Author Contributions Checklist Form

This form documents the artifacts associated with the article (i.e., the data and code supporting the computational findings) and describes how to reproduce the findings.

Part 1: Data

☐ This paper **does not** involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Abstract

The data used in this paper is from a population scale single cell RNA-sequencing study from Perez et al 2022 on eQTLs for patients with the autoimmune disease SLE. The data includes over 1.2 million PBMCs with over 250 total individuals. Our paper focuses on three specified cell types using the annotations used from the original paper: CD4+ T-cells, CD8+ T-cells, and classical monocytes. The raw data is of .h5ad form, but we primarily use .RDS files to use Seurat and focus on cell-type specific signatures.

Availability

☒ Data **are** publicly available
☐ Data **cannot be made** publicly available

If the data are publicly available, see the *Publicly available data* section. Otherwise, see the *Non-publicly available data* section, below.

Publicly available data

☒ Data are available online at:
https://zenodo.org/records/17402494?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjM0MzNkNDE1LWY4ZTctNDVhYi1hODk5LWJmNzhjNzg4MDUxNyIsImRhdGEiOnt9LCJyYW5kb20iOiI1OTVhOGVjZTBkYmZkZjBjMDA2ZTY4ZTBmNmVjN2Q3NiJ9.ONISAR5Zgx5GZ0odRZKmfSKmKTzBUTRyZ250S-hCc18EzXopSVeq12rdOqvJt_VgHZaHObG8x909Sya_aV9CVQ
☐ Data are available as part of the paper's supplementary material.

☐ Data are publicly available by request, following the process described here:

☐ Data are or will be made available through some other mechanism, described here:

Non-publicly available data

Discussion of lack of publicly available data:

Description

File format(s)

☒ CSV or other plain text:
☒ Software-specific binary format (.Rda, Python pickle, etc.):
☒ Standardized binary format (e.g., netCDF, HDF5, etc.):
☐ Other (described here):

```
The file from the original study is of .h5ad format. Because we use
Seurat objects, we end up using .rds and .csv file formats for our
processed objects.
```

Data dictionary

☐ Provided by the authors in the following file(s):
☐ Data file(s) is (are) self-describiing (e.g., netCDF files)
☒ Available at the following URL:

https://github.com/qianzach/sef_deDist

Additional information (optional)

Part 2: Code

Abstract

The code is segmented into multiple .R files. We include code for simulation and real data analysis. For simulation and real data analysis code, we provide a main file declaring each function used.
We also provide .rds, and .csv objects used in analysis.

Description

Code format(s)

☒ Script files
- ☒ R ☒ Python ☐ Matlab
- ☐ Other:

☒ Package
- ☒ R ☒ Python ☐ MATLAB toolbox
- ☐ Other:

☐ Reproducible report
- ☐ R Markdown ☐ Jupyter notebook
- ☒ Other: .R file

☐ Shell script

☐ Other (described here):

We use R in the form of .R files. However, we also use scripts where we directly run on command line/linux, especially for parallel computations.

Supporting software requirements

Version of primary software used

R: 4.5.1, Rstudio 2025.9.1.401

Libraries and dependencies used by the code

R:
dplyr
ggplot2
tidyverse
ggplot2
mvtnorm
MASS
patchwork
grid
parallel
truncnorm
Seurat
pbmcapply
anndata
parallel
doSNOW
foreach
doParallel
Ake
Hmisc
clusterProfiler
enrichplot
org.Hs.eg.db

Python:
Scanpy
numpy

Supporting system/hardware requirements (optional)

Parallelization used

☐ No parallel code used
☒ Multi-core parallelization on a single machine/node
Number of cores used: 2-6
☐ Multi-machine/multi-node parallelization
Number of nodes and cores used:

License

☐ MIT License (default)
☐ BSD
☐ GPL v3.0
☐ Creative Commons
☐ Other (described here):

Additional information (optional)

Part 3: Reproducibility workflow

Scope

The provided workflow reproduces:

☒ Any numbers provided in text in the paper

☒ The computational method(s) presented in the paper (i.e., code is provided that implements the method(s))

☒ All tables and figures in the paper

☐ Selected tables and figures in the paper, as explained and justified here:

Workflow details

Location

The workflow is available:

☐ As part of the paper's supplementary material

☒ In this Git repository: https://github.com/qianzach/sef_deDist

☐ Other:

The data dictionary in the GitHub repository provides details regarding the contents of each file.

Format(s)

☐ Single master code file

☒ Wrapper (shell) script(s)

☒ Self-contained R Markdown file, Jupyter notebook, or other literate programming approach

☐ Text file (e.g., a readme-style file) that documents workflow

☐ Makefile

☐ Other (more detail in 'Instructions' below)

Instructions

Expected run-time

Approximate time needed to reproduce the analyses on a standard desktop machine:

☐ <1 minute
☐ 1-10 minutes
☐ 10-60 minutes
☒ 1-8 hours
☐ >8 hours
☐ Not feasible to run on a desktop machine, as described here:

Preprocessing using SCT at population scale is extremely computationally intensive, even within a specific cell type. Aside from this, the modeling and inferential procedures time needed for one cell type would fall into 10-60 minutes. The permutation script is also computationally expensive.

Additional documentation (optional)

Notes (optional)