RESEARCH Open Access

A machine learning one-class logistic regression model to predict stemness for single cell transcriptomics and spatial omics

Felipe Segato Dezem[1,2,3], Maycon Marção[2,3], Bassem Ben-Cheikh[4], Nadya Nikulina[4], Ayodele Omotoso[5,6], Destiny Burnett[5,6], Priscila Coelho[5,6], Judith Hurley[5,6], Carmen Gomez[5,6], Tien Phan-Everson[7], Giang Ong[7], Luciano Martelotto[8], Zachary R. Lewis[7], Sophia George[5,6], Oliver Braubach[4], Tathiane M. Malta[3] and Jasmine Plummer[1,2,9,10*]

Abstract

Cell annotation is a crucial methodological component to interpreting single cell and spatial omics data. These approaches were developed for single cell analysis but are often biased, manually curated and yet unproven in spatial omics. Here we apply a stemness model for assessing oncogenic states to single cell and spatial omic cancer datasets. This one-class logistic regression machine learning algorithm is used to extract transcriptomic features from non-transformed stem cells to identify dedifferentiated cell states in tumors. We found this method identifies single cell states in metastatic tumor cell populations without the requirement of cell annotation. This machine learning model identified stem-like cell populations not identified in single cell or spatial transcriptomic analysis using existing methods. For the first time, we demonstrate the application of a ML tool across five emerging spatial transcriptomic and proteomic technologies to identify oncogenic stem-like cell types in the tumor microenvironment.

Keywords Single cell, Spatial, Machine learning, Cancer stem, Proteomic, Transcriptomic

Introduction

Traditional bulk RNA sequencing provides an average gene expression profile of a population of cells, which may obscure differences among individual cells [1]. In contrast, single-cell RNA sequencing (scRNAseq) has emerged as a powerful method for understanding cellular heterogeneity and allows the measurement of gene expression profiles in individual cells, providing a high-resolution view of transcriptional variability within a population [2]. Advances in scRNAseq technology and computational analysis methods have enabled the identification of rare cell populations, the discovery of novel cell types, and the characterization of cell state transitions during development, disease progression, and treatment response [3]. Despite the demonstrated utility of scRNAseq in aiding discoveries, it has limitations for biological interpretations. scRNAseq data analysis is based

*Correspondence:
Jasmine Plummer
jasmine.plummer@stjude.org
[1] Center for Spatial Omics, St Jude Children's Research Hospital, Memphis, TN, USA
[2] Department of Developmental Neurobiology, St Jude Children's Research Hospital, Memphis, TN, USA
[3] Department of Clinical Analysis, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirao Preto, University of Sao Paulo, Sao Paulo, SP, Brazil
[4] Akoya Biosciences, The Spatial Biology Company, Marlborough, MA, USA
[5] Department of Obstetrics, Gynecology and Reproductive Sciences, University of Miami Miller School of Medicine, Miami, FL, USA
[6] Sylvester Comprehensive Cancer Center, UHealth Medical Systems, Miami, FL, USA
[7] Nanostring Technologies, Seattle, WA, USA
[8] SAiGENCi, University of Adelaide, Adelaide, Australia
[9] Department of Cellular & Molecular Biology, St Jude Children's Research Hospital, Memphis, TN, USA
[10] Comprehensive Cancer Center, St Jude Children's Research Hospital, Memphis, TN, USA

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

on proper cell type annotation. Cell type annotation is used to represent cells as clusters based on their gene expression profiles using unsupervised learning methods [4–6]. These cell clusters are then annotated using cluster profiles aided by marker gene information. These annotation methods are based on taking previously annotated scRNAseq databases as reference (efforts such as the Human Cell Atlas has been instrumental in aiding this) or using marker genes themselves to annotate cell clusters e.g. CD4 cell clusters are annotated as immune cells, and GFAP positive cells are glial clusters. However, there are many different subtypes of cell identities for each of these cell types that extend beyond the traditional cell phenotyping annotations. The informatic tools and techniques used to identify different cell types, map cellular interactions, and investigate gene expression patterns [7] are heavily biased and limited to inference for the features/gene expression annotations used in the reference datasets [8, 9]. These cell annotations may not reflect the biology seen in the dataset to which it is being applied to. Since each single cell dataset has heterogeneity in the sample processing, sample type and biological treatments, it is difficult to use these methods to match across reference cell annotations or dynamic gene expression profiles. Some methods are developed using the annotated cells from the same dataset to infer the cell types of the remaining but these models require unstructured parameters e.g. number of clusters or are dependent on cell number input requirements [6].

These problems are further exemplified in the spatial omics field. An explosion of technologies has emerged in spatial omics which are revolutionizing our understanding of complex tissues by providing detailed information on the cellular organization and functioning within intact heterogeneous environments [10]. Spatial omics allows researchers to simultaneously analyze multiple molecular features in the same tissue sample, often at a single cell level, to provide a comprehensive view of cellular interactions, gene expression, and cellular environment. A promising application of spatial omics is the integration of scRNAseq with spatial data to study the heterogeneity of intact tissue. The raw data from spatial omics techniques is often in the form of images or matrices that represent the gene expression across a tissue section [11]. The information captured from spatial omics data is typically processed to generate expression profiles for individual cells, which are then suitable for downstream analysis. The conversion of spatial omics data into single-cell expression profiles is a critical step in the analysis of these data hence the same challenges for scRNAseq analysis exist and are further compounded by the laborious task of cell phenotyping. This is often done by manually assigning cell labels based on the gene marker annotation to the spatial data.

Many Machine Learning (ML) models have been developed to address this problem of large datasets and manual curation**.** A growing number of deep learning-based methods have been applied to scRNAseq data analyses to enhance the accuracy and efficiency of the analysis and achieve superior performance [12]. These ML models harness the generalization of cell annotations and the robustness dataset integration. These algorithms can be used to analyze high-dimensional data and identify patterns that may not be apparent with our present tools. In this study we harness ML algorithms to train and recognize patterns in gene expression data with high accuracy at a much faster rate, hence reducing the time needed to analyze large single cell and spatial datasets to provide new insights into the biology of the tissue being studied.

Breast cancer cells that possess stem cell like properties (e.g. self renewal, rapid proliferation) are associated with chemotherapy and radiation resistance, disease recurrence and poorer outcomes [13, 14]. Annotating cell types in scRNAseq data is important for establishing breast cancer severity and progression based on their tumor microenvironments (TME) [15]. ML model generation from bulk tumor RNAseq datasets has been instrumental in identifying breast cancer stemness [16–18] as the field explores various strategies to target breast cancer stem cells to improve treatment. We apply a bulk RNAseq ML model to single cell and sequencing based, probe based and protein based spatial omics datasets to identify the cell types asoociated with stemness in breast cancer. This study highlights the use of ML in single cell and spatial omic analysis to accelerate our understanding of the complex relationships between stemness, gene expression and tissue morphology.

Methods

Data collection for reference sets

The scRNAseq data used for this work was accessed from the Gene ExpressionOmnibus database (GEO; GSE176078 and GSE161529) and ArrayExpress (E-MTAB-6524) [19–21]. We used GSE176078 (dataset 1) which contains 26 samples of primary breast cancer, (11 ER+, 5 HER2+, and 10 TNBC) with major and minor cell types previously annotated. Another cohort GSE161529 (dataset 2) within 32 samples of primary breast cancer (17 ER+, 6 HER2+, 4 BRCA1 pre-neoblastic, 4 TNBC, 4 TNBC/BRCA1, and 1 PR+) was used to validate our main findings of stemnessin breast cancer cells. We then used an induced pluripotent stem cells dataset (E-MTAB-6524) to validate our stemness model in scRNAseq data.

Spatial transcriptomic datasets were retrieved from publicly available datasets from a previously published breast cancer study [19, 22] and the Vizgen website (MERSCOPE) [23]. The spatial proteomic datasets were generated by Akoya Biosciences and Nanostring Technologies. The Phenocycler Fusion dataset was run on a FFPE breast cancer sample from the biorepository at University of Miami Health Science Systems. This tumor underwent pathological review and sections were sent to Akoya Biosciences for processing. qpTIFF files were generated and rendered by the company. All files were processed based on 10X Genomics and Akoya's instructions (see below) and used as input for the stemness model. These datasets were processed in a similar format for gene counts as single cell data and used in the same format as scRNAseq for stemness prediction modeling. The CosMX proteomic dataset was generated and provided by Nanostring Technologies.

Single cell processing and analysis

Samples from both datasets were processed with the same settings using Seurat v4.0 for QC, filtering, normalization, clustering and visualization. The filtering parameters were as follows: nFeature_RNA > 200 < 5000, percent.mt < 10; nCount_RNA > 200. The cells that have mitochondrial genes greater than 10% or have fewer than 200 detected genes were filtered out. A scale factor of 10,000 was used to normalize all the remaining cells. To correct for the batch effect between different samples, the reciprocal principal component analysis (RPCA) method in the Seurat package was applied to integrate the complete data set. The genes enriched in each cluster were identified using FindAllMarkers function in Seurat by applying a Wilcoxon Rank Sum test and then performing multiple test corrections using the Bonferroni method. The multiple-test corrected $P < 0.05$ was used as a cut-off for significance. Samples were normalized using the following settings: normalization.method = "LogNormalize"; scale.factor = 10000 >. The remaining Seurat parameters were default with the exception of: FindVariableFeatures:-selection.method = 'vst', nfeatures = 3000; RunPCA-features = 3000 VariableFeatures; FindNeighbors- reduction = 'pca', dims = 1:20 (20 PCs); FindClusters: resolution 0.5; RunUMAP: reduction = 'pca', dims = 1:20. For additional processing and graphical representation the following R packages were used in sc-type v1.0, dplyr v1.1.1, and ggplot2 v3.4.2.

Spatial data analysis

Visium dataset

Visium breast cancer data from 10X genomics was mapped and demultiplexed using SpaceRanger as per company default parameters. Processing included transforming to read counts, overlaying expression data with H&E tissue images and unsupervised clustering. Specifically, spots classified as artifact and spots with count = 0 were removed. Seurat version 5 was used to perform data normalization with the method SCT transform [24], and data clustering using the Louvain algorithm with multi-level refinement and resolution = 1. For the stemness prediction, we used the normalized expression matrix to calculate spearman correlation with the feature weights from the OCLR model matching genes and scaled to 0–1 format. We then transferred the scores to the Visium seurat object for visualization.

Xenium dataset

This dataset included RNA reads, images, and cell segmentation were directly downloaded from 10X website and run using their recommended Explorer pipeline. Molecules with count = 0 were removed and the same analysis routine described previously for Visium data was employed. We then performed cell-type deconvolution using the RCTD method [25] to predict each cell point present in the xenium dataset using the major cell type annotations from the scRNAseq dataset.

Vizgen dataset

Output files from Vizgen dataset were processed as follows: Molecules with count = 0 were removed, SCT was used to transform and normalize the data, with a clip range of [-10,10]. For cell clustering, SCT filter was used at a lower resolution of 0.3.

PhenoCycler fusion dataset

QuPath v0.4 was used to process the 66-channel qpTIFF image generated from the PhenoCycler instrument. Briefly, we employed the StarDist (arXiv:1806.03535) algorithm for cell segmentation using the DAPI channel and exported the measurements as a text file. We selected measurements in cells and mean intensity of captured signal. The DAPI channel was removed for further analysis. To create a Seurat object, we created a sparse matrix with each channel as rows and each segmented cell as columns, and a metadata matrix with centroid coordinates for each cell. Centroids coordinates were used for visualization. In order to predict stemness score for each cell, we created a new model using 28 genes that intersected the panel used in this assay and genes found in the bulk RNAseq data used to train the OCLR model and employed the same strategy to transfer the scores back to the Seurat object.

CosMX dataset

The Nanostring dataset was segmented and generated by Nanostring Technologies. The input file was provided as a cell count matrix as per their published pipelines. Seurat object was created using the same method used for the Phenocycler dataset.

Stemness prediction modeling

Building on existing methods from bulk RNAseq [16], we extended this iPS/ES model for utilization in single cell RNAseq data. The model parameters were based on the total content of 12 922 genes. The iPS dataset (20 000 cells) [19–21] was used to validate the iPS/ES bulk model in single cell data. Similarly, this model was tested using two large published breast cancer datasets of 80 000 and 180 000 cells respectively. These three datasets were combined as Seurat objects for normalization and extraction of their gene expression. The stemness model was applied to these single cell datasets using a spearman correlation value of both genes weights in the model and gene expression values on the datasets for each cell in the single cell matrix. We scaled all correlations within 0 to 1 calling these scores to create the stemness index. These individual Seurat objects were used for the next stage of analysis where they were used as inputs into the Seurat clusterization pipeline.

Performance of single cell stemness model

To evaluate how dependable our stemness model is for single-cell data, we conducted a bootstrapping analysis to both high and low-quality samples, determined by their RNA features and counts, aiming to check the model's consistency in various scenarios (Supp. Fig. 1). We then analyzed cell type annotations by comparing two datasets. From the first dataset, we took the most expressed cell type markers and matched them with cluster markers in the second dataset. Specifically, for the primary cell types referenced in the dataset 1 paper, we picked the top 30 markers and looked for overlaps in the second dataset, focusing on markers with an average fold change >|0.5|. For cell types like Cancer Cycling, Cancer Basal, and T Cycling, which had the highest stemness scores in dataset 1, we concentrated on the top 10 markers, again seeking overlaps in dataset 2 with a similar fold change criteria. To assist in this comparison, we employed the sc-type software to align cell types from dataset 1 with those in dataset 2. Additionally, we conducted a differential expression analysis to identify distinct marker genes across cell clusters using default parameters on the Seurat FindMarkers function (p-value < 0.05; FoldChange > 0.5). This comparison was specifically between the highest and lowest stemness cells (in the 25th vs. 75th percentiles) for the mentioned cell types across both datasets. Finally, for a broader biological context, we undertook GO pathway enrichment analysis using the clusterProfiler package, version 3.16.26 [26].

Data visualization

All plotting functions used in this study were included in Seurat v4,Seurat v5, ggplot2, RColorBrewer [27] and plotly [28] packages.

Results

A stemness index was calculated for every cell in two independent breast cancer scRNAseq datasets (dataset 1 and dataset 2) by correlating the gene weight derived from the one-class model trained on bulk RNASeq iPSC/ESC cell lines with the normalized gene expression from the scRNAseq datasets and scaling that metric to [0–1] format (Fig. 1). The stemness index was visualized in UMAP formats (Fig. 2A, B) and based on top 75% and bottom 25% of scoring stemness cells, we calculated that 20,171 of 80,682 total cells in dataset 1 present a high stemness phenotype, compared to 45,079 of 180,315 of cells in dataset 2. In the opposite stemness spectrum, 20,171 of cells in dataset 1 present a low stemness score, compared to 45,079 of cells in dataset 2. Both datasets were annotated using the same cell types. The proportion of cells per clinical classification in dataset 1, were representative of ER+, HER2+and Triple-Negative Breast Cancer (TNBC). These categories are also present in dataset 2, with the addition of BRCA1 pre-neoplastic, PR+and TNBC+BRCA1 cells (Fig. 2C). In evaluating the biological relevance of the stemness score, we utilized our established model on an iPS scRNAseq dataset, encompassing a total of 21,599 cells (Fig. 2D). This revealed a median stemness score of 0.724. For context, the stemness scores for datasets 1 and 2 were 0.394 and 0.402, respectively (Fig. 2E). Further delineation of the datasets based on stemness profiles led to the identification of three distinct cell types: Cycling T-Cells: In dataset 1, there were 1,442 cells, representing 1.78% of the total. Meanwhile, dataset 2 contained 13,760 cells, or 7.63% of the total cells. Cancer Cycling Cells: Dataset 1 had 2,631 cells (3.26%), while dataset 2 comprised 12,125 cells (6.72%). Cancer Basal Cells: These represented 2,877 cells (3.56%) in dataset 1 and a notable 20,156 cells (11.1%) in dataset 2. These classifications and their respective distributions can be visualized in Fig. 3A and B. In dataset 1, we observe clusters where cancer cycling and cancer basal cells are clustered together. Cycling T-cells were associated mostly having higher stemness profiles than average in both datasets (Fig. 3A, C). In dataset 2, cycling T-cells were mostly clustered with cancer basal cells and more sparsely separated compared to dataset 1, but with a similar overall stemness index profile (Fig. 3B, C). We see this pattern in both breast cancer datasets.

From the selected cell types we conducted differential expression of gene (DEG) analysis on both datasets. We analyzed the DEGs between the selected cell types

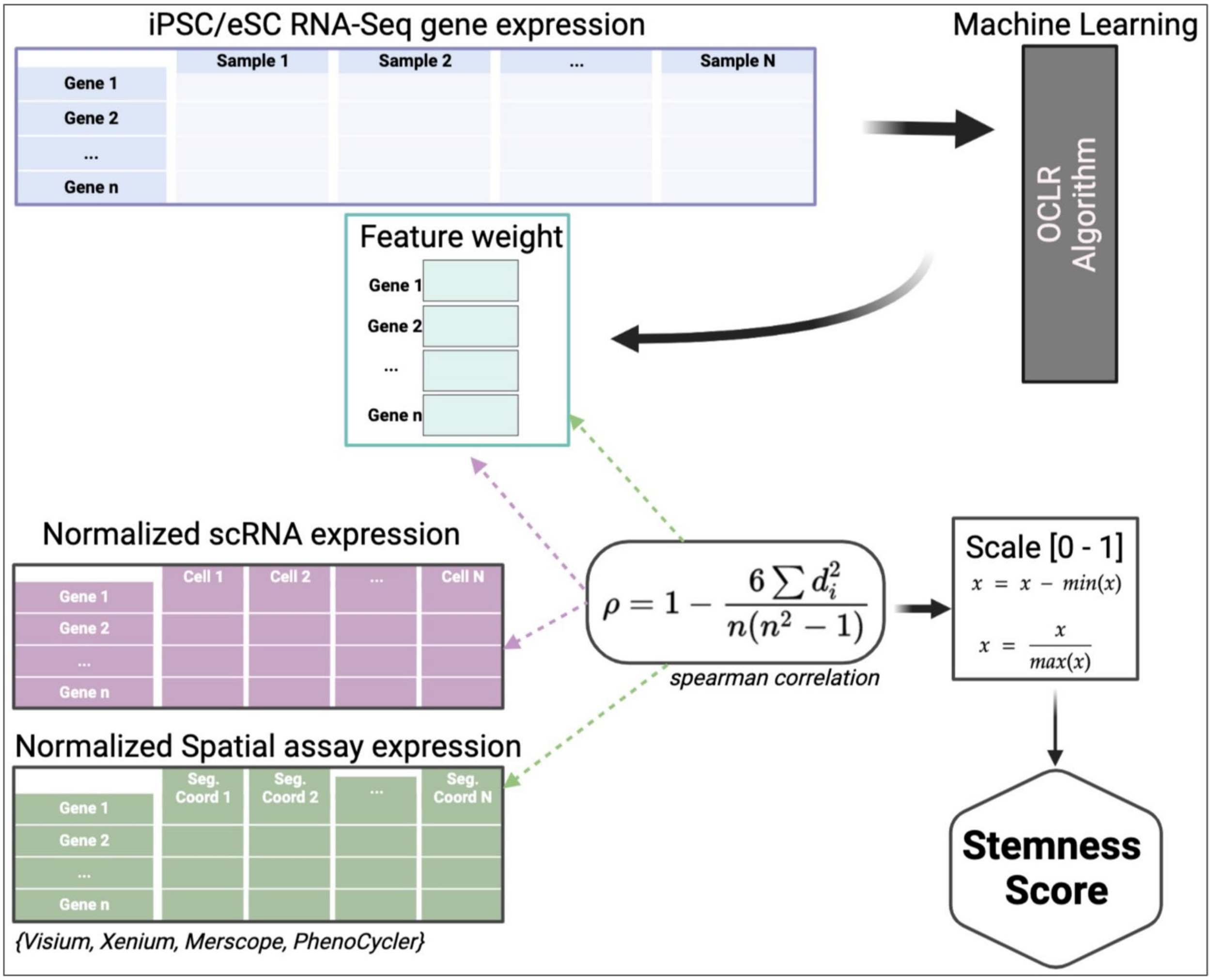

Fig. 1 Graphical abstract describing step-by-step how the OCLR algorithm is employed to predict a stemness score from scRNA ad Spatial Omics datasets

in both datasets. The groups were separated based on their stemness scores, with one group having higher scores and the other having lower scores. To ensure equal group sizes, we retained the same number of cells in both groups. From dataset 1, cancer basal, cancer cycling and cycling T cells have 685, 393 and 230 DEGs respectively. From dataset 2, cancer basal, cancer cycling and cycling T cell had 525, 528 and 632 DEGs respectively. For cancer basal in dataset 1, we identified 336 genes highly expressed and presenting high stemness scores, among the top expressed genes with high stemness is LDHB (Fig. 4A), which has a known role in cancer cell proliferation [29]. For dataset 2, 311 highly expressed genes were identified differentiating top and bottom stemness cells. One of the genes with higher log2 Fold Change is CXCL1 (Fig. 4B) and was shown in previous studies [30] to stimulate invasion and migration in ER negative breast cancer cells. We found 104 genes in common in both datasets, 30.9% and 33.4% of dataset 1 and dataset 2 respectively. We performed gene ontology (GO) enrichment analysis using these gene lists as input for both datasets. In dataset 1, GO terms were enriched ($n = 72$) using the GO Biological Process ontologies. For dataset 2, that number was $n = 91$. The number of overlapped GO terms was $n = 17$ (Fig. 4C). To add another level of confidence that the results seen in dataset 1 and 2 were similar, we performed a semantic similarity analysis [31], which takes the non-overlapping genes in both datasets as input and computes a similarity score based on the same gene ontology. To establish if the DEGs from dataset 1 and 2 were similar beyond their pathway enrichment overlap, we conducted a gene similarity analysis. In this analysis we took the dataset 1 DEGs and dataset 2 DEGs that define stemness to calculate the similarity index [31] to take into account confounding variables in our analysis (Supp. Fig. 3). Our results revealed that the similarity

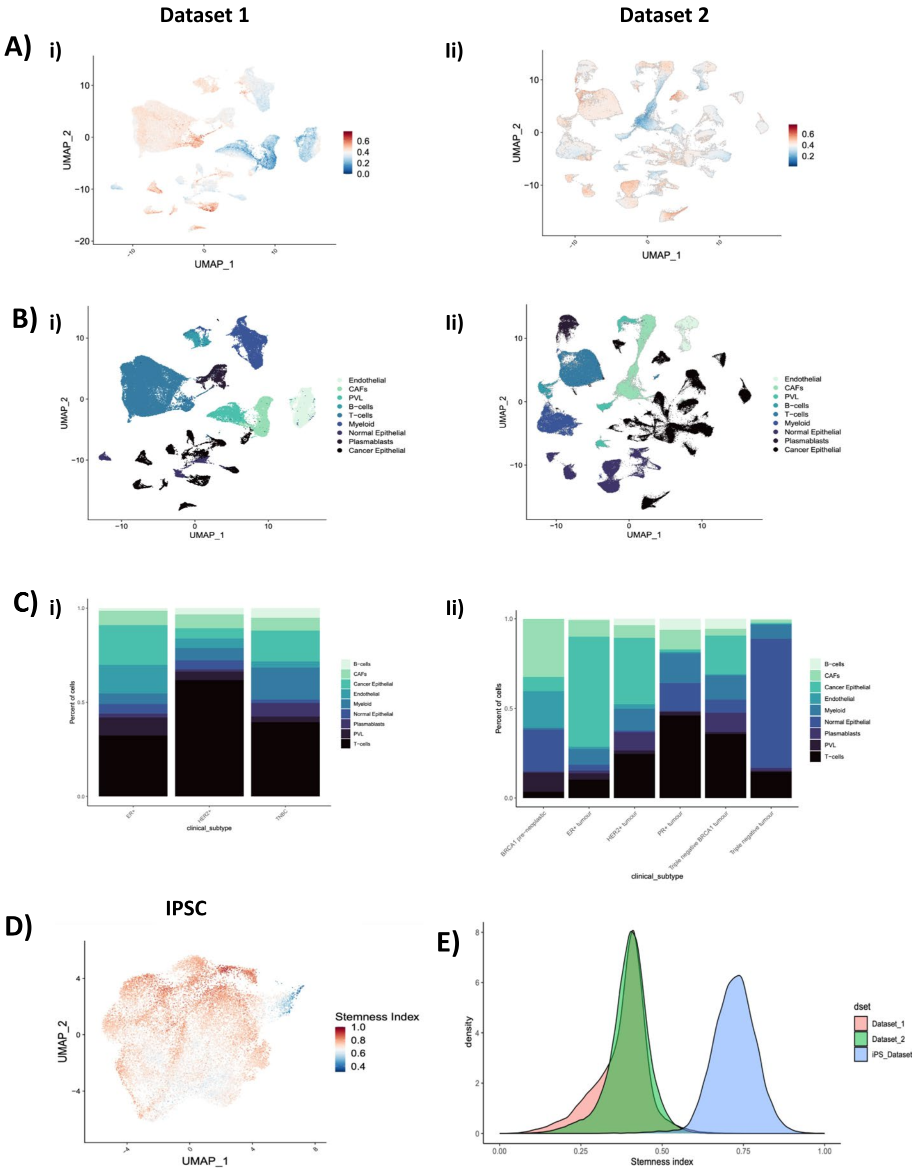

Fig. 2 Stemness model using single cell datasets. **A** UMAP plot of two independent single cell breast cancer datasets (i, ii). Gradient (blue to red) indicates low to high stemness. **B** Cell clustering of major cell types across each dataset. Colors denote the following cell types seen endothelial, cancer associated fibroblasts (CAFs) B cell, T cells, myeloid, normal epithelial, plasmablasts and cancer epithelial cells (from lightest to darkest blue) in dataset 1 (i) and dataset (ii). **C** Cell proportion of major cell type clusterings in each breast cancer histotype in both dataset 1 (i) and dataset 2 (ii). Colors align with cell type identification. **D** UMAP of iPS single cell dataset used for stemness model. Gradient (blue to red) indicates low to high stemness. **E** Stemness index distribution of single cell iPS (blue) compared to breast cancer cells in dataset 1(pink) and dataset 2(green)

scores between the high stemness gene lists were higher across datasets than the regular clusters.

To understand the biological and clinical relevance of these genes identified in the DEG analysis of high stemness scoring cells, we used breast cancer tumor gene expression (bulk RNAseq) and clinical information from TCGA (The Cancer Genome Atlas, $n = 902$) [32]. To correlate gene expression with survival status, we performed a cox regression analysis for each gene with overall survival up to 60 months. We identified 94 genes (27.9%) with a significant ($p < 0.05$) correlation with poor survival in cancer basal cells for dataset 1, and 70 genes (22.5%) for dataset 2. In cancer cycling cells, we found 51 genes (21.4% and 19.5%) correlating with poor survival in both datasets (Fig. 4C) and in cycling T-cells we found 37 genes (19.4%) that correlates with poor survival in dataset 1 and 94 genes (21.5%) in dataset 2.

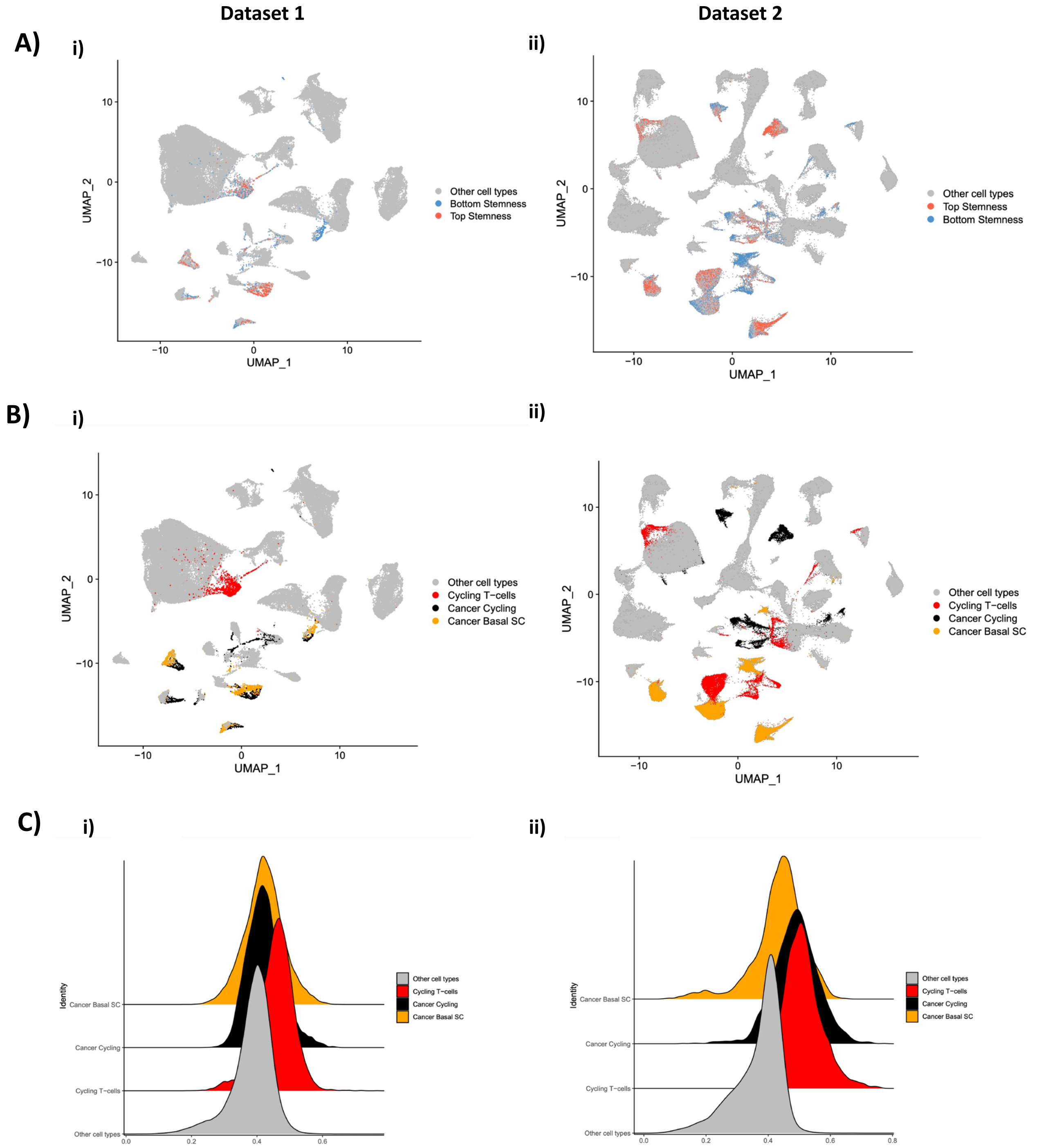

Fig. 3 **A** UMAP plots of stemness model from two breast cancer single cell datasets (i, ii) of the top (red) and bottom (blue) stemness cells. **B** Cancer basal (yellow), cancer cycling (black), and cycling t-cells (red) showing cell types overlapping most top stemness cells in both datasets (i, ii). **C** Stemness distribution in dataset 1 (i) and dataset 2 (ii) of breast cancer single cells with the highest stemness seen in cancer basal (yellow), cancer cycling (black) and cycling t-cells (red) compared to all other cells for both datasets (gray)

We sought to apply this single cell derived stemness model into spatial omics (transcriptomic and proteomics) datasets. Breast cancer tumor samples profiled across five different spatial omics assays were analyzed: one spot based whole-transcriptomics sequencing assay; two high plex in situ probe-based imaging platforms

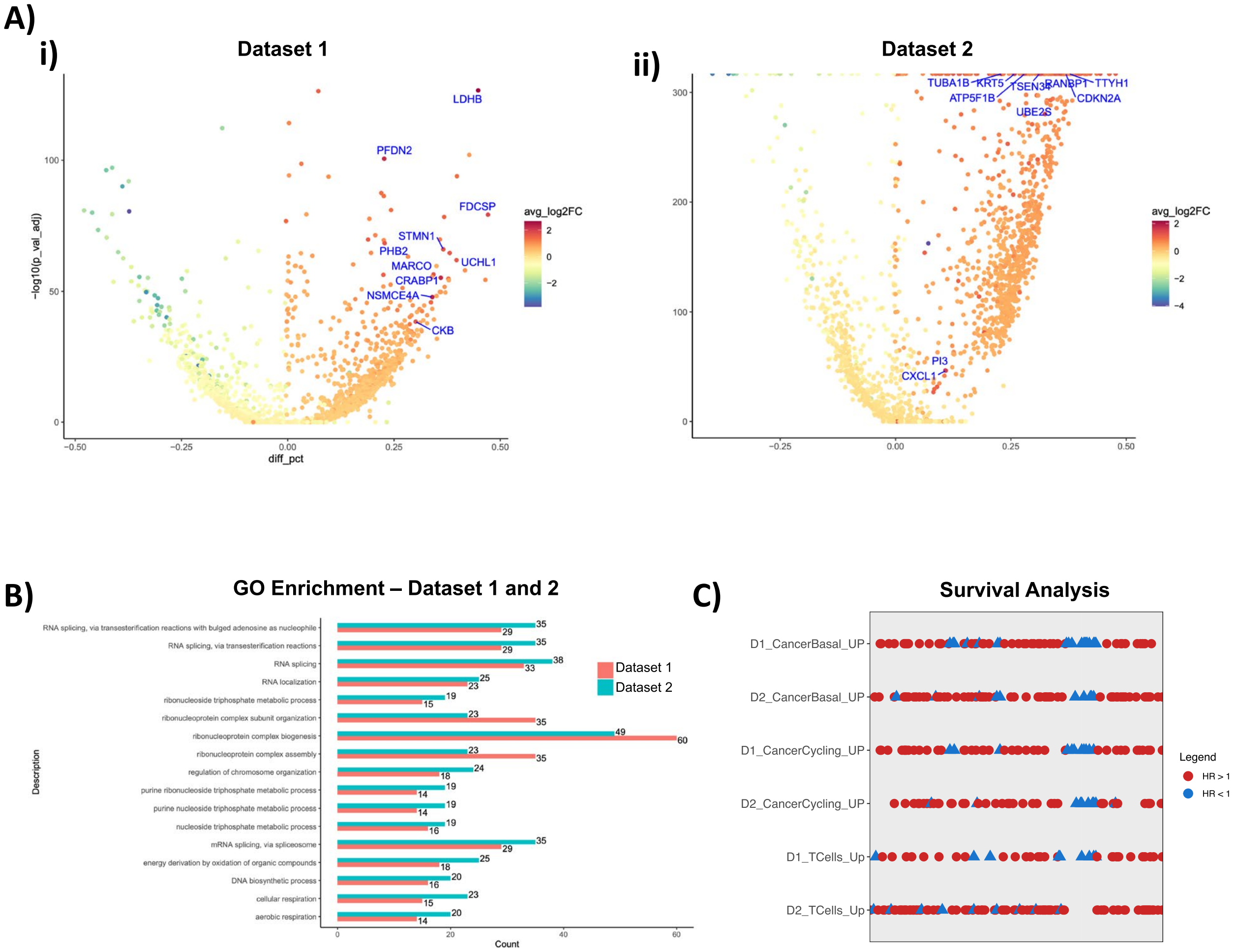

Fig. 4 Gene expression analysis of stemness model **A**) Volcano plots of top and bottom stemness cancer basal cells from dataset 1 (i) and dataset 2 (ii) respectively. Gradient depicts negative logFC (blue) to positive logFC (red). Y axis represents the statistical significance and X axis the different percentage of cells expressing a given gene. Positive different percentage means that top stemness cells have more cells expressing that gene. **B** GO enrichment analysis of the top regulated genes in cancer basal top stemness cells for dataset 1 (red) and dataset 2 (blue). The number of DEG is plotted on the X axis and GO categories the genes are represented in are plotted on the y axis. **C** Survival analysis from TCGA breast cancer cohort of up regulated genes on top stemness cell type across datasets 1 and 2. Rows represent different gene lists according to dataset and cell type. Hazard ratio > 1 in blue, < 1 in red (p-value < 0.05)

and two high plex protein-based platforms. The spatial whole transcriptomics assay (Visium) used spatially barcoded RNAseq to generate high-resolution maps of gene expression patterns in intact tissue sections. We obtained 17 distinct clusters from 4,665 spots and 47,774 features (Fig. 5A). We used 12,342 features overlapping with the original model to generate new weights to perform correlation and generate stemness scores for each spot. The average stemness score of 0.52 for all spots across the whole sample. For spots demarcated by pathologist classification as invasive cancer, the ML model had the highest average of all classified spots presented of 0.55.

We next determined if the ML model was applicable to RNA probe based assays that are not the entire transcriptome. Using a publicly available dataset, we obtained 27 clusters from 889,765 cells profiled from a breast cancer panel (280 genes) on a high plex in situ imaging Xenium platform (Fig. 5B). We predicted stemness scores for each segmented cell as described in the methods section. The average stemness score was 0.546. We used the scRNAseq dataset 1 to predict cell types based on 1) the cell type scoring using the biomarkers present in the antibody panel and 2) the corresponding cell type previously annotated. We created a scoring approach to classify each cluster based on the expression of these markers and labeled by the highest scoring cell type. To interrogate the reliability and reproducibility of the ML stemness model across platform types,

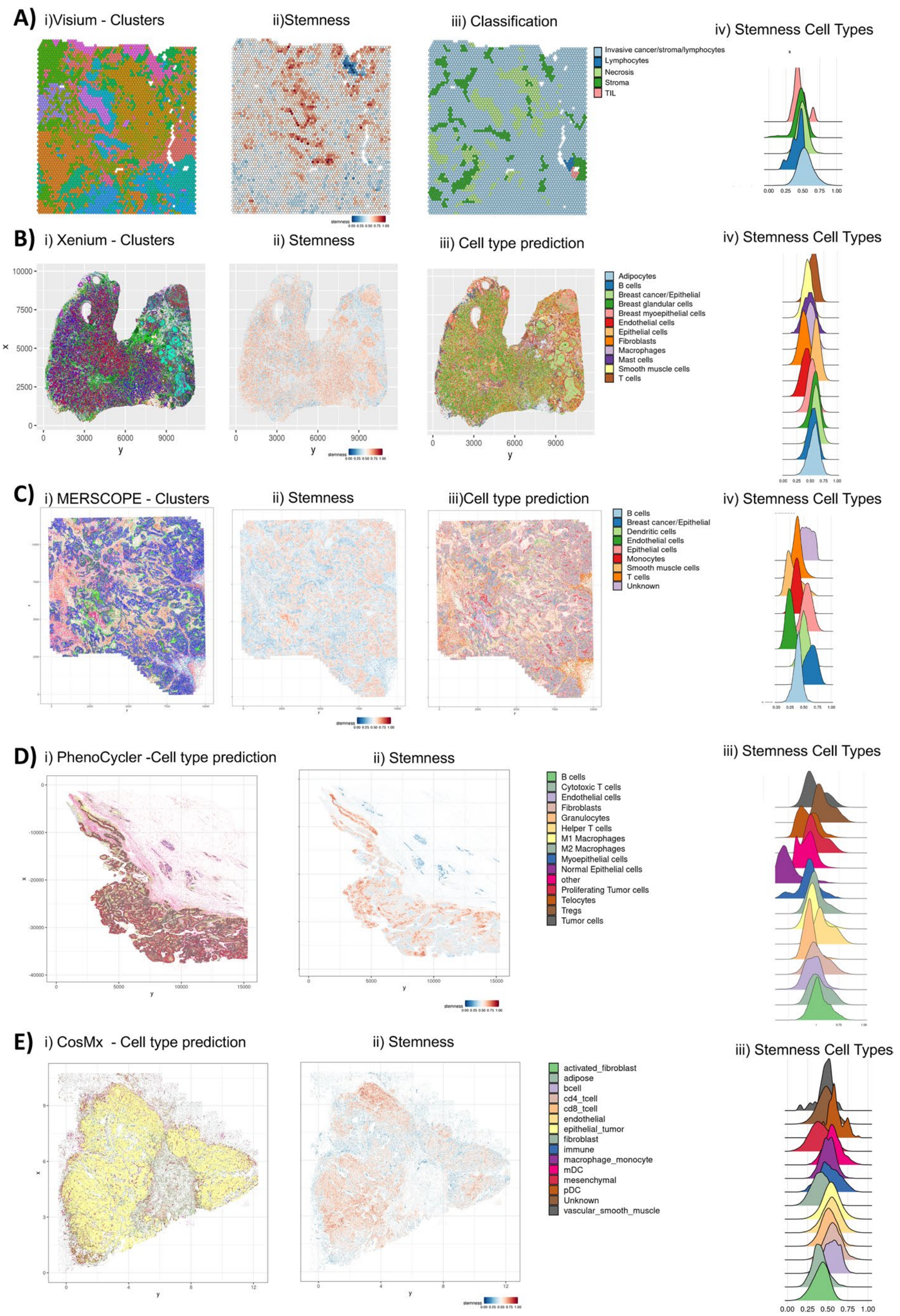

Fig. 5 Spatial analysis of stemness model **A**) Each spot from the section placed on Visium barcoded array is represented with i) a cluster projection ii) stemness score pathology classification for each spot of the Visium sample and distribution plots of stemness score across classified regions. **B** Cluster projection, stemness score, cell type prediction for each segmented cell's centroid location and distribution plot of stemness score per annotated cell types of Xenium sample. **C** Cluster projection, stemness score, cell type annotation for each segmented cell's centroid location and distribution plot of stemness score per annotated cell types of the MERSCOPE sample. **D** Cell type annotation, stemness score for each segmented cell's centroid and distribution plot of stemness score per annotated cell type of the PhenoCycler sample. **E** Cell type annotation, stemness score for each segmented cell's centroid and distribution plot of stemness score per annotated cell type of the CosMX sample

we tested this model in a publicly obtained dataset from another in situ probe based imaging platform. From the breast câncer sample using MERSCOPE technology, we obtained 18 clusters from 710,073 cells profiled with 500 genes (Fig. 5C). We applied the same reference to score clusters and assign cell type annotation in another in situ platform dataset (Xenium). The average stemness score was 0.474 for all cells, and the two highest scoring cell populations were in breast cancer cells and epithelial cells and the lowest cell population stemness score was in fibroblast cells (Fig. 5C).

The ML model showed robustness across three distinct spatial transcriptomics methodologies. We next tested its application to high plex proteomic based imaging platforms. A breast cancer sample was analyzed using an immuno-oncology specific 45 plex antibody panel (Phenocycler Fusion Discovery panel) and 62 plex panel (CosMX). From the Phenocycler Fusion sample, 1,040,049 cells were seen across 56 clusters. Based on marker correlation, we annotated 14 different cell types (e.g. cycling cancer cells), similarly seen in both the single cell and spatial transcriptomic data. The average stemness score was 0.516 and the highest scoring cell populations were tumor cells (0.52), Tregs (0.604), proliferating tumor cells (0.525) and helper T cells (0.603). Consistently, normal epithelial cells had the lowest score of all cell types, with an average of 0.246 (Fig. 5D). For the CosMx sample, 472,763 cells were identified across 14 cell types, the average stemness score was 0.51 and the highest stemness scoring cell types were epithelial tumor (0.542) and endothelial (0.547) (Fig. 5E). Overall, the average stemness score for all different spatial omics assays were similar, and the cell types presenting a higher stemness profile were also similar.

Discussion

Our study demonstrates the utility of ML modeling for agnostic cell type annotation in both scRNAseq and spatial omic data [33]. This ML model overcomes the computational hurdles and challenges of analyzing millions of cells against hundreds of features [11]. An advantage to this ML model for scRNAseq and spatial omics analysis also include faster processing time than laborious manual cell phenotyping which is exponentially longer in spatial omics analysis. We demonstrate its robustness across a variety of spatial omics technologies including, sequencing based, RNA/in situ probe based and protein based platforms. Using our stemness scRNAseq and spatial model approach to breast cancer, we identified cell types within existing cell clusters attributable to cancer stemness which were not previously described. This ML approach can identify scRNAseq features that are predictive of clinical outcome, which can help in patient stratification and personalized medicine.

Breast cancer tumor biology is concordant with the stemness predictions of the model [14, 34] such that cell types known to be involved in more aggressive disease have increased stemness (e.g. proliferating tumor cells). Similarly, cells that are responsible for effective immune response (e.g. helper T cells) are decreased in tumors with high stem. This method also addresses the problems of proper cell annotation such that cells regardless of their annotated cell identity can be properly identified based on their level of stemness. Normal epithelial cells without annotation were identified with no levels of stem, while proliferating tumor cells had high levels of stemness. This model can be more broadly applied to various diseases, treatment conditions and genetic differences compared to existing manual cell curation methods.

The interaction of breast cancer cells with their environment relies on communication between local cues, cancer cells, cancer stem cells, immune cells and stromal cells within the tissue [13]. This TME has clearly been linked to prognosis, recurrence, treatment resistance and outcome [13]. For example, breast cancer treatment targets the proliferative advantage breast cancer cells have compared to adjacent normal cells. Pathological diagnosis and treatment assessment is based on the subtype that the breast cancer tumor cells fall into based on 1) clinical staging, 2) histology, 3) biomarker expression and 4) molecular profiling by gene expression [35–37]. This study demonstrates the TME in breast cancer based on a variety of spatial data and technologies is very consistent with pathological anatomical features in relation to stemness. We present for the first time a ML model that can predict stemness while preserving the spatial context of cell interactions within the tissue. This model highlights the importance of understanding where in the tumor, more aggressive, stem-like cells are situated and the TME that surrounds them. This model preserves single cell identities and can incorporate existing single cell data to inform which cell types are more stem-like, highlighting the adaptability to existing single cell datasets and the integration into emerging spatial omic datasets. This model is also agnostic to which single cell or spatial technology is used to generate the input datasets.

Because the TME is a critical aspect of breast cancer management, this study lends to the importance of harnessing cell identities, composition and interactions from single cell and spatial omics data to improve our understanding of clinical outcomes and treatment. The ease of use and speed of this model makes it highly attractive to generate stemness prediction and discover pathogenicity of the tumor when confronted with analyzing millions of cells at a time. Overall, ML-based methods can provide a powerful toolset for scRNAseq and spatial omics analysis, which can help accelerate our understanding of cancer stemness and its clinical relevance.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09722-6.

Supplementary file 1: Supplemental Figure 1. (A) Stemness on bootstrapped high and poor quality breast cancer samples on dataset 1 (i,ii) and dataset 2 (iii,iv). Each figure represents one sample and the box plot indicates the stemness distribution of a subset of cells that grows in number to the right untill the total cells for that given sample. (B) Rank barplot of number of cells by patient samples for dataset 1 (i) and dataset 2 (ii). **Supplemental Figure 2.** Gene expression analysis of stemness model. Volcano plots of top stemness A) cancer cycling cells from dataset 1 (i) and dataset 2 (ii) respectively. Gradient depicts negative logFC (blue) to positive logFC (red). Y axis represents the statistical significance and X axis the different percentage of cells expressing a given gene. Positive different percentage means that top stemnes cells have more cells expressing that gene. B) GO enrichment analysis of the top regulated genes in cancer cycling cells for dataset 1 (red) and dataset 2 (blue). The number of DEG is plotted on the X axis and GO categories the genes are represented in are plotted on the y axis. C) Volcano plots of top stemness cycling T cells from dataset 1 (i) and dataset 2 (ii) respectively. Gradient depicts negative logFC (blue) to positive logFC (red). D) GO enrichment analysis comparison from the number of up regulated genes in cycling t-cells dataset 1 (red) and dataset 2 (blue). E) Gene similarity score of upregulated genes from cancer basal high stemness cells vs cluster markers genes. On the left, dataset 1 high stemness genes are compared to gene markers from dataset 2. On the right, dataset 2 high stemness genes are compared to gene markers from dataset 1. Gradient depicts less gene similarity (blue) to greater gene similarity (red). **Supplemental Figure 3.** (A) Heatmap of gene similarity analysis of cluster markers in dataset 1 compared to dataset 2 for cancer basal DEGs and (B) heatmap of gene similarity analysis of cluster markers in dataset 2 compared to dataset 1 for the same cell type. **Supplemetal Figure 4.** QC Metrics of Akoya PhenoCycler and Nanostring CosMx samples. A) Violin plot of fluorescence intensity signal (nCount_Akoya) showing the number of proteins detected per cell (nFeature_Akoya). B) Scatter plot of number of fluorescence intensity (x-axis).

Acknowledgements
We acknowledges our funding sources. Thank you for the the publicly available datasets from Vizgen and 10X Genomics and the datasets provided by Akoya and Nanostring. We thank our collaborators and their patients for their time and precious samples.

Authors' contributions
All authors contributed.

Funding
São Paulo Research Foundation (FAPESP), 2018/00583–0 (T.M.M.), 2021/00283–9, 2022/06678–8 (M.M.), OCRA Grant (J.T.P) Chan Zuckerberg Institute Grant (OS00001235) (S.H.L.G and J.T.P.).

Availability of data and materials
Expression Omnibus database (GEO; GSE176078 and GSE161529) and ArrayExpress (E-MTAB-6524).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 June 2023 Accepted: 7 October 2023
Published online: 28 November 2023

References

1. Li X, Wang C-Y. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021;13:36. https://doi.org/10.1038/s41368-021-00146-0.
2. Hegenbarth J-C, Lezzoche G, De Windt LJ, Stoll M. Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. Front Mol Med. 2022;2. https://doi.org/10.3389/fmmed.2022.839338.
3. Jindal A, Gupta P, Jayadeva, Sengupta D. Discovery of rare cells from voluminous single cell expression data. Nat Commun. 2018;9:4719. https://doi.org/10.1038/s41467-018-07234-6.
4. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20:194. https://doi.org/10.1186/s13059-019-1795-z.
5. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525:251–5. https://doi.org/10.1038/nature14966.
6. Liu J, Fan Z, Zhao W, Zhou X. Machine Intelligence in Single-Cell Data Analysis: Advances and New Challenges. Front Genet. 2021;12: 655536. https://doi.org/10.3389/fgene.2021.655536.
7. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21:31. https://doi.org/10.1186/s13059-020-1926-6.
8. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, Baglaenko Y, Brenner M, Loh P-R, Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96. https://doi.org/10.1038/s41592-019-0619-0.
9. Liu J, Gao C, Sodicoff J, Kozareva V, Macosko EZ, Welch JD. Jointly defining cell types from multiple single-cell datasets using LIGER. Nat Protoc. 2020;15:3632–62. https://doi.org/10.1038/s41596-020-0391-8.
10. Park J, Kim J, Lewy T, Rice CM, Elemento O, Rendeiro AF, Mason CE. Spatial omics technologies at multimodal and single cell/subcellular level. Genome Biol. 2022;23:256. https://doi.org/10.1186/s13059-022-02824-6.
11. Atta L, Fan J. Computational challenges and opportunities in spatially resolved transcriptomic data analysis. Nat Commun. 2021;12:5283. https://doi.org/10.1038/s41467-021-25557-9.
12. Hamamoto R, Komatsu M, Takasawa K, Asada K, Kaneko S. Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine. Biomolecules. 2019;10. https://doi.org/10.3390/biom10010062.
13. Li JJ, Tsang JY, Tse GM. Tumor Microenvironment in Breast Cancer-Updates on Therapeutic Implications and Pathologic Assessment. Cancers (Basel). 2021;13. https://doi.org/10.3390/cancers13164233.
14. Song K, Farzaneh M. Signaling pathways governing breast cancer stem cells behavior. Stem Cell Res Ther. 2021;12:245. https://doi.org/10.1186/s13287-021-02321-w.
15. Asada K, Takasawa K, Machino H, Takahashi S, Shinkai N, Bolatkan A, Kobayashi K, Komatsu M, Kaneko S, Okamoto K, et al. Single-Cell Analysis Using Machine Learning Techniques and Its Application to Medical Research. Biomedicines. 2021;9. https://doi.org/10.3390/biomedicines9111513.
16. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. Cell. 2018;173:338-354.e15. https://doi.org/10.1016/j.cell.2018.03.034.
17. Zhang L, Liu Z, Zhu J. In silico screening using bulk and single-cell RNA-seq data identifies RIMS2 as a prognostic marker in basal-like breast cancer: A retrospective study. Medicine (Baltimore). 2021;100: e25414. https://doi.org/10.1097/MD.0000000000025414.
18. Zhang Z, Wang Z-X, Chen Y-X, Wu H-X, Yin L, Zhao Q, Luo H-Y, Zeng Z-L, Qiu M-Z, Xu R-H. Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. Genome Med. 2022;14:45. https://doi.org/10.1186/s13073-022-01050-w.

19. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53:1334–47. https://doi.org/10.1038/s41588-021-00911-1.
20. Pal B, Chen Y, Vaillant F, Capaldo BD, Joyce R, Song X, Bryant VL, Penington JS, Di Stefano L, Tubau Ribera N, et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. EMBO J. 2021;40:e107333. https://doi.org/10.15252/embj.2020107333.
21. Daniszewski M, Nguyen Q, Chy HS, Singh V, Crombie DE, Kulkarni T, Liang HH, Sivakumaran P, Lidgerwood GE, Hernández D, et al. Single-Cell Profiling Identifies Key Pathways Expressed by iPSCs Cultured in Different Commercial Media. iScience. 2018;7:30–39. https://doi.org/10.1016/j.isci.2018.08.016.
22. Janesick A, Shelansky R, Gottscho A, Wagner F, Rouault M, Beliakoff G, Faria de Oliveira M, Kohlway A, Abousoud J, Morrison C, et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. BioRxiv. 2022. https://doi.org/10.1101/2022.10.06.510405.
23. Vizgen. MERSCOPE$^{\mathrm{TM}}$ FFPE Sample Prep Solution. 2022. https://info.vizgen.com/merscope-ffpe-access. Accessed 25 Aug 2023.
24. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296. https://doi.org/10.1186/s13059-019-1874-1.
25. Cable DM, Murray E, Zou LS, Goeva A, Macosko EZ, Chen F, Irizarry RA. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40:517–26. https://doi.org/10.1038/s41587-021-00830-w.
26. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141.
27. Harrower M, Brewer CA. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. The British Cartographic Society 2003. Cartographic J. 2003;40(1):27–37. https://www.cs.rpi.edu/~cutler/classes/visualization/S18/papers/colorbrewer.pdf.
28. Sievert, C. plotly: Create Interactive Web Graphics via "plotly.js" (CRAN). 2021.
29. Mishra D, Banerjee D. Lactate Dehydrogenases as Metabolic Links between Tumor and Stroma in the Tumor Microenvironment. Cancers (Basel). 2019;11. https://doi.org/10.3390/cancers11060750.
30. Yang C, Yu H, Chen R, Tao K, Jian L, Peng M, Li X, Liu M, Liu S. CXCL1 stimulates migration and invasion in ER negative breast cancer cells via activation of the ERK/MMP2/9 signaling axis. Int J Oncol. 2019;55:684–96. https://doi.org/10.3892/ijo.2019.4840.
31. Yu G. Gene ontology semantic similarity analysis using gosemsim. Methods Mol Biol. 2020;2117:207–15. https://doi.org/10.1007/978-1-0716-0301-7_11.
32. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20. https://doi.org/10.1038/ng.2764.
33. Yuan G-C, Cai L, Elowitz M, Enver T, Fan G, Guo G, Irizarry R, Kharchenko P, Kim J, Orkin S, et al. Challenges and emerging directions in single-cell analysis. Genome Biol. 2017;18:84. https://doi.org/10.1186/s13059-017-1218-y.
34. Zhang X, Powell K, Li L. Breast cancer stem cells: biomarkers, identification and isolation methods, regulating mechanisms, cellular origin, and beyond. Cancers (Basel). 2020;12. https://doi.org/10.3390/cancers12123765.
35. Marra A, Trapani D, Viale G, Criscitiello C, Curigliano G. Practical classification of triple-negative breast cancer: intratumoral heterogeneity, mechanisms of drug resistance, and novel therapies. NPJ Breast Cancer. 2020;6:54. https://doi.org/10.1038/s41523-020-00197-2.
36. Eliyatkın N, Yalçın E, Zengel B, Aktaş S, Vardar E. Molecular Classification of Breast Carcinoma: From Traditional, Old-Fashioned Way to A New Age, and A New Way. J Breast Health. 2015;2013(11):59–66. https://doi.org/10.5152/tjbh.2015.1669.
37. Tsang JYS, Tse GM. Molecular classification of breast cancer. Adv Anat Pathol. 2020;27:27–35. https://doi.org/10.1097/PAP.0000000000000232.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC