to word algebra from natural language processing (NLP). Comparisons with existing methods demonstrated that SCellBOW provides superior clustering results and exhibits robust performance across a wide range of datasets. Importantly, a distinguishing feature of SCellBOW compared to other tools is its ability to assign risk scores to specific cancer cell clusters. Using SCellBOW, the authors identified a new group of prostate cancer cells characterized by a highly aggressive and dedifferentiated phenotype.*
>
> *Strengths:*
>
> *The application of natural language processing (NLP) to single-cell RNA sequencing (scRNA-seq) datasets is both smart and insightful. Encoding gene expression levels as word frequencies is a creative way to apply text analysis techniques to biological data. When combined with transfer learning, this approach enhances our ability to describe the heterogeneity of different cells, offering a novel method for understanding the biological behavior of individual cells and surpassing the capabilities of existing cell clustering methods. Moreover, the ability of the package to predict risk, particularly within cancer datasets, significantly expands the potential applications.*
>
> *Major concerns:*
>
> *Given the promising nature of this tool, it would be beneficial for the authors to test the risk-stratification functionality on other types of tumors with high heterogeneity, such as liver and pancreatic cancers, which currently lack clinically relevant and well-recognized*

> *stratification methods. Additionally, it would be worthwhile to investigate how the tool could be applied to spatial transcriptomics by analyzing cell embeddings from different layers within these tissue*

(1) Our selection of glioblastoma and breast cancer for this study was primarily driven by the focus on extensively studied and well-defined cancer types. To demonstrate the effectiveness of our model, we tested it on advanced prostate cancer, which currently lacks clinically relevant and well-recognized stratification methods. This application to metastatic prostate cancer serves as a proof of concept, illustrating our model's potential to provide valuable insights into cancer types where established stratification approaches are limited or absent. However, as suggested by the Reviewer, we will try to incorporate results for liver cancer, subject to the availability of adequate data for model building.

(2) Regarding the application of our tool to spatial transcriptomics, we have already analyzed data from Digital Spatial Profiling (DSP). The article is already quite complex and involved, and we are afraid the inclusion of spatial transcriptomics may amount to a significant extension of the method. To this end, although we will discuss the future possibilities, we will skip the method validity check on spatial transcriptomics data.

https://doi.org/10.7554/eLife.98469.1.sa0