SC2MeNetDrug: A computational tool to uncover inter-cell signaling targets and identify relevant drugs based on single cell RNA-seq data

Jiarui Feng[1,2], S. Peter Goedegebuure[5,6], Amanda Zeng[1], Ye Bi[4,6], Ting Wang[5], Philip Payne[1], Li Ding[5,6], David DeNardo[5,6], William Hawkins[4,6], Ryan C. Fields[4,6], Fuhai Li[1,7#]

[1]Institute for Informatics (I2), Washington University School of Medicine, [2] Department of Computer Science and Engineering, [3]Department of Neurology, [4]Department of medicine, [5]Department of Surgery, [6]Siteman Cancer Center, [7]Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA. [#]Correspondence, Email: Fuhai.Li@wustl.edu

Abstract

Single cell RNA sequencing (scRNA-seq) is a powerful technology to investigate the transcriptional programs in stromal, immune and tumor cells or neuron cells within the tumor or Alzheimer's Disease (AD) brain microenvironment (ME) or niche. Cell-cell communications within ME play important roles in disease progression and immunotherapy response, and are novel and critical therapeutic targets. Though many tools of scRNA-seq analysis have been developed to investigate the heterogeneity and sub-populations of cells, few were designed for uncovering cell-cell communications of ME and predict the potentially effective drugs to inhibit the communications. Moreover, the data analysis processes of discovering signaling communication networks and effective drugs using scRNA-seq data are complex and involving a set of critical analysis processes and external supportive data resources, which are difficult for researchers who have no strong computational background and training in scRNA-seq data analysis. To address

these challenges, in this study, we developed a novel computational tool, SC2MeNetDrug (https://fuhaililab.github.io/sc2MeNetDrug/). It was specifically designed using scRNA-seq data to identify cell types within MEs, uncover the dysfunctional signaling pathways within individual cell types, inter-cell signaling communications, and predict effective drugs that can potentially disrupt cell-cell signaling communications. SC2MeNetDrug provided a user-friendly graphical user interface to encapsulate the data analysis modules, which can facilitate the scRNA-seq data based-discovery of novel inter-cell signaling communications and novel therapeutic regimens.

Introduction

Tumor-stroma communication within the tumor microenvironment (TME) plays an important role in tumor development and responses to both conventional- and immune-based therapies. For example, immunotherapy in pancreatic cancer treatment has not been successful[1]. One possible cause of immunotherapy resistance is the abundance of stromal cells and tumor signaling communications in Pancreatic ductal adenocarcinoma (PDAC) tumor microenvironments[1]. Such immunosuppressive cells include tumor associated macrophages (TAMs), myeloid-derived suppressor cells (MDSCs), regulatory T cells (Tregs), as well as cancer associated fibroblasts (CAFs)[1,2,3,4,5,6]. Moreover, CAFs were recently reported to be able to regulate the invasive epithelial-to-mesenchymal transition (EMT) and proliferative (PRO) phenotypes of PDAC[7]. This indicates that stroma-tumor communication in PDAC tumor microenvironments play a critical role in immunotherapy resistance. Thus, stroma-tumor signaling communications are potential targets to improve drug or immunotherapy response in cancer treatment. The inhibition of signaling communication between TAMs and PDAC cells via the Colony Stimulating Factor 1 (CSF1) (ligand secreted by PDAC) and CSF1R (receptor on TAM) can reprogram

TAMs, and the synergistic combination of TAM-tumor signaling inhibition with the immune checkpoint blockade[8] can improve the immunotherapy response. In another study, the inhibition of signaling communication between CAF and PDAC via CXCL12 (ligand secreted by CAF) and CXCR4 (receptor on PDAC) was shown to improve immunotherapy response[5].

Recent advances in single cell RNA sequencing (scRNA-seq) create a powerful technology to analyze the genetic and functional heterogeneity of stromal and tumor cells (e.g., TAM, CAF and T cells) within tumor microenvironments[9,10,7]. Though many tools and studies reported to have discovered the heterogeneity and sub-populations of cells, few studies[11] have been designed to investigate the cell-cell communication using sc-RNAseq data. For example, the CCCExplorer[12,13] was first developed for uncovering the potential tumor and stroma cell communication using microarray and bulk RNA data on a small set of curated ligand-receptor interactions. CellPhoneDB[14] provided a repository of ligands, receptors and their interactions using the novel computational ligand-receptor interaction prediction approaches. NicheNet[15] was the latest software tool that integrates the large set of ligand-receptor interactions from CellPhoneDB, and it accepted the pre-analyzed scRNA-seq data. However, the computational modules of inferring the dysfunctional signaling networks, and predict potentially effective drugs inhibiting the dysfunctional signaling networks and cell-cell communications are not available in these tools.

Specifically, compared with the existing tools, novel computational models and tools that solve the following challenges are in high demand to 1) provide an end-to-end model that can take the raw scRNA-seq data as input, analyze, annotate and display the scRNA-seq data, 2) uncover dysfunctional signaling network within individual cells, and uncover complex signaling communications among multiple stromal and tumor cells; 3) identify effective drugs and drug combinations that disrupt the cell-cell communications, like stroma-tumor, to improve the targeted and immunotherapy response. Moreover, 4) a user-

friendly interactive graphical user interface (GUI) is helpful and critical for biomedical researchers because these analyses are highly composite complex and involve a set of computational analysis processes and integration of external supportive data resources that require visualization by non-bioinformatics experts to functionalize the complex data. To resolve the aforementioned challenges, in this study, we developed a novel computational tool: **SC2MeNetDrug** (scRNA-seq based modeling to discover disease microenvironment signaling communication networks and drugs targeting on the cell-cell signaling communications). SC2MeNetDrug provided a user-friendly graphical user interface to encapsulate the data analysis modules, which can facilitate the scRNA-seq data based-discovery of novel inter-cell signaling communications and novel therapeutic regimens.

Results

Overview of SC2MeNetDrug

Fig. 1 summarizes the **SC2MeNetDrug** tool. The input of SC2MeNetDrug is the raw counts of genes from single cell RNA-seq (scRNA-seq) data of different experimental conditions or samples, e.g., normal tissues vs disease tissues. The output of the tool includes the annotation of cell types, dysfunctional signaling networks within individual cells, intercellular signaling communications, and drugs that can potentially inhibit dysfunctional signaling pathways and intercellular signaling communications. Specifically, there are 3 major modules: scRNA-seq pre-analysis module that consists of the quality control, normalization, imputation, dimension reduction, visualization, cell clustering and cell type annotation; and ***iCSC*** (**i**nter-**c**ell **s**ignaling **c**ommunication discovery) module that uncovers the activated signaling pathways and gene ontology (GO) terms within individual cell types, and uncovers the cell-cell signaling communications within the disease ME; and ***dCSC*** (**d**rug prediction for disrupting **c**ell **s**ignaling **c**ommunication) module that

identify and predict the potentially effective drugs, based on drug-target and revere gene signature, to disrupt the cell signaling communications. All the data analysis and modeling were designed in the modular format, which can be upgraded or replaced conveniently to select the best practice models. As an example, we applied the SC2MeNetDrug model to a cohort of pancreatic ductal adenocarcinoma (PDAC) scRNA-seq data, and demonstrate the functionality of the tool. The detailed introduction to the downloading, installation, analysis modules, and examples, as well as the video tutorials for each analysis module were provided at: https://fuhaililab.github.io/sc2MeNetDrug/

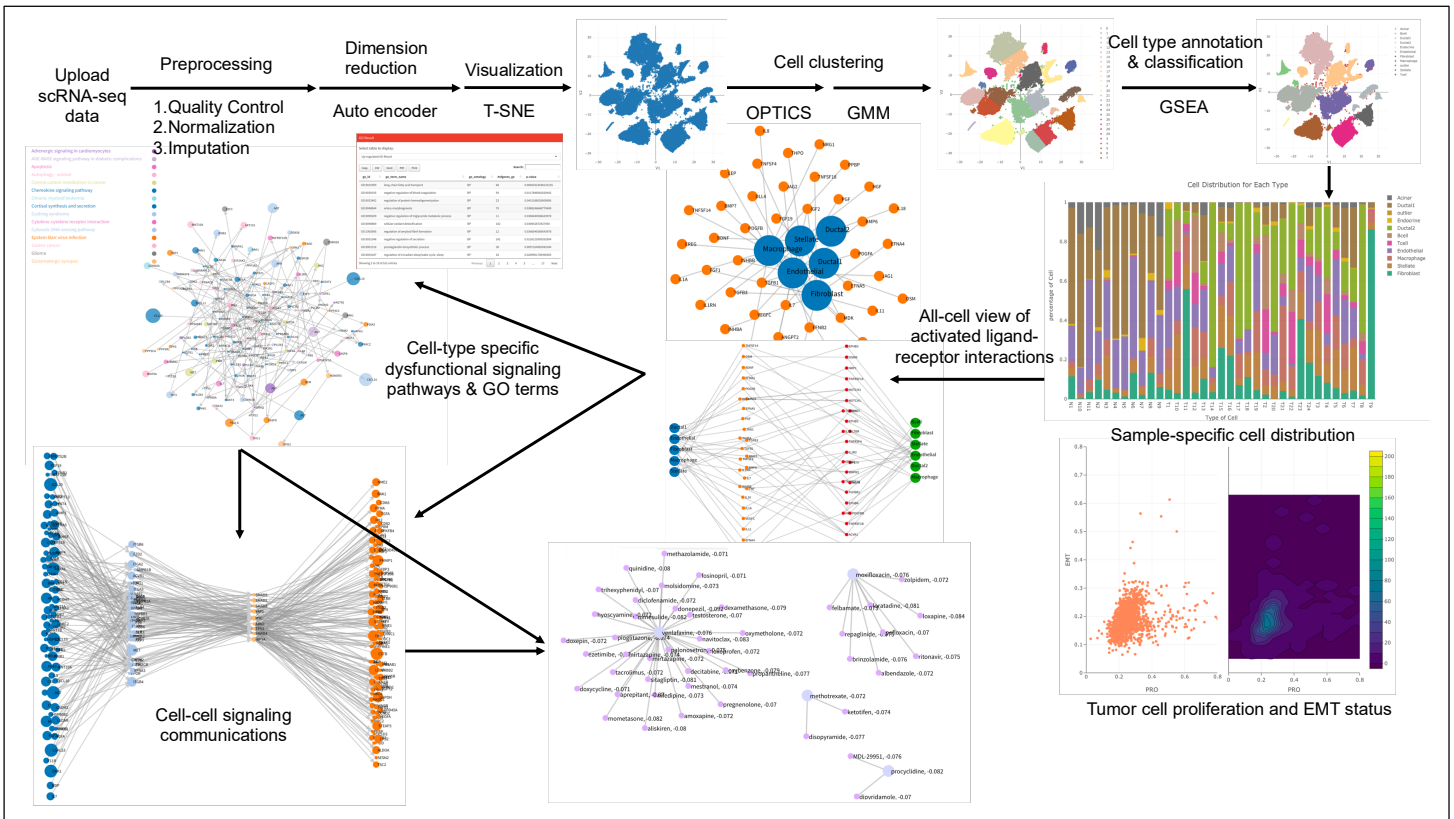

Figure 1: Overview of SC2MeNetDrug.

The scRNA-seq data pre-analysis module

There are many great scRNA-seq tools publicly available[16] for a for different analyses, e.g., quality control, imputation, dimension reduction, clustering and cell type annotation. It is often confusing and hard to select the right tools or best practice pipelines for a given specific project[16], especially for clinical and research investigators without bioinformatics expertise. Moreover, it is not trivial to use and integrate the results derived from these different computational tools. To address this challenge, we implemented the scRNA-seq pre-analysis module, which is a pipeline that includes quality control, normalization, imputation (using the methods in the Seurat package[17]), dimension reduction (using the auto-encoder), t-SNE visualization, clustering (using the OPTICS[18] and gaussian mixture model (GMM) models) and cell type annotation (classification using the gene set enrichment analysis (GSEA) model[19]).

Figure 2: Marker gene sets to annotate diverse cell types.

Both mice and human scRNA-seq data can be analyzed. A large set of biomarker genes were collected[10,20–22] to support different research projects, like cancer cell, immune cells, AD neuron cells (see **Fig. 2**). We will keep updating the marker genes sets. Moreover, we provided a function to enable users to upload new or user-defined marker gene sets. Then the annotation classifiers based on these selected cell types and corresponding marker genes sets will be built automatically for the cell type annotation analysis (see **Fig. 3**). Also, the distribution (percentage) of individual cell types in each sample will be displayed, and the Epithelial–mesenchymal transition (EMT) and PRO (proliferation) scores of each sample will be calculated (see **Fig 4**).

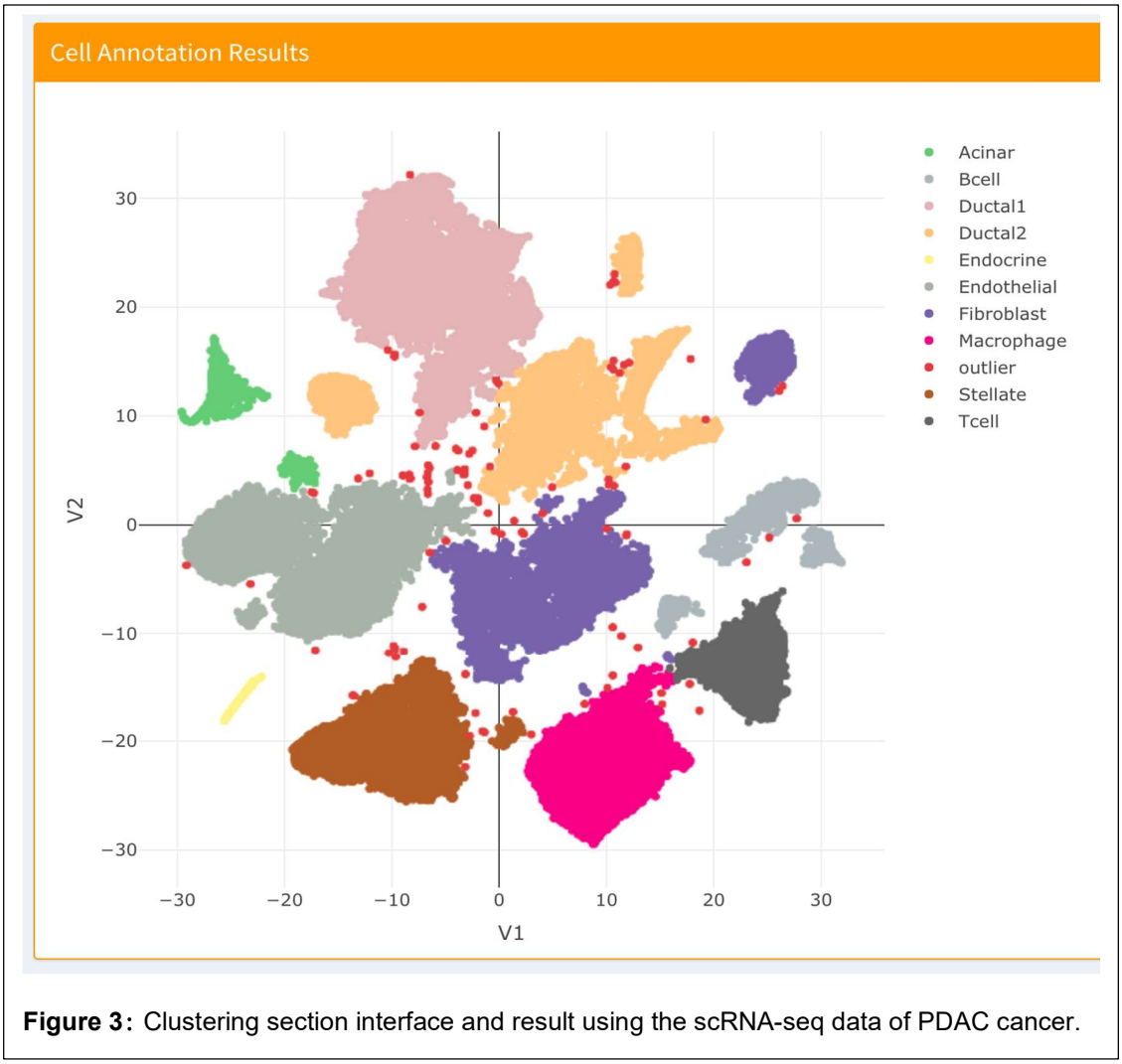

Figure 3: Clustering section interface and result using the scRNA-seq data of PDAC cancer.

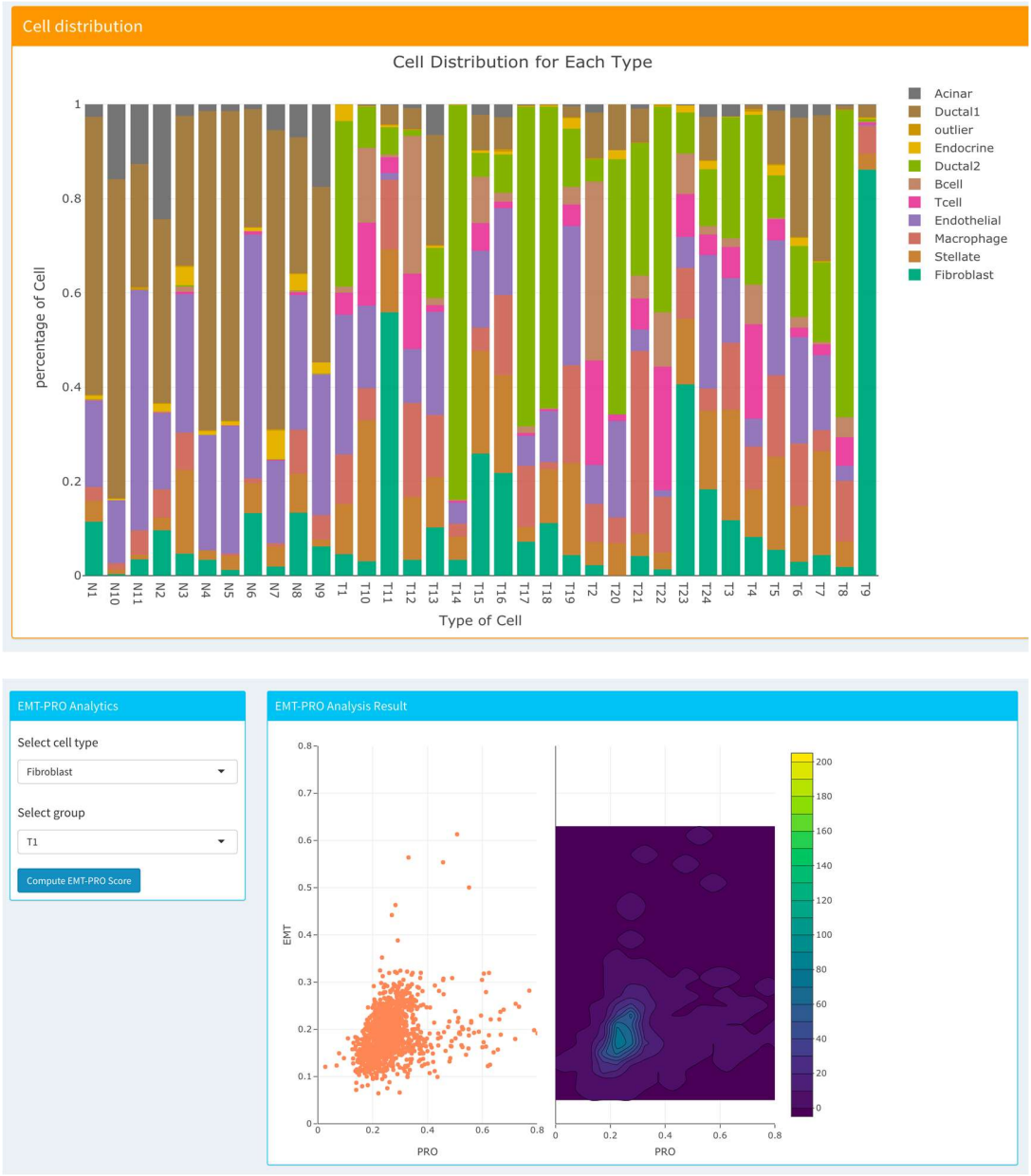

Figure 4: The distribution (percentage) of individual cell types in each sample, and the Epithelial–mesenchymal transition (EMT) and PRO (proliferation) scores of each sample.

Uncover the dysfunctional signaling pathways in individual cell types using the iCSC module

Uncovering the dysfunctional signaling pathways within individual cell types, and cell-cell signaling communications, as novel therapeutic targets, are the highly needed functions. The SC2MeNetDrug provided functions to facilitate the network analysis. Specifically, after the cell type annotation, the differentially expressed genes in each cell type between two

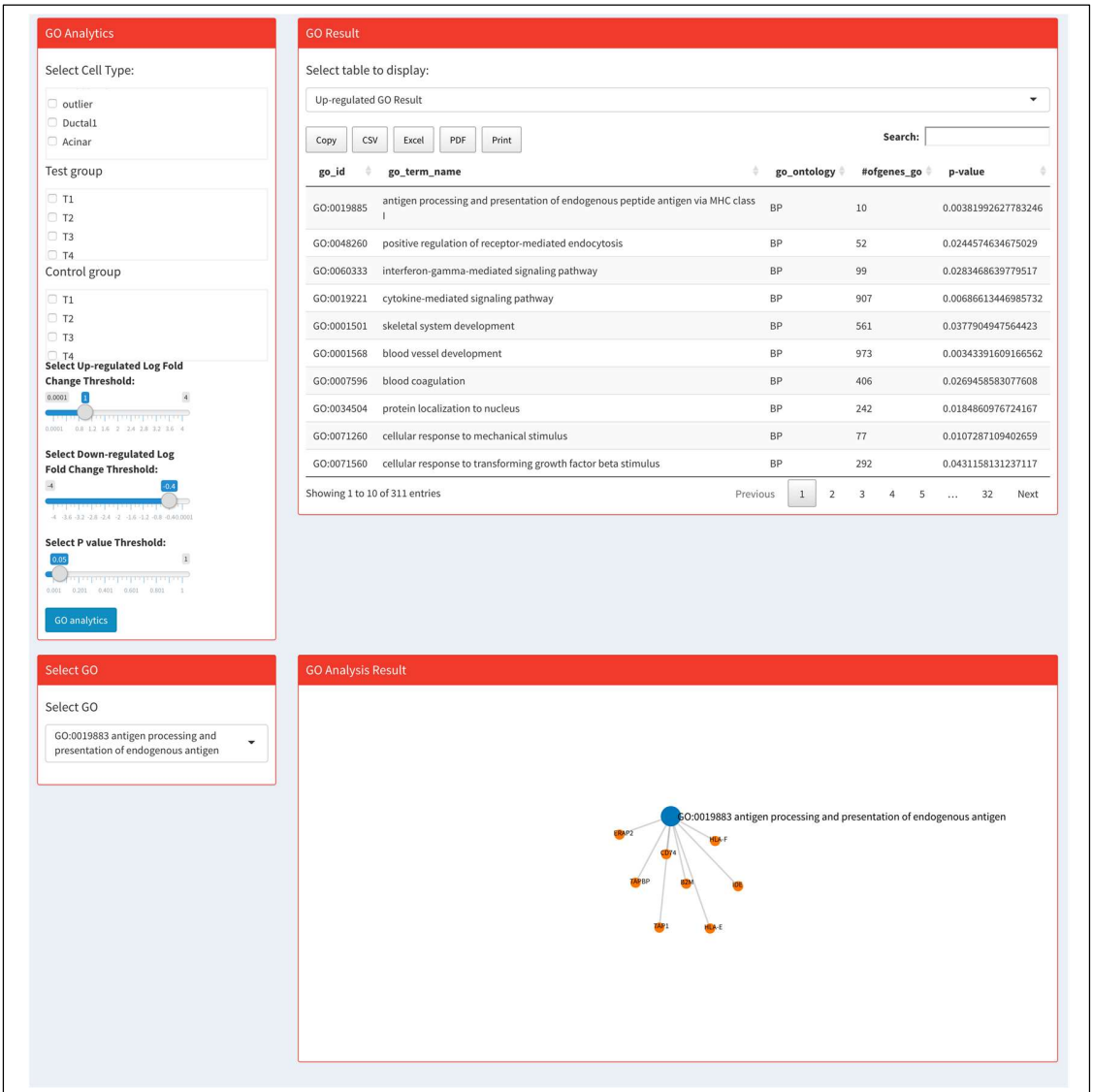

Figure 5: Gene ontology (GO) term analysis to identify the activated biological processes.

different experimental conditions, for example the immunotherapy responder vs non-responder, male vs female, or tumor cells co-cultured with macrophage vs no macrophages, can be calculated. A function was developed to enable the selection of samples and conditions of interest for the differential gene expression analysis. Based on the differentially expressed genes within individual cell types, the gene ontology (GO) enrichment analysis can be identified (see **Fig. 5**).

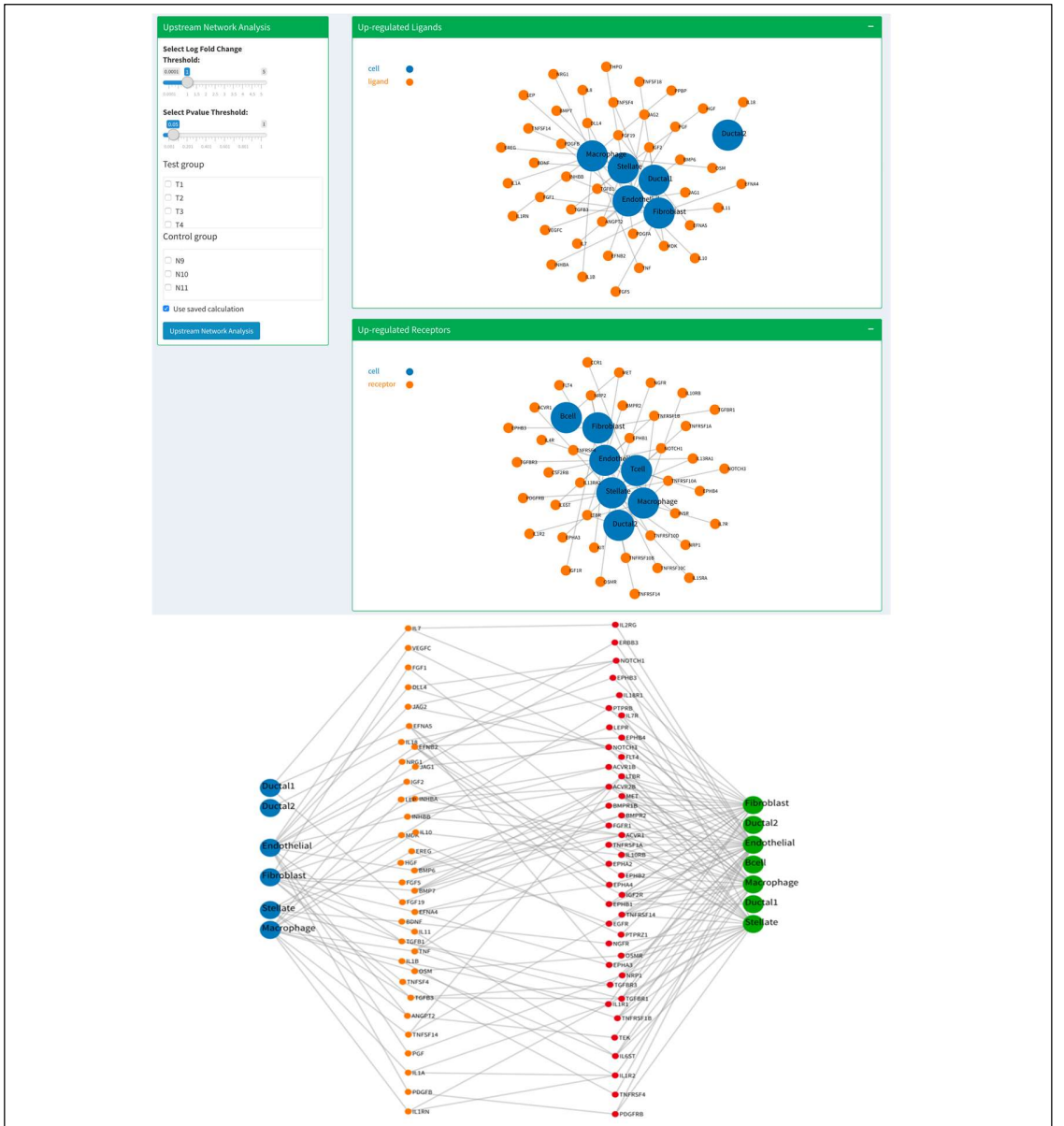

Figure 6: Up-regulated ligands and receptors (upper) and potential ligand-receptor mediated signaling communications among cells within PDAC ME (lower).

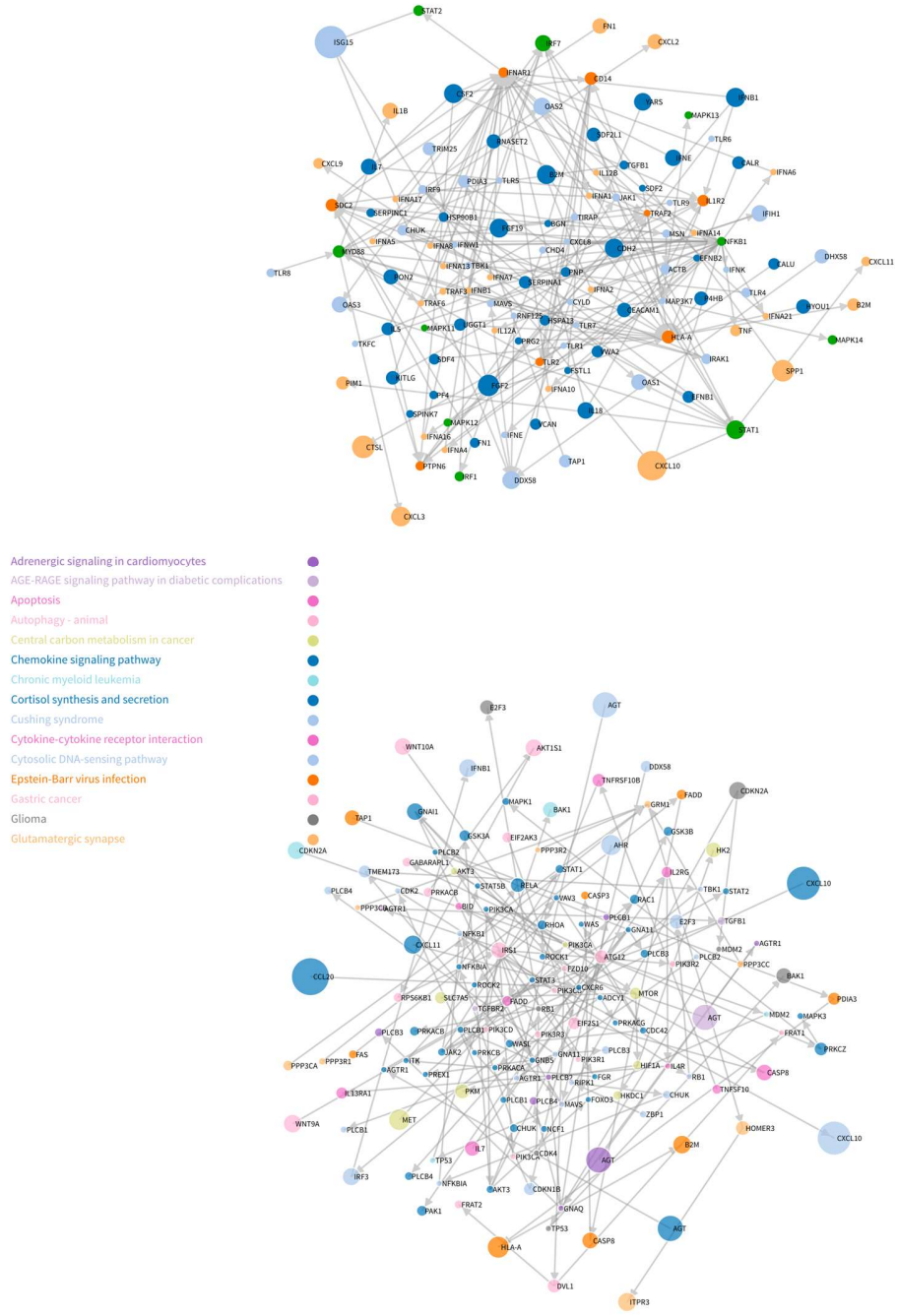

Figure 7: Activated signaling network (upper) between macrophage (secreting ligands) and ductal tumor cells (with activated receptors), and whole activated signaling pathways within tumor cells in a PDAC patient.

expressed (fold change >= *Tc) – up-regulated receptors (*fold change >= *Td)* will be identified as the potential signaling communications inter-cell types within the tumor or other disease ME (see **Fig. 6**). To further investigate the signaling communications among two types of cells, (i.e. macrophage and tumor cells), the network analysis function can be applied to uncover the activated signaling pathways in individual cell types (see **Fig. 7**)

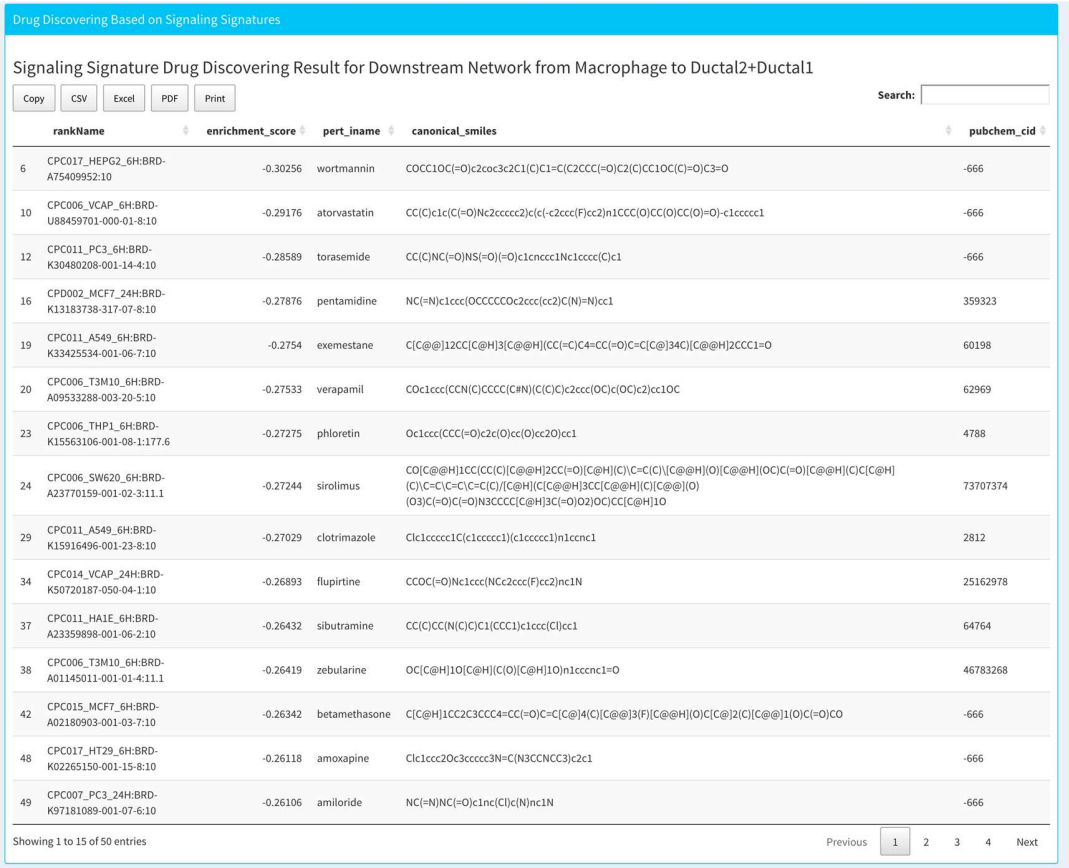

Figure 8: Top drugs that can potentially inhibit the activated signaling communication pathways within individual cell types based on the CMAP data.

Predict drugs inhibiting signaling communications using the dCSC model

To identify drugs that can potentially inhibit the down-stream signaling pathways, the computational model, the ***dCSC*** model was developed, which is designed to integrate the down-stream signaling network, drug-target (derived from DrugBank[23] database) and

reverse gene signature data available from connectivity map (CMAP) database[24]. It is relatively straight forward to identify drugs with targets on the down-stream signaling network based on the drug-target information. For the CMAP data[24], the uncovered down-stream signaling network will be used as a signature of the Gene set enrichment analysis (GSEA) to identify drugs that can potentially inhibit the expression of genes in the network (see **Fig. 8)**. To further understand the relationship of the selected drugs, drug clustering based on chemical structures was conducted to identify therapeutics with similar targets or mechanism of action (see **Fig. 9**).

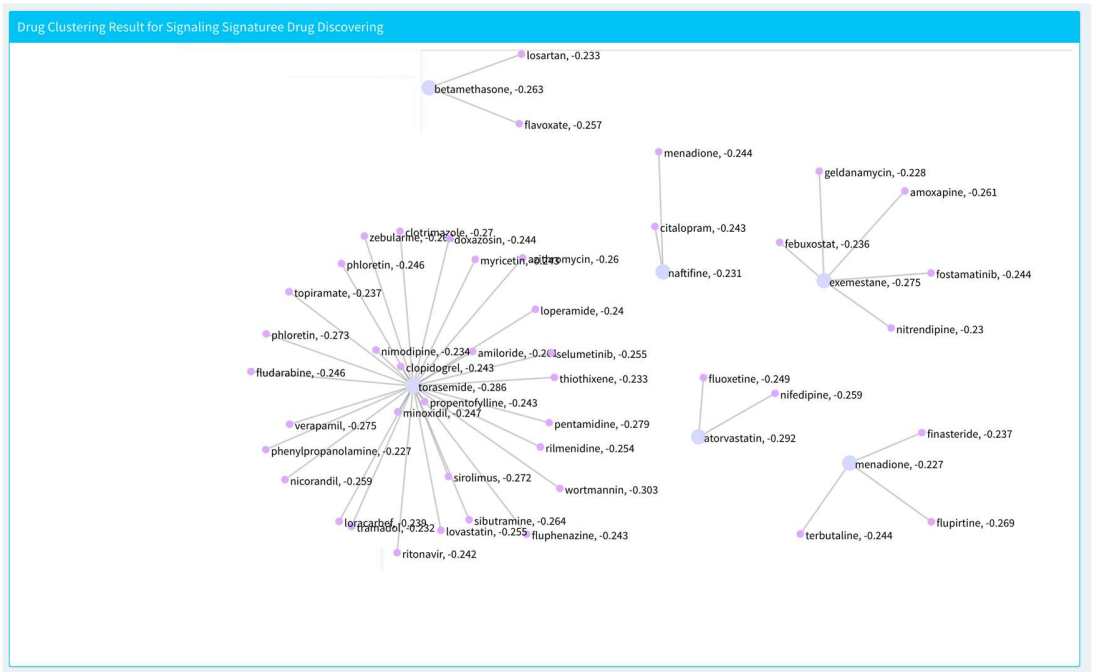

Figure 9: Drug clusters of the top 50 FDA approved drugs based on their chemical structures.

In summary, the **SC2MeNetDrug** tool provides 1) graphical user interface (GUIs), enabling the interactive analysis of scRNA-seq data, 2) models to uncover complex signaling communications among multiple stromal, immune and tumor (or other disease) cells; and 3) models to predict drugs that can potentially disrupt the stroma-tumor

communications to improve the immunotherapy response. It can facilitate the studies of uncovering novel mechanism of inter-cell communications, and identify novel therapies targeting signaling interactions in tumor and disease ME, improve drug and immunotherapy responses in tumor treatment or other complex diseases, like AD. Furthermore, the interface of the tool is designed for seamless use by physicians and translational investigators without formal bioinformatics training in order to functionalize the tool for integration into biomedical research. The tool was also packed the tool with all required dependent libraries in a Docker container, which can be directly used without installing any additional libraries.

Discussion

Single cell RNA sequencing (scRNA-seq) is a powerful technology to investigate the transcriptional programs in stromal, immune and tumor cells or neuron cells within tumor or brain microenvironment (ME) or niche. Cell-cell interactions and communications within ME play important roles in disease progression and immunotherapy response, and are novel and critical therapeutic targets. However, it is challenging, for many researchers without solid training in computational data analysis and scRNA-seq data analysis, because the data analysis pipelines usually consist of diverse and complex analysis modules, and the integrative analysis of diverse and heterogenous external data resources. There is a lack of easy-use tools with complete and integrative computational modules for uncovering cell-cell communications of ME and predict the potentially effective drugs to inhibit the communications, although many tools of scRNA-seq analysis have been developed to investigate the heterogeneity and sub-populations of cells. In this study, we developed a novel computational tool, SC2MeNetDrug (https://fuhaililab.github.io/sc2MeNetDrug/) to address these challenges. Specifically, the advantages of the tool are as follows. First, it is a tool specifically designed for scRNA-seq

data analysis to identify cell types within MEs, uncover the dysfunctional signaling pathways within individual cell types, inter-cell signaling communications, and predict effective drugs that can potentially disrupt cell-cell signaling communications. Second, the analysis modules in the analysis pipelines were separated with pre-designed interfaces. Users can develop and update novel data analysis modules, and easily replace the updated modules back to the data analysis pipeline. In another word, users or scientists with different expertise can conveniently replace user-specific data analysis modules just by following the input and output of individual modules, like network inference, cell type identification, cell clustering, drug prediction, in the data analysis pipeline. Third, it provides a user-friendly graphical user interface (GUI), encapsulating the data analysis modules, which requires no coding and programming and can facilitate the scRNA-seq data analysis in an interactive manner.

Conclusion

In this study, we developed a novel computational tool, SC2MeNetDrug (https://fuhaililab.github.io/sc2MeNetDrug/), which is specifically designed, with user-friendly GUI for interactive scRNA-seq data analysis for the purpose of uncovering cell-cell communications of ME, and predicting the potentially effective drugs to perturb the cell-cell communications within disease ME.

Methodology

PDAC data resource

The PDAC data was downloaded from Genome Sequence Archive under project PRJCA00106310. There was a total of 57530 cell samples and 24003 genes. The data was generated from 24 PDAC tumor samples and 11 control, untreated pancreas samples.

Quality control

Quality control is done in several steps. First, cells with total counts less than the threshold will be removed. The threshold is computed by 0.012*(number of genes). Then, cells with expressed genes (at least 1 read) less than threshold will be removed. The threshold is computed by 0.012*(number of genes). Next, cells with an abnormally high ratio of counts mapping to 34 mitochondrial genes (relative to the total number of genes) will be removed. To be specific, we have soft and hard thresholds to discover abnormal cells. The total mitochondrial expression ratio is computed by:

$$\frac{Total\ counts\ in\ mitochondrially\ encoded\ genes}{Total\ counts\ in\ all\ genes}$$

Then, a soft threshold is applied by using K-means clustering algorithm to cluster cells based on mitochondrial expression ratio, the number of clusters is set to k=2. If one cluster with a lower mean mitochondrial expression ratio has the number of cells larger than 5 times the number of cells in another cluster, we keep the cells in this cluster and remove the others. If both clusters have a mean mitochondrial expression ratio less than 0.02, we will keep all the cells. Otherwise, we will apply a hard threshold. The 98% quantile of mitochondrial expression ratio of the whole dataset is obtained, and if the ratio is larger than 0.09, we set the threshold as 0.09, otherwise, we set the threshold as this ratio. Finally, we remove all the cells that have a mitochondrial expression ratio larger than the threshold. The fourth step of quality control is to remove all mitochondrial encoded genes.

Normalization

To normalize scRNA-seq read count data, the scaling read count value for gene X in one cell sample is calculated using the following formula:

$$scaled\ expression\ for\ gene\ X = \frac{Read\ count\ for\ gene\ X}{Total\ count\ of\ cell\ sample} \times 10000$$

Then, the data is transformed to log space using the natural logarithm. This is done by the *NormalizeData* function in the Seurat package[25].

Imputation

Imputation is done by the *runALRA* function in the Seurat package with default parameters. The method[17] is to compute the K-rank approximation to A_norm and adjust it according to the error distribution learned from the negative values.

Dimension reduction

To reduce the data dimensions, we designed a two-step method to extract features from high dimensional gene space. In the first step, we select the top 2048 variable genes. Variable genes are selected using local polynomial regression to fits the relationship of log(variance) and log(mean). The gene expression values are then standardized using the observed mean and expected variance (given by the fitted line). Gene expression variance is then calculated on the standardized values after clipping. This is done using the *FindVariableFeatures* function with *selection.method* set as *vst* in the Seurat package. Next, auto encoder is used to reduce the dimensions from 2048 to 64. First, min-max normalization based on genes is used to normalize data in these 2048 genes. Then, the data is trained by an auto encoder model. The structure of auto encoder is described in the following: In the encoder part, we have four dense layers with output dimensions 1024, 512, 128 and 64. After each dense layer we add a batch normalization layer to speed up convergence. After the second and third dense layer, we add a dropout layer with drop out percentage 0.2 and 0.3. In the decoder part, we have four dense layers with output dimensions 128,512,1024 and 2048. After each dense layer, we add a batch normalization

layer. The activation function of each layer is relu. In the training part, we set loss function to MSE, optimizer to "adam", epoch to 15, and batch size to 128. This is done by the Keras package in R. In the second step, T-SNE[26] were used to further reduce the 64-dimensional data to 2 dimensions. The iteration time and optimal number of neighbors (perplexity) can be chosen by the user. T-SNE is done using Rtsne R package.

Cell clustering

In the clustering part, we designed a two-step method to cluster single cells to different groups. The first step is the main clustering step. We use OPTICS[18] algorithm in the R package dbscan to cluster data. The upper limit of the epsilon neighborhood size (*eps*), number of minimum points in the eps region (*minPts*), and threshold to identify clusters (*eps_cl*) is chosen by the user. The results of the main clustering may not be favorable if the cluster shape isn't well-defined. Thus, it is hard to find sub-groups and sub-types based on main clustering results. To address this, we use Gaussian Mixture Model(GMM) in the R package Mclust[27] to further cluster data in each main cluster after main clustering. Each main cluster group will be analyzed to see which ones have potential sub-clusters. If potential sub-clusters are identified, we will use GMM to find the sub-cluster in the group. The number of clusters in GMM is set based on the size of main cluster.

Biomarker gene sets

In total, we collected 56 cell type and biomarker genes from several sources[10,20–22]. In the marker gene table, a value of 1 indicates that the gene is a marker gene specific cell type, and a value of 0 indicates that it is not. We also specified classical cell type sets for Alzheimer's disease and Pancreatic Cancer based on published articles[10,21]. The user could easily select these cell types by clicking the corresponding button. We also provide

the user with the ability to modify and add their own marker genes for better analysis; the user can add, delete and modify existing marker gene tables.

Cell type annotation

In the cell annotation part, we use Gene Set Enrichment Analysis (GSEA)[28] to annotate cell types for every cluster. First, user should select candidate cell types and corresponding marker genes in the Biomarker gene section. Then, for every cluster, application computes log fold change for cluster N by:

$$log\ fold\ change\ for\ cluster\ N = mean\ expression\ for\ cluster\ N - mean\ expression\ for\ other\ cells$$

Then we rank the genes based on fold change and calculate the enrichment score of marker gene sets for every cell type user selected. Finally, cell type with the largest enrichment score will be selected as the type of this group. However, if none of cell types have a positive enrichment score, the cluster will be annotated as unknown. user can choose between main clustering results and sub-clustering results for cell annotation.

Cell distribution plots

Once user gets classification results or uploads gene expression data, the application can calculate the percentage of each cell type in each sample group. If user doesn't provide sample group information, the application will simply calculate the percentage of each cell type in the whole dataset.

Epithelial-mesenchymal transition (EMT) and proliferation (PRO) analysis

EMT-PRO analysis in SC2NetDrug was analyzed by computing mean expressions for the selected design and cell type of EMT and PRO-related genes. The HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION database was chosen for

EMT related marker genes and the HALLMARK_E2F_TARGETS database was chosen for PRO-related marker genes. After user selects the design and cell type, min-max normalization is used to normalize the whole dataset based on the genes. Then the intersecting genes in the EMT and PRO-related marker gene sets are selected and a mean score of all EMT and PRO-related genes are calculated and labeled as the EMT and PRO scores, respectively.

Ligands and receptors data resources

We collected ligand-receptor data from several sources: (1) Database of Ligand-Receptor Partners(DLRP)[29] with 175 unique ligands, 133 unique receptors and 470 unique interactions (2) Ligand-receptor interaction data sources in NicheNet[15] with 1737 unique ligands, 1925 unique receptors and 12659 unique interactions (3) cell-cell interactions database in baderLab. We selected all the proteins to be annotated as ligands named “Ligand” or “ECM/Ligand” and all the proteins to be annotated as recepters named “Receptor” or “ECM/Receptor”. Then we selected all the interactions including the chosen ligands and receptors. There are 1104 unique ligands, 924 unique receptors and 16833 unique interactions. In total, there are 1424 unique ligands, 1214 unique receptors and 27291 unique interactions.

Ligand-receptor mediated signaling interactions(Upstream Network).

Upstream network analysis is used to discover up-regulated ligands, receptors and potential ligand-receptor signaling interactions. First, user need to specify the log fold change threshold, p value threshold, the group or design user want to analyze. The up-regulated ligands and receptors are discovered using the following steps. First, the differential expression genes are calculated based on two tests, the first being the Wilcoxon rank sum test and the second being the Likelihood-ratio test[30]. The genes that

have log fold changes larger than the threshold and adjusted p-values (from the two tests) less than the threshold will be selected as differentially expressed genes. The test is done by the *FindMarkers* function in the Seurat package with parameters set as *test.use=wilcox* and *test.use=bimod* for the two tests respectively. After the differentially expressed genes for all cell types in the dataset designed by the user are identified, the ligands and receptors are found by searching for all differentially expressed genes in our ligands-receptors database. Finally, up-stream interaction networks are generated by searching for all the discovered ligand-receptor interactions in our ligands-receptors database. To be specific, four networks will be generated: the up-regulated ligands to expressed receptors network, expressed ligands to up-regulated receptors network, up-regulated ligands to up-regulated receptors network, and combined network. Up-regulated ligands and receptors are ligands and receptors that have log fold changes and adjusted p-values for two tests that satisfy the user's settings. Expressed ligands and receptors are ligands and receptors that have log fold changes larger than 0. The combined network is then combined with the up-regulated and the expressed ligands and receptors.

Gene ontology (GO) term enrichment analysis

To obtain the gene-gene ontology (GO)[31] term information, the R libraries, org.Hs.eg.db and GO.db were used. The Fisher's exact test was used to identify the statistically activated/enriched GOs based on the up-regulated genes and the genes in each GO term.

Inter-Cell Communication (Downstream Network) Analysis

The inter-cell communication analysis in SC2NetDrug is done by several steps. First, differential genes in each cell type are discovered using the Wilcoxon rank sum test and the Likelihood-ratio test[30]. The genes that have log fold changes larger than the threshold and adjusted p-values (for both tests) less than the threshold will be selected as

differentially expressed genes. The tests are done by the *FindMarkers* function in the Seurat package with parameters set as *test.use=wilcox* and *test.use=bimod* for the two tests, respectively. Next, ligands, receptors and transcript factors are discovered using the ligands-receptors interaction database and the transcript factor-target interaction database.

To uncover the down-stream signaling of ligand-receptor of interest, a computational model, ***iCSC*** (**i**nter-**c**ell **s**ignaling **c**ommunication discovery using **sc**RNA-seq), was developed. Specifically, 2 background signaling resources were used: KEGG[32] signaling pathways (curated) and STRING[33] (general protein-protein interactions). For KEGG signaling pathways, the shortest paths starting from the given receptors to all the target genes (without out-signaling) were identified, denoted as $p_{i,j} = (g_i, g_{k1}, g_{k2}, \ldots, g_j)$, *where* g_i *is the receptor,* g_j is the target gene, and g_{km}, m=1, 2, …, are the genes on the shortest paths between g_i and g_j on the KEGG signaling pathways. Then an activation score for each path, $p_{i,j}$, was defined as: $s_{ij} = {\sum_{g_m \in p_{ij}} fc(g_m)}/{n}$, where ***fc(.)*** is the fold change calculator, and n is the number of genes on the signaling path. Then signaling paths with activation scores greater than a given threshold will be selected to generate the inter-cell communication network of the ligand-receptor of interest.

For STRING background signaling network, there are much more genes (nodes) and interactions (edges) than KEGG signaling. Thus, the above model for KEGG does not work for STRING. Herein, we proposed a novel down-stream signaling network discovery model. Specifically, let $G_0^i = \langle R_i, \emptyset \rangle$ denote the initialized down-stream signaling network of receptor R_i. The update of the down-stream signaling is defined as: $G_{t+1}^i = f\left(G_t^i, G_B, V_{k1}, k2\right)$, where G_t^i and G_B is the current down-stream and background (STRING) signaling networks respectively. The edge, e_{ij} (protein interactions between g_i and g_j) of background signaling network, G_B, is weighted as: $w\left(e_{ij}\right) = \frac{1}{abs(fc(g_i)} + \frac{1}{abs(fc(g_j)}$. V_{k1} is a

vector including k1 candidate genes (based on the absolute fold change in the decreasing order) to be investigated and added to the down-stream signaling network. For any gene, $g_k \in node(G_t^i)$, the shortest paths from g_k to the k1 candidate genes in V_{k1}, will be calculated. Then, an activation score for each path, $p_{k,j}$, was defined as: $s_{kj} = \sum_{g_m \in p_{ij}} fc(g_m)/n$, where *fc(.)* is the fold change calculator and n is the number of genes on the signaling path. If $n > k2$, the signaling paths will be discarded. In another word, the parameters k1 and k2 decide the search width and depth. Finally, the signaling path with highest activation score will be added to the down-stream signaling network. The process will be conducted iteratively until it reaches a network size limit, e.g., N nodes. The down-stream signaling network is generated by combining the down-stream signaling networks of all receptors: $G_1 = \cup_i G_t^i$.

Drug-target information derived from DrugBank

We collected 6650 drugs from the drug bank database and corresponding target genes. After the down-stream signaling network is generated, the drugs for genes in the network is discovered by looking through each gene in the network and searching for drugs that target this gene in drug bank database.

Connectivity Map data

Drug discovering based on signaling signatures using Connectivity Map data, which seeks to enable the discovery of functional connections between drugs, genes and diseases through analysis of patterns induced by common gene-expression changes. User can find CMAP data in National Center for Biotechnology Information database under dataset GSE92742. Before doing the analysis, user need to download corresponding data from website and we provide function to generate drug rank matrix based on data.

Drug discovering based on signaling signatures

The procedure of drug discovering is following: After the up-regulated genes for each cell group in the cell-cell communication part are obtained, the application will use GSEA and the drug rank matrix to discover potential drugs for each group. First, the application will calculate the enrichment score of up-regulated gene sets for each drug in each group. Then, the top K drugs with the lowest enrichment scores will be selected as potential drugs, where K is the number of top drugs selected by user.

Drug clustering based on GSEA scores in CMAP

After the top drug is identified, Affinity Propagation Clustering[34] will be used to cluster top drugs. First, a similarity matrix will be constructed for the top drugs. Given that the number of top drugs is K, the dimensions of the matrix will be K*K. The similarity score for drug i to drug j will be computed by the following process: select the top 150 up-regulated genes and top 150 down-regulated genes for drug i to use as the gene set. Then, compute the GSEA score for drug j using the drug rank matrix and the gene set from drug i. The enrichment score will be used as the similarity score for drug i to drug j. After the similarity matrix is constructed, it will be used to do AP clustering, which is done using the R package apcluster.

Drug clustering based on chemical structures

To clustering drugs discovered by targets, we use the chemical structure of each drugs[35]. First, the SMILES information of drugs is used to generate drug object for each drug, this is done by *parse.smiles* function in rcdk R package. Next, the fingerprint of drug is computed using *get.fingerprint* function in fingerprint R package. Based on fingerprint of

drugs, the similarity between drugs is computed using Tanimoto index. The formulation of Tanimoto index is follow:

$$S_{A,B} = \frac{c}{a + b - c}$$

Where $S_{A,B}$ is the similarity between drug *A* and drug *B*. *a* is number of bits in drug *A* and *b* is number of bits in drug *B*. *c* is number of bits in both two drugs. This is done by *fp.sim.matrix* function in R package fingerprint and set parameter method as *tanimoto*.

Author contributions

FL conceived the project, JF and FL conducted the model, tool development and data analysis and wrote the manuscript. SPG, YB, AZ, DD, WH, RF revised the manuscript. SPG, TW, PP, LD, DD, WH, RF discussed the results.

Acknowledgment

This study was partially supported by the Children's Discovery Institute (CDI) M-II-2019-802 to Li.

References

1. Torphy, R. J., Zhu, Y. & Schulick, R. D. Immunotherapy for pancreatic cancer: Barriers and breakthroughs. *Ann. Gastroenterol. Surg.* **2**, 274–281 (2018).
2. Kurahara, H. *et al.* Significance of M2-polarized tumor-associated macrophage in pancreatic cancer. *J. Surg. Res.* (2011) doi:10.1016/j.jss.2009.05.026.
3. Hutcheson, J. *et al.* Immunologic and metabolic features of pancreatic ductal adenocarcinoma define prognostic subtypes of disease. *Clin. Cancer Res.* (2016) doi:10.1158/1078-0432.CCR-15-1883.

4. Kraman, M. *et al.* Suppression of antitumor immunity by stromal cells expressing fibroblast activation protein-α. *Science (80-.).* (2010) doi:10.1126/science.1195300.

5. Feig, C. *et al.* Targeting CXCL12 from FAP-expressing carcinoma-associated fibroblasts synergizes with anti-PD-L1 immunotherapy in pancreatic cancer. *Proc. Natl. Acad. Sci. U. S. A.* (2013) doi:10.1073/pnas.1320318110.

6. Nywening, T. M. *et al.* Targeting both tumour-associated CXCR2+ neutrophils and CCR2+ macrophages disrupts myeloid recruitment and improves chemotherapeutic responses in pancreatic ductal adenocarcinoma. *Gut* (2018) doi:10.1136/gutjnl-2017-313738.

7. Ligorio, M. *et al.* Stromal Microenvironment Shapes the Intratumoral Architecture of Pancreatic Cancer. *Cell* **178**, 160-175.e27 (2019).

8. Zhu, Y. *et al.* CSF1/CSF1R Blockade Reprograms Tumor-Infiltrating Macrophages and Improves Response to T-cell Checkpoint Immunotherapy in Pancreatic Cancer Models. *Cancer Res.* **74**, 5057 LP – 5069 (2014).

9. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **50**, (2018).

10. Peng, J. *et al.* Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* (2019) doi:10.1038/s41422-019-0195-y.

11. Wang, S., Karikomi, M., MacLean, A. L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz204.

12. Choi, H. *et al.* Transcriptome Analysis of Individual Stromal Cell Populations Identifies Stroma-Tumor Crosstalk in Mouse Lung Cancer Model. *Cell Rep.* **10**, 1187–1201 (2015).

13. Leung, C. S. *et al.* Systematic Identification of Druggable Epithelial–Stromal Crosstalk Signaling Networks in Ovarian Cancer. (2018) doi:10.1093/jnci/djy097.

14. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

15. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).

16. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* (2019) doi:10.15252/msb.20188746.

17. Linderman, G. C., Zhao, J. & Kluger, Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* 397588 (2018) doi:10.1101/397588.

18. Ankerst, M., Breunig, M., Kriegel, H.-P. & Sander, J. *OPTICS: Ordering Points to Identify the Clustering Structure*. *Sigmod Record* vol. 28 (1999).

19. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

20. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

21. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

22. Kumar, M. P. *et al.* Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. *Cell Rep.* **25**, 1458-1468.e4 (2018).

23. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).

24. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452 (2017).

25. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* (2019) doi:10.1016/j.cell.2019.05.031.

26. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2625 (2008).

27. Scrucca, L., Fop, M., Murphy, T. & Raftery, A. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J.* **8**, 205–233 (2016).

28. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545 LP – 15550 (2005).

29. Graeber, T. G. & Eisenberg, D. Bioinformatic identification of potential autocrine signaling loops in cancers from gene expression profiles. *Nat. Genet.* **29**, 295–300 (2001).

30. McDavid, A. *et al.* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461–467 (2013).

31. Gene Ontology Consortium, T. *et al.* Gene Ontology: tool for the unification of biology NIH Public Access Author Manuscript. *Nat Genet* **25**, 25–29 (2000).

32. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28 (1999) doi:10.1093/nar/27.1.29.

33. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

34. Frey, B. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).

35. Voicu, A., Duteanu, N., Voicu, M., Vlad, D. & Dumitrascu, V. The rcdk and cluster R packages applied to drug candidate selection. *J. Cheminform.* **12**, 3 (2020).