Bioinformatics, 38(20), 2022, 4745–4753
https://doi.org/10.1093/bioinformatics/btac590
Advance Access Publication Date: 30 August 2022
Original Paper

Gene expression

scFeatures: multi-view representations of single-cell and spatial data for disease outcome prediction

Yue Cao [1,2], Yingxin Lin[1,2], Ellis Patrick [1,2,3], Pengyi Yang [1,2,3,*,†] and Jean Yee Hwa Yang[1,2,4,*,†]

[1]Charles Perkins Centre, The University of Sydney, Sydney, NSW 2006, Australia, [2]School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia, [3]Computational Systems Biology Group, Children's Medical Research Institute, Westmead, NSW 2145, Australia and [4]Laboratory of Data Discovery for Health Limited (D24H), Science Park, Hong Kong SAR, China

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Olga Vitek

Received on May 12, 2022; revised on July 21, 2022; editorial decision on August 16, 2022; accepted on August 28, 2022

Abstract

Motivation: With the recent surge of large-cohort scale single cell research, it is of critical importance that analytical methods can fully utilize the comprehensive characterization of cellular systems that single cell technologies produce to provide insights into samples from individuals. Currently, there is little consensus on the best ways to compress information from the complex data structures of these technologies to summary statistics that represent each sample (e.g. individuals).

Results: Here, we present scFeatures, an approach that creates interpretable cellular and molecular representations of single-cell and spatial data at the sample level. We demonstrate that summarizing a broad collection of features at the sample level is both important for understanding underlying disease mechanisms in different experimental studies and for accurately classifying disease status of individuals.

Availability and implementation: scFeatures is publicly available as an R package at https://github.com/SydneyBioX/scFeatures. All data used in this study are publicly available with accession ID reported in the Section 2.

Contact: jean.yang@sydney.edu.au or pengyi.yang@sydney.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent single-cell or near single-cell resolution omics technologies such as spatial transcriptomics enable the discovery of cell- and cell type specific knowledge and have transformed our understanding of biological systems, including diseases (Longo *et al.*, 2021). Key to the exploration of such data are the ability to untangle and extract useful information from their high feature dimensions (Yang *et al.*, 2021) and uncover hidden insights. A plethora of computational methods has been developed on this front, with the main focus on individual cell analysis (Stegle *et al.*, 2015), such as cell type identity (Abdelaal *et al.*, 2019; Kim *et al.*, 2021) and pseudotime ordering within a lineage (Saelens *et al.*, 2019). While these tools enable characterization of individual cells, there is a lack of tools that allow for the representation of individual samples based on their cellular characteristics and the investigation of how these cellular properties are driving disease outcomes. With the recent surge of multi-condition and multi-sample single-cell studies on large sample cohort (Lin *et al.*, 2020), the next frontier of research is on representing and characterizing cellular properties at the sample (e.g. individual patient) level for linking such information with the disease outcome.

Creating a representation of each sample from the collection of sequenced cells is a crucial step for subsequent analysis as successful modelling and interpretation of disease outcomes requires biologically relevant learning features from the data. While using the original expression matrix as the input to various models could inform the change in transcriptomics level across disease conditions, the ability to represent the data with other layers of information is critical for uncovering additional insights given the complex and non-linear relationships among the feature dimensions (e.g. interaction of genes, gene networks and pathways). The single-cell field has a wealth of tools for data exploration (Wu and Zhang, 2020) which enables the exploration of biology underlying the individuals. Most current tools are not specifically designed to derive a set of features that can be used to represent an individual. Yet, with careful adaptation, a number of approaches can be used to construct novel molecular representations of individual samples. Cell-cell interactions tools (Armingol *et al.*, 2021; Jin and Ramos, 2022) for example, calculate cell type specific signalling scores between pairs of ligand and receptor molecules. The interaction scores can be used to represent the intercellular communications of cells and cell types in a sample.

Another example is gene set enrichment analysis (Maleki *et al.*, 2020) which infers the pathway enrichment score of individual cells. By summarizing the scores across cell types, a cell type specific representation of the pathway enrichment of each sample can be constructed.

To this end, we develop scFeatures, a tool that generates a large collection of interpretable molecular representations for individual samples in single-cell omics data, which can be readily used by any machine learning algorithms to perform disease outcome prediction and drive biological discovery. Together, scFeatures generates features across six categories representing different molecular views of cellular characteristics. These include (i) cell type proportions, (ii) cell type specific gene expressions, (iii) cell type specific pathway expressions, (iv) cell type specific cell-cell interaction (CCI) scores, (v) overall aggregated gene expressions and (vi) spatial metrics. The different types of features constructed thereby enable a more comprehensive multi-view representation of the expression data. Based on the generated features, scFeatures produces an HTML report containing visual summaries of features most associated with conditions. In a collection of 17 published single-cell RNA-seq, single-cell spatial proteomics and spatial transcriptomics datasets, scFeatures reveal different feature types are useful for predicting the disease outcomes in different datasets. Furthermore, through examining the selected features in two case studies, scFeatures uncovers cell types important to ulcerative colitis and stratified individuals with distinct survival outcomes in a triple negative breast cancer dataset. Together, these results demonstrate that scFeatures enables data-driven feature generation (or feature engineering) and facilitates unbiased identification of feature types most perturbed by the disease conditions.

2 Materials and methods

2.1. Data collection and processing

2.1.1 scRNA-seq

To demonstrate scFeatures on scRNA-seq data, we collected data from four published studies and curated a total of 15 datasets from the studies. The data are described in detail below:

Six Ulcerative Colitis datasets: The UC data (Smillie *et al.*, 2019) sequenced healthy control, inflamed and non-inflamed colon biopsies from multiple individuals. The data was retrieved from Single Cell Portal with accession ID SCP259. We subset the data into epithelial, stromal cells and immune subsets according to the original publication, resulting in the following six datasets:

- UC healthy versus non-inflamed (Epi)
- UC healthy versus non-inflamed (Fib)
- UC healthy versus non-inflamed (Imm)
- UC inflamed versus non-inflamed (Epi)
- UC inflamed versus non-inflamed (Fib)
- UC inflamed versus non-inflamed (Imm)

where Epi stands for epithelial, Fib stands for stromal and Imm stands for immune subsets. Inflamed, non-inflamed and healthy are conditions of interest.

Six lung datasets: The lung data (Adams *et al.*, 2020) sequenced healthy control, idiopathic pulmonary fibrosis (IPF) and chronic obstructive pulmonary disease (COPD) biopsies from multiple individuals. The data was retrieved from Gene Expression Omnibus (GEO) with accession ID GSE136831. We subset the data into epithelial, stromal cells and immune subsets according to the original publication, resulting in the following datasets:

- Lung healthy versus IPF (Epi)
- Lung healthy versus IPF (Fib)
- Lung healthy versus IPF (Imm)
- Lung healthy versus COPD (Epi)
- Lung healthy versus COPD (Fib)
- Lung healthy versus COPD (Imm)

where healthy, IPF and COPD are conditions of interest.

Two melanoma data (Sade-Feldman *et al.*, 2019) sequenced immune cells from tumour biopsies of melanoma patients before and after treatment with immune checkpoint therapy. The data was retrieved from GEO with accession ID GSE120575. We subset the data into pre-treatment and post-treatment datasets. The conditions of interest in both datasets are non-responding and responding.

The COVID dataset (Schulte-Schrepping *et al.*, 2020) sequenced peripheral blood mononuclear cells (PBMC) from COVID-19 individuals. The data was retrieved from European Genome-phenome Archive (EGA) with accession ID EGAS00001004571. We subset the original data into mild and severe individuals and consider the mild and severe disease stage as the conditions of interest.

2.1.2 Spatial proteomics

The *triple negative breast cancer dataset* (Keren *et al.*, 2019) measured the patient's protein expression using MIBI-TOF (multiplexed ion beam imaging by time of flight) technology. Data was obtained from https://mibi-share.ionpath.com.

2.1.3 Spatial transcriptomics

The amyotrophic lateral sclerosis dataset (Maniatis *et al.*, 2019) sequenced lumbar spinal cord tissue of ALS and control mice at varying time points using the spatial transcriptomics technology. The data was retrieved from GEO with accession ID GSE120374. We used the subset of data sequenced at the disease onset time point.

2.2 Implementation of feature types

We generated 17 feature types that can be broadly categorized into six categories: (i) cell type proportions, (ii) cell type specific gene expressions, (iii) cell type specific pathway expressions, (iv) cell type specific CCI scores, (v) overall aggregated gene expressions and (vi) spatial metrics. All feature types except for the overall aggregated gene expressions category have different implementations for scRNA-seq and spatial data to better leverage the characteristics of different data types and the implementation details are described in Supplementary Table S2.

For spot-based spatial transcriptomics, we performed the following additional processing to allow certain feature types to be applicable. First, since the cell type specific feature categories require cell type information while the spot in spot-based data contains a mixed population of multiple cells, we used Seurat's TransferData function to predict the cell type probability of each spot. A published scRNA-seq data on mouse spinal cord with cell type labels was used as the reference (Sathyamurthy *et al.*, 2018). Then, given that each spot contains an unknown number of cells that varies between spots, we weighted the contribution of each spot to the generated features by the relative number of cells it contains. We used library size as an estimate of the relative number of cells, motivated by a study that found a high correlation between the number of cells and library size of spots (Saiselet *et al.*, 2020). To calculate the relative number of cells, we binned the log2 transformed total library size of cells into 100 bins and assigned each spot a relative number of cells ranging between 1 and 100 according to its bin. The cell type probability of each spot together with the relative number of cells were used in the implementation of feature types for spatial transcriptomics.

2.3 Correlation between features and feature types

Given scFeatures constructs a standard matrix of samples by features, we can readily compute the Pearson's correlation between individual features as shown in Supplementary Figure S2. We subsampled 100 features from feature types that have more than 100 features to avoid the correlation plot being dominated by feature types with greater number of features.

To summarize the correlation between pairs of feature types as shown in Figure 2b, the following approach was taken. First, we calculated the Pearson's correlation between all features from a pair of feature types, such as proportion raw and gene mean celltype. This is repeated for each pairwise combination of feature types for each dataset. Then we subsampled 1000 values from the correlation values to reduce the computational burden of plotting. For ease of visual interpretation, the absolute values of the correlation values were taken.

These correlation values were further summarized in Figure 2c by taking the average correlation values, followed by hierarchical clustering to cluster the feature types.

2.4 Classification and survival analysis using generated features

In scFeatures, we provide functionality to perform classification and survival analysis for the convenience of users. The classification function builds upon the functions in the classification package classifyR (Strbenac *et al.*, 2015) that was published by our group earlier. We used the random forest model, set the number of folds to three, performed 50 cross-validation and calculated F1 score. classifyR has an in-built feature selection function. We used the default setting that uses the feature selected from the random forest model built on the training set to evaluate on the test set. These were also the settings used to report the classification performance in this study and can be specified by the user. The only exception is that 100 repeats of cross-validation were performed instead of 50 to obtain a more stable feature importance score for the case study on the 'UC healthy versus non-inflamed (Fib)' dataset.

For survival analysis, we used a Cox proportional-hazards model provided in the rms R package. By default, we set the number of folds to three, performed 50 cross-validation and calculated C-index. Note that as the Cox model is not designed to take in a large number of features at once, unlike a typical classification model, we input one feature from the generated feature type at a time for building the Cox model. The best C-index was reported as the performance for the feature type.

2.5 Complementarity of the generated features

To explore the complementarity of the generated features, we compared the classification accuracy of using features from individual feature types with using the combination of features from all feature types. In detail, we used the classification model described above, which is trained on all feature space to derive the feature importance. We then identified the top eight features from each feature type and combined them into the 'combined feature set'. This set contains 96 features (8 features × 12 feature types) for the ALS dataset and 104 features (8 features × 13 feature types) for the other 15 datasets. The triple negative breast cancer dataset was excluded from this analysis as Cox proportional-hazards model is not designed to take in a large number of features at once. For fair comparison with the individual feature type, we used the top 100 features from each individual feature type. For feature types with less than 100 features, i.e. 'proportion raw' and 'proportion logit', we used all features. We used the random forest model, set the number of folds to three, performed 50 cross-validation and recorded the F1 score.

2.6 Feature importance score

The runTests function in ClassifyR outputs the features selected by the classification model. Since repeated cross-validation was performed, this generated one set of included features for each cross-validation process. Based on all the derived sets, the frequency of inclusion was considered as the 'feature importance score' of each feature.

For the cell type specific feature category, given that each feature is associated with a cell type, it is also of interest to aggregate the feature importance score associated with each cell type. We approached this by summing the feature importance score of all features associated with a cell type, then dividing by the number of features constructed for that particular cell type to adjust for the difference in the number of features per cell type. The final score was considered the feature importance score of each cell type.

2.7 Speed and memory usage

To benchmark the scalability of the 17 feature types, we used the UC inflamed versus non-inflamed (Imm) dataset and took random samples to construct datasets with 1000, 2000, 3000, 5000, 10 000, 20 000, 30 000, 50 000, 70 000 and 100 000 cells. Each dataset contains the same 15 individuals and the same 15 cell types.

For the purpose of evaluating the feature types designed for spot-based data which require each spot to be associated with a cell type probability vector, we treated each cell as a 'spot' and randomly created a cell type probability vector for each cell. Similarly, for the purpose of evaluating the feature types under the category of spatial metrics which require spatial coordinates of each cell, we randomly assigned a pair of x and y-coordinates to each cell. In addition, the cell type probability and number of cells in each spot were randomly generated to represent such data.

Runtime was measured using the built-in Sys.time function in R. Memory was measured by recording the peak resident set size, which measures the peak amount of memory that a process consumes across all cores. All code was run in parallel using 8 cores three times and the average measurements were taken. All processes were carried out using a research server with dual Intel(R) Xeon(R) Gold 6148 Processor with 40 cores and 768 GB of memory.

3 Results

3.1 scFeatures performs multi-view feature engineering for single-cell and spatial data

We propose scFeatures, a new multi-view feature engineering framework that creates an interpretable representation of cellular level features for each individual sample from a given single-cell or spot-based expression dataset (Fig. 1a). To capture the wide range of cellular information for sample classification (e.g. diseased versus healthy individuals) using single-cell data, we implemented an extensive collection of algorithms to extract over 50 000 interpretable features from a given dataset. These features, spanning a total of 17 types, are motivated by established analytical approaches in a broad range of single-cell literature and can be broadly grouped into six distinct categories including (i) cell type proportions, (ii) cell type specific gene expressions, (iii) cell type specific pathway expressions, (iv) cell type specific CCI scores, (v) overall aggregated gene expressions and (vi) spatial metrics (Fig. 1b). These collections of constructed features can then be used for various downstream analyses such as disease outcome prediction, biomarker selection, survival analysis and enable the identification of interpretable features and feature types associated with disease conditions.

The six feature categories represent different 'views' of the single-cell information. Specifically, category I captures cell type proportion information in which the proportion of cell types for each sample and the ratio of proportions between two cell types are measured. Category II represents cell type specific gene expression and examines the expression of sets of genes or proteins in each cell type. We implemented different approaches for representing this information, including average expression, proportion of expression and correlation of expressions. In category III, which calculates cell type specific pathway scores, by default the 50 hallmark pathways in the Molecular Signatures Database (MSigDB) (Liberzon *et al.*, 2015; Subramanian *et al.*, 2005) were used to generate various features such as the average expression of each pathway in each cell type. Category IV contains the CCI scores, which measure the probability of ligand-receptor interaction based on the expression values. Category V is designed to recreate the bulk expression by aggregating the expression across cells. Category VI is designed specifically for spatial data type and includes classical metrics for identifying spatial patterns. For all feature categories except category V, the values are summarized at per cell type level, for example, feature x cell

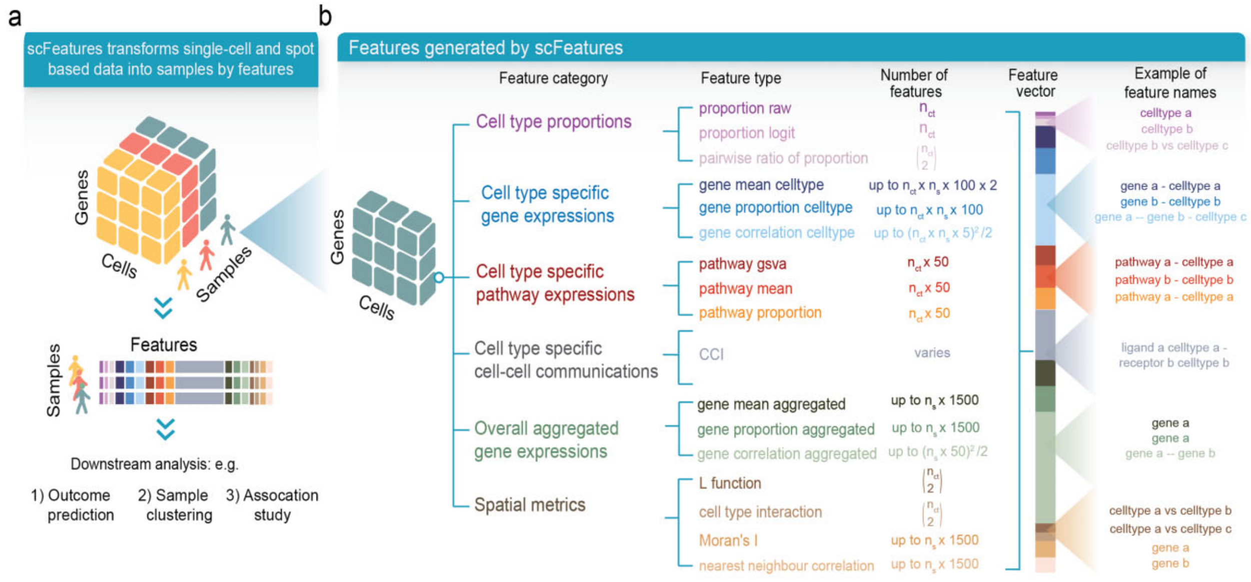


Fig. 1. Overview of scFeatures. (**a**) The input for scFeatures is an omics dataset containing multiple samples such as individuals and cell type labels. scFeatures extracts different views of the data, thereby transforming the gene by cell matrix into a vector of features for each sample. (**b**) scFeatures constructs 17 feature types that can be broadly classified into six categories. Each feature type consists of multiple individual features. For example, for 'gene mean celltype', 100 features are generated by default per cell type (nct) per sample (ns) (see Section 2). Examples of feature names from each feature type are given to illustrate the data format

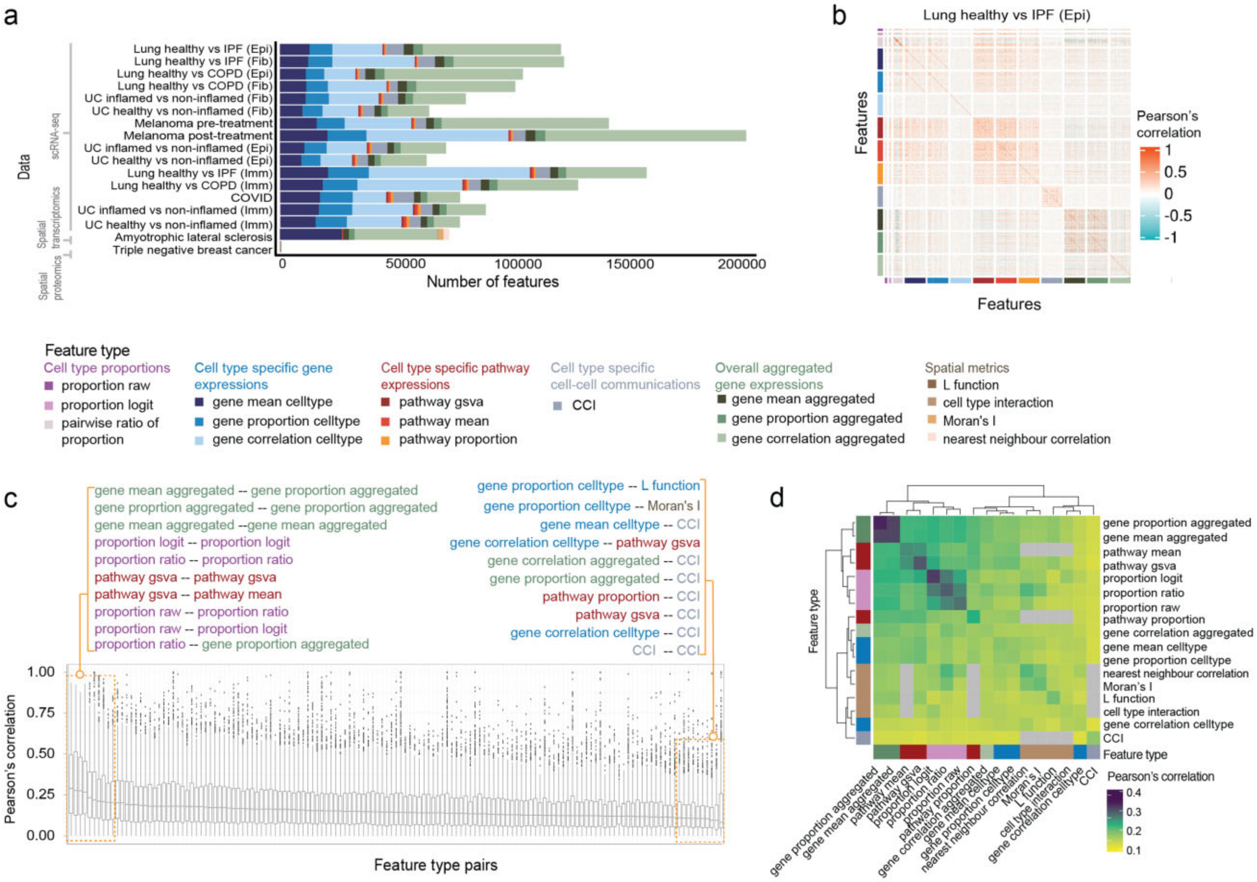


Fig. 2. Characteristics of the features generated by scFeatures. (**a**) Compositional barchart showing the number of features generated by scFeatures for each dataset. Datasets are first ordered by data types, and then by the number of cell types. (**b**) Correlation plot showing Pearson's correlation of features on the 'Lung healthy versus IPF (Epi)' dataset as a representative example. The features are colour labelled by feature types for ease of interpretation. (**c**) Boxplots summarizing the correlation between pairs of features across all datasets (see Section 2). Texts highlight the 10 most and 10 least correlated feature types pairs, coloured according to their feature category. (**d**) Hierarchical clustering of the average correlation between feature types. Heatmap is colour labelled by feature category for ease of interpretation

type *a* and feature *y* cell type *b*, which then forms the vector of molecular representation containing over 50 000 features for each sample. The implementation details can be found in Supplementary Table S2.

scFeatures extracts interpretable features from data generated by scRNA-seq, spatial proteomics and spatial transcriptomics (Table 1). In particular, spatial transcriptomics data, a spot-based technique in which the expression value of each spot is based on a small population of cells, often contains cells from multiple cell types in each spot. We developed several novel ways to adapt the 13 feature types to spot-based data whenever possible; this collection of spatial metrics considers the properties of spot-based technology and reveals cell type specific features in spot-based data. For example, spot-based data precludes direct application of cell type proportion computation since each spot includes an unknown number of cells, while cell type percentage estimation requires individual cell counts for each cell type. To overcome this issue, we estimated the number of cells in each spot using the library size of that location based on the association between the two values. Supplementary Table S2 provides more documentation on the implementation details on the adaptation of feature types from single-cell RNA-sequencing to spot-based technologies.

3.2 scFeatures generates a large collection of diverse features and is scalable to large datasets

To demonstrate the characteristics of the feature representation, we applied scFeatures to 17 datasets measured using scRNA-seq, spatial proteomics and spatial transcriptomics data (Supplementary Table S1). For typical scRNA-seq data, scFeatures generated over 50 000 features (Fig. 2a). As expected, the number of features generated was mostly associated with the number of cell types in the dataset and not with other data characteristics, including the number of genes and number of cells (Supplementary Fig. S1).

To explore the diversity of the features generated from scFeatures, we first examined the correlation between the features across 17 datasets (Fig. 2a, Supplementary Fig. S2). By summarizing the correlation values between every pairwise combination of feature types (Supplementary Fig. S3), we observed that overall the feature types were poorly correlated, with the median correlation ranging from 0.1 to 0.3 (Fig. 2b). Hierarchical clustering of the correlations revealed that the higher correlation was observed between certain feature types from the same feature category (Fig. 2c and d). For example, the 'gene mean aggregated' and 'gene proportion aggregated' from the aggregated gene expression category had high correlation within each of the feature types and between the feature types pair. This is consistent with our expectation of some degree of co-expression linked with disease conditions.

To further examine the complementarity of the feature types, we compared the performance of individual feature types with the combination of features across feature types (Supplementary Fig. S4). The ability to accurately classify disease outcomes was used as the evaluation metric (see Section 2). We found the combination of features in general performed better than most of the individual feature types and achieved the best classification performance in 11 out of the 16 datasets, suggesting the complementarity of the feature types.

We next benchmarked the runtime and memory requirements of the feature types on single-cell scRNA-seq (Supplementary Fig. S5a), spatial proteomics (Supplementary Fig. S5b), as well as spot-based spatial transcriptomics datasets (Supplementary Fig. S5c) for evaluating both the single-cell RNA-sequencing implementation and the spot-based implementation. All datasets contain 1000 to 100 000 cells. On the largest datasets with 100 000 cells, the majority of feature types took less than a minute to compute when executed on eight cores, demonstrating that scFeatures is highly scalable to large datasets. As expected, there was some trade-off between processing time and memory. As a result of parallel computation over eight cores, some feature types required more than 10GB of RAM in total; however, users can run on a single core to reduce the memory requirement.

Table 1. List of features generated by scFeatures

	Feature category																
	Cell type specific proportions			Cell type specific gene expressions			Cell type specific pathway expressions			Cell–cell interaction scores	Overall aggregated gene expressions			Spatial metrics			
	Feature type																
Application	Proportion raw	Proportion logit	Pairwise ratio of proportion	Gene mean celltype	Gene proportion celltype	Gene correlation celltype	Pathway mean	Pathway gsva	Pathway proportion	CCI	Gene mean aggregated	Gene mean proportion	Gene mean correlation	L function	Cell type interaction	Moran's I	Nearest neighbour correlation
scRNA-seq data	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓				
Spatial proteomics data	✓	✓	✓	✓	✓	✓					✓	✓	✓	✓	✓	✓	✓
Single cell transcriptome data	✓	✓	✓	✓			✓				✓	✓	✓	✓	✓	✓	✓

Note: scRNA-seq data are single-cell based and representative technologies are 10× and Smart-seq. Spatial proteomics data are single-cell based and representative technologies are MIBI-TOF, IMC and CODEX. Single-cell transcriptome data are spot based and representative technologies are ST, Visium and MALDI.

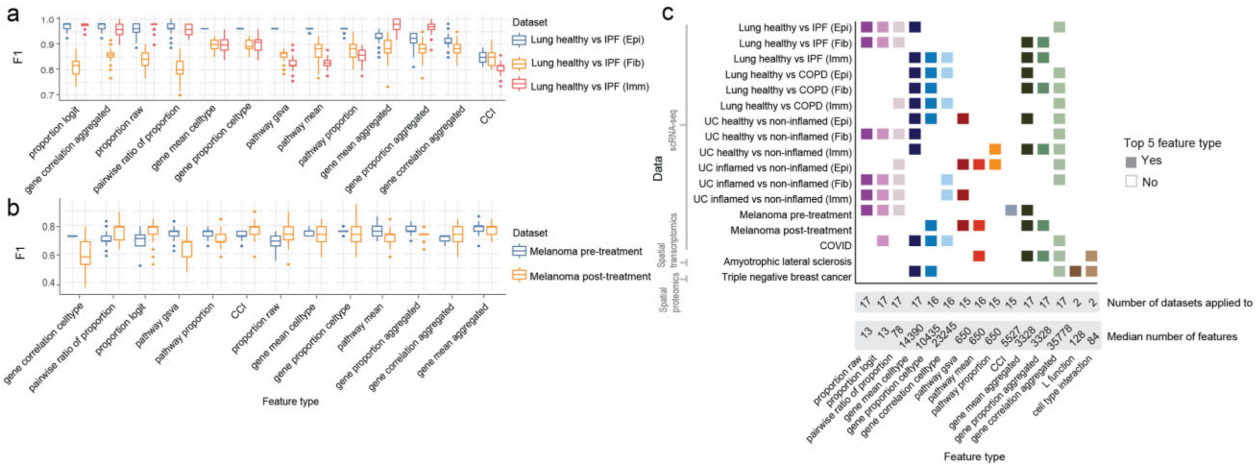


Fig. 3. Performance of feature types on patient outcomes. (**a**) The epithelial, fibroblast and immune subsets of healthy and IPF individuals, where the outcome of interest is classifying healthy and IPF status. The feature types are ordered by their F1 scores on the epithelial subset. (**b**) Pre-treatment and post-treatment melanoma patients, where the outcome of interest is classifying therapy responders and non-responders. The feature types are ordered by the difference of the F1 scores between the two datasets. (**c**) For each of the 17 datasets, the squares denote the top five feature types with the highest F1 scores

3.3 The most informative feature types differ between different datasets

We hypothesized that distinct feature types would be informative for different datasets since each dataset comprises samples with varying characteristics and disease outcomes. Several datasets were used where each feature type was evaluated on its ability to predict disease outcome and the observations were in alignment with our hypothesis. First, we used a lung disease dataset collection (Supplementary Table S1) where the cells were split into the epithelial, immune and fibroblast subsets and the outcome of interest was to classify the individuals into healthy or idiopathic pulmonary fibrosis (IPF). In Figure 3a, we visualized the classification performance of the feature types on the three subsets and ordered the feature types according to their performance in the epithelial subsets. This reveals that feature types related to cell type proportions (i.e. ‘proportion ratio’, ‘proportion logit’ and ‘proportion raw’) achieved the highest accuracy in the epithelial subset (Fig. 3a). In contrast, the performance of feature types on the immune and fibroblast subsets did not follow the same trend as on the epithelial subset, demonstrating that different feature types are useful for the three datasets.

Similar observations were also found in the melanoma pre-treatment dataset and melanoma post-treatment dataset (Supplementary Table S1) where the question of interest was classifying non-responders and responders. This revealed that proportion features (i.e. ‘proportion raw’ and ‘proportion logit’) more accurately classified individuals in the post-treatment dataset than in the pre-treatment dataset, while pathway features (i.e. ‘pathway gsva’ and ‘pathway proportion’) provided higher classification accuracy for pre-treated individuals (Fig. 3).

We then examined 17 datasets (Supplementary Fig. S6) and highlighted the five informative feature types for each dataset (Fig. 3c) for a more comprehensive assessment of the performance of the feature types. Across the 17 datasets tested, ‘gene mean celltype’, which examines expression in cell type specific manner, occurred in 10 datasets as the top five informative feature types. This is perhaps not surprising, as it elucidates the power of single-cell technology to profile the cell type specific expression to uncover changes in response to diseases. Across the spatial datasets, we saw feature types in the spatial feature category appeared as the top five informative feature types, indicating the effectiveness of this category to capture spatial information and the potential of spatial data modality to offer complementary information. Altogether, these findings highlight that different feature types are useful for exploring disease mechanisms in different datasets and even in different subsets of the same dataset, as seen by the pre- and post-treatment melanoma datasets and the lung disease dataset subset by cell types, and argue for the need for a diverse compendium of feature types for such analyses.

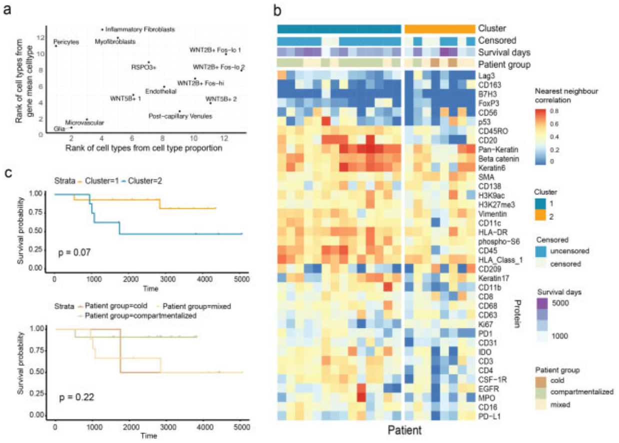


Fig. 4. Selected features generated on the ‘UC healthy versus non-inflamed (Fib)’ dataset and the ‘triple negative breast cancer’ datasets. (**a**) Scatterplot of cell type rank for the feature type ‘cell type proportion’ and ‘gene mean celltype’. (**b**) Heatmap showing the clustering result using the nearest neighbour correlation. (**c**) Kaplan–Meier plot of individuals stratified by the clustering output (top) and stratified by patient groups defined in the original study (bottom)

3.4 scFeatures provides interpretable insight into disease outcomes from scRNA-seq data

To illustrate that scFeatures provides interpretable features for the understanding of diseases, we applied scFeatures to the ‘UC healthy versus non-inflamed (Fib)’ dataset (Smillie *et al.*, 2019). This scRNA-seq dataset compares fibroblast cells of non-inflamed biopsies from ulcerative colitis (UC) individuals with biopsies from healthy individuals. We focused on the two top performing feature types of ‘gene mean celltype’ and ‘proportion raw’ based on the classification model performance from the previous section (Fig. 3c) and discovered different sets of cell types were important to the two feature types. In particular, for the feature type based on cell type specific gene expression (denoted by ‘gene mean celltype’), the fourth-ranked cell type according to feature importance score (see Section 2) was WNT5B+ 2 (Fig. 4a). This cell type was ranked as the 11th cell type in terms of the differences in cell type proportion (denoted by ‘proportion raw’) (Supplementary Fig. S7), indicating that while the gene expression was different between disease outcomes, the proportion of cell types was similar. In contrast, glia was ranked first in terms of gene expression and second in terms of cell type proportion. These two feature types offer different perspectives from the same data and reveal distinct collections of cell types where one group is more concerned with changes in expression and the other collection is more concerned with changes in proportion. It would have been challenging or impossible to accurately disentangle the contributions of cell type percentage and cell specific gene expression in classical bulk gene expression data. These observations not only highlight the necessity of single-cell research, but also emphasize the importance of evaluating various feature types, as generated by scFeatures.

3.5 scFeatures uncovers data features associated with survival outcomes from spatial proteomics

To demonstrate the utility of scFeatures at extracting spatial information, we applied scFeatures to a spatial proteomics dataset of tumours from triple negative breast cancer individuals (Supplementary Table S1). The question of interest is classifying tumours based on cellular organization into distinct types that are associated with patient survival. The original study defined three tumour groups based on mixing scores, where a 'cold group' is identified by low immune infiltrate, a 'compartmentalised group' is identified by compartments formed by almost entirely of either tumour or immune cells, and a 'mixed group' is when there is no clear boundary separating the tumour and immune cells.

The nearest neighbour correlation is a feature type in scFeatures that was created primarily to capture spatial co-expression patterns. It computes the correlation of a cell's protein expression with that of its nearest neighbour. Therefore, spatial organization of cells, such as whether tumour cells are next to immune cells would affect the correlation of protein expression of cells with neighbouring cells. To construct this feature type, we used scFeatures on selected 'triple negative breast cancer' samples from the dataset and clustered the resulting features (Fig. 4b). Survival analysis using the Kaplan–Meier Curve revealed differences between survival outcomes of individuals from the two clusters (*P*-value of 0.07, Fig. 4c), compared to the patient group defined in the original study with *P*-value of 0.22. This suggests that the new patient subgroup found by scFeatures has greater association with the survival outcomes and demonstrates the ability of the spatial feature category to represent spatial organizations and uncover novel patterns in the data.

3.6 scFeatures automatically generates an HTML file that reports features most associated with conditions to facilitate interpretable discoveries

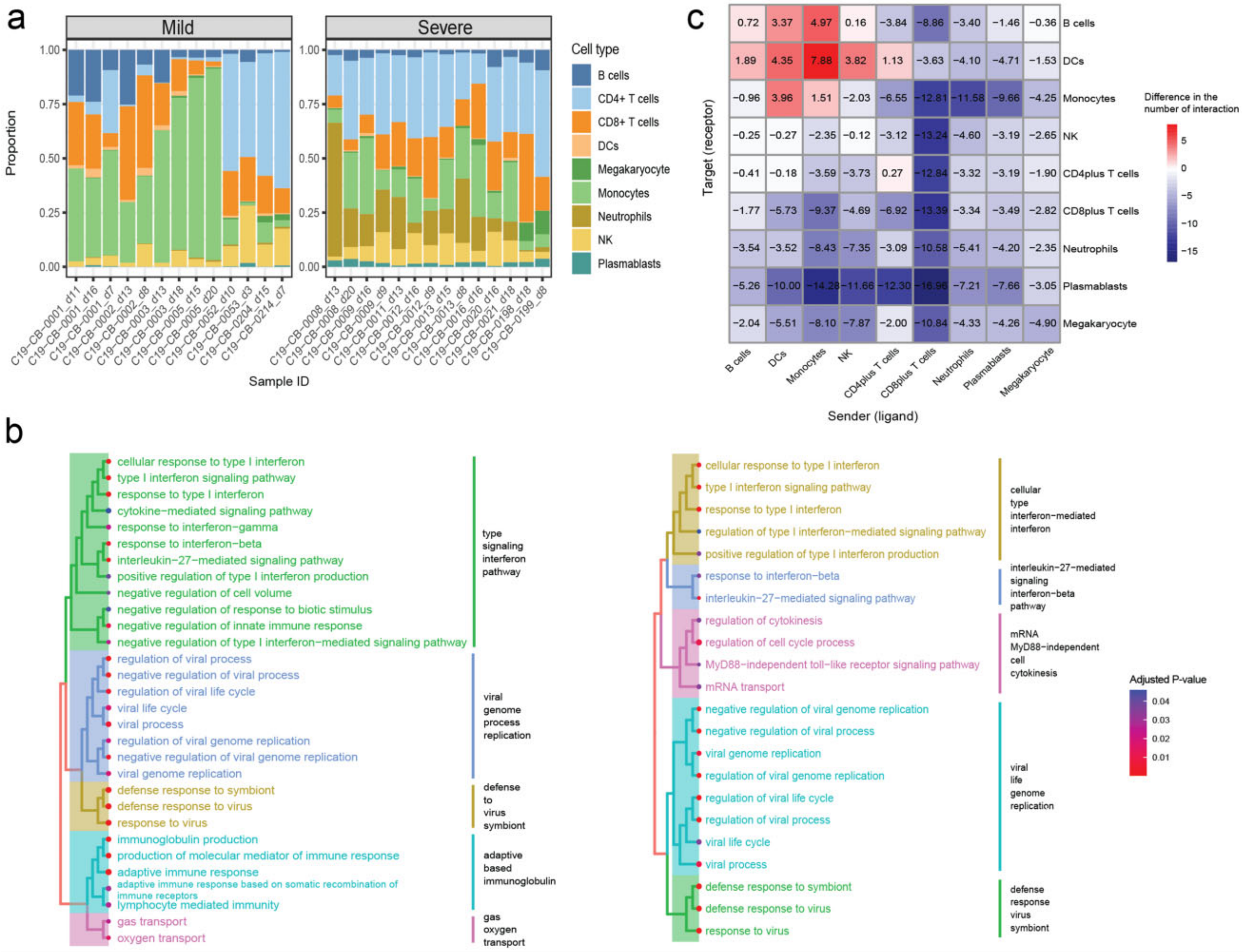
One of the most commonly investigated questions by researchers is what features are most associated with disease conditions. We implemented a function within scFeatures that takes generated features as input and automatically performs a series of association studies for each feature type, producing an HTML report as the output. An example of a comprehensive HTML report can be found on our Github (https://github.com/SydneyBioX/scFeatures). The HTML report includes a variety of visual summaries to aid the downstream interpretation of features. Here, we used the 'COVID' dataset to identify features associated with disease severity and illustrate a selected panel of visual summaries (Fig. 5). The composition plot visualized the features from 'cell type proportion raw' (Fig. 5a) and revealed that many cell types underwent drastic change between mild and severe conditions. The pathway enrichment plots (Fig. 5b) summarized that, in the rare cell type plasmablasts, genes associated with severe condition were enriched in immune pathways. Heatmap is used to visualize the difference between conditions that can be expressed numerically. The heatmap on feature type 'CCI' revealed that the cell-cell interactions in most pairs of cell types increased in severe patients compared to mild patients (Fig. 5c). Overall, the association study and visual summaries provided by the HTML facilitate a more focused exploration of features for further analysis.

4 Discussion and conclusion

In summary, scFeatures creates a multi-view molecular representation of individuals by generating tens of thousands of interpretable features based on single-cell and spot-based spatial data. The innovation and motivation of scFeatures lie in the generation of various literature motivated and biologically relevant feature vectors for phenotype disease modelling and disease prediction. We have designed 17 feature types across six categories based on a broad range of analytical approaches in literature from cell type specific gene expression to measures of cell-cell (ligand-receptor co-expression) interaction and demonstrated that the feature types are diverse with low correlation amongst them. We illustrated scFeatures on scRNA-seq data from ulcerative colitis and discovered a number of features linked with disease characteristics. scFeatures is also able to extract spatial features from a triple negative breast cancer proteomics data, resulting in the stratification of tumours that are more strongly related to survival outcomes than the original study's subgroups. Through the automatic report generation that highlights features most associated with disease, scFeatures supports ease of feature exploration.

The features vector generated by scFeatures can be used for a broader set of downstream applications and is not limited to the ones illustrated in the case studies. For example, given the feature vectors are generated at the sample level, this provides the opportunity for the exploration of differential patient responses to diseases due to heterogeneity between individuals. Even amongst those recorded as responders to treatment, the extent of response and the change at omics level vary between individuals. The feature vector can be subjected to latent class analysis, which has typically been applied on single-cell level to explore cellular diversity (Cheng *et al.*, 2019; Buettner *et al.*, 2017) and to enable detection of subpopulations in the cohort, as well as the biology driving patient heterogeneity. Given that scFeatures creates a representation for each patient, this also enables the integrative analysis of patients across multiple datasets to increase the power of analysis and to expand the range of questions that can be asked. Batch correction methods, such as scMerge (Lin *et al.*, 2019) and Harmony (Korsunsky *et al.*, 2019), may be needed in this case to remove the unwanted technical variation due to datasets.

The multiple feature types generated by scFeatures can be considered as multiple views of the data and as such, lead naturally to multi-view learning. This is one of the many collections of methods that perform integration across multiple feature classes to enhance model performance. There exist many approaches for data integration (Li *et al.*, 2018), from the simple concatenation of features from all feature types into a single vector as the input, to incorporating and optimizing the procedure within the model training process. While current multi-view learning in bioinformatics typically refers to the use of multiple omics obtained from the same sample (Nguyen and Wang, 2020), we envisage the generation of multiple feature types by scFeatures opens new opportunities for multi-view learning for single omic type.scFeatures is currently designed to perform feature engineering for single-cell RNA-seq, spatial proteomics and spatial transcriptomics data, but the framework is not limited to these platforms. Taking chromatin accessibility as an example, a commonly used analysis strategy is assigning genes based on nearby peaks, thereby converting the peak matrix to a matrix of gene activity scores similar to gene expressions (Baek and Lee, 2020). Using this approach, all feature types designed for scRNA-seq are then applicable to chromatin accessibility data. In future, we plan to extend scFeatures to other single-cell omics such as single-cell DNA methylation, single-cell chromatin accessibility and single-cell genomics, leveraging the common analytical approach in these omics and constructing specific feature types. For chromatin accessibility, the co-accessibility between pairs of peaks, which is used to predict



	B cells	DCs	Monocytes	NK	CD4plus T cells	CD8plus T cells	Neutrophils	Plasmablasts	Megakaryocyte
B cells	0.72	3.37	4.97	0.16	−3.84	−8.86	−3.40	−1.46	−0.36
DCs	1.89	4.35	7.88	3.82	1.13	−3.63	−4.10	−4.71	−1.53
Monocytes	−0.96	3.96	1.51	−2.03	−6.55	−12.81	−11.58	−9.66	−4.25
NK	−0.25	−0.27	−2.35	−0.12	−3.12	−13.24	−4.60	−3.19	−2.65
CD4plus T cells	−0.41	−0.18	−3.59	−3.73	0.27	−12.84	−3.32	−3.19	−1.90
CD8plus T cells	−1.77	−5.73	−9.37	−4.69	−6.92	−13.39	−3.34	−3.49	−2.82
Neutrophils	−3.54	−3.52	−8.43	−7.35	−3.09	−10.58	−5.41	−4.20	−2.35
Plasmablasts	−5.26	−10.00	−14.28	−11.66	−12.30	−16.96	−7.21	−7.66	−3.05
Megakaryocyte	−2.04	−5.51	−8.10	−7.87	−2.00	−10.84	−4.33	−4.26	−4.90

Fig. 5. Selected visualization summaries from the HTML report. The 'COVID' dataset containing mild and severe COVID-19 patients was used to show a subset of visualization summaries from the association analysis report. (**a**) Composition plot of the 'cell type proportion raw' features in mild and severe patients. (**b**) Enriched pathways of the top 200 features associated with the plasmoblasts of severe patients. Pathway enrichment in the left plot was calculated based on features from 'cell type specific mean expression'. Pathway enrichment in the right plot was calculated based on features from 'cell type specific mean proportion'. Similar pathway terms were grouped by hierarchical clustering. (**c**) Heatmap shows the difference in the number of CCI features between mild and severe patients. Positive number indicates more interactions in the mild patients and negative number indicates more interactions in the severe patients

cis-regulatory interactions, can be constructed and stored as a vector for each sample. The correlation values between transcription factors (TF) motifs can be readily constructed as another class of feature representation vector, and can be used to identify the modules of TF motifs affected in disease state.

With the recent surge of cohort based single-cell studies and the number of tools for characterizing individual cells, there is an increased demand for defining samples in a study based on their cellular characterization to guide better understanding of disease and health. Here, we present scFeatures, a tool that provides a multi-view extraction of molecular features from single-cell and spot-based spatial data to characterize cellular features of each individual. scFeatures efficiently extracts collections of interpretable features from large-scale data and derives biological insights in both scRNA-seq and spatial data. We envision that scFeatures, a public R package available at https://github.com/SydneyBioX/scFeatures, will facilitate better understanding of single-cell data from a sample (i.e. patient) perspective and the signatures underlying disease conditions from different angles.

Acknowledgments

The authors would like to thank all their colleagues at The University of Sydney, School of Mathematics and Statistics, for their intellectual engagement and constructive feedback. In particular, we would like to acknowledge Shila Ghazanfar for her valuable feedback toward the manuscript.

Author contributions

J.Y.H.Y., P.Y. and Y.C conceived the study. Y.C. performed the experiments with input from Y.L. and E.P. and interpreted the results with input from all authors. All authors wrote, read and approved the final manuscript.

Funding

This study was made possible in part by Australia National Health and Medical Research Council (NHMRC) Investigator Grant [APP1173469 to P.Y.]; AIR@innoHK programme of the Innovation and Technology Commission of Hong Kong, Australia NHMRC Career Developmental Fellowship [APP1111338 to P.Y. and J.Y.H.Y.]; Australian Research Council Discovery Early Career Researcher Award [DE200100944] funded by the Australian Government. to E.P.; Research Training Program Tuition Fee Offset and University of Sydney Postgraduate Award Stipend Scholarship to Y.C.; Research Training Program Tuition Fee Offset and Stipend Scholarship and Chen Family Research Scholarship to Y.L.

Conflict of Interest: none declared.

Data availability

All data used in this study are publicly available. The accession links are reported in the Section 2.

Code availability

scFeatures is publicly available as an R package at https://github.com/SydneyBioX/scFeatures.

References

Abdelaal,T. *et al.* (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.*, **20**, 194.

Adams,T.S. *et al.* (2020) Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.*, **6**, eaba1983.

Armingol,E. *et al.* (2021) Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet.*, **22**, 71–88.

Baek,S. and Lee,I. (2020) Single-cell ATAC sequencing analysis: from data preprocessing to hypothesis generation. *Comput. Struct. Biotechnol. J.*, **18**, 1429–1439.

Buettner,F. *et al.* (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.*, **18**, 212.

Cheng,C. *et al.* (2019) Latent cellular analysis robustly reveals subtle diversity in large-scale single-cell RNA-seq data. *Nucleic Acids Res.*, **47**, e143.

Jin,S. and Ramos,R. (2022) Computational exploration of cellular communication in skin from emerging single-cell and spatial transcriptomic data. *Biochem. Soc. Trans.*, **50**, 297–308.

Keren,L. *et al.* (2019) MIBI-TOF: a multiplexed imaging platform relates cellular phenotypes and tissue structure. *Sci. Adv.*, **5**, eaax5851.

Kim,H.J. *et al.* (2021) Uncovering cell identity through differential stability with Cepo. *Nat. Comput. Sci.*, **1**, 784–790.

Korsunsky,I. *et al.* (2019) Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods*, **16**, 1289–1296.

Li,Y. *et al.* (2018) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*, **19**, 325–340.

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

Lin,W.N. *et al.* (2020) The role of single-cell technology in the study and control of infectious diseases. *Cells*, **9**, 1440.

Lin,Y. *et al.* (2019) scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci. USA*, **116**, 9775–9784.

Longo,S.K. *et al.* (2021) Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.*, **22**, 627–644.

Maleki,F. *et al.* (2020) Gene set analysis: challenges, opportunities, and future research. *Front. Genet.*, **11**, 654.

Maniatis,S. *et al.* (2019) Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. *Science*, **364**, 89–93.

Nguyen,N.D. and Wang,D. (2020) Multiview learning for understanding functional multiomics. *PLoS Comput. Biol.*, **16**, e1007677.

Sade-Feldman,M. *et al.* (2019) Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, **176**, 404.

Saelens,W. *et al.* (2019) A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, **37**, 547–554.

Saiselet,M. *et al.* (2020) Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *J. Mol. Cell Biol.*, **12**, 906–908.

Sathyamurthy,A. *et al.* (2018) Massively parallel single nucleus transcriptional profiling defines spinal cord neurons and their activity during behavior. *Cell Rep.*, **22**, 2216–2225.

Schulte-Schrepping,J. *et al.*; Deutsche COVID-19 OMICS Initiative (DeCOI). (2020) Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell*, **182**, 1419–1440.e23.

Smillie,C.S. *et al.* (2019) Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, **178**, 714–730.e22.

Stegle,O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133–145.

Strbenac,D. *et al.* (2015) ClassifyR: an R package for performance assessment of classification with applications to transcriptomics. *Bioinformatics*, **31**, 1851–1853.

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

Wu,Y. and Zhang,K. (2020) Tools for the analysis of high-dimensional single-cell RNA sequencing data. *Nat. Rev. Nephrol.*, **16**, 408–421.

Yang,P. *et al.* (2021) Feature selection revisited in the single-cell era. *Genome Biol.*, **22**, 321.