PLOS ONE

RESEARCH ARTICLE

TMExplorer: A tumour microenvironment single-cell RNAseq database and search tool

Erik Christensen[1,2☯], **Alaine Naidas**[2,3☯], **David Chen**[2,3☯], **Mia Husic**[4], **Parisa Shooshtari**[1,2,3,5,6*]

1 Department of Computer Science, University of Western Ontario, London, ON, Canada, **2** Children Health Research Institute, Victoria Research Labs, London, ON, Canada, **3** Department of Pathology and Lab Medicine, University of Western Ontario, London, ON, Canada, **4** Genetics and Genome Biology, SickKids Research Institute, Toronto, ON, Canada, **5** Lawson Health Research Institute, London, ON, Canada, **6** Ontario Institute for Cancer Research, Toronto, ON, Canada

☯ These authors contributed equally to this work.
* pshoosh@uwo.ca

OPEN ACCESS

Citation: Christensen E, Naidas A, Chen D, Husic M, Shooshtari P (2022) TMExplorer: A tumour microenvironment single-cell RNAseq database and search tool. PLoS ONE 17(9): e0272302. https://doi.org/10.1371/journal.pone.0272302

Editor: Sumitra Deb, Virginia Commonwealth University, UNITED STATES

Received: March 9, 2022

Accepted: July 17, 2022

Published: September 9, 2022

Data Availability Statement: All included datasets are from free, public sources. These datasets are all referenced in the current paper and in the metadata of our package. Additionally, we have hosted the consistently processed data on figshare at the following link: https://figshare.com/projects/TMExplorer_A_Tumour_Microenvironment_Single-cell_RNAseq_Database_and_Search_Tool/101471. All source code of TMExplorer is freely available on Github (https://github.com/shooshtarilab/TMExplorer), and as an R BioConductor package at the following link: https://bioconductor.org/packages/release/data/

Abstract

Motivation

The tumour microenvironment (TME) contains various cells including stromal fibroblasts, immune and malignant cells, and its composition can be elucidated using single-cell RNA sequencing (scRNA-seq). scRNA-seq datasets from several cancer types are available, yet we lack a comprehensive database to collect and present related TME data in an easily accessible format.

Results

We therefore built a TME scRNA-seq database, and created the R package TMExplorer to facilitate investigation of the TME. TMExplorer provides an interface to easily access all available datasets and their metadata. The users can search for datasets using a thorough range of characteristics. The TMExplorer allows for examination of the TME using scRNA-seq in a way that is streamlined and allows for easy integration into already existing scRNA-seq analysis pipelines.

Introduction

Single-cell RNA sequencing (scRNA-seq) is a new technology that has emerged as an important tool to measure gene expression for individual cells, enabling the examination of cellular heterogeneity and tissue composition with incredible precision. This has been particularly applicable in cancer research for the study of tumour composition, heterogeneity and phenotype, all of which are directly impacted by the tumour-microenivronment (TME). TMEs are composed of different stromal and cancer cell types whose interactions likely dictate different aspects of tumour behaviour, such as metastasis [1–4]. Combined with scRNA-seq analysis methods, scRNA-seq enables us to dissect the TME into individual cells and investigate the

experiment/html/TMExplorer.html. Users wishing to have their data added to TMExplorer can open an issue at the following link with a link to their study and data along with a brief description and we will review it for inclusion. https://github.com/shooshtarilab/TMExplorer/issues.

Funding: This project was supported by the internal research funds from Lawson Health Research Institute (PS) (Grant # R-20-303) and Schulich School of Medicine and Dentistry (PS). PS is supported by the Children's Health Research Institute, the Early Investigator Award from the Ontario Institute for Cancer Research (OICR), and the Government of Canada through Natural Sciences and Engineering Research Council (NSERC) (PS) (Grant # DGECR-2021-00298). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

different cell subpopulations that exist. Such investigations into the TME are becoming increasingly important, as tumour composition and heterogeneity can influence cancer progression and the outcome of cancer therapy [1, 4–9].

With the advancement of scRNA-seq in cancer research, the number of TME datasets that are generated continues to increase, yet they can be difficult to access. Raw sequence reads generated by scRNAseq can be shared through online archives, such as the Sequence Read Archive (SRA) [10], however they exist as large files that require further processing to be analyzed, making data access a challenge. Already processed scRNA-seq data containing gene expression information can be accessed through online archives, such as the Genome Expression Omnibus (GEO) [11], and can be more easily downloaded for use in one's own analysis. Furthermore, to manage the growing abundance of publicly available scRNA-seq data, proper quality control and curation of datasets must be done [12, 13]. Currently, several online databases offer curated collections of public scRNA-seq datasets, such as PanglaoDB [12], scRNASeqDB [13], JingleBells [14] and the Single Cell Portal created by the Broad Institute of MIT and Harvard [15]. Most existing scRNA-seq databases include a mixture of samples from normal tissues and tissues affected by cancer or other diseases [12–14], while others focus primarily on samples from normal tissues [16, 17]. A recently published toolkit called CReSCENT [18] contains only cancer scRNA-seq data, however it mainly acts as a cancer data analysis pipeline rather than a database. A comprehensive database for the collection and sharing of TME scRNA-seq datasets from a range of tumour types does not yet exist, and researchers interested in using publicly available TME data must search through several databases to collect relevant datasets for their study. A database of TME scRNA-seq samples will thus streamline the data collection steps required for researching cancer at a single-cell level, lowering the barrier for entry to this type of study.

It is likewise important that scRNA-seq databases are designed to facilitate streamlined data collection and analysis. This can include a search tool that allows users to select datasets based on desired characteristics. While existing databases include search tools, they provide few options in characteristics users are able to search for and often require users to browse through a metadata table prior to selecting datasets of interest. Furthermore, they are designed as web-based tools, and thus are not intended to be integrated into workflows [12–15]. Workflow integration would enable users to access data directly in their pipelines, thus automating the data collection process and increasing analysis efficiency. A scRNA-seq database that is provided as an R-package and contains a comprehensive search tool which allows users to select datasets based on a wider variety of characteristics would make the data collection process easier for researchers.

Here, we present a curated collection of tumour scRNA-seq datasets made available as an R-package called TMExplorer. TMExplorer contains publicly available scRNA-seq datasets specific to TMEs from various tumour types collected from different scRNA-seq studies [1–3, 5–9, 19–53] and online databases [11, 54]. In addition to gene expression data, TMExplorer contains the corresponding cell type annotations and gene-signature information for several datasets, and provides a search tool that enables users to search for multiple datasets according to 13 different characteristics (Table 1). When selecting datasets, users can review the metadata table first or they can retrieve datasets that match specific criteria without having to browse through the metadata table. While online databases require users to download a given dataset prior to use, TMExplorer allows users to access and search available datasets within R. Users can thus input the data directly into existing pipelines with only a few commands. Each dataset can be used directly within R as a *SingleCellExperiment* object, or exported as a gene expression matrix in multiple formats for use with other applications. Users interested in validating scRNA-seq analysis algorithms, as they apply to TME data, can easily access this information

Table 1. A list of search parameters that can be passed to queryTME in order to filter the available datasets.

Search Parameter	Description
geo_accession	Search by GEO accession
score_type	Search by type of score available
has_signatures	Search by presence of cell type signature gene sets
has_truth	Search by presence of cell type annotations
tumour_type	Search by type of tumour
author	Search by first author
journal	Search by publication journal
year	Search by publication year
pmid	Search by PMID
sequence_tech	Search by sequencing technology
organism	Search by source organism
sparse	Return expression in sparse matrices
download_format	Specify a list of score formats to download. Additional formats will be stored in altExps

https://doi.org/10.1371/journal.pone.0272302.t001

through TMExplorer and incorporate it into their pipelines. Altogether, TMExplorer makes it easier for researchers to access and share TME scRNA-seq datasets, facilitating the study of TMEs at the single-cell level in the field of cancer research.

Materials and methods

Data collection

In order to collect the datasets, we searched the National Center for Biotechnology Information (NCBI) [55] for relevant scRNA-seq studies using the following keywords: single cell RNA sequencing, tumour, cancer, tumour microenvironment, and malignant. We then carefully reviewed the published literature and any associated data to confirm if they matched our criteria. Datasets were included in our data collection if they were publicly available as processed data, were generated by scRNA-seq and if they consisted of TME expression data. A total of 48 datasets originating from different types of human and mouse tumours were collected from online sources such as the NCBI's Gene Expression Omnibus (GEO) [11], ArrayExpress [54], and Github [56]. Out of the 48 datasets we collected, 44 datasets originated from human tumours and 4 datasets originated from mouse tumours (Fig 1). Descriptions of the collected datasets are provided in Table 2. Metadata for each dataset, such as tumour type and number of cells sequenced, were collected from descriptions in the corresponding publications and/or from the online sources that the datasets were obtained from. If publicly available, we also retrieved cell-type annotations and/or gene signature information that accompanied the datasets. All data is hosted on FigShare [57] under the TMExplorer project.

Data curation

Datasets found on GEO often contain extra information such as Ensemble ID or chromosome region in additional rows or columns. We modified all datasets using R to ensure they followed a similar genes-by-cells format with the gene column serving as an index. If any dataset is published as separate samples, samples are merged into a single file with a suffix identifying the sample appended to cell IDs, so users may separate the samples and perform batch correction if necessary. Having a similar format for datasets reduces the preprocessing required to use this data in other analysis pipelines.

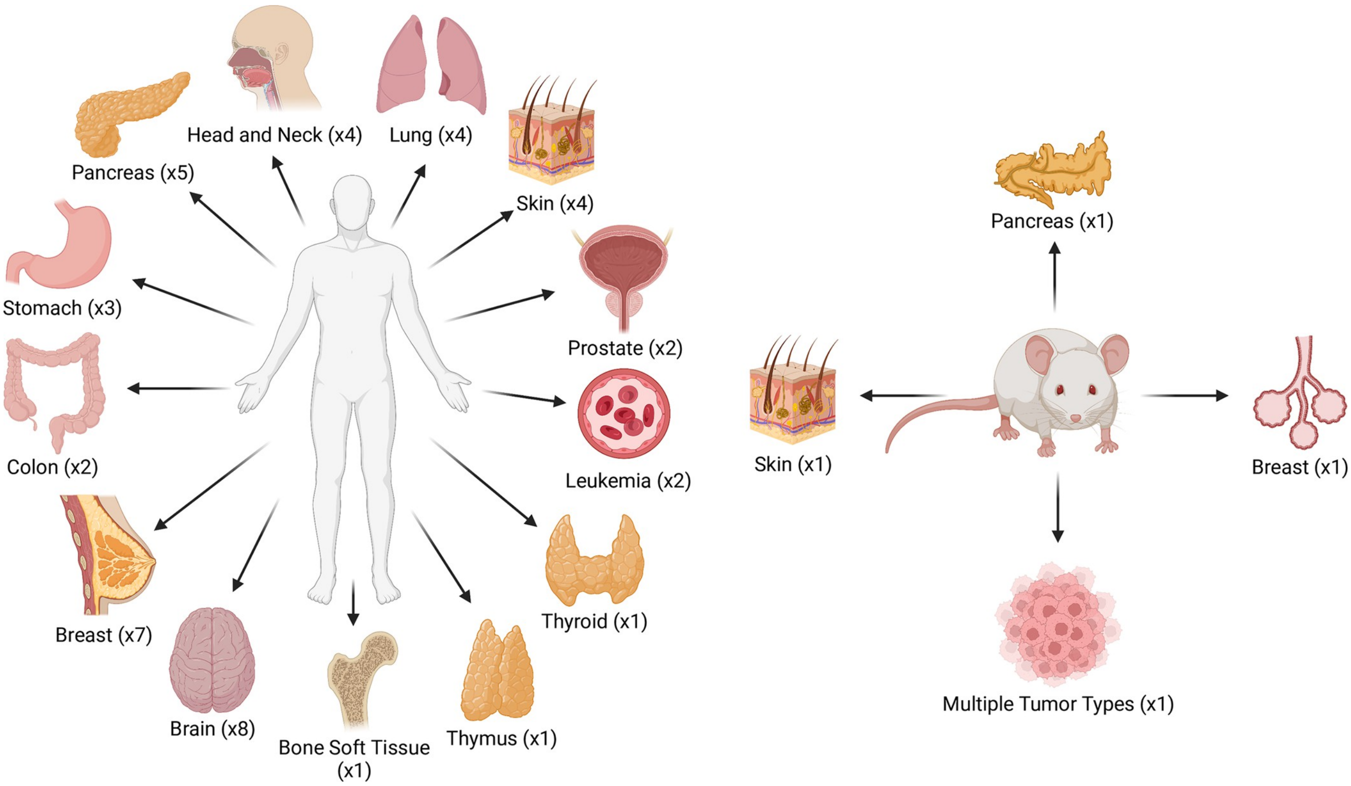


Fig 1. A visualisation of the various tissue types included in TMExplorer. TMExplorer includes 48 TME scRNA-seq datasets from 26 different human cancer types from 13 different sites and 4 different mouse cancer types. TMExplorer is generalizable and extendable, and the new datasets are added to the database as they become available. Fig 1 is created with BioRender.com.

https://doi.org/10.1371/journal.pone.0272302.g001

There are three main components to each dataset in our database: (1) gene-by-cell expression matrices; (2) cell type labels; and (3) gene signatures. The cell type labels are R *dataframes* with two columns; one contains every cell barcode present in the expression matrix, and the other one contains that cell's type. The gene signatures are stored in R *dataframes* containing one column per cell type, with a list of genes that are differentially expressed by that cell type and reported in the original paper in which each dataset was first introduced. All data for each dataset is accessible within a single object in order to make it as easy to use as possible.

Since R BioConductor has existing infrastructure for working with scRNA-seq data [58, 59], we used it as the platform to build our package upon. In order to maintain compatibility with existing Bioconductor software, we return all datasets as *SingleCellExperiment* objects [59]. Fig 2 shows the structure of a *SingleCellExperiment* object, where the expression data is stored as a named assay, cell type labels (if present) are stored under *colData()*, and all other information is stored in a metadata list.

- **Expression Data:** Named assays allow certain formats to be easily accessed with *getter* functions such as *counts()* and *tpm()*, while other formats can still be accessed with the *assay() getter* function [59]. All *SingleCellExperiments* have one assay named according to the type of score (e.g. Counts and TPM) represented in that object. Calling *assay()* returns an expression matrix with rows of genes and columns of cells.
- **Cell Type Labels:** *ColData* stores metadata for the columns in the assay matrix. In our case this refers to the cell type annotations, if they are available. *ColData* is a dataframe that

Table 2. List of tumor microenvironment scRNA-seq datasets included in TMExplorer.

Dataset	Cancer type	Sequencing Technology	Number of tumors	Number of cells	Number of genes	Annotation available?	Gene signature available?
Patel *et al*. Science 2014	Glioblastoma	SMART-seq	5 human primary glioblastoma tumors	1,456	5,796	Yes	No
Tirosh *et al*. Science 2016	Metastatic melanoma	SMART-seq 2	19 human melanoma tumors	4,645	23,686	Yes	Yes
Tirosh *et al*. Nature 2016	Oligodendroglioma	SMART-seq 2	6 human IDH-mutant oligodendroglioma tumors	4,347	23,686	Yes	Yes
Venteicher *et al*. Science 2017	Astrocytoma	SMART-seq2	10 human IDH-mutant astrocytoma tumors	6,341	23,686	No	No
Li *et al*. Nature Genetics 2017	Colorectal cancer	Fluidigm C1	11 human primary colorectal cancer tumors	375	57,241	Yes	Yes
Chung *et al*. Nature Communications 2017	Breast cancer	Fluidigm C1	11 human primary breast cancer tumors	563	57,915	Yes	Yes
Puram *et al*. Cell 2017	Head and neck squamous cell carcinoma	SMART-seq 2	18 human primary oral cavity tumors and 5 lymph node metastases	5,902	21,884	Yes	Yes
Giustacchini *et al*. Nature Medicine 2017	Chronic myeloid leukemia	SMART-seq2	20 human bone marrow aspirates	2,287	23,384	No	No
Filbin *et al*. Science 2018	H3 K27M-mutant glioma	SMART-seq2	6 human primary H3K27M-glioma tumors	4,058	23,686	Yes	No
Jerby-Arnon *et al*. Cell 2018	Melanoma	SMART-seq2	33 human melanoma tumors	7,186	23,686	Yes	Yes
VanGalen *et al*. Cell 2019	Acute myeloid leukemia	Seq-Well	40 human bone marrow aspirates	23,383	27,899	No	No
Ting *et al*. Cell Reports 2014	Pancreatic cancer	Tang Protocol	5 mice with pancreatic cancer, 1 mouse embryonic fibroblast cell line, 1 mouse pancreatic cancer cell line, 1 control mouse	187	29,018	No	No
Miyamoto *et al*. Science 2015	Prostate cancer	ABI SOLiD	18 patients with metastatic prostate cancer, 4 patients with localized prostate cancer, 12 bulk primary prostate tumors, 4 prostate cancer cell lines	169	21696	No	No
Jordan *et al*. Nature 2016	Breast cancer	Truseq	2 ER+/HER2- breast cancer patients, 14 triple negative breast cancer patients	74	23,368	No	No
Azizi *et al*. Cell 2018	Breast cancer	InDrop	8 human breast carcinomas	46,016	14,875	No	No
Lambrechts *et al*. Nature Medicine 2018	Non-small cell lung carcinoma	10x Genomics	5 human non metastatic lung squamous carcinoma tumors	51,775	22,533	Yes	Yes
Davidson *et al*. Cell Reports 2018	Melanoma	SMART-seq2	Mouse tumors	6,422	26,946	No	No
Peng *et al*. Cell Resesarch 2019	Pancreatic ductal adenocarcinoma	10x Genomics	24 human primary pancreatic ductal adenocarcinoma tumors, 11 control pancreases	57,530	24,005	Yes	Yes
Darmanis *et al*. Cell Reports 2017	Glioblastoma	Smart-seq2	4 human glioblastoma tumors	3,589	23,465	No	No
Kumar *et al*. Cell Reports 2018	Mixed cancer: Melanoma, breast mammary carcinoma, Lewis lung carcinoma, colon carcinoma, fibrosarcoma	10x Genomics	1 mouse melanoma tumor, 1 mouse breast mammary carcinoma tumor, 1 mouse Lewis lung carcinoma tumor, 2 different mouse colon carcinoma tumors, 1 mouse fibrosarcoma tumor	10,473	27,998	No	No

(*Continued*)

Table 2. (Continued)

Dataset	Cancer type	Sequencing Technology	Number of tumors	Number of cells	Number of genes	Annotation available?	Gene signature available?
Zhao *et al*. BMC Medical Genomics 2019	Glioblastoma	Fluidigm C1	1 human glioblastoma cancer cell line, 1 normal neural stem cell line	134	21,209	No	No
Chen *et al*. Cell Research 2020	Nasopharyngeal carcinoma	10x Genomics	15 human nasopharyngeal carcinoma tumors	48,584	24,720	Yes	No
Lin *et al*. Genome Medicine 2020	Pancreatic ductal adenocarcinoma	10x Genomics	16 human pancreatic ductal adenocarcinoma tumors	14,926	22,217	No	Yes
Gillen *et al*. Cell Reports 2020	Ependymoma	10x Genomics	26 human ependymoma tumors	18,500	23,580	Yes	No
Zhang *et al*. Cell Reports 2019	Gastric cancer	10x Genomics	13 human gastric tumors	56,440	22,910	No	No
Yeo *et al*. Elife 2020	Breast cancer	10x Genomics	4 mouse breast cancer tumours	13,745	31,053	No	No
Gao *et al*. Nature Biotechnology 2019	Anaplastic thyroid cancer	10x Genomics	5 human anaplastic thyroid tumors	19,568	33,540	No	No
Gao *et al*. Nature Biotechnology 2019	Breast ductal carcinoma	10x Genomics	1 human breast ductal carcinoma tumor	1,480	33,694	No	No
Gao *et al*. Nature Biotechnology 2019	Triple negative breast cancer	10x Genomics	3 human triple negative breast cancer tumors	2,663	33,964	No	No
Gao *et al*. Nature Biotechnology 2019	Triple negative breast cancer	10x Genomics	2 human triple negative breast cancer tumors	6,281	33,538	No	No
Gao *et al*. Nature Biotechnology 2019	Breast invasive ductal carcinoma	10x Genomics	2 human breast invasive ductal carcinoma tumors	6,209	33,540	No	No
Paulson *et al*. Nature Communications 2018	Merkel cell carcinoma	10x Genomics	2 human primary merkel cell carcinoma tumors	25,066	11,072	No	No
Bautista *et al*. Nature Communications 2021	Thymic cancer	10x Genomics	7 human primary thymic cancer tumors	74,780	33,694	No	Yes
Paulson *et al*. Nature Communications 2018	Merkel cell carcinoma	10x Genomics	2 primary merkel cell carcinoma tumors from 1 human patient at 2 timepoints	7,432	21,861	No	No
Kim *et al*. Genome Biology 2015	Lung adenocarcinoma	SMART-seq	2 primary human lung adenocarcinoma tumors	201	57,820	No	No
Aynaud *et al*. Cell Reports 2020	Ewing sarcoma	10x Genomics	3 Ewing sarcoma patient-derived xenografts samples	97	56,764	No	No
Song *et al*. Nature Communications 2022	Prostate cancer	Seq-Well S^3	6 prostate biopsies from 3 different patients, 4 radical prostatectomies with tumor-only samples from 4 patients, and 4 radical prostatectomies with matched normal samples from 4 patients	53765	19,665	No	Yes
Liu *et al*. Nature Communications 2021	Nasopharyngeal carcinoma	10x Genomics	10 human nasopharyngeal carcinoma tumor-blood paired samples	176,447	20,930	No	Yes
Kurten *et al*. Nature Communications 2021	Head and neck squamous cell carcinoma	10x Genomics	18 primary human head and neck squamous cell carcinoma tumors	61,221	33,545	No	Yes
Gojo *et al*. Cancer Cell 2020	Ependymoma	SMART-seq2	20 fresh surgical tumor specimens from 18 ependymoma patients, eight patient-derived cell models, and two patient-derived xenograft models	6,739	20,447	Yes	Yes

(*Continued*)

Table 2. (Continued)

Dataset	Cancer type	Sequencing Technology	Number of tumors	Number of cells	Number of genes	Annotation available?	Gene signature available?
Zhang *et al*. Cell 2020	Colon cancer	SMART-seq2	18 primary human colorectal cancer tumors	43,817	13,538	No	Yes
Steele *et al*. Nature Cancer 2021	Pancreatic ductal adenocarcinoma	10x Genomics	16 primary human pancreatic ductal adenocarcinoma tumors	55,652	32,738	No	No
Lee *et al*. Clinical Cancer Research 2021	Pancreatic ductal adenocarcinoma	10x Genomics	16 metastatic human pancreatic ductal adenocarcinoma tumors	17,889	33,694	No	Yes
Moncada *et al*. Nature Biotechnology 2020	Pancreatic ductal adenocarcinoma	inDROP	11 primary human pancreatic cancer tumors	19,738	4,343	No	No
Wu *et al*. Nature Communications 2021	Non-small cell lung cancer	10x Genomics	42 primary human non-small cell lung cancer tumors	89,887	29,527	No	No
Kim *et al*. NPJ Precision Oncology 2022	Gastric cancer	10x Genomics	47 patient biopsies consisting of 24 gastric cancer lesions and 23 adjacent normals	13,113	8,705	No	No
Kumar *et al*. Cancer Discovery 2022	Gastric cancer	10x Genomics	48 primary human gastric cancer tumors	158,641	26,571	No	Yes
Kim *et al*. Nature Communications 2020	Lung adenocarcinoma	10x Genomics	11 tumour, 11 distant normal lung, 10 normal lymph node, and 10 metastatic brain tissue samples from patients without prior treatment. 7 metastatic lymph node and 4 lung tumour tissue samples from advanced stage patients.	208,506	29,634	Yes	Yes

https://doi.org/10.1371/journal.pone.0272302.t002

always has one row for every column in the assay matrix, ensuring that there is a label present for every cell. If the cell type is not available for a given cell, it is labelled as "unknown".

- **Metadata:** The metadata list serves to store any other information that does not fit into a preexisting attribute of the *SingleCellExperiment* object, and is accessed with the *metadata()* function. This named list contains the signature gene sets, available score types, tumour and host organism type, sequencing technology, author, and all other descriptive information as strings. All information that is available in the metadata table can be accessed by calling the query function of TMExplorer (i.e. *queryTME)* with the *metadata_only* parameter set to true.

Metadata

After collecting the datasets, corresponding metadata was compiled into a table which serves as the core of the package (S1 Fig). The metadata table contains information such as GEO accession, author, journal, year, PMID, sequencing technology, expression score type(s), source organism, type of cancer, number of patients, tumours, cells and genes, and the database that the data was obtained from (S1 Table). All items in the metadata table were chosen as either entities that distinguish one dataset from the others or criteria that may make a dataset or group of datasets interesting to researchers (e.g. a specific tumour type or availability of cell labels or gene signatures). Users can view the available data using the metadata table and decide which dataset best fits their needs.

Database query

TMExplorer provides a query function (i.e. *queryTME)* that users can employ in order to select multiple datasets based on their desired characteristics (S2 Fig). For example, users can select

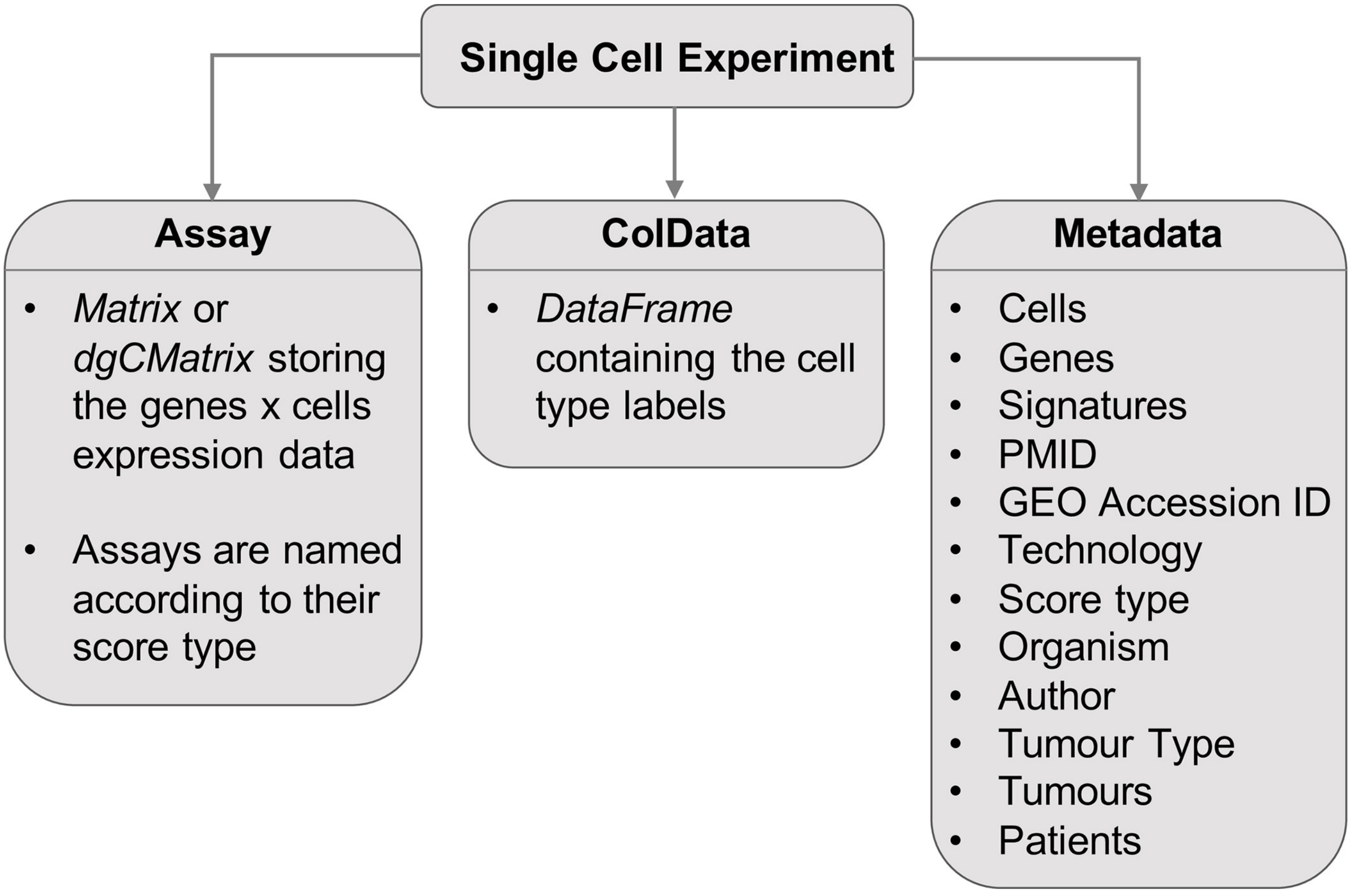


Fig 2. The format of the *SingleCellExperiment* objects containing TME datasets. The *Assay* is a matrix or *dgCMatrix* containing the gene expression table, named according to the type of score (i.e. an Assay containing raw counts would be named "Counts"); *colData* is a *DataFrame* with the number of rows equal to the number of columns in the Assay and describes the cells in the dataset; Metadata is a named list of additional metadata objects describing the dataset. A *SingleCellExperiment* object may contain one or more *AltExps*, which are nested *SingleCellExperiment* objects containing a different score type in the Assay.

https://doi.org/10.1371/journal.pone.0272302.g002

specific studies by PMID or GEO accession, or filter subsets by sequencing technology, whether cell type labels or cell type signature gene sets are available etc. Sequencing technology, score type, organism, tumour type, and year were all chosen as search parameters because they represent differences in the type of data and make it easier to find data that fits the needs of different studies. Some datasets may publish multiple tumour types under the same study. TMExplorer is able to handle this by having multiple rows of different datasets from the same study. In these cases, users will need to provide multiple search parameters to select a single row, for instance the GEO accession and tumour type for a study that contains multiple cancers. We have also made it possible to search for datasets for which the cell labels and gene signatures are available. This facilitates developing and testing algorithms that require specific types of dataset information. For example, testing cell classification algorithms requires cell labels that can be used as a gold standard, and many existing algorithms require gene signatures that represent the cell types in the dataset [60, 61].

Alternative experiments

For several datasets, gene expressions are available in multiple score types including raw counts and normalized data by FPKM, TPM or CPM. In order to store each dataset in multiple score types, we used nested *SingleCellExperiments* objects with the alternative experiments (*altExps*) concept. Alternative experiments are guaranteed to have the same dimensions as the primary object, but can be kept separate for use in other pipelines [59]. This allows users to download multiple types of scoring for use in different steps of analysis while still being able to access each dataset through a single object. Being able to download multiple score types allows our datasets to be used in a variety of algorithms that require a specific type of score, and keeping them separated as nested objects prevents accidentally applying an algorithm to the wrong score type.

Dense vs. sparse data formats

In order to reduce the memory requirements for working with large datasets, expression data is optionally available as a sparse matrix. We implemented sparse matrices using the *dgCMatrix* class from R Matrix [62]. This reduces memory usage by only storing non-zero expression values. With sparse matrices, the memory required to store a dataset is reduced by as much as 8 Gb for a dataset with 51,775 cells and 22,533 genes. It should be noted that not all software packages are compatible with sparse matrices, and converting large datasets from sparse to dense may crash R on machines with low memory. Thus users should confirm that their algorithms support sparse matrices before using them. By default, TMExplorer returns dense matrices to avoid these problems.

Exporting data in multiple formats

Several tools for scRNA-seq analysis are written in R and therefore a *SingleCellExperiment* object can easily be incorporated into these pipelines and tools. However, many other analysis tools are written in Python or as webapps [18, 63, 64]. To facilitate the use of TMExplorer with these tools, we wrote a function *saveTME* that writes individual TME datasets to disk as CSV or Matrix Market files, depending on whether data was loaded as dense or sparse matrices by *queryTME*, respectively. *SaveTME* takes a *SingleCellExperiment* object and a path to an output directory as parameters and saves the gene expression matrix, cell type labels, and cell type signature gene sets to disk. The resulting files can then be converted as needed and used in other applications.

Adding new datasets

We keep TMExplorer updated with new datasets as they get published. Additionally, users of the package doing their own novel research will have access to an issue template on Github where they can submit their data for inclusion. The interested users will need to provide their scRNA-seq data as raw counts or normalized data and the corresponding metadata. Since TMExplorer is open source, those same users can create a fork of the repository and build it from source with their own data for pre-publication work. Users wishing to fork the repository for their own data need only replace or add to the metadata table used by the package and update any documentation or function names to reflect the new data. If users are adding new TME data, no functions need to be changed since this package already uses TME data. Those users who are interested in adopting the package for other types of single-cell sequencing data (such as sc-ATAC seq) can do so by changing documentation and functions to reflect the new data type.

Results

Overview of the TMExplorer package

To make it as easy as possible to integrate TMExplorer into other pipelines, all interactions with the package are done directly in R. Here, the *queryTME* function serves as the primary

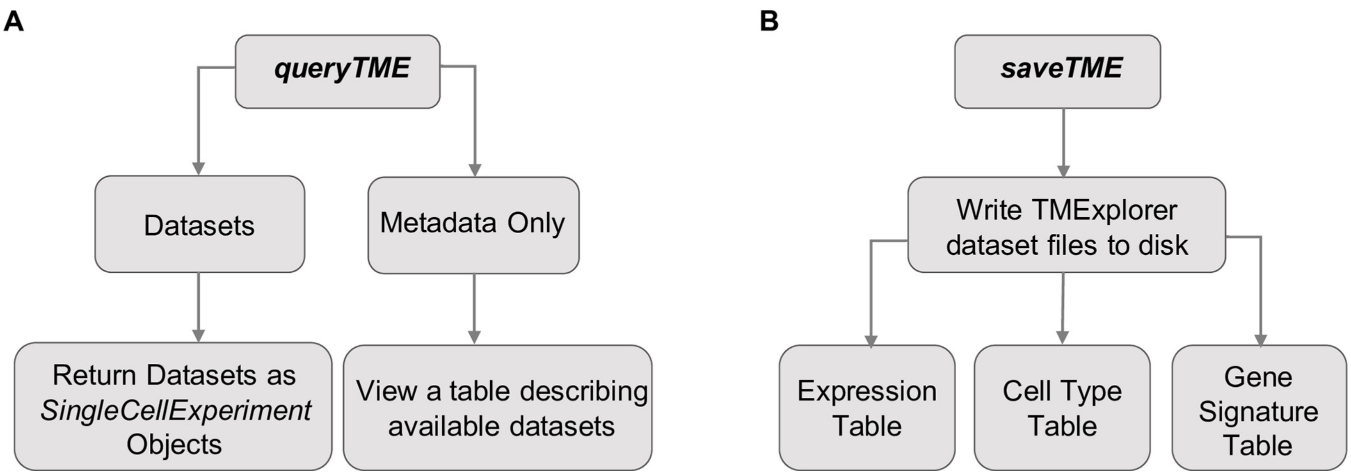

A
queryTME
Datasets
Metadata Only
Return Datasets as *SingleCellExperiment* Objects
View a table describing available datasets
B
saveTME
Write TMExplorer dataset files to disk
Expression Table
Cell Type Table
Gene Signature Table

Fig 3. An overview of the main functions of TMExplorer. A. *queryTME* allows users to search and return datasets in either a descriptive table or as a list of *SingleCellExperiment* objects for analysis. **B.** *saveTME* allows users to write datasets to disk. For each dataset written to disk, up to three files are created; a table storing the expression data as either a CSV or matrix market file, depending on whether a dense or sparse matrix is passed to the function; a table containing the cells and their truth label, if available; and a table containing the cell type signature gene sets, if available.

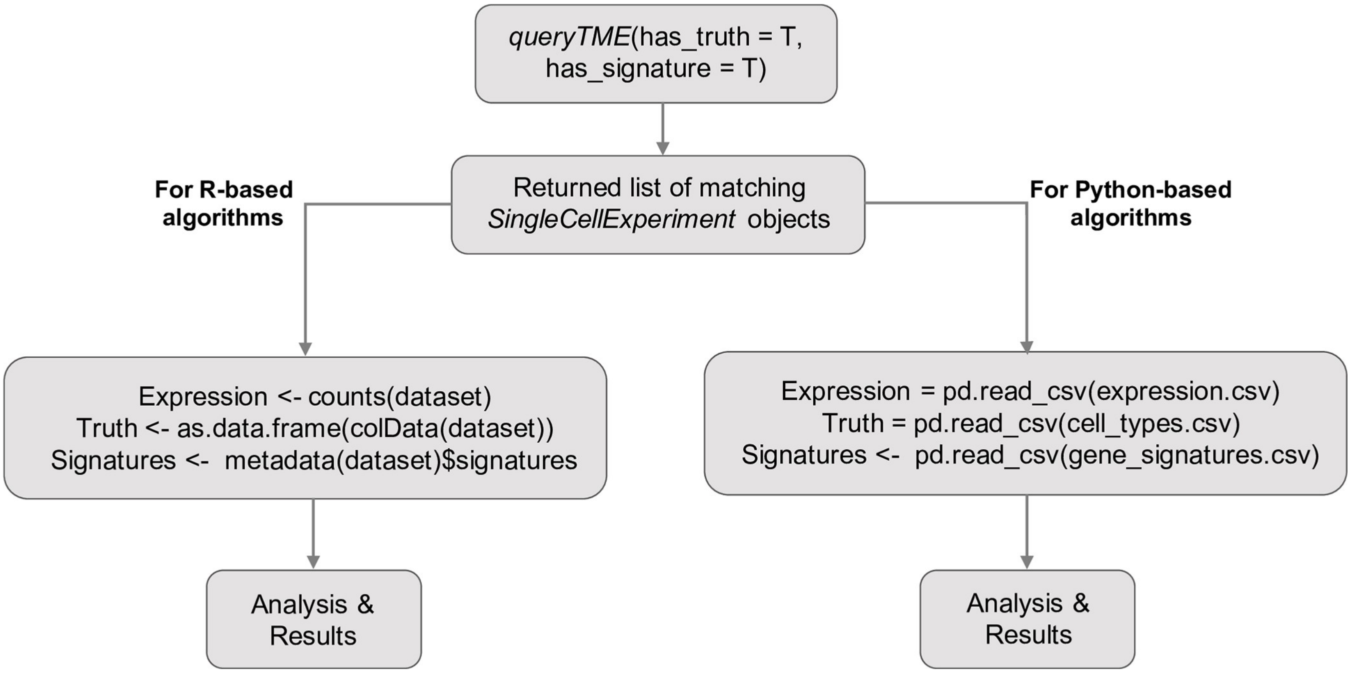
https://doi.org/10.1371/journal.pone.0272302.g003

interface for the package, allowing users to view the metadata for all available datasets, or select a subset of datasets according to descriptive criteria (Fig 3A). *queryTME* provides a set of parameters (Table 1) used to select a subset of datasets according to characteristics. To review the available datasets, the *metadata_only* parameter should be set to *TRUE* when querying the package, and a table describing the datasets will be returned instead of the datasets themselves. The search parameters can be used to find relevant data without requiring users to review the metadata table first, lowering the barrier for use. For example, users looking for a certain type of cancer, such as melanoma, can search using *queryTME(tumour_type =“Melanoma”)* without needing to first examine the metadata for datasets containing melanoma cancers.

After querying the database, a list of *SingleCellExperiment* objects is returned. The objects in this list can then be passed to any other algorithms that accept a *SingleCellExperiment* object, sparse *dgCMatrix*, or dense gene expression matrix for inclusion in a pipeline (Fig 4). Alternatively, the *saveTME* function can be used to write the returned data to disk for further manipulation or use in applications outside of R (Fig 3B). Fig 4 shows how *saveTME* can be used to save data for analysis in Python. In order to maintain consistency, the returned value is always a list of results, whether or not multiple datasets match the query.

TMExplorer database contents

TMExplorer is a curated collection of TME scRNA-seq datasets that have been made available as an R-package. We created TMExplorer to improve accessibility and sharing of tumour scRNA-seq data. It acts as a single-entry point to various tumour scRNA-seq datasets for users interested in studying gene expression of the TME at the single-cell level. Fig 5 provides a summary of TMExplorer contents at the time of publication. Currently, the collection contains 48 datasets, including 44 datasets derived from human tumours and 4 datasets derived from mouse tumours (Fig 5A). This comprises 28 different cancer types including leukemia [22, 24], breast [2, 27, 28, 36], colorectal [3, 45], glioblastoma [7, 9, 31], glioma [23], head and neck [20, 47, 48], astrocytoma [21], oligodendroglioma [8], melanoma [1, 5, 19], lung carcinoma [6, 38, 51], non-small cell lung carcinoma [41], pancreatic [25, 29, 33, 42–44] prostate [26, 49], gastric [39, 40], merkel cell carcinoma [53], thymic [52], ewing sarcoma [50], ependymoma



```
queryTME(has_truth = T,
has_signature = T)
```

```
Expression <- counts(dataset)
Truth <- as.data.frame(colData(dataset))
Signatures <-  metadata(dataset)$signatures
```

```
Expression = pd.read_csv(expression.csv)
Truth = pd.read_csv(cell_types.csv)
Signatures <-  pd.read_csv(gene_signatures.csv)
```

Fig 4. An example workflow of using TMExplorer to obtain datasets for the downstream analysis using Python and R. Users start by using *queryTME* to return all datasets that have cell type labels and cell type signature gene sets, which will get a list of matching datasets contained in *SingleCellExperiment* objects. Then, for R based algorithms, users can pass the *SingleCellExperiments* directly if that is supported, or users can pass the individual components required. For Python based algorithms, *saveTME* can be used to save the files for each dataset to disk, which can then be opened in Python for analysis.

https://doi.org/10.1371/journal.pone.0272302.g004

[46] (Fig 5E). 13 out of 28 cancer types have more than one associated dataset (Fig 5E). Also, 6 out of 48 datasets are from rare cancers (incidence rate of < 6 in a million persons), including merkel cell carcinoma [53], thymic carcinoma [52], ewing sarcoma [50], astrocytoma [21], oligodendroglioma [8], and ependymoma [46]. Numbers of cells and genes vary across datasets and fall within the range of 4,343–57,915 genes and 74–208,506 cells (Fig 5F and S1 Table). The datasets are sequenced by different sequencing technologies including 10x Genomics, SMART-seq2 and Fluidigm C1 (Fig 5B). Each dataset is provided as processed gene expression data, and are provided either as raw counts or normalized data (e.g. TPM, RPKM, and RPM) (Fig 5D). We did not include raw scRNA-seq data (i.e. FASTQ files) in our collection because these files tend to be very large and can be accessed through the SRA, if available. Out of the 48 datasets, cell-type annotations are also provided for 16 datasets and gene signature information is provided for 18 datasets (Fig 5C), so that users may access and use this information in their analyses. Also, for 10 datasets both cell type annotations and gene signatures are available (Fig 5C). Users can browse through the available datasets using the metadata table and then choose which dataset(s) they would like to analyze. Users can also save the datasets for use outside of R, for instance in Python or web-based analysis pipelines.

TMExplorer search capability

An important feature of TMExplorer is that it acts as both a database and search tool that can be easily implemented in one's own workflow. Some other currently available scRNA-seq databases have a search function, but cannot be easily integrated into workflows because they are web-based [12–15]. Currently available R-based scRNA-seq databases lack built in search tools, requiring users to access vignettes to see the available data before it can be retrieved for

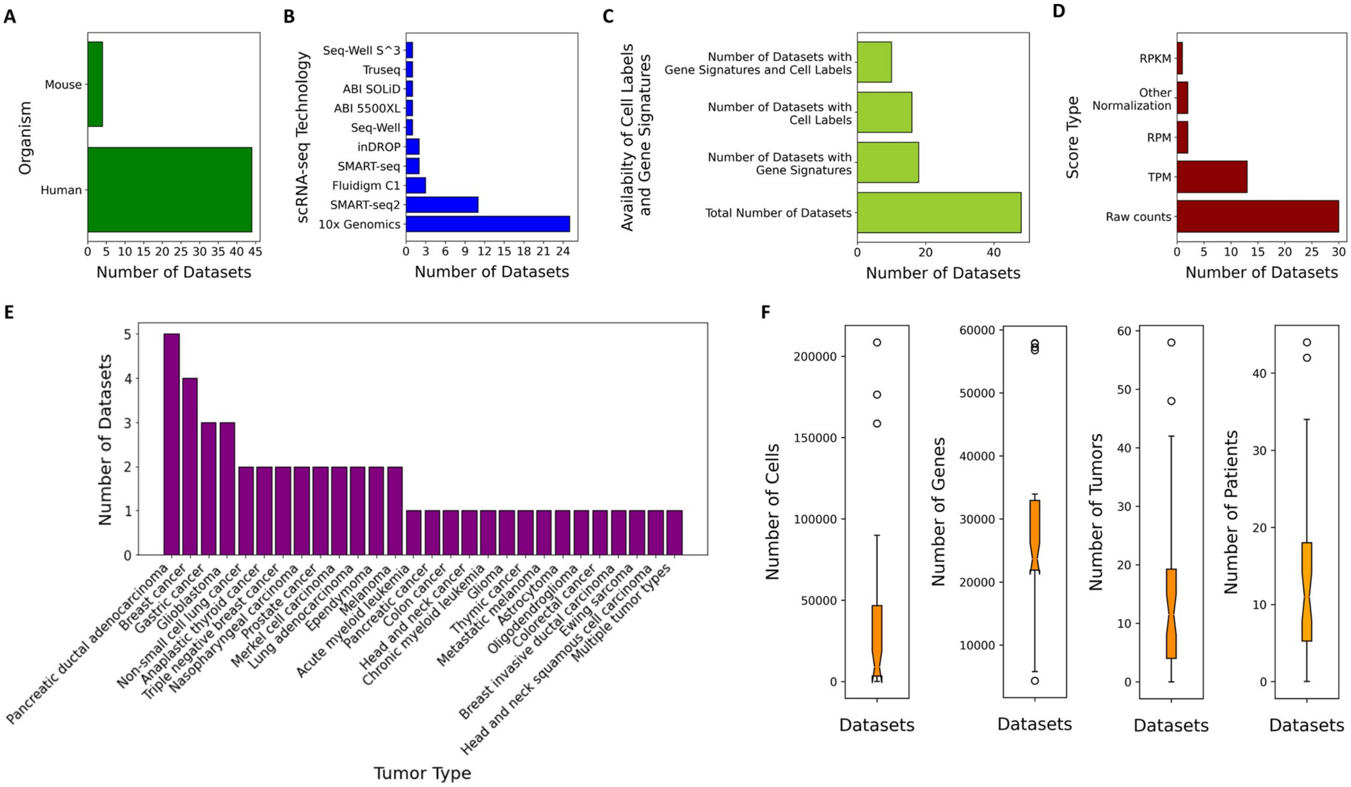


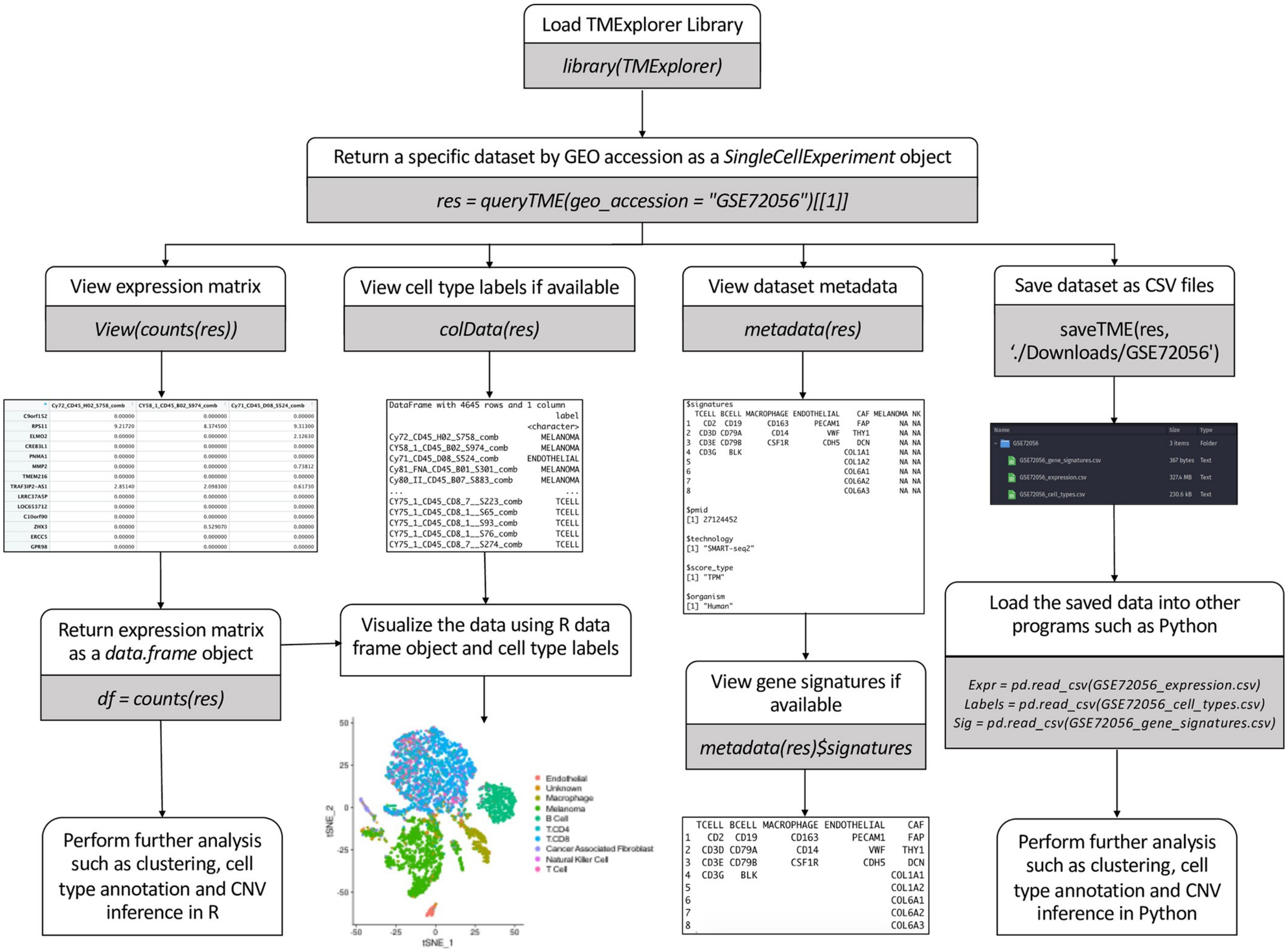
Fig 5. A summary of TMExplorer contents. Here, we provide a summary of the number of humans and mice datasets in TMExplorer (A); the number of datasets generated by various sequencing technologies (B); the number of datasets for which cell type labels and gene signatures are available (C); and the distributions of score types of different datasets (D) and the tumour types (E). In addition, boxplots of the number of cells, genes, tumours and patients across different datasets are provided (F).

https://doi.org/10.1371/journal.pone.0272302.g005

use in a pipeline [16]. TMExplorer provides a search tool that allows users to search for datasets that fit their needs by tumour type, sequencing technology, source organism, and more (Table 1), all from the R command line. This makes TMExplorer an improvement over both R-based and web-based databases because users are able to browse and query data from the same console they are using for analysis. By including a search tool and database in a single package, TMExplorer provides a single point of entry to include TME scRNA-seq data into data analysis pipelines. In Fig 6, we provide a flowchart that shows various steps involved in querying TMExplorer, obtaining the datasets of interest, saving them on the local machines, and performing further analysis in R or other programs.

Case studies

In this section, we bring two example applications where TMExplorer can be used to facilitate data analysis. In the *case study 1*, we show how TMExplorer can be combined with automated cell-type identification algorithms to identify different cell types in TME scRNA-seq data. Here, we also show how users can return datasets with both the signature gene sets and gold standard annotations needed for testing cell-type identification. In *case study 2*, we show how TMExplorer can be integrated with the algorithms for inferencing copy-number variations in individual cells and facilitate the separation of malignant and non-malignant cells in multiple tumour scRNA-seq datasets of the same cancer type.


Load TMExplorer Library
Return a specific dataset by GEO accession as a *SingleCellExperiment* object
View expression matrix
View cell type labels if available
View dataset metadata
Save dataset as CSV files
Return expression matrix as a *data.frame* object
Visualize the data using R data frame object and cell type labels
View gene signatures if available
Load the saved data into other programs such as Python
Perform further analysis such as clustering, cell type annotation and CNV inference in R
Perform further analysis such as clustering, cell type annotation and CNV inference in Python
Endothelial
Unknown
Macrophage
Melanoma
B Cell
T.CD4
T.CD8
Cancer Associated Fibroblast
Natural Killer Cell
T Cell
tSNE_1
tSNE_2
50
25
0
-25
-50
Name
Size
Type

```
library(TMExplorer)
```

```
res = queryTME(geo_accession = "GSE72056")[[1]]
```

```
View(counts(res))
```

	Cy72_CD45_H02_S758_comb	CY58_1_CD45_B02_S974_comb	Cy71_CD45_D08_S524_comb
C9orf152	0.00000	0.000000	0.00000
RPS11	9.21720	8.374500	9.31300
ELMO2	0.00000	0.000000	2.12630
CREB3L1	0.00000	0.000000	0.00000
PNMA1	0.00000	0.000000	0.00000
MMP2	0.00000	0.000000	0.73812
TMEM216	0.00000	0.000000	0.00000
TRAF3IP2-AS1	2.85140	2.098300	0.61730
LRRC37A5P	0.00000	0.000000	0.00000
LOC653712	0.00000	0.000000	0.00000
C10orf90	0.00000	0.000000	0.00000
ZHX3	0.00000	0.529070	0.00000
ERCC5	0.00000	0.000000	0.00000
GPR98	0.00000	0.000000	0.00000

```
df = counts(res)
```

```
colData(res)
```

```
DataFrame with 4645 rows and 1 column
                                     label
                               <character>
Cy72_CD45_H02_S758_comb           MELANOMA
CY58_1_CD45_B02_S974_comb         MELANOMA
Cy71_CD45_D08_S524_comb        ENDOTHELIAL
Cy81_FNA_CD45_B01_S301_comb       MELANOMA
Cy80_II_CD45_B07_S883_comb        MELANOMA
...                                    ...
CY75_1_CD45_CD8_7__S223_comb         TCELL
CY75_1_CD45_CD8_1__S65_comb          TCELL
CY75_1_CD45_CD8_1__S93_comb          TCELL
CY75_1_CD45_CD8_1__S76_comb          TCELL
CY75_1_CD45_CD8_7__S274_comb         TCELL
```

```
metadata(res)
```

```
$signatures
  TCELL BCELL MACROPHAGE ENDOTHELIAL    CAF MELANOMA NK
1   CD2  CD19      CD163      PECAM1    FAP       NA NA
2  CD3D CD79A       CD14         VWF   THY1       NA NA
3  CD3E CD79B      CSF1R        CDH5    DCN       NA NA
4  CD3G   BLK                        COL1A1       NA NA
5                                    COL1A2       NA NA
6                                    COL6A1       NA NA
7                                    COL6A2       NA NA
8                                    COL6A3       NA NA

$pmid
[1] 27124452

$technology
[1] "SMART-seq2"

$score_type
[1] "TPM"

$organism
[1] "Human"
```

```
metadata(res)$signatures
```

```
  TCELL BCELL MACROPHAGE ENDOTHELIAL    CAF
1   CD2  CD19      CD163      PECAM1    FAP
2  CD3D CD79A       CD14         VWF   THY1
3  CD3E CD79B      CSF1R        CDH5    DCN
4  CD3G   BLK                        COL1A1
5                                    COL1A2
6                                    COL6A1
7                                    COL6A2
8                                    COL6A3
```

```
saveTME(res,
'./Downloads/GSE72056')
```

Name	Size	Type
GSE72056	3 items	Folder
GSE72056_gene_signatures.csv	367 bytes	Text
GSE72056_expression.csv	327.4 MB	Text
GSE72056_cell_types.csv	230.6 kB	Text

```
Expr = pd.read_csv(GSE72056_expression.csv)
Labels = pd.read_csv(GSE72056_cell_types.csv)
Sig = pd.read_csv(GSE72056_gene_signatures.csv)
```

Fig 6. A flowchart of data query and analysis using TMExplorer. TMExplorer provides a search and analysis capability, where users can look up and return their datasets of interest, view the expression matrix, cell type labels and metadata including gene signatures (if available) and continue by either using R for data visualization and analysis, or save the datasets in CSV format to be analyzed by their programming language of choice (e.g. Python).

https://doi.org/10.1371/journal.pone.0272302.g006

Case study 1: Identifying different cell types in TME scRNA-seq data. Often, when using TME scRNA-seq data, we are interested in the cellular composition of the dataset. In order to find this, automated cell type identification algorithms are used. This is usually done by first clustering the cells, and then assigning appropriate cell type labels to each cluster [65]. In Fig 7, we show how TMExplorer can be combined with a clustering method (e.g. Seurat [66]) and a cluster labelling method (e.g. GSVA [61]) to create a workflow for the identification of cell populations within a dataset. Seurat requires only the gene expression matrix to perform clustering, but GSVA requires a list of cell-type signature gene sets in addition to the expression matrix. TMExplorer can return all of the datasets that have signature gene sets available using *queryTME(has_signatures = TRUE)*. If after identifying the cell types within a dataset, users want to assess the performance of their workflow by comparing the automated annotations to those reported alongside the dataset, the *has_truth = TRUE* parameter can be added to *queryTME* to only return datasets that have gold standard labels available. Seurat and GSVA can be replaced by any other tool that accepts a *SingleCellExperiment* object or a matrix of gene expression values, providing flexibility for users to incorporate TMExplorer into their own workflows.

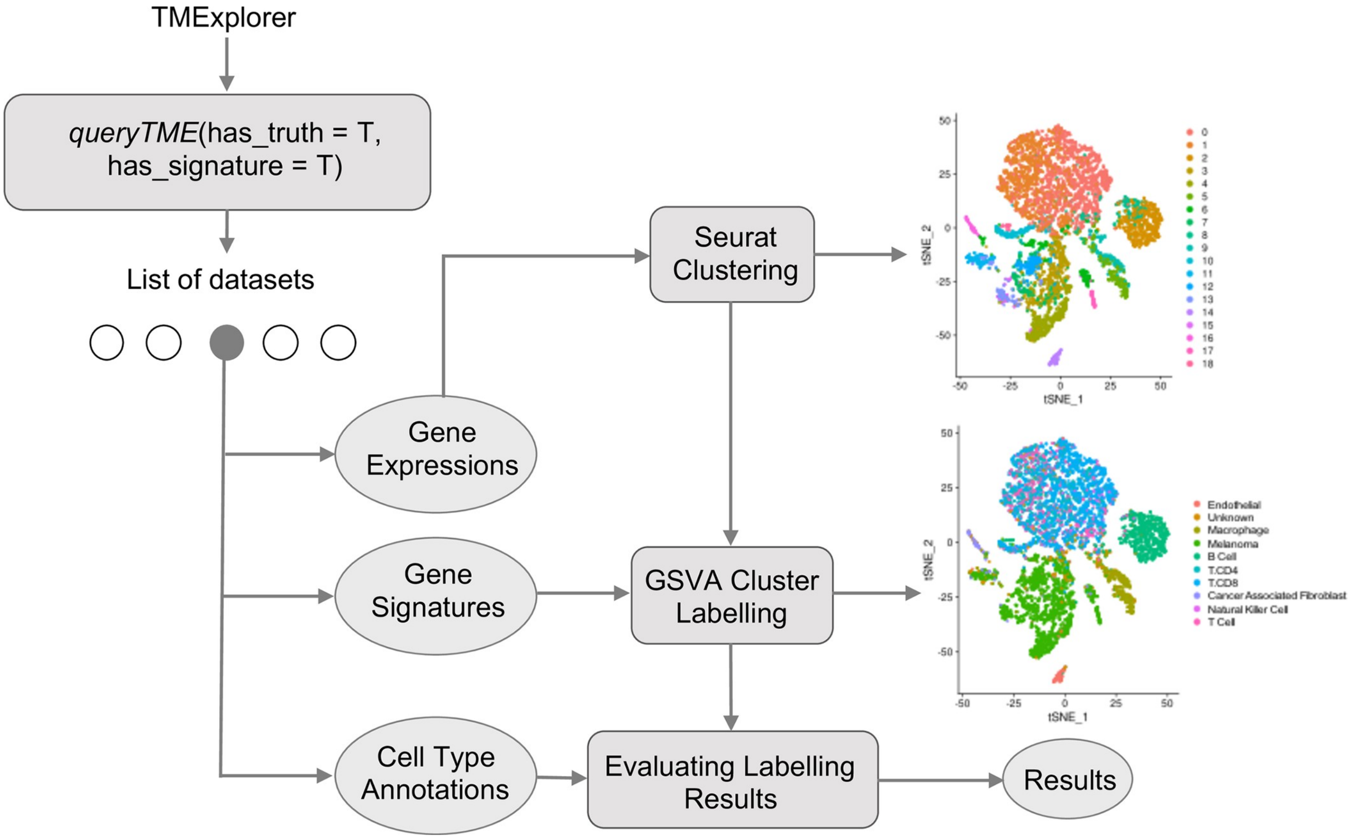
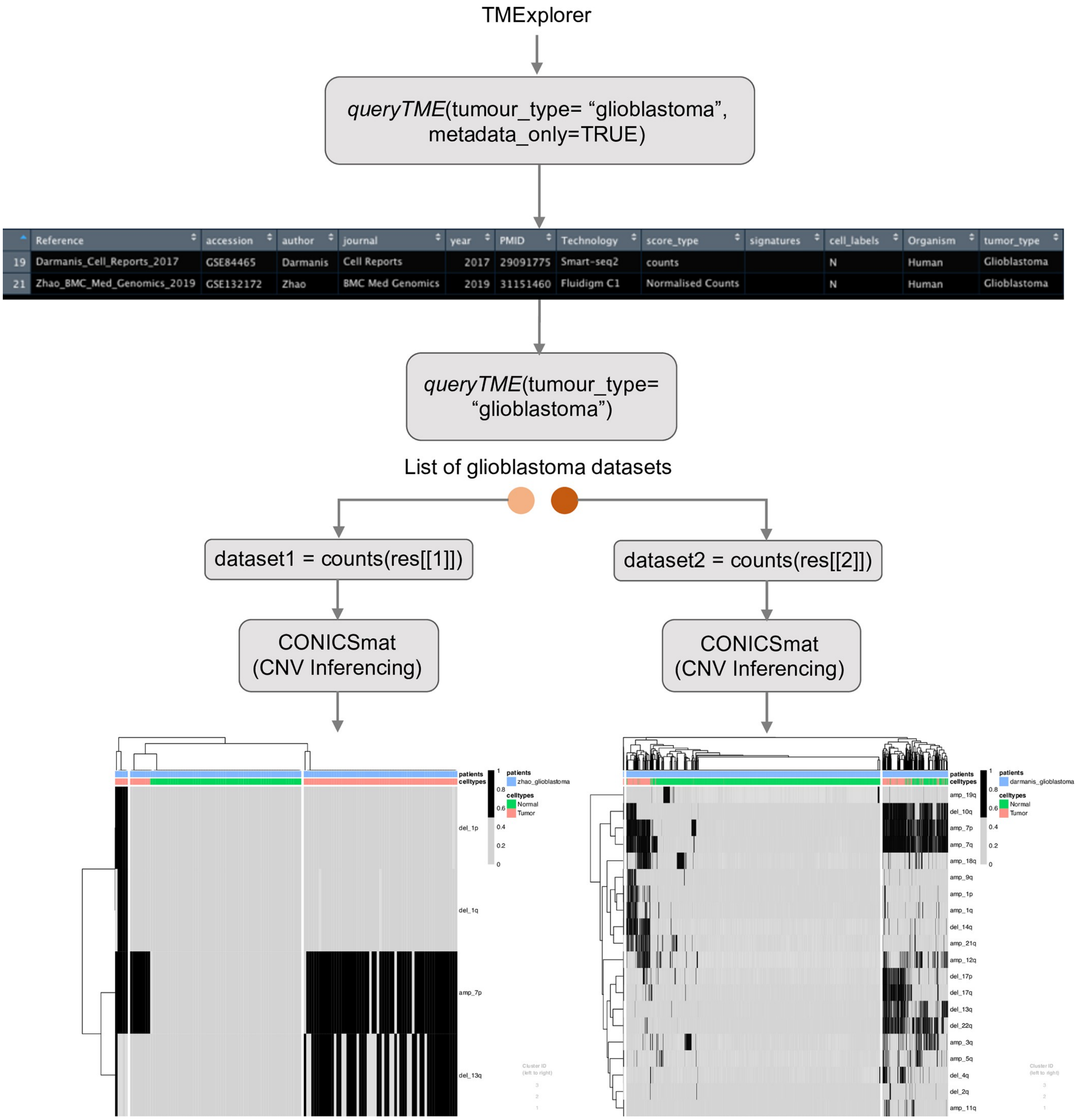


Fig 7. A case study on using TMExplorer to identify cell types. A case study showing how TMExplorer can be used in order to obtain datasets for cell cluster labelling via Seurat and GSVA. *queryTME* can be used to return those datasets which have both gene signatures and cell type annotations required for testing the automated identification of cell types. The expression data can be passed to Seurat for cell clustering, and the gene signatures can be used by GSVA to identify the cell types in Seurat's clusters. Finally, the cell type annotations can be used as the truth labels to measure the performance of the results obtained by Seurat clustering followed by GSVA.

https://doi.org/10.1371/journal.pone.0272302.g007

Case study 2: Inferencing copy number variations in multiple datasets of the same cancer type. Single cell sequencing is an important tool that enables the dissection of TMEs into malignant and non-malignant cells. Researchers interested in comparing the tumour composition across different datasets of a specific cancer type would have to collect datasets from different sources prior to application of separation methods. With TMExplorer, users can easily access multiple datasets of a specific tumour type, as well as the accompanying cell type annotations and/or gene signature information from one location, thus avoiding inconsistencies when acquiring data from different databases. TMExplorer can be easily incorporated with other packages into workflows for the analysis of scRNA-seq data, therefore enabling users to access and use the data entirely within R.

Fig 8 displays an example workflow that uses *queryTME(tumour_type = "Glioblastoma")* to retrieve datasets of a specific cancer type (i.e. glioblastoma) for use in the downstream analysis. In this example, we retrieved glioblastoma datasets from the TMExplorer database as *SingleCellExperiment* objects and converted them to gene expression count data matrices. We then applied a copy number variation (CNV) inferencing method called CONICSmat [67] to each of the datasets individually, and generated heatmaps displaying the inferred CNV patterns. This allowed us to separate malignant and non-malignant cells considering their long-range CNV patterns. The proportion of malignant and non-malignant cells and the patterns of CNV across the different datasets can then be compared.



TMExplorer

queryTME(tumour_type= “glioblastoma”, metadata_only=TRUE)

	Reference	accession	author	journal	year	PMID	Technology	score_type	signatures	cell_labels	Organism	tumor_type
19	Darmanis_Cell_Reports_2017	GSE84465	Darmanis	Cell Reports	2017	29091775	Smart-seq2	counts		N	Human	Glioblastoma
21	Zhao_BMC_Med_Genomics_2019	GSE132172	Zhao	BMC Med Genomics	2019	31151460	Fluidigm C1	Normalised Counts		N	Human	Glioblastoma

queryTME(tumour_type= “glioblastoma”)

List of glioblastoma datasets

dataset1 = counts(res[[1]])

dataset2 = counts(res[[2]])

CONICSmat (CNV Inferencing)

CONICSmat (CNV Inferencing)

Fig 8. A case study on using TMExplorer for inferring CNVs. A case study showing how TMExplorer can be used to obtain multiple datasets for a specific tumour type, to be used with CNV-based separation methods, such as CONICSmat. *QueryTME* returns datasets of a specific tumour type, such as Glioblastoma. These datasets can then be inputted directly into large-scale CNV inferencing methods, such as CONICSmat.

https://doi.org/10.1371/journal.pone.0272302.g008

Discussion

The emergence of single-cell RNA sequencing has enabled the study of tumour composition and phenotype. With the increasing use of scRNA-seq in cancer research, scRNA-seq data from TMEs continues to be generated and published. In order to streamline the data collection process for researchers interested in studying the TME, we created a curated database of TME scRNA-seq datasets, made available as an R-package called TMExplorer. Here we have built a database using a variety of cancers from multiple sources. We searched NCBI [11, 55] for TME scRNA-seq datasets that contain gene expression data, as well as comprehensive metadata such as tumour type, sequencing technology, cell type annotations, and gene signatures. In total, 48 datasets representing 26 different human cancer types and 4 different mouse cancer types are represented, along with their cell type annotations and cell type signature gene sets if they were available.

TMExplorer addresses a gap in currently available scRNA-seq databases by providing a focused, easily accessible database as an R package. TMExplorer has several advantages over other currently available scRNA-seq databases, the most prominent being:

1. Existing curated scRNA-seq databases consist of mostly normal tissue or non-cancer data and relatively few cancer datasets. To allow researchers to easily locate and access TME scRNA-seq data, we curated publicly available TME datasets and made them available in a database accessible as an R package. With TMExplorer, researchers can access all publicly available TME scRNA-seq datasets from a single location and can also return multiple datasets that match their desired criteria with a single command.

2. TMExplorer provides a variety of search parameters (Table 1) that can be used to return a subset of the available data that matches specific criteria. The parameters were designed so that users can search for matching datasets without having to first view a list of all available datasets, making it easier and faster to access data of interest.

3. The majority of existing scRNA-seq databases can only be accessed online as web-based tools and are not easily incorporated into pipelines for analysis of scRNA-seq data. Since many researchers use R or Python for their analyses, we chose to provide TMExplorer as an R-package so that it may be easily integrated into existing pipelines.

4. Some analyses require more than just gene expression information, and TMExplorer provides cell type annotations and cell type signature gene sets alongside gene expression matrices, where they are available. This facilitates the use of a wider range of analysis methods without requiring additional work from the researchers.

We regularly maintain TMExplorer and add new datasets to our database as they get published. Additionally, we have provided an issue template and vignette on GitHub showing how users can process their data and submit it for inclusion in the package. Users who have found new published datasets or sequenced their own should read the formatting instructions and open a new issue using our template. The users who want their dataset to be included in TMExplorer need to provide a description of the dataset, a link to the source for the dataset, a link to the dataset files that will be added to the package, and the completed metadata table. TMExplorer is generalisable to many other sources, including both single-cell and bulk sequencing data. We have recently worked on adopting it for the scATAC-seq data in scATAC.Explorer BioConductor package [68].

In summary, TMExplorer allows researchers to easily access, share and integrate TME scRNA-seq data into their own analysis pipelines. TMExplorer can be used to access data needed for the validation of new algorithms and to allow researchers interested in the tumour microenvironment to study specific types of cancer.

Supporting information

S1 Fig. Viewing the TMExplorer metadata and documentation. Users can view the TMExplorer database metadata of scRNA-seq datasets, interact with the metadata as a dataframe

object, and view the TMExplorer documentation of function arguments included in the package.
(TIF)

S2 Fig. Example searchable parameters to filter scRNA-seq datasets. A set of searchable parameters can be used to filter scRNA-seq datasets. The users can search for specific datasets using user-specified parameters, and return one specific dataset as a SingleCellExperiment object for downstream analysis.
(TIF)

S1 Table. Metadata of TMExplorer. The metadata table contains information such as GEO accession, author, journal, year, PMID, sequencing technology, expression score type(s), source organism, type of cancer, number of patients, tumours, cells and genes, and the database that the data was obtained from.
(XLSX)

Author Contributions

Conceptualization: Parisa Shooshtari.

Data curation: Erik Christensen, Alaine Naidas, David Chen.

Formal analysis: Erik Christensen, David Chen.

Funding acquisition: Parisa Shooshtari.

Investigation: Parisa Shooshtari.

Methodology: Erik Christensen, Alaine Naidas, Parisa Shooshtari.

Resources: Alaine Naidas, Parisa Shooshtari.

Software: Erik Christensen, David Chen.

Supervision: Parisa Shooshtari.

Validation: Erik Christensen, David Chen.

Visualization: Erik Christensen, Alaine Naidas, David Chen, Parisa Shooshtari.

Writing – original draft: Erik Christensen, Alaine Naidas, Parisa Shooshtari.

Writing – review & editing: David Chen, Mia Husic.

References

1. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science [Internet]. 2016; Available from: http://dx.doi.org/10.1126/science.aad0501
2. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat Commun [Internet]. 2017; Available from: https://doi.org/10.1038/ncomms15081 PMID: 28474673
3. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet [Internet]. 2017; Available from: https://doi.org/10.1038/ng.3818 PMID: 28319088
4. Tirosh I, Suvà ML. Deciphering Human Tumor Biology by Single-Cell Expression Profiling. Annual Review of Cancer Biology [Internet]. 2019; Available from: http://dx.doi.org/10.1146/annurev-cancerbio-030518-055609

5. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC, et al. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell [Internet]. 2018; Available from: https://doi.org/10.1016/j.cell.2018.09.006 PMID: 30388455

6. Lambrechts D, Wauters E, Boeckx B, Aibar S, Nittner D, Burton O, et al. Phenotype molding of stromal cells in the lung tumor microenvironment. Nat Med [Internet]. 2018; Available from: https://doi.org/10.1038/s41591-018-0096-5 PMID: 29988129

7. Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, Samghababi P, et al. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. Cell Rep [Internet]. 2017; Available from: https://doi.org/10.1016/j.celrep.2017.10.030 PMID: 29091775

8. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature [Internet]. 2016; Available from: https://doi.org/10.1038/nature20123 PMID: 27806376

9. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science [Internet]. 2014; Available from: https://doi.org/10.1126/science.1254257 PMID: 24925914

10. Shumway M, Cochrane G, Sugawara H. Archiving next generation sequencing data. Nucleic Acids Res. 2010 Jan; 38(Database issue):D870–1. https://doi.org/10.1093/nar/gkp1078 PMID: 19965774

11. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1; 30(1):207–10. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295

12. Franzén O, Gan LM, Björkegren JLM. Panglao DB: A web server for exploration of mouse and human single-cell RNA sequencing data. Database [Internet]. 2019; Available from: http://dx.doi.org/10.1093/database/baz046

13. Cao Y, Zhu J, Han G, Jia P, Zhao Z. scRNASeqDB: a database for gene expression profiling in human single cell by RNA-seq. bioRxiv [Internet]. 2017; Available from: http://dx.doi.org/10.1101/104810

14. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T. JingleBells: A Repository of Immune-Related Single-Cell RNA–Sequencing Datasets. The Journal of Immunology [Internet]. 2017; Available from: http://dx.doi.org/10.4049/jimmunol.1700272

15. The Broad Institute of MIT and Harvard. Single Cell Portal [Internet]. Single Cell Portal. 2019 [cited 2020 Aug 20]. Available from: https://singlecell.broadinstitute.org/single_cell

16. Risso D, Cole M. scRNAseq: A collection of public single-cell RNA-seq datasets. R package version. 2016; 1(0).

17. Wang Z, Feng X, Li SC. SCDevDB: A database for insights into single-cell gene expression profiles during human developmental processes. Front Genet [Internet]. 2019; Available from: http://dx.doi.org/10.3389/fgene.2019.00903

18. Mohanraj S, Díaz-Mejía JJ, Pham MD, Elrick H, Husić M, Rashid S, et al. CReSCENT: CanceR Single Cell ExpressioN Toolkit. Nucleic Acids Res [Internet]. 2020; Available from: https://doi.org/10.1093/nar/gkaa437 PMID: 32479601

19. Davidson S, Efremova M, Riedel A, Mahata B, Pramanik J, Huuhtanen J, et al. Single-cell RNA sequencing reveals a dynamic stromal niche within the evolving tumour microenvironment. bioRxiv [Internet]. 2018; Available from: http://dx.doi.org/10.1101/467225

20. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell [Internet]. 2017; Available from: https://doi.org/10.1016/j.cell.2017.10.044 PMID: 29198524

21. Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science [Internet]. 2017; Available from: http://dx.doi.org/10.1126/science.aai8478

22. Giustacchini A, Thongjuea S, Barkas N, Woll PS, Povinelli BJ, Booth CAG, et al. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. Nat Med [Internet]. 2017; Available from: https://doi.org/10.1038/nm.4336 PMID: 28504724

23. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante LE, Mathewson ND, et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science [Internet]. 2018; Available from: https://doi.org/10.1126/science.aao4750 PMID: 29674595

24. van Galen P, Hovestadt V, Wadsworth MH, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell [Internet]. 2019; Available from: https://doi.org/10.1016/j.cell.2019.01.031 PMID: 30827681

25. Ting DT, Wittner BS, Ligorio M, Vincent Jordan N, Shah AM, Miyamoto DT, et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. Cell Rep [Internet]. 2014; Available from: https://doi.org/10.1016/j.celrep.2014.08.029 PMID: 25242334

26. Miyamoto DT, Zheng Y, Wittner BS, Lee RJ, Zhu H, Broderick KT, et al. RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science [Internet]. 2015; Available from: https://doi.org/10.1126/science.aab0917 PMID: 26383955

27. Jordan NV, Bardia A, Wittner BS, Benes C, Ligorio M, Zheng Y, et al. HER2 expression identifies dynamic functional states within circulating breast cancer cells. Nature [Internet]. 2016; Available from: https://doi.org/10.1038/nature19328 PMID: 27556950

28. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell [Internet]. 2018; Available from: https://doi.org/10.1016/j.cell.2018.05.060 PMID: 29961579

29. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Res [Internet]. 2019; Available from: http://dx.doi.org/10.1038/s41422-019-0195-y

30. Kumar MP, Du J, Lagoudas G, Jiao Y, Sawyer A, Drummond DC, et al. Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics. Cell Rep [Internet]. 2018; Available from: https://doi.org/10.1016/j.celrep.2018.10.047 PMID: 30404002

31. Zhao Y, Carter R, Natarajan S, Varn FS, Compton DA, Gawad C, et al. Single-cell RNA sequencing reveals the impact of chromosomal instability on glioblastoma cancer stem cells. BMC Med Genomics [Internet]. 2019; Available from: https://doi.org/10.1186/s12920-019-0532-5 PMID: 31151460

32. Chen Y-P, Yin J-H, Li W-F, Li H-J, Chen D-P, Zhang C-J, et al. Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma. Cell Res. 2020 Nov; 30(11):1024–42. https://doi.org/10.1038/s41422-020-0374-x PMID: 32686767

33. Lin W, Noel P, Borazanci EH, Lee J, Amini A, Han IW, et al. Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. Genome Med. 2020 Sep 29; 12(1):80. https://doi.org/10.1186/s13073-020-00776-9 PMID: 32988401

34. Gillen AE, Riemondy KA, Amani V, Griesinger AM, Gilani A, Venkataraman S, et al. Single-Cell RNA Sequencing of Childhood Ependymoma Reveals Neoplastic Cell Subpopulations That Impact Molecular Classification and Etiology. Cell Rep. 2020 Aug 11; 32(6):108023. https://doi.org/10.1016/j.celrep.2020.108023 PMID: 32783945

35. Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, et al. Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. Cell Rep. 2019 May 7; 27 (6):1934–47.e5. https://doi.org/10.1016/j.celrep.2019.04.052 PMID: 31067475

36. Yeo SK, Zhu X, Okamoto T, Hao M, Wang C, Lu P, et al. Single-cell RNA-sequencing reveals distinct patterns of cell state heterogeneity in mouse models of breast cancer. Elife [Internet]. 2020 Aug 25;9. Available from: https://doi.org/10.7554/eLife.58810 PMID: 32840210

37. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol. 2021 May; 39(5):599–608. https://doi.org/10.1038/s41587-020-00795-2 PMID: 33462507

38. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020; 11: 2285. https://doi.org/10.1038/s41467-020-16164-1 PMID: 32385277

39. Kumar V, Ramnarayanan K, Sundar R, Padmanabhan N, Srivastava S, Koiwa M, et al. Single-Cell Atlas of Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs in Gastric Cancer. Cancer Discov. 2022; 12: 670–691. https://doi.org/10.1158/2159-8290.CD-21-0683 PMID: 34642171

40. Kim J, Park C, Kim KH, Kim EH, Kim H, Woo JK, et al. Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity. NPJ Precis Oncol. 2022; 6: 9. https://doi.org/10.1038/s41698-022-00251-1 PMID: 35087207

41. Wu F, Fan J, He Y, Xiong A, Yu J, Li Y, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. Nat Commun. 2021; 12: 2540. https://doi.org/10.1038/s41467-021-22801-0 PMID: 33953163

42. Moncada R, Barkley D, Wagner F, Chiodin M, Devlin JC, Baron M, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. Nat Biotechnol. 2020; 38: 333–342. https://doi.org/10.1038/s41587-019-0392-8 PMID: 31932730

43. Lee JJ, Bernard V, Semaan A, Monberg ME, Huang J, Stephens BM, et al. Elucidation of Tumor-Stromal Heterogeneity and the Ligand-Receptor Interactome by Single-Cell Transcriptomics in Real-world Pancreatic Cancer Biopsies. Clin Cancer Res. 2021; 27: 5912–5921. https://doi.org/10.1158/1078-0432.CCR-20-3925 PMID: 34426439

44. Steele NG, Carpenter ES, Kemp SB, Sirihorachai VR, The S, Delrosario L, et al. Multimodal Mapping of the Tumor and Peripheral Blood Immune Landscape in Human Pancreatic Cancer. Nat Cancer. 2020; 1: 1097–1112. https://doi.org/10.1038/s43018-020-00121-4 PMID: 34296197

45. Zhang L, Li Z, Skrzypczynska KM, Fang Q, Zhang W, O'Brien SA, et al. Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell. 2020; 181: 442–459.e29. https://doi.org/10.1016/j.cell.2020.03.048 PMID: 32302573

46. Gojo J, Englinger B, Jiang L, Hübner JM, Shaw ML, Hack OA, et al. Single-Cell RNA-Seq Reveals Cellular Hierarchies and Impaired Developmental Trajectories in Pediatric Ependymoma. Cancer Cell. 2020; 38: 44–59.e9. https://doi.org/10.1016/j.ccell.2020.06.004 PMID: 32663469

47. Kürten CHL, Kulkarni A, Cillo AR, Santos PM, Roble AK, Onkar S, et al. Investigating immune and non-immune cell interactions in head and neck tumors by single-cell RNA sequencing. Nat Commun. 2021; 12: 7338. https://doi.org/10.1038/s41467-021-27619-4 PMID: 34921143

48. Liu Y, He S, Wang X-L, Peng W, Chen Q-Y, Chi D-M, et al. Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. Nat Commun. 2021; 12: 741. https://doi.org/10.1038/s41467-021-21043-4 PMID: 33531485

49. Song H, Weinstein HNW, Allegakoen P, Wadsworth MH 2nd, Xie J, Yang H, et al. Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. Nat Commun. 2022; 13: 141. https://doi.org/10.1038/s41467-021-27322-4 PMID: 35013146

50. Aynaud M-M, Mirabeau O, Gruel N, Grossetête S, Boeva V, Durand S, et al. Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. Cell Rep. 2020; 30: 1767–1779.e6. https://doi.org/10.1016/j.celrep.2020.01.049 PMID: 32049009

51. Kim K-T, Lee HW, Lee H-O, Kim SC, Seo YJ, Chung W, et al. Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome Biol. 2015; 16: 127. https://doi.org/10.1186/s13059-015-0692-3 PMID: 26084335

52. Bautista JL, Cramer NT, Miller CN, Chavez J, Berrios DI, Byrnes LE, et al. Single-cell transcriptional profiling of human thymic stroma uncovers novel cellular heterogeneity in the thymic medulla. Nat Commun. 2021; 12: 1096. https://doi.org/10.1038/s41467-021-21346-6 PMID: 33597545

53. Paulson KG, Voillet V, McAfee MS, Hunter DS, Wagener FD, Perdicchio M, et al. Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. Nat Commun. 2018; 9: 3868. https://doi.org/10.1038/s41467-018-06300-3 PMID: 30250229

54. Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, et al. ArrayExpress update—From bulk to single-cell expression data. Nucleic Acids Res [Internet]. 2019; Available from: https://doi.org/10.1093/nar/gky964 PMID: 30357387

55. National Center for Biotechnology Information (NCBI) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 1988 [cited 2020 Aug 20]. Available from: https://www.ncbi.nlm.nih.gov/

56. GitHub I. GitHub [Internet]. Github. 2007 [cited 2020 Aug 21]. Available from: https://github.com/

57. Singh J. FigShare. J Pharmacol Pharmacother. 2011 Apr; 2(2):138–9. https://doi.org/10.4103/0976-500X.81919 PMID: 21772785

58. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015 Feb; 12(2):115–21. https://doi.org/10.1038/nmeth.3252 PMID: 25633503

59. Lun A, Risso D, Korthauer K. SingleCellExperiment: S4 classes for single cell data. R package version. 2019; 1(1).

60. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015 May; 12(5):453–7. https://doi.org/10.1038/nmeth.3337 PMID: 25822800

61. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013 Jan 16; 14:7. https://doi.org/10.1186/1471-2105-14-7 PMID: 23323831

62. Maechler M. Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2–18. 2019.

63. Franzén O, Björkegren JLM. alona: a web server for single-cell RNA-seq analysis. Bioinformatics. 2020 Jun 1; 36(12):3910–2. https://doi.org/10.1093/bioinformatics/btaa269 PMID: 32324845

64. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2; 46(W1):W537–44. https://doi.org/10.1093/nar/gky379 PMID: 29790989

65. Diaz-Mejia JJ, Meng EC, Pico AR, MacParland SA, Ketela T, Pugh TJ, et al. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. F1000Res [Internet]. 2019 Mar 15;8. Available from: https://doi.org/10.12688/f1000research.18490.3 PMID: 31508207

66. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13; 177(7):1888–902.e21. https://doi.org/10.1016/j.cell.2019.05.031 PMID: 31178118

67. Müller S, Cho A, Liu SJ, Lim DA, Diaz A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. Bioinformatics. 2018 Apr 20; 34(18):3217–9. https://doi.org/10.1093/bioinformatics/bty316 PMID: 29897414

68. Gibson-Khademi A, Christensen E, Shooshtari P. scATAC.Explorer: A Collection of Single-cell ATAC Sequencing Datasets and Corresponding Metadata. R package version 1.2.0. 2022.