scfetch: an R package to access and format single-cell RNA sequencing datasets from public repositories

Yabing Song[1], Jiaxin Gao[2,*], Jianbin Wang[1,*]

[1]School of Life Sciences, Tsinghua University, Beijing, China.

[2]State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China

*To whom correspondence should be addressed.

Abstract

Summary: Downloading and reanalyzing the existing single-cell RNA sequencing (scRNA-seq) datasets is an efficient method to gain clues or new insights. However, there is no tool to access diverse scRNA-seq datasets (*fastq*/*bam* files, count matrices and processed objects) distributed in various repositories, consider features of datasets from different scRNA-seq protocols, and prepare for downstream analysis. Here, we present *scfetch*, an R package to download diverse scRNA-seq datasets from SRA, GEO, PanglaoDB, UCSC Cell Browser, Zenodo and CELLxGENE, and load the downloaded datasets to *Seurat*. *scfetch* supports scRNA-seq datasets generated by different protocols such as 10x Genomics and Smart-seq2. Besides, *scfetch* enables users to convert formats between different scRNA-seq objects, including *SeuratObject*, *Anndata*, *SingleCellExperiment*, *CellDataSet*/*cell_data_set* and *loom*.

Furthermore, *scfetch* also supports downloading *fastq*/*bam* files and count matrices of bulk RNA-seq from SRA and GEO.

Availability and Implementation: The *scfetch* package and vignettes are freely available at https://github.com/showteeth/scfetch and https://showteeth.github.io/scfetch/.

Contact: gaojx@im.ac.cn, jianbinwang@tsinghua.edu.cn.

Supplementary information: Supplementary data are appended.

1 Introduction

In recent years, single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool to reveal cellular heterogeneity and fundamental characteristics of gene expression (Haque *et al.* 2017). As a result, thousands of scRNA-seq datasets have been generated and uploaded to public repositories. Downloading and reanalyzing the existing scRNA-seq datasets provides an alternative method to gain clues or new insights when there are some difficulties in performing the experiments. For example, rare cell types identification (Cheng *et al.* 2023) and pan-cancer single-cell analysis (Zheng *et al.* 2021) usually require a large number of samples and cells. The existing scRNA-seq datasets have three main types: *fastq*/*bam* files, count matrices and processed objects (e.g. *rds* and *h5ad* files), which are stored in different repositories. The *fastq*/*bam* files are mainly stored in Sequence Read Archive (SRA) (Katz *et al.* 2022), the count matrices can be found in Gene Expression Omnibus (GEO) (Barrett *et al.* 2012) and some single-cell databases (e.g. UCSC Cell Browser (Speir *et al.* 2021) and PanglaoDB (Franzén *et*

al. 2019)), and Zenodo and CELLxGENE (https://cellxgene.cziscience.com/) provide the processed objects. Furthermore, as many scRNA-seq protocols and preprocessing tools have been developed, the datasets they generated and stored in public repositories differ. For instance, the *sra* files of 10x Genomics (10x) (Zheng *et al.* 2017) datasets stored in SRA usually contain read1, read2 and index reads, while those of Smart-seq2 (Picelli *et al.* 2014) datasets only include read1 and read2; the count matrix output by *CellRanger* consists of barcode, matrix and feature files, while other preprocessing tools usually output a single count matrix file. The multiple types of datasets, their distribution in various repositories and the differences in scRNA-seq strategies are challenging for scRNA-seq datasets download and downstream analysis.

Some tools have been developed for downloading datasets from GEO and SRA, such as *GEOquery* (Davis and Meltzer 2007) and *GEOfetch* (Khoroshevskyi *et al.* 2023). However, there are few annotations (e.g. cell type annotation and dimension reduction) stored in GEO and SRA, and these tools can't load and integrate the downloaded data for downstream analysis, lack support for the features of datasets generated by 10x, which is the dominant scRNA-seq protocol. For example, the content of *sra* files mentioned above, the custom tags of *bam* files and specially developed tool to convert *bam* files to *fastq* files. *rPanglaoDB* (Osorio *et al.* 2022) can be used to download scRNA-seq count matrices and annotations from PanglaoDB, but PanglaoDB only contains human and mouse datasets, has limited sample numbers, and is no longer maintained. All the above tools only cover partial aspects of diverse scRNA-seq datasets. Besides, there are many scRNA-seq analysis tools and each with its strengths

and object format, such as *Seurat* (Satija *et al.* 2015), *Scanpy* (Wolf *et al.* 2018) and *Monocle* (Qiu *et al.* 2017). Comprehensive scRNA-seq analysis typically requires the combination of multiple tools, it's necessary to convert formats between these tools.

Here, we present *scfetch*, an R package to access diverse scRNA-seq datasets from various public repositories, load the downloaded data to *Seurat*, and convert formats between different scRNA-seq objects. It can be used to (i) efficiently obtain *fastq* files with provided GEO accession numbers and conditionally format them to 10x standard style; (ii) download *bam* files, including normal and 10x original *bam* files, and convert them to *fastq* files; (iii) download dataset metadata, count matrices and various annotations from GEO, PanglaoDB and UCSC Cell Browser, and load them to *Seurat*; (iv) download processed scRNA-seq objects from Zenodo and CELLxGENE, including *rds*, *h5ad* and *loom* files; (v) convert formats between different scRNA-seq objects, including *SeuratObject*, *Anndata*, *SingleCellExperiment*, *CellDataSet*/*cell_data_set* and *loom*.

2 Implementation

The architecture of *scfetch* is shown in Fig. 1. There are four categories: download *fastq*/*bam* files, download count matrices, download scRNA-seq objects (Fig. 1A), and convert formats between different scRNA-seq objects (Fig. 1B). The detailed function information and key tools used are list in Supplementary Table S1.

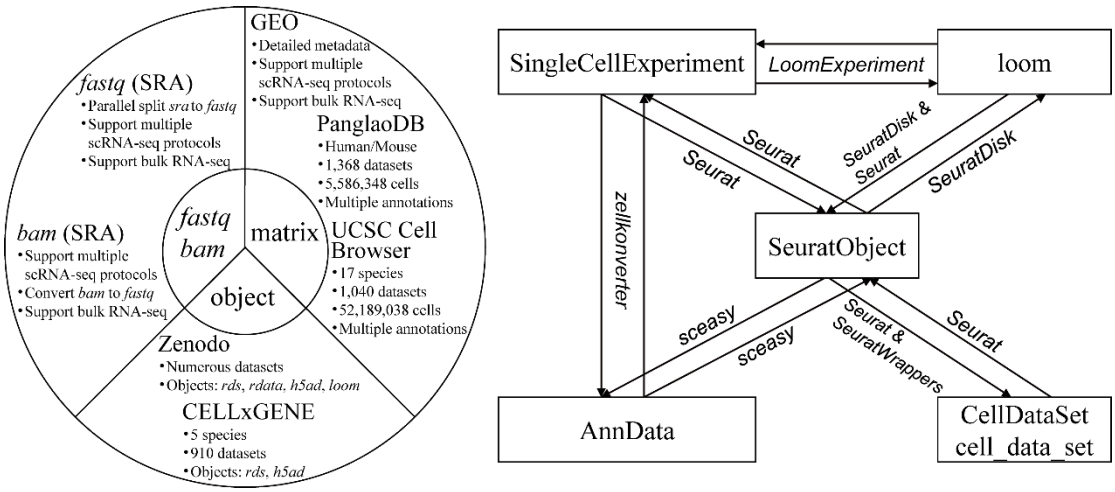

Fig. 1. The architecture of *scfetch*. (**A**) Supported types of scRNA-seq datasets (the inner circle), corresponding public repositories and a brief profile of features provided by *scfetch* for these repositories (the outer circle). (**B**) Supported formats conversion between different scRNA-seq objects, and tools used in the conversion (in *italics*).

Download *fastq*/*bam* files. *scfetch* will first extract all runs under given GEO accession numbers. For *fastq* files, *scfetch* will download the corresponding *sra* files and then split them into *fastq* files in parallel. When the datasets are from 10x, the *--split-files* parameter is added and at least two *fastq* files are generated. *scfetch* will automatically identify read1 and read2 based on read length, and rename them according to the 10x required format. As for datasets from other scRNA-seq protocols, the *--split-3* parameter is used. For *bam* files, if the *bam* files are from 10x, *scfetch* will download the original *bam* files (with custom tags) directly and convert them to 10x-style *fastq* files with *bamtofastq* developed by 10x. Otherwise, *scfetch* will download *sra* files first, then extract *bam* files from the *sra* files, and finally convert the *bam* files to *fastq* files with *samtools* (Li *et al.* 2009). For bulk RNA-seq, the *fastq*/*bam* files can also be

downloaded with the non-10x mode mentioned above. Examples are shown in Supplementary Text 1.

Download count matrices. *scfetch* will first extract dataset metadata such as the detailed description, source/tissue, organism, protocol and related publication. Users can filter datasets based on these attributes, and the filtered metadata will be used as subsequent input. When downloading count matrices from GEO, *scfetch* will try to extract the count matrix from *ExpressionSet* first, if the count matrix is empty or contains non-integer values, *scfetch* will download supplementary files. For the supplementary files output by *CellRanger*, *scfetch* will categorize the files based on sample names. For the supplementary files composed of matrix of every single cell or every bulk sample, *scfetch* will create a matrix containing all cells or all samples. For PanglaoDB, in addition to downloading scRNA-seq count matrices, it also contains cell type annotation and cell type composition, *scfetch* enables users to extract these data using *rPanglaoDB*. For UCSC Cell Browser, *scfetch* will access the scRNA-seq count matrices online instead of downloading them to save time and storage space. UCSC Cell Browser also contains cell type annotation, cell type composition and dimensional reduction information, *scfetch* enables users to extract all these data. With available scRNA-seq count matrices (multiple files output by *CellRanger* or a single matrix), *scfetch* will create *SeuratObjects*, add available dimensional reduction information to the *SeuratObjects*, and merge multiple *SeuratObjects* if applicable. Examples are shown in Supplementary Text 2.

Download scRNA-seq objects. For scRNA-seq objects stored in Zenodo, *scfetch* enables users to extract dataset metadata with given DOIs. Then users can choose to provide DOIs or filtered metadata, the file extensions such as *rds*, *rdata*, *loom* and *h5ad* to download the objects. For scRNA-seq objects stored in CELLxGENE, *scfetch* will first extract the metadata of all available datasets, then users can filter the metadata based on given attributes. With filtered metadata and provided file extensions, *scfetch* enables users to download the objects. Examples are shown in Supplementary Text 3.

Convert formats between different scRNA-seq objects. *scfetch* supports formats conversion (i) between *SeuratObject* and *Anndata*, *SingleCellExperiment*, *CellDataSet*/*cell_data_set*, *loom*; (ii) between *SingleCellExperiment* and *loom*; (iii) between *SingleCellExperiment* and *Anndata*. The conversion implemented in *scfetch* has two main advantages: (i) one-step conversion between different objects to avoid unnecessary information loss; (ii) tools used for the conversion are developed by the team of the source/destination object as far as possible. For example, *scfetch* uses *SeuratDisk* to convert *SeuratObject* to *loom*. The detailed tools used can be found in Fig. 1B and Supplementary Table S1. Examples are shown in Supplementary Text 4.

3 Conclusion

scfetch is an R package dedicated to facilitating researchers to access scRNA-seq *fastq*/*bam* files, count matrices and processed objects from SRA, GEO, PanglaoDB, UCSC Cell Browser, Zenodo and CELLxGENE, and load count matrices and annotations to *Seurat*. It supports scRNA-seq datasets generated by multiple scRNA-

seq protocols. Furthermore, *scfetch* enables users to convert formats between different scRNA-seq objects, including *SeuratObject*, *Anndata*, *SingleCellExperiment*, *CellDataSet*/*cell_data_set* and *loom*. In addition to accessing scRNA-seq datasets, *scfetch* can also be used to download *fastq*/*bam* files and count matrices of bulk RNA-seq from SRA and GEO.

Funding

This work was supported by the National Natural Science Foundation of China [22050004]; and the start-up funds from Institute of Microbiology of Chinese Academy of Sciences.

Conflict of Interest: none declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

References

Barrett T, Wilhite SE, Ledoux P et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2012;41:D991-D5.

Cheng Y, Fan X, Zhang J et al. A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data. *Commun Biol* 2023;6:545.

Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 2007;23:1846-7.

162 Franzén O, Gan L-M, Björkegren JL. PanglaoDB: a web server for exploration of
163 mouse and human single-cell RNA sequencing data. *Database* 2019;2019:baz046.
164 Haque A, Engel J, Teichmann SA et al. A practical guide to single-cell RNA-sequencing
165 for biomedical research and clinical applications. *Genome Med* 2017;9:1-12.
166 Katz K, Shutov O, Lapoint R et al. The sequence read archive: a decade more of
167 explosive growth. *Nucleic Acids Res* 2022;50:D387-D90.
168 Khoroshevskyi O, LeRoy N, Reuter VP et al. GEOfetch: a command-line tool for
169 downloading data and standardized metadata from GEO and SRA. *Bioinformatics*
170 2023;39:btad069.
171 Li H, Handsaker B, Wysoker A et al. The sequence alignment/map format and
172 SAMtools. *Bioinformatics* 2009;25:2078-9.
173 Osorio D, Kuijjer ML, Cai JJ. rPanglaoDB: an R package to download and merge
174 labeled single-cell RNA-seq data from the PanglaoDB database. *Bioinformatics*
175 2022;38:580-2.
176 Picelli S, Faridani OR, Björklund ÅK et al. Full-length RNA-seq from single cells using
177 Smart-seq2. *Nat Protoc* 2014;9:171-81.
178 Qiu X, Mao Q, Tang Y et al. Reversed graph embedding resolves complex single-cell
179 trajectories. *Nat Methods* 2017;14:979-82.
180 Satija R, Farrell JA, Gennert D et al. Spatial reconstruction of single-cell gene
181 expression data. *Nat Biotechnol* 2015;33:495-502.
182 Speir ML, Bhaduri A, Markov NS et al. UCSC cell browser: visualize your single-cell
183 data. *Bioinformatics* 2021;37:4578-80.

Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19:1-5.

Zheng GX, Terry JM, Belgrader P et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;8:14049.

Zheng L, Qin S, Si W et al. Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* 2021;374:abe6474.

Supplementary Table S1: The functions provided by *scfetch* and key tools involved.

Type	Function	Usage	Key tools
Download and format *fastq*	ExtractRun	Extract runs with GEO accession number or GSM number	GEOquery, GEOfastq
	DownloadSRA	Download *sra* files	SRA Toolkit
	SplitSRA	Split *sra* files into *fastq* files and format *fastq* files to 10x standard style	SRA Toolkit, parallel-fastq-dump
Download *bam* and convert to *fastq*	DownloadBam	Download *bam* files, support 10x original *bam* files and normal *bam* files	SRA Toolkit
	Bam2Fastq	Convert *bam* files to *fastq* files	bamtofastq, samtools
Download matrices and load to Seurat	ExtractGEOMeta	Extract dataset metadata from GEO	GEOquery
	ParseGEO	Download matrix from GEO and load to Seurat	GEOquery, Seurat
	ExtractPanglaoDBMeta	Extract dataset metadata from PandlaoDB	rPanglaoDB
	ExtractPanglaoDBComposition	Extract cell type composition of PanglaoDB datasets	rPanglaoDB
	ParsePanglaoDB	Download matrix from PandlaoDB and load to Seurat	rPanglaoDB, Seurat
	ShowCBDatasets	Show all available datasets in UCSC Cell Browser	utils, jsonlite
	ExtractCBDatasets	Extract UCSC Cell Browser datasets with attributes	NA
	ExtractCBComposition	Extract cell type composition of UCSC Cell Browser datasets	jsonlite
	ParseCBDatasets	Download UCSC Cell Browser datasets and load to Seurat	data.table, Seurat, Matrix
Download scRNA-seq objects	ExtractZenodoMeta	Extract dataset metadata from Zenodo with DOIs	curl, jsonlite
	ParseZenodo	Download *rds* , *rdata* , *h5ad* , *loom* from Zenodo with DOIs	curl, utils, jsonlite, parallel
	ShowCELLxGENEDatasets	Show all available datasets in CELLxGENE	curl, jsonlite
	ExtractCELLxGENEMeta	Extract CELLxGENE datasets with attributes	curl, jsonlite
	ParseCELLxGENE	Download *rds* , *rdata* , *h5* , *h5ad* , *loom* from CELLxGENE	httr, utils, jsonlite, parallel
Convert between scRNA-seq objects	ExportSeurat	Convert *SeuratObject* to *AnnData* , *SingleCellExperiment* , *CellDataSet* /*cell_data_set* and *loom*	sceasy, Seurat, SeuratWrappers, SeuratDisk, reticulate
	ImportSeurat	Convert *AnnData* , *SingleCellExperiment* , *CellDataSet* /*cell_data_set* and *loom* to *SeuratObject*	sceasy, Seurat, SeuratDisk, reticulate
	SCEAnnData	Convert between *SingleCellExperiment* and *AnnData*	SingleCellExperiment, zellkonverter
	SCELoom	Convert between *SingleCellExperiment* and *loom*	LoomExperiment
Summarize datasets	StatDBAttribute	Summarize datasets in PandlaoDB, UCSC Cell Browser and CELLxGENE based on attributes	NA