# Latent representation of single-cell transcriptomes enables algebraic operations on cellular phenotypes

Namrata Bhattacharya[1,5,8], Anja Rockstroh[1,8], Sanket Suhas Deshpande[4], Sam Koshy Thomas[10], Smriti Chawla[2], Pierre Solomon[7], Cynthia Fourgeux[7], Gaurav Ahuja[4], Brett G. Hollier[1,8], Himanshu Kumar[3], Antoine Roquilly[7], Jeremie Poschmann[7], Melanie Lehman[1,9], Colleen C. Nelson[1,8*], Debarka Sengupta[4,5,6*]

1. 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
2. Center for Computational Biomedicine, Harvard Medical School, Boston, MA-02115, USA
3. Laboratory of Immunology and Infectious Disease Biology, Department of Biological Sciences, Indian Institute of Science Education and Research (IISER), Bhopal, India
4. Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India
5. Department of Computer Science and Engineering, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India
6. Centre for Artificial Intelligence, Indraprastha Institute of Information Technology-Delhi (IIIT-Delhi), Okhla, Phase III, New Delhi-110020, India.
7. Nantes Université, CHU Nantes, INSERM, Center for Research in Transplantation and Translational Immunology, UMR, 1064, Nantes, France
8. Translational Research Institute, Princess Alexandra Hospital, Woolloongabba, Queensland-4102, Australia
9. Vancouver Prostate Centre, Department of Urologic Sciences, University of British Columbia, Vancouver, Canada
10. School of Mathematical Sciences, The University of Adelaide, North Terrace, Adelaide, SA-5005, Australia

**Corresponding authors: {debarka@iiitd.ac.in, colleen.nelson@qut.edu.au}.*

# Abstract

Robust characterization of cellular phenotypes from single-cell gene expression data is of paramount importance in understanding complex biological systems and diseases. Single-cell RNA-seq (scRNA-seq) datasets are inherently noisy due to small amounts of starting RNA. Over the last few years, several methods have been developed to make single-cell analysis fast and efficient. Most of these methods are based on statistical and machine learning principles. In the current work, we describe SCellBOW, which encodes single-cell expression vectors as documents, thereby enabling the application of powerful language models. Beyond the identification of robust cell type clusters, our algorithm provides a latent representation of single-cells in a manner that captures the 'semantics' associated with cellular phenotypes. These representations, *aka* embeddings, allow algebraic operations such as '+' and '-'. We use this hitherto unexplored utility to stratify cancer clones in terms of their aggressiveness and contribution to disease prognosis. Further, the application of SCellBOW to a scRNA-seq dataset comprising human splenocytes and matched peripheral blood mononuclear cells (~5000 cells) identifies unknown cell states that bear significance in advancing our understanding of spleen biology.

Keywords - Doc2vec, Single-cell RNA-seq, Transfer learning, Clustering, Phenotype algebra

# Main

The advent of commercially available high throughput single-cell omics platforms has led to a genomic data singularity, where the projected rate of genomic data accumulation is set to outpace the growth of supercomputer power efficiency within the next few years. Off-the-shelf clustering approaches from the field of machine learning and pattern recognition are not entirely suitable for optimal analysis of single-cell RNA-seq (scRNA-seq) data, given its unique statistical properties and requirements of normalization. This is primarily attributable to the noisy nature of the data, with variable numbers of expression dropouts and other factors influencing cell-to-cell variabilities. Traditionally scRNA-seq analysis methods have been built for standalone data analysis use cases. Seurat[12], Scanpy[3], and dropClust[4,5] are among these. DESC[6] is another such method that uses AutoEncoder-based unsupervised embeddings of cells. Most of these methods use Principal Component Analysis (PCA) or dispersion-based gene selection and network-based clustering algorithms as main workhorses. With a sharp increase in transcriptomics data, a recent trend has been observed in the emergence of transfer learning (TL) methods that allow cross-study learning[7].

Presently, the vast majority of single-cell data analysis is study specific and standalone, rarely reaping the benefit of relevant existing data. To foster such extraneous learning from independent datasets, transfer learning approaches have been envisaged. These are particularly useful in analyzing datasets of small sample sizes. A number of these methods are supervised since they require cell type annotations of the source data to learn embeddings (low-dimensional vector representation of the cells) for cells in the target scRNA-seq data. ItClust[8] is an example of such a method. ItClust uses AutoEncoder for transferable representation learning of single-cells. Using source data annotations as a reference limits the identification of new cell types in the target data. There are other methods, such as ProjectR[9], and coupleCoC[10], that are unsupervised in nature. For example, ProjectR uses PCA, non-negative matrix factorization (NMF), and correlation to implement transfer learning. Notably, it does not require cell type annotation. It is essential to realize that exploratory studies involving single-cell expression profiling require unsupervised analysis of the data. Supervised TL methods are not so valuable for these scenarios.

A rather unexplored area in this context is language models. In the last few years, the fields of single-cell genomics and language models have grown concurrently yet independently until the recently observed diffusion between the two. Key breakthroughs in language modeling include the following: 1. Shallow neural networks such as Word2vec[11], sentence2vec[12], and Doc2vec[13] for numeric vector representation of words, sentences, and documents, respectively; 2. Deeper transformer architectures for the same tasks. Some examples include BERT[14] by Google and GPT-3[15] by OpenAI. These methods have found widespread use cases in genomics, especially in the fields of biomedical literature mining (e.g., BioBERT[16], BioReader[17]) and representation learning from nucleotide sequences (e.g., DNABERT[18], LOGO[19]). A growing number of NLP-based methods have recently been applied to scRNA-seq data analyses and achieved superior performance. For example, scBERT[20] performs cell type annotation using a variant of BERT. Exceiver[21] uses attention-based transformers to output context-specific gene-gene similarities. ScETM[22] utilizes transferable neural-network-based encoders to integrate scRNA-seq batches and infer the cell type mixture.

Transformer-based encoding models require large training data sizes and high-end compute infrastructure. Further, transformers are useful for sequences as inputs. Genes do not have any inherent ordering. In typical scenarios, they are treated as a collection or class of molecules. Even though the genes have regulatory dependencies, analytical methods are often not prejudiced on the same. The proposed single-cell embedding technique SCellBOW, on the contrary, uses Doc2vec, which is a bag-of-words model, and therefore is independent of any strict ordering of genes. We demonstrated the utility of bag-of-words-like modeling of single-cell expression data by testing the method on multiple scRNA-seq datasets with known cell type annotations. Application of SCellBOW to an in-house scRNA-seq dataset comprising human peripheral blood mononuclear cells (PBMC) and matched splenocytes revealed hitherto unknown cell states associated with B cell subtypes. In line with word algebra, which allows exploration of word analogies (i.e., algebraic operations on word vectors, e.g., word vector associated with *royal* when added with *man,* brings it closer to *king*), cellular embeddings show biologically intuitive outcomes under algebraic operations. We exploited this feature to introduce phenotype algebra and applied the same in attributing prognostic value to cancer cell subpopulations identified from scRNA-seq studies.

# Results

## Overview of SCellBOW

Single-cell expression studies have become ubiquitous in biomedical research. Serious limitations still exist in terms of data quality and sample size. scRNA-seq, in its crude form, embodies significant noise and technical biases. These include excessive amounts of dropout events due to low RNA input[23–25]. Given the sharp increase in public scRNA-seq data availability, the single-cell computational method development community has proactively investigated the scope of transfer learning, such that analysis of small-sized scRNA-seq datasets can benefit from the inheritance of pre-trained models that ingested large volumes of scRNA-seq data from independent studies. Analysis of scRNA-seq data has been shown to benefit from such transfer-learning approaches[26,27]. The proposed approach, SCellBOW, is an adoption of the popular natural language model, Doc2vec, in single-cell representation learning (**Fig. 1a**). In addition to enabling transfer learning, SCellBOW also unlocks the new possibility of enabling algebraic operations of cellular phenotypes within its vector space. Below we highlight key constructions and benefits of this new approach.

Doc2vec learns embeddings from documents, where each document is considered a bag of words. SCellBOW represents cells as documents. Gene names are considered as words, and their expression levels are encoded by word frequencies. This encoding is intuitive and produces meaningful and robust embeddings of cells. This encoding can be leveraged for applying other language models to single-cell expression analysis. Doc2vec uses a neural network to learn numeric embeddings associated with individual cells such that cellular similarities observed in the gene expression space are preserved. Doc2vec produces document embeddings by attempting to predict a word's existence, given its context words from a given document. Words (genes) constitute the vocabulary of the model. Although Doc2vec is not meant for transfer learning, our experiments involving the repurposing of a pre-trained, modified Doc2vec model (*aka* SCellBOW) establish promising transfer learning use cases (**Supplementary Note 1 and Supplementary Fig. 1**). In this work, we show several applications across

diverse tissue types, where SCellBOW, trained on large independent datasets, improves single-cell embedding quality and clustering outcomes.

In summary, SCellBOW involves learning in two phases, i.e., the pre-training and the fine-tuning phases. The pre-training phase involves training a shallow source network that extracts the gene expression patterns from a relatively large scRNA-seq dataset. The parameters estimated at training are repurposed during the fine-tuning phase based on the target data, which might represent a rather specialized task. Our results allude to improvement in scRNA-seq analysis outcomes, even with small sample sizes and the presence of multiple covariates.

Word embeddings generated by neural networks such as Word2vec are known to preserve linear semantic relationships across words[28]. For example, the word embedding associated with 'queen' is highly similar to the result of the algebraic expression: *king* - *man* + *woman*. Document or sentence-level embeddings also display similar properties. We empirically show the retention of such relationships in the context of cellular phenotypes of cancer cells.

## SCellBOW performs competitively compared to state-of-the-art single-cell analysis methods

To evaluate the strength of SCellBOW, we benchmarked the method with five existing packages - (1) Scanpy[3], (2) Seurat[2], (3) DESC[6], (4) ItClust[8], and (5) ProjectR[9] (**Supplementary Table 3 for more details**). For this, multiple scRNA-seq datasets were used from independent studies and processed using diverse technologies. Among these, Seurat and Scanpy are the two most used single-cell analysis software, whereas DESC leverages deep neural networks. ItClust and ProjectR are TL methods. ItClust performs supervised transfer learning using a stacked auto-encoder. ProjectR uses the K-means algorithm on top of principal component analysis (PCA)[29]. ItClust selects the resolution with the highest silhouette score. We also compared the performance of SCellBOW with an additional transfer learning-based tool, scETM[22]. For the objective evaluation of performance, we used an adjusted Rand index (ARI) and normalized mutual information (NMI) for comparing clusters with known cell annotations. To measure the signal-to-noise ratio of low-dimensional single-cell embeddings, we computed the Silhouette index (SI) for both clusters, as well as for cell type annotations. For all methods, we tracked ARI, NMI, and SI. For all methods, except for ProjectR and ItClust, which are resolution-independent, we also computed ARI for different values of resolution ranging between 0.2 to 2.0.

We used a number of single-cell datasets to evaluate the algorithms cited above. We constructed four problem instances, each with a pair of single-cell expression datasets, of which one was used for self-supervised model training (herein referred to as the source data) and the other for model fine-tuning and analysis of the clustering outcomes (referred to as the target data). The target data, in all cases, had associated cell type annotations, which we used for qualitative assessment of the clustering outcomes. Here we provide brief descriptions of the datasets – a. The first problem instance comprised a source single-cell dataset of non-cancerous prostate tissues from prostate cancer patients (120,300 cells)[30] and a target dataset consisting of cells from healthy prostate tissues (28,606 cells)[30,31]. The target data were annotated based on Fluorescence Activated Cell Sorting (FACS) enriched cell populations. b. The second problem instance comprised a large PBMC dataset (68,579 cells)[32] 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. c. In the case of the third problem instance, as source data, we considered pancreatic cells from three independent studies processed with different single-cell profiling technologies[33–35]. The target data used in this case was from an independent study, processed with a different technology[36]. The target data was annotated using expression profiles derived from FACS enriched cell-subpopulations (**Fig. 1, Supplementary Fig. 2-5, Supplementary Table 4, and Supplementary Methods**). We compared the performance of scETM on the PBMC and pancreas datasets; however, we excluded the normal prostate dataset due to underlying limitations (**Supplementary Note 2 and Supplementary Fig. 2**).

For the prostate and PBMC datasets, SCellBOW produced the highest ARI scores across most notches of the resolution spectrum (0.2 to 2.0), whereas, on the pancreas dataset, ProjectR yielded the best ARI for the majority of resolution settings (**Fig. 2a-c**). This could be attributed to the privileged setting of the number of clusters as the number of annotated cell types present in the data. For further deterministic evaluation of the different methods across the datasets, we set 1.0 as the default resolution (**Fig. 2d-f and Supplementary Table 5**). ARI, NMI, and SI thus obtained were found to be consistent in the cases of SCellBoW, Seurat, and Scanpy. Fluctuations were observed in the remaining algorithms, including ones enabling transfer learning (ItClust, ProjectR). In the normal prostate dataset, SCellBOW achieved the highest ARI, NMI, and second-highest SI (cluster) at resolution 1.0 (**Fig. 2d**). However, we observed small improvements in Seurat and ProjectR only in terms of the SI (cell type). Known cell types such as basal epithelial (BE), luminal epithelial (LE), endothelial (Endo), and smooth muscle (SM) cells were grouped into homogeneous clusters by SCellBOW (**Fig. 1b and Supplementary Fig. 3**). We observed that the majority of fibroblasts and endothelial cells were mapped by SCellBOW to single clusters, unlike Scanpy and ItClust. In the PBMC dataset, SCellBOW yielded the highest ARI, NMI, and SI (cell type) (**Fig. 2e**). However, the SI (cluster) for SCellBOW was found to be lower than other tools. Apparently, SCellBOW retained similar cell types (for example, T cell subpopulations) at close proximity in the gene expression embedding space. In contrast, tools like DESC and ItClust scattered them at disparate locations against the PBMC lineage hierarchy (**Fig. 1c and Supplementary Fig. 4**). In the pancreatic islet dataset, the NMI and ARI offered by SCellBOW are the second highest after ProjectR (**Fig. 2f**). One possible reason for ProjectR outperforming SCellBOW is that ProjectR uses K-means, with K set as the number of cell types (which will be unknown in real life use-cases). Higher SI (cluster) was obtained from Seurat and ItClust. However, SCellBOW returned distinct, spatially segregated clusters for known cell types such as the Acinar, Beta, Delta, Ductal, Gamma, and Mast cells, as compared to ItClust and Seurat (which scatters the cell types across multiple clusters) (**Fig. 1d, and Supplementary Fig. 5**). SCellBOW yielded the highest SI (cell type), same as Seurat, as compared to other tools. Also, the benchmarking tools other than ProjectR mixed the majority of cell types in different clusters.

Notably, SCellBOW, unlike ItClust[37] and scETM[22], does not limit model training to genes intersecting between the source and the target datasets. Given the shallow architecture of the neural network, SCellBOW is faster than other studied TL methods (**Supplementary Table 6**).

## SCellBOW robustly portrays heterogeneity in PBMCs and splenocytes

The prevalence of single-cell expression profiling of PBMCs makes it one of the best-studied tissue systems in humans. SCellBOW-based analysis of PBMCs, with near-ground-truth cell type annotation, confirmed its ability to adequately decipher the underlying cellular heterogeneity. Analyses reported in the previous sections confirm the robustness of SCellBOW as compared to some of the prominent single-cell analysis pipelines, including some prominent transfer-learning methods. To this end, we applied SCellBOW to investigate the heterogeneity of splenocytes in contrast with the already well-characterized PBMCs. Of note, the spleen is the largest secondary lymphoid organ in the body and orchestrates a wide range of immunological functions[38]. The physical organization of the spleen allows it to filter blood-borne pathogens and antigens[39]. 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) (**Fig. 3a**). We used CITE-seq with hashtag oligos (HTO) to pool the eight samples into a single scRNA-seq experiment[40]. After post-sequencing quality control, we were left with a total of 4,819 cells. We annotated the single-cells using Azimuth[41], with occasional manual interventions (**Supplementary Note 3)**. 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 for single-cell analysis. SCellBOW yielded clusters largely coherent with the independently done cell type annotation using Azimuth (**Fig. 3b and Supplementary Fig. 6**). 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. The majority of CD4+ T cells map to CL0 and CL9 (**Fig. 3e**). 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. 3f**). While Scanpy performed reasonably well, a few 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[41] **(more details in Supplementary Note 4 and Supplementary Fig. 7g-h)**. 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. Further, we quantitatively evaluated SCellBOW and the rest of the benchmarking methods by measuring several metrics – ARI, NMI, and SI (cluster and cell type) (**Fig. 3c**). While most methods did reasonably well, SCellBOW offered an edge.

Differential expression analysis between immune cells originating from the spleen and peripheral blood exhibited spleen-specific elevated expression of dissociation-associated genes (*JUNB*, *FOSB*, *JUND*, *JUN*) and stress response heat shock genes (*HSP*, *HSP90AB1*, *HSPD1*, *DNAJB1*) (**Fig. 3g, h and Supplementary Data 1**). Upregulation of dissociation and stress response-associated gene families can be attributed to the additional dissociation step during sample preparation, which is exclusive to the splenocytes[42,43]. Furthermore, B cell differentiation-related genes such as *VPREB3*, *MS4A1*, and *CD83* were found upregulated among splenocytes as compared to PBMC, suggesting the dominance of various stages of development, differentiation, and maturation of B cells in the spleen. Notably, the

spleen has different structures, such as primary follicles and the germinal center within the white pulp where various stages of B cells reside. We also spotted spleen-specific upregulation of nuclear receptors (*NR4A1*, *NR4A2*) and B cell receptors (BCR) (*BANK1*, *CD79A*, *CD79B*), indicating that B cells are inherently different between circulation and the spleen and likely undergo additional maturation steps in the spleen.

We also observed spleen-specific expression of B cell inhibitory co-receptors (*CD22*, *CD72*, *LY9*), suggesting that multipotent progenitors reside in the spleen. These multipotent progenitors of the spleen stem cells could be differentiating to myeloid and erythroid lineage cells in addition to differentiation of the lymphoid lineage cells (primarily to B cells and to some extent to T cells)[44].

B cells are capable of developing a wide range of effector functions, including antibody secretion via differentiating to plasma cells, antigen processing and presentation to T helper cells, cytokine production, and generation of immunological memory. The mature spleen plays an important role in B cell development and antigen-dependent maturation because it is the site of terminal differentiation for developing B cells that arrive from the bone marrow known as Transitional B cells of type 1 (T1), and these cells develop to transitional B cells of type 2 (T2) and reside to the primary follicles of the spleen[45]. Naive B cells enter the spleen from the circulation, where they differentiate into either B effectors that differentiate into antibody-secreting plasma cells or B memory cells[46]. The prominent role of the spleen in B cell differentiation and activation inspired us to delve deeper into compositional and phenotypic differences associated with B cell subtypes across peripheral blood and spleen. Based on annotation, as expected, we noted enrichment of B effectors in the spleen (**Fig. 3i**). SCellBOW embedding could spatially segregate B effectors, B naive cells, and plasmablasts, although effectors and naive cells were not split into two distinct clusters (**Fig 3j**).

Further analysis of B cell subtypes showed higher expression of the dissociation genes and stress response of genes in spleen B effector and B naive cells as compared to their peripheral blood counterpart (**Fig. 3k, l)**. B naive has higher expression of actin genes (*ACTB*, *ACTG1*) as compared to B effectors suggesting the naive B cells are more dynamic for differentiation to any of B cell subtypes, such as the effector or memory[47]. Notably, the mutation in these genes results in the development of large B-cell lymphoma. In the case of the B effector, we observed that expression of *CD27*, *HLADQA1*, *CD83*, *PRMT1*, and *CD69* are significantly higher in the mature effector B cells[48,49]. Collectively, our analysis very well corroborates with known B cell development biology under physiological conditions.

## SCellBOW facilitates subclonal survival-risk attribution

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 limited number of past studies have reported an association between malignant/stromal cell composition in tumor microenvironments and patient survival[50]. This has sparked debate about the utility of gene expression in the profiling of single tumor

cells. To date, there is no method that can estimate the relative aggressiveness of different malignant cell subtypes, where survival information can be considered a surrogate for aggressiveness. SCellBOW helps achieve this by enabling algebraic operations (subtraction or addition between two expression vectors or phenotypes) in the embedding space (**Fig. 4a**). We refer to this as *phenotype algebra*. The ability to perform algebraic operations in cellular embedding vector space is imparted by the use of word embedding methods as the backbone of SCellBOW. Vector representation of words learned on large text corpora enables algebraic operations. Below is an example.

$$king - male + female = queen$$

We utilized this ability to find an association between the numeric vector, representing $Total\ tumor - A\ specific\ malignant\ cell\ cluster$ with patient survival, which immediately opens up a way to infer the aggressiveness of a particular cancer cell subtype (obtained by clustering single-cells). Subtracting an aggressive phenotype (a specific malignant cell subtype) 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.

To explore phenotype algebra, three independent datasets are required. Out of the three datasets, two are scRNA-seq datasets used for transfer learning (source and target data), and one bulk RNA-seq dataset with patient-survival information, which is used to train the survival prediction model (**Methods**). For ease of reference, we stick to the following nomenclature – a. The scRNA-seq dataset used for model pre-training is referred to as the *source data*, b. The *target data* comprises malignant cell subtypes under investigation; c. Survival *data* comprises tumor bulk RNA-seq samples, with survival information acquired through patient follow-up. Cancer-specific data from The Cancer Genome Atlas (TCGA) are an example.

As proof of concept, we first validated our approach on Glioblastoma multiforme (GBM), which has been studied widely employing single-cell technologies (**Fig. 4b-g**). GBM has three well-characterized malignant subtypes: proneural (*Pro*), classical (*Cla*), and mesenchymal (*Mes*)[51,52]. As source data, we used the GBM scRNA-seq data from Neftel *et al.*[53]. Our survival data consisted of 613 bulk GBM samples from the TCGA[54]. 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[55] and used it as our target data (**Supplementary Methods**). Notably, embeddings for single-cell transcriptomes from target data and bulk RNA-seq samples from survival data were determined concurrently by calibrating the pre-trained model. A survival prediction model was trained on the embeddings of the survival data using Random Survival Forest[56]. We conjectured that for the most aggressive malignant cell subtype, the *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., $Mes \succ Cla \succ Pro$, where *Mes* succeeds the rest of the subtypes in aggressiveness[57–59] (**Fig. 4g**). More complex queries can be formulated such as *Total tumor - (Mes+Cla),* which would indicate a tumor consisting of *Pro* cells alone. This, hypothetically, represents the most favorable scenario, as testified by the phenotype algebra[52,60].

We performed a similar benchmarking on well-established PAM50[61] based breast cancer subtypes (BRCA) – luminal A (*LumA*), luminal B (*LumB*), HER2-enriched (*Her2*), basal-like (*Basal*), and normal-like (*Normal*) (**Fig 4h-j**). We used 24,271 cancer cells from Wu *et al.*[62] as the *source data*, 545

single-cell samples from Zhou *et al.*[63,64] as the *target data,* and 1,079 bulk BRCA samples from TCGA[54] as the *survival data.* SCellBOW predicted survival risks for the different subtypes were generally in agreement with the clinical grading of the PAM50 subtypes[65,66], with *LumA* generally having the best prognosis, followed by *LumB*, *Her2*; whereas *Basal* and *Normal* were assigned worse prognosis[67–69] (**Fig. 4j**).

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 +/-.

## Phenotype algebra enables subclonal grading of metastatic prostate cancer cells

Prostate cancer originates from secretory prostate epithelial cells in which differentiation and function of the cells are driven by the androgen receptor (AR), a transcription factor regulated in response to androgens[70–72]. The primary form of treatment for metastatic prostate cancer is androgen-targeted therapies (ATT), which are most effective in well-differentiated prostate cancer cells with high AR activity[72,73]. Despite treatment responses of varying duration, eventually, cancer will emerge as metastatic castration-resistant prostate cancer (mCRPC)[74]. With further prolonged and more potent ATTs, prostate cancer cells adapt to escape reliance on AR, with low AR activity or loss of AR expression[75]. Loss of differentiation pressure results in altered states of lineage plasticity in prostate tumors[76,77]. The most well-accepted and defined phenotype is treatment-induced neuroendocrine prostate cancer which is highly aggressive and often presents with visceral metastases and no available effective therapies[76,78,79]. However, it is increasingly apparent that additional prostate cancer phenotypes arise through lineage plasticity of malignant cells that can likewise be described by other gene expression patterns. These include malignant states with 'androgen receptor signaling low' or 'double-negative' phenotypes likely representing intermediate or transitory states of the progression trajectory from high AR activity to neuroendocrine prostate cancer[80].

Here, we performed a pooled analysis of scRNA-seq *target data* consisting of 836 malignant cells derived from tumors collected from 11 patients with mCRPC data from He *et al.*[81]. We initially classified cells into three categories - AR activity high (*ARAH*), AR activity low (*ARAL*), and neuroendocrine (*NE*) using K-means clustering on GSVA scores based on the expression of well-characterized androgen receptor (ARPC) or neuroendocrine (NEPC) genes[80] (**Fig. 5a, Supplementary Fig. 8, and Supplementary Methods**). We subsequently assessed the relative aggressiveness of these high-level categories using phenotype algebra (**Fig. 5c**). As *source data* for transfer learning, we used the pre-trained model built on Karthaus *et al.*[30] dataset. For the phenotype algebra model, we used *survival data* from 81 advanced metastatic prostate cancer patient samples from Abida *et al.*[82]. As expected, the predicted risk score was lowest when the aggressive *NE* subtype was subtracted from the *Total tumor*[83]. In contrast, the removal of the *ARAH* subtype from the pseudo-bulk had much less impact on the predicted risk score. Subtracting *ARAL* resulted in a predicted risk score between *ARAH* and *NE* which supports the hypothesis of *ARAL* representing an intermediate state between *ARAH* and *NE* [80].

Grouping prostate cancer cells into these three high-level categories is an oversimplified view of the actual heterogeneity of advanced prostate cancer biology. As demonstrated in the previous sections,

SCellBOW could be utilized for the discovery of novel subtypes based on the single-cell expression profiles, followed by assigning a grading of relative aggressiveness to these clusters using phenotype algebra. To apply this concept, we subsequently used SCellBOW for unsupervised clustering of the 836 malignant cells from He *et al.*[81] (**Fig. 5d**), followed by cluster-wise relative risk prediction (**Fig. 5e**) using the metastatic prostate cancer phenotype algebra model.

This approach enables a novel and more refined view of lineage plasticity states and features defining aggressiveness during prostate cancer progression beyond the traditional categorization into *ARAH*/*ARAL* and *NE* (**Fig. 5f**). 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 (**Fig. 5g**) and that other, yet under-explored biological programs play additional important roles.

To further elucidate these altered cellular programs, we performed a cluster-wise functional gene set enrichment analysis (**Supplementary Methods and Supplementary Data 2**). While CL3 and CL1 resemble most closely the traditional ARAH subtype (**Fig. g**), our phenotype algebra model predicted clusters CL6 and CL7 as the least aggressive (**Fig. e**). The prominence of immune pathways in those two clusters may point towards the involvement of the tumor microenvironment [84]. Despite CL4 having the strongest NEPC features, CL2 was predicted to convey an even higher risk. While CL2 shared some traits of neuroendocrine biology, it was further characterized by a high expression of genes known to be androgen repressed and activated by ATTs (**Fig. h, i**). This is in line with the hypothesis that prolonged exposure to ATTs can give rise to aggressive subtypes in response to treatment[85–87]. The expression pattern of a gene set of 'efflux transmembrane transporters' matched most closely Kendall's tau 0.93) the predicted risk ranking of the clusters (**Fig. j**). This gene set contains well-characterized ABC transporter molecules (ABCC1, ABCC4, and ABCB6) involved in mediating drug resistance and the SLC26A6 transporter, an emerging oncogene associated with poor survival[88–90]. Hence, SCellBOW clustering combined with phenotype algebra can provide novel insights into factettes of prostate cancer biology that can be used to delineate features of lineage plasticity and aid in better classification of molecular phenotypes with clinical significance.

# Discussion

In this work, we introduce SCellBOW, a distributed bag-of-words-based transfer learning model for single-cell data analysis and phenotype algebra to estimate the relative aggressiveness of the cancer cell subtype. SCellBOW is inspired by Doc2vec, which assigns a small cosine distance to a pair of similar documents. SCellBOW treats a cell as a document, a specific mRNA as a word, and its expression as the word frequency. SCellBOW allows self-supervised pre-training on large-scale gene expression data. The pre-trained model can then be fine-tuned based on expression vectors from data of interest. The entire SCellBOW framework is independent of any supervision, i.e., the availability of cell-type annotations for any of the source and destination datasets.

We compared our approach to popular single-cell analysis software as well as some recently proposed transfer learning approaches. An array of scRNA-seq data was used for the same, including healthy tissues and malignancies. 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, TL 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, processed using varied single-cell technologies. Notably, SCellBOW, unlike ItClust[37] and scETM[22], does not limit model training to genes intersecting between the source and the target datasets. Given the shallow architecture of the neural network, SCellBOW is faster than other studied TL methods (**Supplementary Table 6**).

We applied SCellBOW to compare observed B-cell states across matched PBMCs and splenocytes. Notably, a higher number of genes were found upregulated in the splenocytes (2,348 for B effector and 414 for B naive cells) as compared to PBMC (5 for B effector and 27 for B naive), indicating a wider gene expression variety in the splenocytes. This can be attributed to diverse activation programs and the incubation of developmental cell states of immune cells in the spleen[38]. Elevated expression of tissue dissociation-related genes was observed in splenocytes. Genes upregulated in the B cells from the spleen were found to be involved in cell-cell adhesion, activation of the B cell receptor signaling pathway, and differentiation of B cells during hematopoiesis[39,91]. The B cell-intrinsic genes that were upregulated in PBMCs were involved in protection from oxidative stress and DNA damage repairs[92].

As discussed earlier, SCellBOW-based vector representation of cellular transcriptomes preserves their phenotypic relationships in a vector space. This feature can be exploited in innovative ways. 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 plus and minus could derive clinically meaningful outcomes. For example, subtracting mesenchymal phenotype from the whole tumor causes the largest improvement in the survival risk in a GBM patient. In the case of metastatic prostate cancer, we observed a misalignment from the general perception that subtype NE has the worst prognosis. Our results pointed towards the existence of a more aggressive and dedifferentiated subclone of prostate cancer in the lineage plasticity continuum. Further gene set enrichment analysis on this specific malignant subtype yielded genes associated with advanced prostate cancer correlating to poor patient survival and a positive metastatic status (**Supplementary methods**). Such a tool can empower medical oncologists and oncology researchers to decide personalized treatment regimes for cancer patients.

# Online methods

## The architecture of the bag-of-words model

SCellBOW generates a low-dimensional embedding of single-cells based on their gene expression patterns. It does this by first preprocessing the gene expression data, which typically involves normalizing the data and removing low-expressed or non-variable genes. After preprocessing, SCellBOW creates a document on the number of RNA transcripts (i.e., *copies of an mRNA molecule*) associated with the genes in a cell. Finally, SCellBOW uses these documents to train a Doc2vec[13] model to generate numerical vectors (i.e., embeddings) latently representing the cells under investigation. The preprocessing part of the pipeline is similar to that used in Seurat and Scanpy for downstream analysis.

SCellBOW learned neuronal weights are transferable. This is particularly useful when a limited amount of labeled data is available for the second task. The model can use the knowledge acquired from the

first task to warm start the learning process on the second task. In the context of SCellBOW, the transfer learning approach is used to generate word embeddings for single-cell RNA-seq data.

Doc2vec is an extension to Word2vec[93]. In Word2vec, the training corpus is a sequence of words $W = \{w_1, w_2, .., w_T\}$ that belong to a vocabulary $V$. A variant of Word2vec is the continuous bag-of-words (CBOW). The training objective of the Word2vec CBOW 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:

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

The probability can be modeled using the hierarchical softmax function:

$$Pr(w_t \mid w_{t-n}, .., w_{t+n}) = \frac{e^{y_{w_t}}}{\sum_i \quad e^{y_i}}, \qquad (2)$$

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. Similar to Word2vec, the training corpus in Doc2vec contains a sequence of $k$ training documents $D = \{D_1, D_2, .., D_k\}$. The words within each document are treated as shared among all documents, i.e., the same word can appear in multiple documents, and the model takes this into account when learning the document embeddings. Doc2vec introduces an additional document vector in addition to the vectors for the individual words it contains. This document vector captures the overall meaning and context of the document and is trained to capture the relationships between the words in the document. One specific variant of Doc2vec is the distributed memory model of paragraph vectors (PV-DM) is similar to the CBOW approach in Word2vec. This pre-trained model can be applied to a new corpus of documents.

During transfer learning, the pre-trained model is used to infer vector representations (also known as embeddings) for new documents. This is typically done by using the document vectors and softmax weights of the pre-trained model as initialization for the new documents. The parameters of the pre-trained model are usually updated reluctantly to better fit the new data[13].

# Data preprocessing and quality control

SCellBOW follows the same preprocessing procedure for both the source and target data. The first step in the preprocessing procedure is to filter the cells and genes in the data to eliminate low-quality cells and genes that could interfere with downstream analyses. To do this, we used *filter_cells* and *filter_genes* functions, respectively, from the Scanpy package[3] (**Supplementary Table 4**). After filtering the cells and genes, SCellBOW performs CPM normalization on the gene expression using the *normalize_total* function to account for the differences in the total number of counts across cells[94].

$$CPM = \frac{mapped\ read \times 10^6}{total\ reads}, \qquad (3)$$

where *total reads* is the total number of mapped reads of a sample, and *mapped reads* is the number of reads mapped to a selected gene[94]. We perform log transformation using the *log1p* function on the expression data to alleviate the effect of extreme values in the data matrices. The log transformation is often performed on UMI data by using the formula $\log_{10}(\text{UMI}+1)$. Highly variable genes are selected using the *highly_variable_genes* function. Only the remaining cells and genes are used for further downstream analysis. To ensure that the variance of highly expressed genes does not dominate the downstream analysis, we perform standardization of the data using the *scale* function (similar to z-score scaling)[95]. SCellBOW can handle data in different formats, including UMI count, FPKM, and TPM. The UMI count data follow the same preprocessing procedure as above. For TPM and FPKM, we omit the normalization step since both the data are length normalized.

## Creating a corpus from source and target data

Let $E \in R^{G \times C}$ be the gene expression matrix obtained from a scRNA-seq experiment, in which $G$ corresponds to genes and $C$ corresponds to cells. The gene expression matrix is analogous to the term-frequency matrix, which represents the frequency of different terms 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 in a single-cell. Here, a specific mRNA is a word, and its copies detected in a cell are its frequency. SCellBOW considers each entry in the matrix as the logarithmically scaled gene frequency across cells:

$$TF(f_{g,c}) = \begin{cases} \log(1+f_{g,c}), & if\ f_{g,c}>0 \\ 0, & otherwise \end{cases}, \qquad (4)$$

where $f_{g,c}$ is the number of RNA transcripts of a gene $g \in G$ in a cell $c \in C$. To generate the embeddings, SCellBOW takes as input a source dataset and a target dataset. The source dataset is used to initialize the neural network model weights, while the target dataset includes the cells that need to be clustered. Before generating the embeddings, SCellBOW scales the gene expression data by applying the *MinMaxScaler* function from the scikit-learn[96] package to scale each feature within a range [0, 10].

$$X_{scaled_value} = int\left(\frac{X-minimum}{maximum-minimum}\right), \qquad (5)$$

where scaling is applied to each entry with *maximum*=10 and *minimum*=0. SCellBOW takes only the integer part of the scaled value. This scaling helps control the number of gene names copied in the document. To create the corpus, SCellBOW copies the gene name as many times as the gene is expressed in each cell and shuffles the genes in each cell to spread them uniformly with other genes. The resulting gene names act as the terms in the document. To build the vocabulary from a sequence of documents, a tag number is associated with each document of the training corpus using the *TaggedDocument* function in the Gensim package[97]. For each word in the documents, a token is assigned using a module called tokenize with a *word_tokenize* method that splits a text into tokens[98].

## The training and fine-tuning architecture of the neural network

As scRNA-seq data are high-dimensional, instead of clustering directly in the original data space, SCellBOW produces a low-dimensional fixed-length embedding of a single cell transcriptome using transfer learning, by transforming the data from $E \in R^{G \times C}$ feature space to $V \in R^{d \times C}$ latent feature space of dimension d, with d "G.

To generate the embeddings, SCellBOW first pre-trains a Doc2vec model using the *doc2vec* function in the Gensim[13,97] library. SCellBOW trains the Doc2vec model using a source scRNA-seq dataset $E_{src}$ with $C_{src}$ cells $\times$ $G_{src}$ genes. The Doc2vec model is trained using the PV-DM method[93]. The size of the embedding is a hyperparameter of the Doc2vec model, and SCellBOW sets it to 300 dimensions.

SCellBOW fine-tunes the weights learned by the pre-trained model from $E_{src}$ using a target dataset $E_{trg}$with $C_{trg}$ cells $\times$ $G_{trg}$ genes. This allows the neural network to converge faster compared to starting from randomly initialized weights. SCellBOW retrains the Doc2vec model with the target dataset $E_{trg}$ 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 in $E_{trg}$ 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 that the network can map the target data into a low dimensional embedding space $R^d$; i.e., $E \rightarrow V$,where $V \in R^{d \times C}$. The resulting embeddings can be used for various downstream analyses of the single-cell data, such as clustering or visualization of cell types.

## Visualization of 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.

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

After generating low-dimensional embeddings for the cells using SCellBOW, the next step is to identify clusters or groups of cells with similar gene expression patterns. To identify the clusters, SCellBOW applies the Leiden algorithm[99] to the similarity matrix using the *leiden* function in the Scanpy package. SCellBOW uses the UMAP[100] visualization from Scanpy to further reduce the dimensionality of the similarity matrix and visualize the clusters within a two-dimensional space.

## Phenotype algebra in subclonal survival risk attribution

SCellBOW uses the '+' and '-' operations to compare and analyze cellular phenotypes in a way that is not possible with traditional methods. For example, the '-' operation can be used to calculate the difference between a specific cellular phenotype from the *Total tumor*. This could potentially identify the contribution of that phenotype to patient survival. SCellBOW estimates the relative aggressiveness of different malignant phenotypes based on survival risk. SCellBOW performs phenotype algebra either on the pre-defined cancer subtypes or known SCellBOW clusters.

To perform phenotype algebra, SCellBOW requires a bulk RNA-seq cancer dataset with survival information, in addition to a source scRNA-seq dataset and a target scRNA-seq dataset. SCellBOW uses the source dataset to build the pre-trained model. In the survival dataset, we excluded cases without follow-up time or survival status and cases with clinical information but no corresponding RNA-seq data. The target scRNA-seq dataset is used to identify the cell types or clusters and their relative ranking based on malignancy.

To account for unequal cell numbers per group, SCellBOW balances the target dataset using SMOTE[101] from the imblearn[102] python library. This confirms that the output of phenotype algebra is not confounded by the cell type proportions. Thereafter, SCellBOW constructs different vectors for the risk prediction task. These vectors are based on the average gene expression of different groups of cells, that is, different cell types or clusters. SCellBOW constructs a reference vector by taking the average of all malignant cells (referred to as *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 constructs a query vector by taking the average of the gene expression of the cells in that individual group. In addition, SCellBOW can also concatenate the query vectors of different groups in order to test their aggressiveness at a combined state. For example,

$$risk_{a,b} = T - a + b \quad , \qquad (7)$$

where $T$ is the *Total tumor,* $a$ and $b$ are query vectors for two different cellular phenotypes. Followed by this, SCelBOW concatenates the gene expression data from bulk RNA-seq and averages expressions of both reference and query vectors based on common genes using the function *concatenate* function in the AnnData[103] python package. The resulting combined dataset is then passed to the pre-trained model, which maps it to a lower-dimensional embedding space.

To predict the survival risk of the different groups, a random survival forest (RSF)[56] machine learning model is trained using the bulk RNA-seq embeddings with the survival information. SCellBOW subtracts each query vector from the reference vector and fits the difference of these vectors as input to the RSF model to predict the risk score. SCellBOW computes the predicted risk score using the *RandomSurvivalForest* function from the *scikit-survival*[104] python package with *n_estimators* = 1000. SCellBOW trains 50 bootstrapped RSF models using 80% of the training set for each of the iterations. The phenotype algebra method 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.

# Sample collection

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, then digested with Collagenase and DNAse. 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 of PBMC-spleen data

CITE-seq using Hashtag Oligos (HTO) was then performed to pool samples into a single 10X Genomics channel for scRNA-seq as described in Abidi *et al.*[40]. HTO binding was carried out according to the manufacturer's guidance (BioLegend, Total-seq B) (**Supplementary Table 7**). Basically, 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 < 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 (10X Genomics Inc. Pleasanton, CA). We then followed the supplier's protocol CG000185, Rev C, until library generation. 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 of PBMC-spleen data

310pM of pooled libraries were sequenced on a NOVAseq6000 instrument with an S1 (v1) flow-cell (Illumina Inc. San Diego, CA) at the genomic platform core facility on the basis of a placement agreement (Nantes, IRS-Un, France). 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. A total of 6,296 cells were recovered with CellRanger, and their gene expression matrices were loaded on R.

## Post-sequencing quality control and data preprocessing of PBMC-spleen data

We subsequently performed the downstream analysis in R using Seurat 3.0. 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[2,105]. We subsequently eliminated doublet HTOs (maximum HTO count > 1) 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. In the preprocessing step, we eliminated cells with gene counts <200 and genes if the number of cells expressing this gene is <3 (**Supplementary Note 3)**. Post gene and cell filtering, the gene expression levels for each cell were normalized and log transformed by Scanpy preprocessing methods. The top 3,000 HVGs were used directly for the pipeline of SCellBOW.

We performed automatic cell annotation based on RNA data, using the Seurat-based Azimuth using human PBMC as the reference atlas[1]. 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. 7d and Supplementary Table 8**).

## Data availability

The PBMC-spleen CITE-seq scRNA-seq expression data generated in this study is available in the GEO with accession number GSE221007 (**Token- ohufwegafhebjyp**). Details of the public datasets analyzed in this paper are described in **Supplementary Table 4**.

## Code availability

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


## Acknowledgments:

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 PBMC-spleen 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.

## Competing interests

DS and GA are stockholders at CareOnco BioTech. Pvt. Ltd. DS is a stockholder at GenterpretR Inc. The remaining authors declare no competing interests.

# Figures

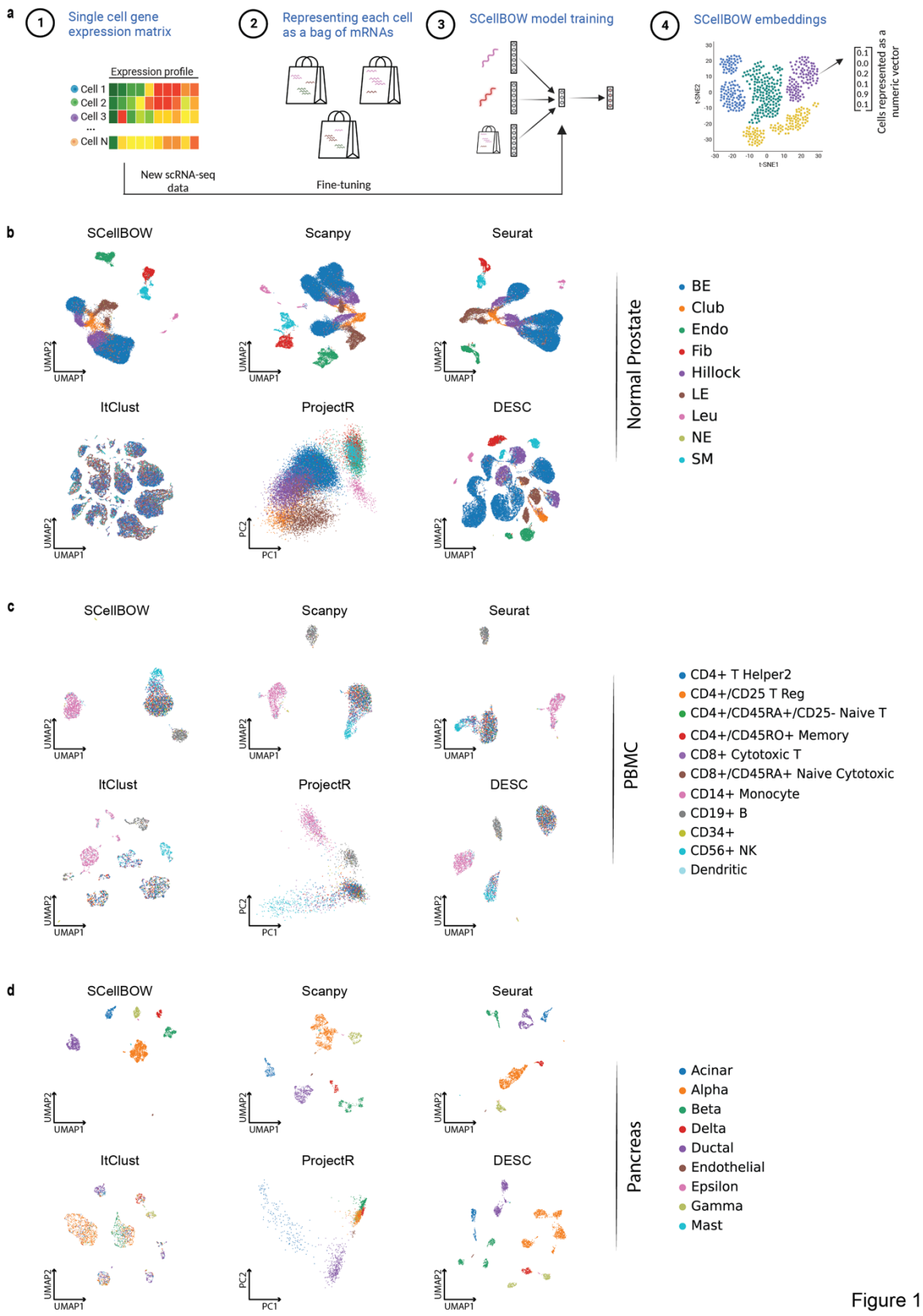


Figure 1

**Fig. 1. SCellBOW workflow. a, Schematic overview of SCellBOW workflow for clustering. The first step was creating a corpus from the gene expression matrix, where cell where genes are analogous to words, and cells are analogous to documents. The second step is updating the vocabulary of the pre-trained Doc2vec neural network model and retraining on the target dataset. The third step is clustering on embeddings generated from the neural network. b, The UMAP plots showing embedding of SCellBOW compared to different existing methods benchmarked on normal prostate, peripheral blood mononuclear cells (PBMC), and pancreas datasets. The coordinates of all the plots are colored by true cell types.**

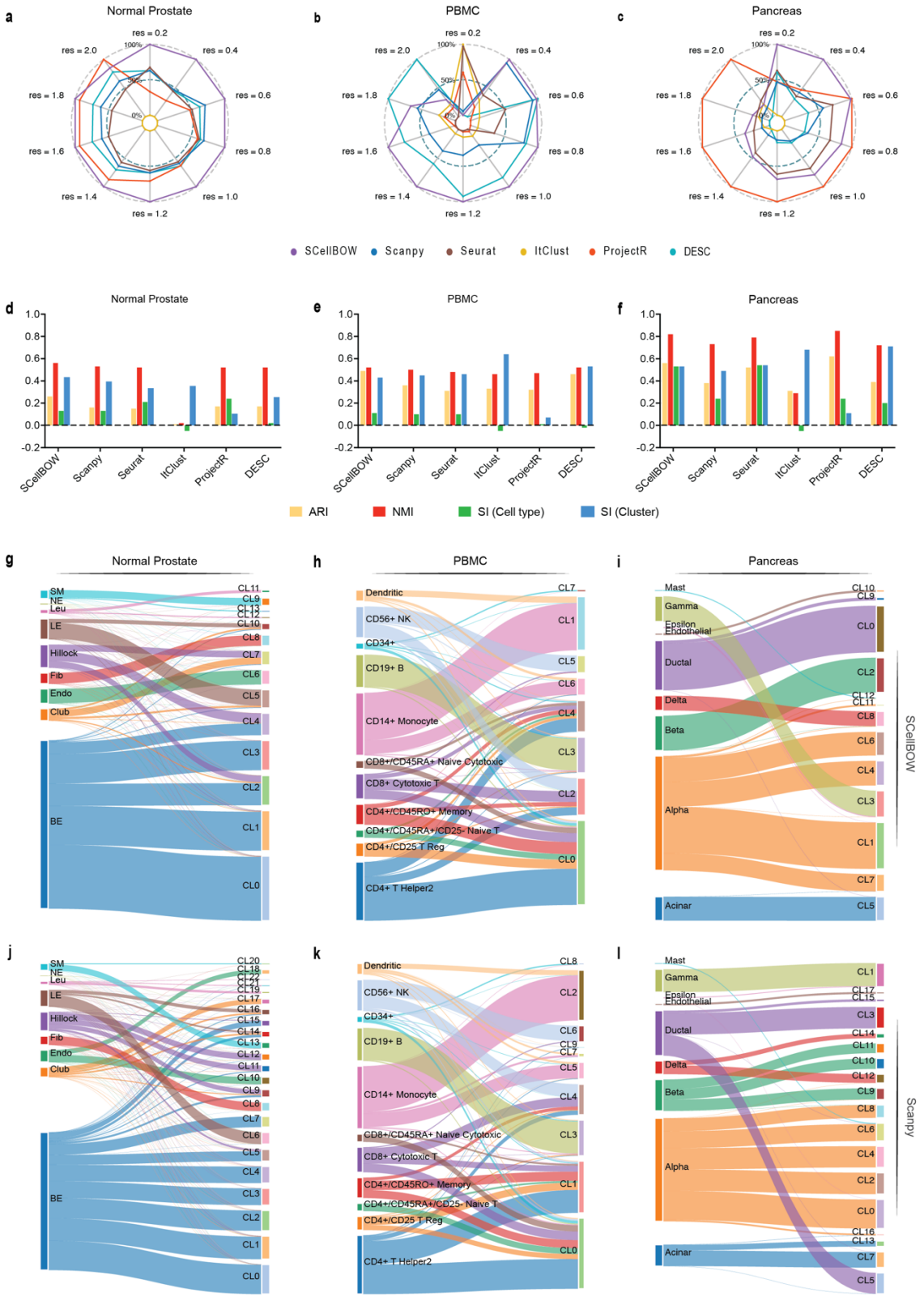
a
Normal Prostate
b
PBMC
c
Pancreas
res = 0.2
res = 0.4
res = 0.6
res = 0.8
res = 1.0
res = 1.2
res = 1.4
res = 1.6
res = 1.8
res = 2.0
100%
50%
0%
SCellBOW
Scanpy
Seurat
ItClust
ProjectR
DESC
d
Normal Prostate
e
PBMC
f
Pancreas
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
ARI
NMI
SI (Cell type)
SI (Cluster)
g
Normal Prostate
h
PBMC
i
Pancreas
SM
NE
Leu
LE
Hillock
Fib
Endo
Club
BE
Dendritic
CD56+ NK
CD34+
CD19+ B
CD14+ Monocyte
CD8+/CD45RA+ Naive Cytotoxic
CD8+ Cytotoxic T
CD4+/CD45RO+ Memory
CD4+/CD45RA+/CD25- Naive T
CD4+/CD25 T Reg
CD4+ T Helper2
Mast
Gamma
Epsilon
Endothelial
Ductal
Delta
Beta
Alpha
Acinar
SCellBOW
j
k
l
Scanpy

Figure 2

**Fig. 2. Clustering accuracy by SCellBOW compared to five benchmarking methods. a, Radial plot showing the percentage of contribution towards ARI for different methods across different resolutions ranged from 0.2 to 2.0 in steps of 0.2. Note that ItClust and ProjectR are resolution-independent methods; thus, their ARI is kept constant across all the resolutions. b, Bar plot showing ARI, NMI Silhouette Index (cluster and cell type). The default resolution was set to 1.0. c, Alluvial plots of normal prostate, PBMC, and pancreas clustering results for the SCellBOW and Scanpy. The resolution for all resolution-dependent algorithms was set to default 1.0.**

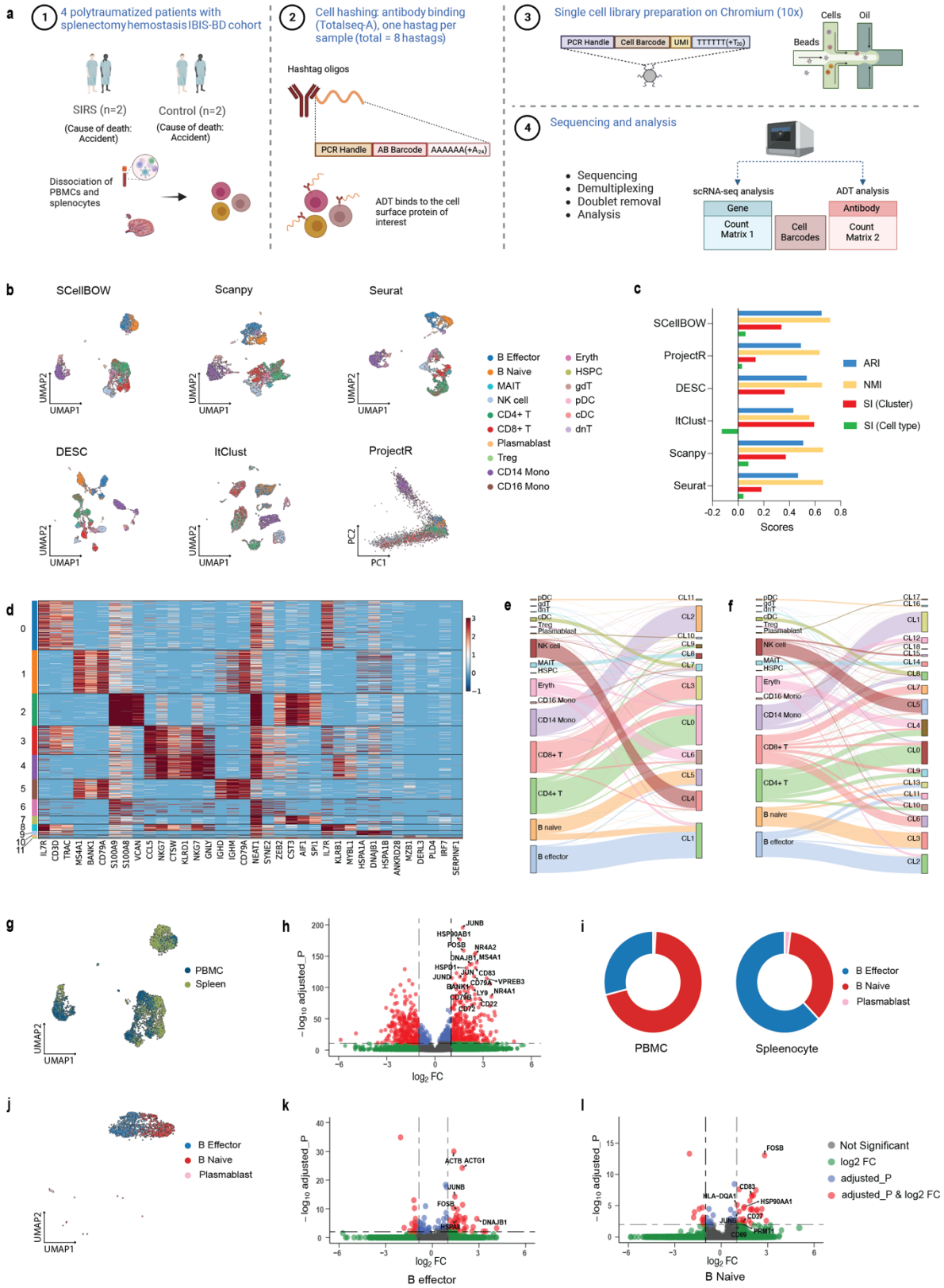
a
1 4 polytraumatized patients with splenectomy hemostasis IBIS-BD cohort
SIRS (n=2)
(Cause of death: Accident)
Control (n=2)
(Cause of death: Accident)
Dissociation of PBMCs and splenocytes
2 Cell hashing: antibody binding (Totalseq-A), one hastag per sample (total = 8 hastags)
Hashtag oligos
PCR Handle
AB Barcode
AAAAAA(+A24)
ADT binds to the cell surface protein of interest
3 Single cell library preparation on Chromium (10x)
PCR Handle
Cell Barcode
UMI
TTTTTT(+T30)
Cells
Oil
Beads
4 Sequencing and analysis
Sequencing
Demultiplexing
Doublet removal
Analysis
scRNA-seq analysis
ADT analysis
Gene
Count Matrix 1
Cell Barcodes
Antibody
Count Matrix 2
b
SCellBOW
Scanpy
Seurat
DESC
ItClust
ProjectR
UMAP1
UMAP2
PC1
PC2
B Effector
B Naive
MAIT
NK cell
CD4+ T
CD8+ T
Plasmablast
Treg
CD14 Mono
CD16 Mono
Eryth
HSPC
gdT
pDC
cDC
dnT
c
SCellBOW
ProjectR
DESC
ItClust
Scanpy
Seurat
ARI
NMI
SI (Cluster)
SI (Cell type)
Scores
d
e
f
g
PBMC
Spleen
h
log2 FC
i
PBMC
Spleenocyte
B Effector
B Naive
Plasmablast
j
k
B effector
l
B Naive
Not Significant
log2 FC
adjusted_P
adjusted_P & log2 FC

Figure 3

**Fig. 3. Evaluation of compartmental differences of PBMC and spleen using SCellBOW. a, An experiment schematic diagram highlighting the sites of the organs for tissue collection and sample processing. In this CITE-seq experiment, PBMCs and splenocytes were collected, followed by high-throughput sequencing and downstream analyses. b, The UMAP plots showing the embedding of SCellBOW compared to different existing methods. The coordinates of all the plots are colored by cell type annotation results using Azimuth. c, Bar plot showing ARI, NMI Silhouette Index (cluster and cell type) at resolution 1.0. d, Heatmap showing the top differentially expressed genes (x-axis) between each SCellBOW cluster (y-axis) and all other cells, tested with a Wilcoxon rank-sum test. e, Alluvial plots of clustering results for the SCellBOW mapped to cell type result using Azimuth. The resolution of SCellBOW was set to 1.0. f, Alluvial plots of clustering results for Scanpy mapped to cell type result using Azimuth at resolution 1.0. g, UMAP plots of the cells colored by their tissue source. h, Volcano plot showing the differential genes (red dots) in the spleen and PBMC for B cells (P-value <0.05, FDR <0.01). i, Donut plot showing the compositional difference in the proportion of B cell subtypes (B naive, B effector, and plasmablast) in PBMC and spleen. j, UMAP plot showing the embedding of SCellBOW colored by B cell subtypes. k-l, Volcano plot showing the differential genes (red dots) in the spleen and PBMC for B effector (k) and B naive cells (l) (P-value <0.05, FDR <0.01).**

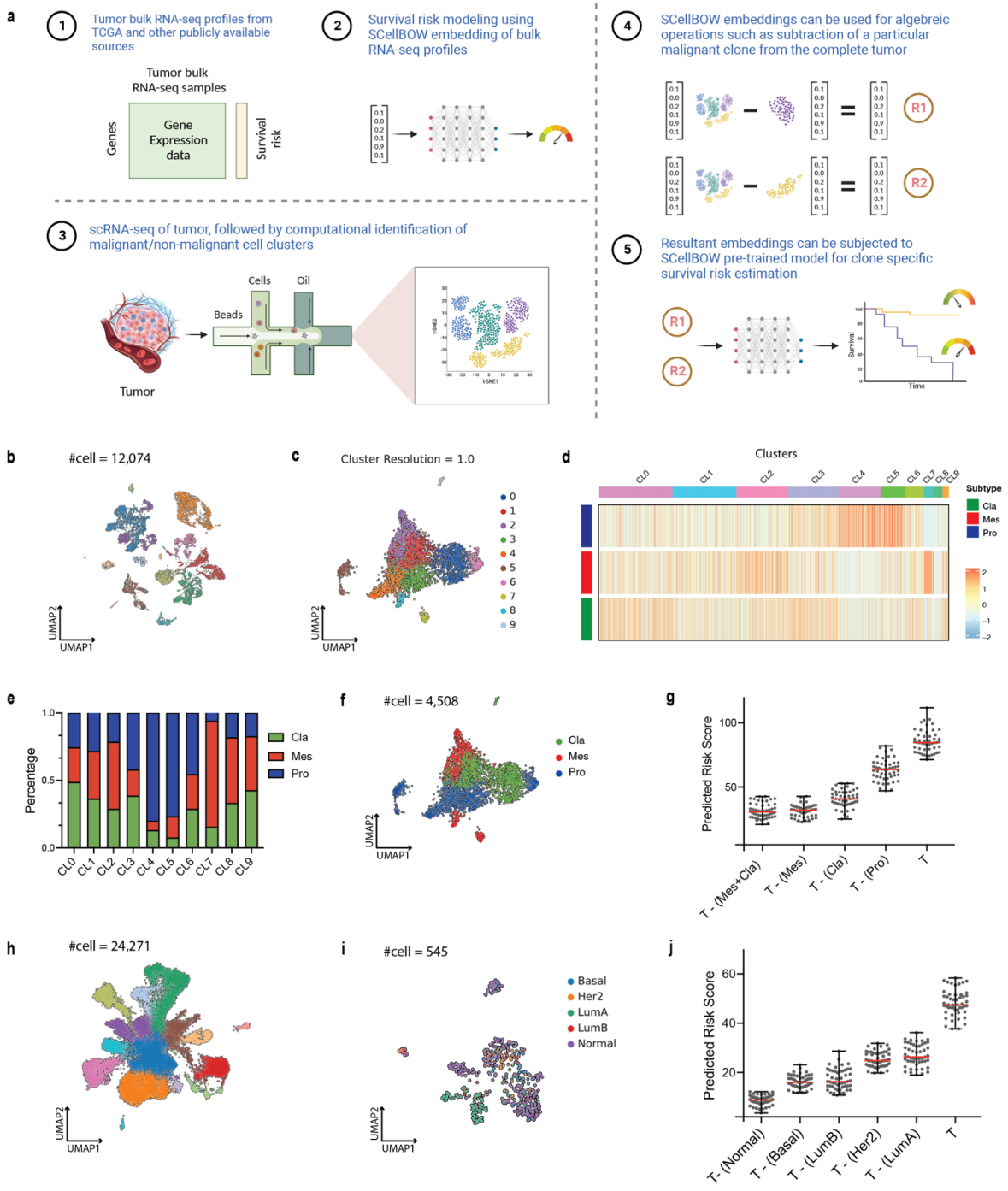
a
1
Tumor bulk RNA-seq profiles from TCGA and other publicly available sources
Tumor bulk RNA-seq samples
Genes
Gene Expression data
Survival risk
2
Survival risk modeling using SCellBOW embedding of bulk RNA-seq profiles
3
scRNA-seq of tumor, followed by computational identification of malignant/non-malignant cell clusters
Tumor
Beads
Cells
Oil
t-SNE1
t-SNE2
4
SCellBOW embeddings can be used for algebreic operations such as subtraction of a particular malignant clone from the complete tumor
R1
R2
5
Resultant embeddings can be subjected to SCellBOW pre-trained model for clone specific survival risk estimation
Survival
Time
b
#cell = 12,074
UMAP1
UMAP2
c
Cluster Resolution = 1.0
0
1
2
3
4
5
6
7
8
9
d
Clusters
CL0
CL1
CL2
CL3
CL4
CL5
CL6
CL7
CL8
CL9
Subtype
Cla
Mes
Pro
e
Percentage
1.0
0.5
0.0
Cla
Mes
Pro
f
#cell = 4,508
g
Predicted Risk Score
100
50
T - (Mes+Cla)
T - (Mes)
T - (Cla)
T - (Pro)
T
h
#cell = 24,271
i
#cell = 545
Basal
Her2
LumA
LumB
Normal
j
Predicted Risk Score
60
40
20
T- (Normal)
T - (Basal)
T - (LumB)
T - (Her2)
T - (LumA)
T

Figure 4

**Fig. 4. Subclonal survival risk inference. a, Schematic overview of SCellBOW phenotype algebra workflow. The first step was creating vectors for reference and queries. The second step was subtracting the query vector from the reference to calculate the predicted risk score by bootstrapping. The third step was stratifying the median of groups ordered by predicted risk score. b, UMAP plot showing the embedding of the Glioblastoma (GBM) source dataset. c, UMAP plot showing SCellBOW embedding of GBM dataset, where cells are colored by SCellBOW clusters at resolution 1.0. d, Heatmap showing GSVA score for three molecular subtypes of GBM: Classical (Cla), Mesenchymal (Mes), and Proneural (Pro), grouped by SCellBOW clusters at resolution 1.0. The *total tumor* is denoted by *T*. e, Bar plot showing the cell type composition of SCellBOW clusters on the GBM dataset. f, UMAP plot showing the embedding of the Glioblastoma (GBM) target dataset colored by assigned cell type. g, Phenotype algebra-based risk scores for GBM molecular subtypes. h. UMAP plot showing the embedding of Breast cancer (BRCA) source dataset. i, UMAP plot showing the embedding of Breast cancer (BRCA) target dataset colored by PAM50 molecular subtype. j, Phenotype algebra-based risk scores of PAM50 molecular subtypes of BRCA. The *total tumor* is denoted by *T*.**

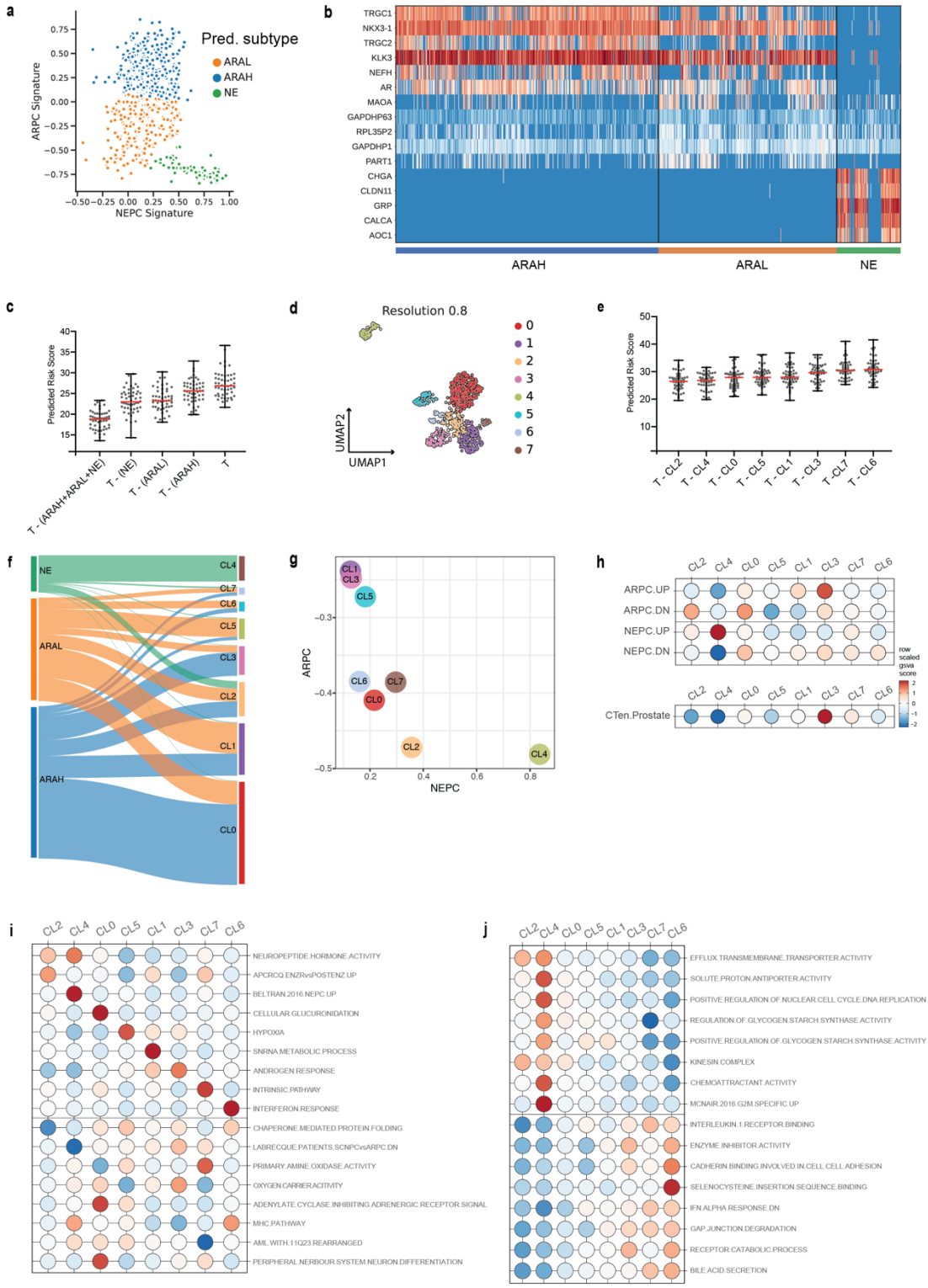
a
Pred. subtype
ARAL
ARAH
NE
ARPC Signature
NEPC Signature
b
TRGC1
NKX3-1
TRGC2
KLK3
NEFH
AR
MAOA
GAPDHP63
RPL35P2
GAPDHP1
PART1
CHGA
CLDN11
GRP
CALCA
AOC1
ARAH
ARAL
NE
c
Predicted Risk Score
T - (ARAH+ARAL+NE)
T - (NE)
T - (ARAL)
T - (ARAH)
T
d
Resolution 0.8
UMAP2
UMAP1
e
Predicted Risk Score
T - CL2
T - CL4
T - CL0
T - CL5
T - CL1
T - CL3
T - CL7
T - CL6
f
NE
ARAL
ARAH
CL4
CL7
CL6
CL5
CL3
CL2
CL1
CL0
g
ARPC
NEPC
h
ARPC.UP
ARPC.DN
NEPC.UP
NEPC.DN
CTen.Prostate
row scaled gsva score
i
NEUROPEPTIDE.HORMONE.ACTIVITY
APCRCQ.ENZRvsPOSTENZ.UP
BELTRAN.2016.NEPC.UP
CELLULAR.GLUCURONIDATION
HYPOXIA
SNRNA.METABOLIC.PROCESS
ANDROGEN.RESPONSE
INTRINSIC.PATHWAY
INTERFERON.RESPONSE
CHAPERONE.MEDIATED.PROTEIN.FOLDING
LABRECQUE.PATIENTS.SCNPCvsARPC.DN
PRIMARY.AMINE.OXIDASE.ACTIVITY
OXYGEN.CARRIER.ACITIVITY
ADENYLATE.CYCLASE.INHIBITING.ADRENERGIC.RECEPTOR.SIGNAL
MHC.PATHWAY
AML.WITH.11Q23.REARRANGED
PERIPHERAL.NERBOUR.SYSTEM.NEURON.DIFFERENTIATION
j
EFFLUX.TRANSMEMBRANE.TRANSPORTER.ACTIVITY
SOLUTE.PROTON.ANTIPORTER.ACTIVITY
POSITIVE.REGULATION.OF.NUCLEAR.CELL.CYCLE.DNA.REPLICATION
REGULATION.OF.GLYCOGEN.STARCH.SYNTHASE.ACTIVITY
POSITIVE.REGULATION.OF.GLYCOGEN.STARCH.SYNTHASE.ACTIVITY
KINESIN.COMPLEX
CHEMOATTRACTANT.ACTIVITY
MCNAIR.2016.G2M.SPECIFIC.UP
INTERLEUKIN.1.RECEPTOR.BINDING
ENZYME.INHIBITOR.ACTIVITY
CADHERIN.BINDING.INVOLVED.IN.CELL.CELL.ADHESION
SELENOCYSTEINE.INSERTION.SEQUENCE.BINDING
IFN.ALPHA.RESPONSE.DN
GAP.JUNCTION.DEGRADATION
RECEPTOR.CATABOLIC.PROCESS
BILE.ACID.SECRETION

Figure 5

**Fig. 5. Phenotype algebra of metastatic prostate cancer data. a, 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. b, Heatmap showing the top differentially expressed genes (x-axis) between each high-level category (y-axis) and all other cells, tested with a Wilcoxon rank-sum test. c, Phenotype algebra-based risk scores for the three high-level categories. The *total tumor* is denoted by *T*. d, UMAP plot for projection of embeddings with coloring based on the SCellBOW clusters at resolution 0.8. e. Phenotype algebra-based cluster-wise risk scores. The *total tumor* is denoted by *T*. f, Alluvial plot illustrating how cells from the three high-level groups distribute across the SCellBOW clusters. g, Centroids of the SCellBOW clusters when extrapolated onto the ARPC vs. NEPC axes. h, Bubble plot of row-scaled GSVA scores for custom curated gene sets containing androgen activated (ARPC.UP) and repressed (ARPC.DN), neuroendocrine up (NEPC.UP) and downregulated (NEPC.DN) genes, as well as a 'prostate tissue signature' gene set from the CTen database[106]. i, Gene sets describing the biology of the SCellBOW clusters, custom selection of the top-up (upper panel) and top-down (lower panel) regulated gene sets per cluster, based on differential expression analysis of GSVA scores of one cluster against all others. j, Top ten gene sets positively (upper panel) and negatively (lower panel) correlated with the rank order of the phenotype algebra risk prediction.**

# References


1. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
2. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
3. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
4. Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S. & Sengupta, D. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Res.* **46**, e36–e36 (2018).
5. Sinha, D., Sinha, P., Saha, R., Bandyopadhyay, S. & Sengupta, D. Improved dropClust R package with integrative analysis support for scRNA-seq data. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz823.
6. Li, X. *et al.* Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 2338 (2020).
7. Big data analytics in single-cell transcriptomics: Five grand opportunities. *Paperpile* https://paperpile.com/app/p/b0e9d5fa-e20a-007e-ad96-e77baaf0721f.
8. Hu, J. *et al.* Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat Mach Intell* **2**, 607–618 (2020).
9. Stein-O'Brien, G. L. *et al.* Decomposing Cell Identity for Transfer Learning across Cellular Measurements, Platforms, Tissues, and Species. *Cell Syst* **12**, 203 (2021).
10. Zeng, P., Wangwu, J. & Lin, Z. Coupled co-clustering-based unsupervised transfer learning for the integrative analysis of single-cell genomic data. *Brief. Bioinform.* **22**, (2021).
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. (2013) doi:10.48550/arXiv.1310.4546.

12. Jebari, C., Cobo, M. J. & Herrera-Viedma, E. A New Approach for Implicit Citation Extraction. *Intelligent Data Engineering and Automated Learning – IDEAL 2018* 121–129 (2018).

13. Le, Q. & Mikolov, T. Distributed Representations of Sentences and Documents. in *Proceedings of the 31st International Conference on Machine Learning* (eds. Xing, E. P. & Jebara, T.) vol. 32 1188–1196 (PMLR, 2014).

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

15. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* 1877–1901 (2020).

16. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

17. Simon, C. *et al.* BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* **19**, 57 (2019).

18. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab083.

19. Yang, M. *et al.* Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution. *Nucleic Acids Res.* **50**, e81 (2022).

20. Yang, F. *et al.* 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 (2022).

21. Connell, W., Khan, U. & Keiser, M. J. A single-cell gene expression language model. *arXiv [q-bio.QM]* (2022).

22. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. Commun.* **12**, 1–15 (2021).

23. Gong, W., Kwak, I.-Y., Pota, P., Koyano-Nakagawa, N. & Garry, D. J. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* **19**, 220 (2018).

24. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1169 (2020).

25. Mongia, A., Sengupta, D. & Majumdar, A. McImpute: Matrix Completion Based Imputation for Single Cell RNA-seq Data. *Front. Genet.* **10**, 9 (2019).

26. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* (2021) doi:10.1038/s41587-021-01001-7.

27. Mieth, B. *et al.* Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-Seq data. *Sci. Rep.* **9**, 20353 (2019).

28. Allen, C. & Hospedales, T. Analogies Explained: Towards Understanding Word Embeddings. in *Proceedings of the 36th International Conference on Machine Learning* (eds. Chaudhuri, K. & Salakhutdinov, R.) vol. 97 223–231 (PMLR, 09--15 Jun 2019).

29. Pearson, K. *On Lines and Planes of Closest Fit to Systems of Points in Space*. (1901).

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

31. Henry, G. H. *et al.* A Cellular Anatomy of the Normal Adult Human Prostate and Prostatic Urethra. *Cell Rep.* **25**, 3530–3542.e5 (2018).

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

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

34. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385–394.e3 (2016).

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

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

37. Hu, J. *et al.* Iterative transfer learning with neural network for clustering and cell type

classification in single-cell RNA-seq analysis. *Nature Machine Intelligence* **2**, 607–618 (2020).

38. Lewis, S. M., Williams, A. & Eisenbarth, S. C. Structure-function of the immune system in the spleen. *Science immunology* **4**, (2019).
39. Bronte, V. & Pittet, M. J. The spleen in local and systemic regulation of immunity. *Immunity* **39**, 806 (2013).
40. Abidi, A. *et al.* Characterization of Rat ILCs Reveals ILC2 as the Dominant Intestinal Subset. *Front. Immunol.* **11**, 255 (2020).
41. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
42. Denisenko, E. *et al.* Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biol.* **21**, 1–25 (2020).
43. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).
44. Tsubata, T. CD22 and CD72 are inhibitory receptors dominantly expressed in B lymphocytes and regulate systemic autoimmune diseases. *Zeitschrift für Rheumatologie* **76**, 10–13 (2016).
45. Nolte, M. A. *et al.* B Cells Are Crucial for Both Development and Maintenance of the Splenic Marginal Zone. *The Journal of Immunology* **172**, 3620–3627 (2004).
46. Tsai, D.-Y., Hung, K.-H., Chang, C.-W. & Lin, K.-I. Regulatory mechanisms of B cell responses and the implication in B cell-related diseases. *J. Biomed. Sci.* **26**, 1–13 (2019).
47. Liu, C. *et al.* Analyzing actin dynamics during the activation of the B cell receptor in live B cells. *Biochemical and Biophysical Research Communications* vol. 427 202–206 Preprint at https://doi.org/10.1016/j.bbrc.2012.09.046 (2012).
48. Zandvoort, A. *et al.* CD27 expression in the human splenic marginal zone: the infant marginal zone is populated by naive B cells. *Tissue Antigens* **58**, (2001).
49. Infantino, S. *et al.* Arginine methylation catalyzed by PRMT1 is required for B cell activation and differentiation. *Nat. Commun.* **8**, 1–12 (2017).

50. Dwivedi, B. & Bhasin, M. Survival Genie: A Web Portal for Single-Cell Data, Gene-Ratio, and Cell Composition-Based Survival Analyses. *Blood* **138**, 276 (2021).

51. Tang, Y. *et al.* Identification of five important genes to predict glioblastoma subtypes. *Neurooncol Adv* **3**, vdab144 (2021).

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

53. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835–849.e21 (2019).

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

55. Couturier, C. P. *et al.* Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.* **11**, 3406 (2020).

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

57. Azam, Z., To, S.-S. T. & Tannous, B. A. Mesenchymal Transformation: The Rosetta Stone of Glioblastoma Pathogenesis and Therapy Resistance. *Adv. Sci.* **7**, 2002015 (2020).

58. Kim, Y. *et al.* Perspective of mesenchymal transformation in glioblastoma. *Acta Neuropathol Commun* **9**, 50 (2021).

59. Grundy, T. J. *et al.* Mechanosensitive expression of the mesenchymal subtype marker connective tissue growth factor in glioblastoma. *Sci. Rep.* **12**, 14982 (2022).

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

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

62. Wu, S. Z. *et al.* Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *The EMBO Journal* vol. 39 Preprint at https://doi.org/10.15252/embj.2019104063

(2020).

63. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, 3588 (2018).

64. Zhou, S. *et al.* 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 (2021).

65. Yin, L., Duan, J.-J., Bian, X.-W. & Yu, S.-C. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res.* **22**, 1–13 (2020).

66. Mathews, J. C. *et al.* Robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes. *NPJ Breast Cancer* **5**, 30 (2019).

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

68. Ahn, H. J., Jung, S. J., Kim, T. H., Oh, M. K. & Yoon, H.-K. Differences in Clinical Outcomes between Luminal A and B Type Breast Cancers according to the St. Gallen Consensus 2013. *Journal of Breast Cancer* vol. 18 149 Preprint at https://doi.org/10.4048/jbc.2015.18.2.149 (2015).

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

70. Dai, C., Heemers, H. & Sharifi, N. Androgen Signaling in Prostate Cancer. *Cold Spring Harb. Perspect. Med.* **7**, (2017).

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

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

73. Lonergan, P. E. & Tindall, D. J. Androgen receptor signaling in prostate cancer development and progression. *J. Carcinog.* **10**, 20 (2011).

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

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

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

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

78. Yamada, Y. & Beltran, H. Clinical and Biological Features of Neuroendocrine Prostate Cancer. *Current Oncology Reports* vol. 23 Preprint at https://doi.org/10.1007/s11912-020-01003-9 (2021).

79. Conteduca, V. *et al.* Clinical features of neuroendocrine prostate cancer. *Eur. J. Cancer* **121**, 7–18 (2019).

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

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

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

83. Wang, H. T. *et al.* 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* vol. 32 3383–3390 Preprint at https://doi.org/10.1200/jco.2013.54.3553 (2014).

84. Kwon, J. T. W., Bryant, R. J. & Parkes, E. E. The tumor microenvironment and immune

responses in prostate cancer patients. *Endocrine-Related Cancer* vol. 28 T95–T107 Preprint at https://doi.org/10.1530/erc-21-0149 (2021).

85. Miller, D. R., Ingersoll, M. A., Teply, B. A. & Lin, M.-F. Targeting treatment options for castration-resistant prostate cancer. *Am J Clin Exp Urol* **9**, 101–120 (2021).
86. Kushwaha, P. P. & Gupta, S. New insights for drug resistance in metastatic castration-resistant prostate cancer. *Cancer Drug Resist* **5**, 846–849 (2022).
87. Karantanos, T., Corn, P. G. & Thompson, T. C. Prostate cancer progression after androgen deprivation therapy: mechanisms of castrate resistance and novel therapeutic approaches. *Oncogene* **32**, 5501–5511 (2013).
88. Vaidyanathan, V. *et al.* Environmental factors and risk of aggressive prostate cancer among a population of New Zealand men - a genotypic approach. *Mol. Biosyst.* **13**, 681–698 (2017).
89. Domenichini, A., Adamska, A. & Falasca, M. ABC transporters as cancer drivers: Potential functions in cancer development. *Biochim. Biophys. Acta Gen. Subj.* **1863**, 52–60 (2019).
90. Fletcher, J. I., Haber, M., Henderson, M. J. & Norris, M. D. ABC transporters in cancer: more than just drug efflux pumps. *Nat. Rev. Cancer* **10**, 147–156 (2010).
91. Loder, B. F. *et al.* B Cell Development in the Spleen Takes Place in Discrete Steps and Is Determined by the Quality of B Cell Receptor–Derived Signals. *J. Exp. Med.* **190**, 75–90 (1999).
92. Oberacker, T. *et al.* Enhanced expression of thioredoxin-interacting-protein regulates oxidative DNA damage and aging. *FEBS Lett.* **592**, 2297 (2018).
93. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv [cs.CL]* (2013).
94. Vasighizaker, A., Danda, S. & Rueda, L. Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data. *Sci. Rep.* **12**, 1–16 (2022).
95. Cao, X. H., Stojkovic, I. & Obradovic, Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics* **17**, (2016).
96. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning*

*research* **12**, 2825–2830 (2011).

97. Rehurek, R. & Sojka, P. Gensim--python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* **3**, (2011).
98. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ('O'Reilly Media, Inc.', 2009).
99. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
100. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
101. Huang, P. J. *Classification of Imbalanced Data Using Synthetic Over-sampling Techniques*. (2015).
102. Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *arXiv [cs.LG]* (2016).
103. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Alexander Wolf, F. anndata: Annotated data. *bioRxiv* 2021.12.16.473007 (2021) doi:10.1101/2021.12.16.473007.
104. Pölsterl, S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. *J. Mach. Learn. Res.* **21**, 1–6 (2020).
105. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
106. Shoemaker, J. E. *et al.* CTen: a web-based platform for identifying enriched cell types from heterogeneous microarray data. *BMC Genomics* **13**, 460 (2012).