Comparative integration of single-cell RNA-seq

CSAMA 2023 project

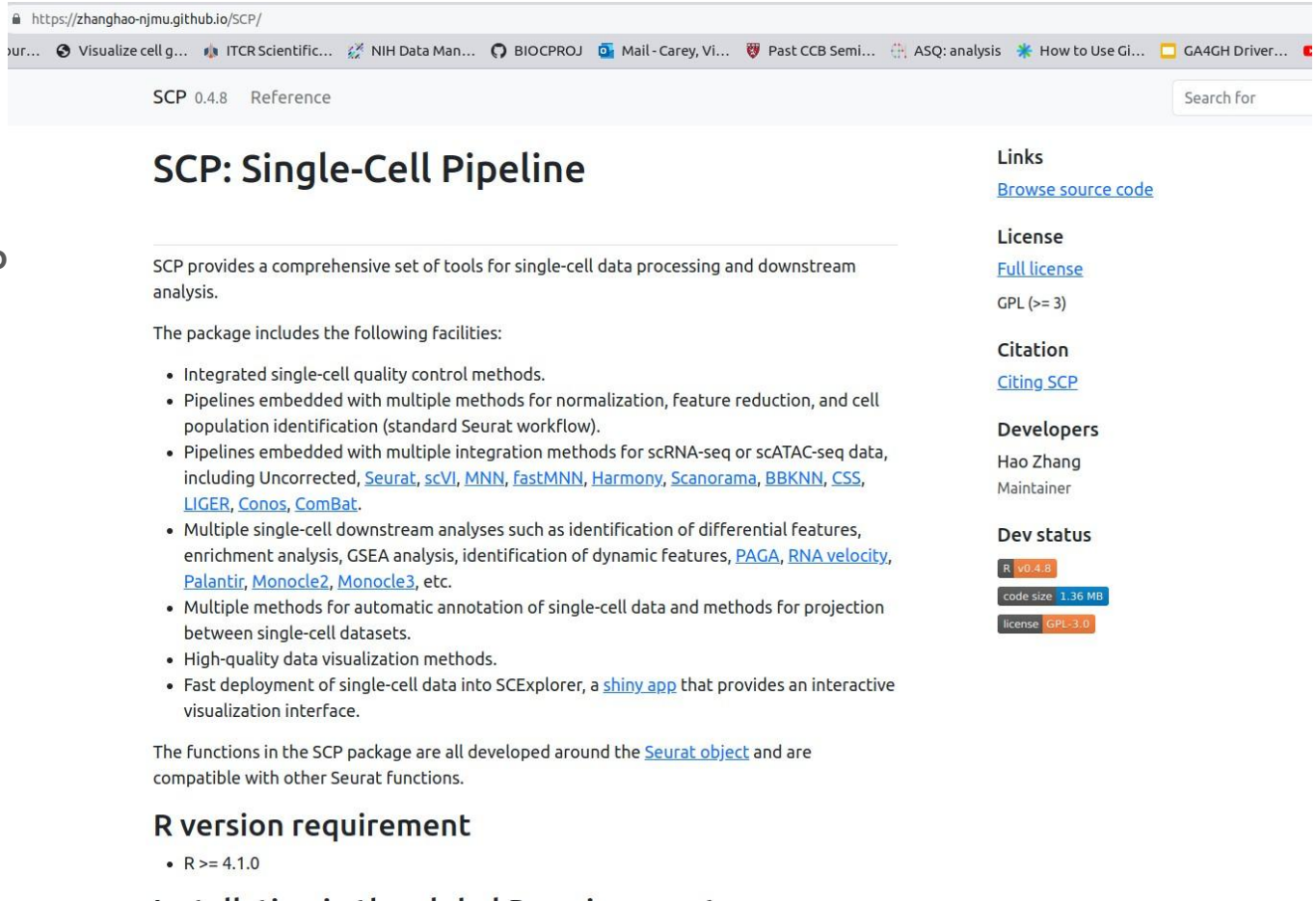
Sebastian Gornik, Yuchen Xiang, Vince Carey

Discussion on Monday night

- "Comparative evaluation of preprocessing or integration"
- SCP has produced convenient interfaces to
  - harmony
  - bbknn
  - seurat
  - scanorama
  - LIGER
  - scvi …
- Arabadopsis and related plant "stem cells"
- Challenging to "align" different samples
- Consider "methods sweep" - cluster numbers, quality via RG PCA regression

How can you proceed?

We want to conveniently use Hao Zhang's "SCP - it handles software acquisition, data structure harmonization



Software solution: Docker container (Dockerfile gist)

```
FROM ubuntu:jammy

LABEL org.label-schema.license="GPL-2.0" \
      org.label-schema.vcs-url="https://github.com/rocker-org/" \
      org.label-schema.vendor="Rocker Project" \
      maintainer="Dirk Eddelbuettel <edd@debian.org>"

## Set a default user. Available via runtime flag `--user docker`
## Add user to 'staff' group, granting them write privileges to /usr/local/lib/R/site.library
## User should also have & own a home directory (for rstudio or linked volumes to work properly).
RUN useradd -s /bin/bash -m docker && usermod -a -G staff docker
RUN apt update -qq && apt install --yes --no-install-recommends wget     ca-certificates gnupg
RUN wget -q -O- https://eddelbuettel.github.io/r2u/assets/dirk_eddelbuettel_key.asc     | tee -a /etc/apt/trusted.gpg.d/cranapt_key.asc
RUN echo "deb [arch=amd64] https://r2u.stat.illinois.edu/ubuntu jammy main"      > /etc/apt/sources.list.d/cranapt.list
RUN apt update -qq
RUN wget -q -O- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc     | tee -a /etc/apt/trusted.gpg.d/cran_ubuntu_key.asc
RUN echo "deb [arch=amd64] https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/"     > /etc/apt/sources.list.d/cran_r.list
RUN apt-key adv --keyserver keyserver.ubuntu.com --recv-keys     67C2D66C4B1D4339 51716619E084DAB9
RUN apt update -qq
RUN DEBIAN_FRONTEND=noninteractive apt install --yes --no-install-recommends     r-base-core \
                 r-base-dev \
                 r-recommended
RUN apt install --yes --no-install-recommends python3-dbus
RUN apt install --yes --no-install-recommends python3-gi
RUN apt install --yes --no-install-recommends python3-apt
RUN ## Then install bspm (as root) and enable it, and enable a speed optimization
RUN Rscript -e 'install.packages("bspm")'
RUN RHOME=$(R RHOME)
RUN echo "suppressMessages(bspm::enable())" >> ${RHOME}/etc/Rprofile.site
RUN echo "options(bspm.version.check=FALSE)" >> ${RHOME}/etc/Rprofile.site
RUN echo "suppressMessages(bspm::enable())" >> ${HOME}/.Rprofile
RUN Rscript -e 'install.packages("Matrix")'
RUN Rscript -e 'install.packages("devtools")'
RUN Rscript -e 'Sys.setenv(GITHUB_PAT=""); devtools::install_github("zhanghao-njmu/SCP")'
RUN Rscript -e 'SCP::PrepareEnv()'
RUN Rscript -e 'install.packages("httpgd")'  # good for graphics
```

Big advantage from Dirk Eddelbuettel's r2u to get fully functional R + thousands of installable binaries in minutes

docker pull vjcitn/csamascp:0.0.2

Two target methods chosen arbitrarily

nature protocols

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Sign up for alerts | RSS feed

nature > nature protocols > protocols > article

Protocol | Published: 12 October 2020

Jointly defining cell types from multiple single-cell datasets using LIGER

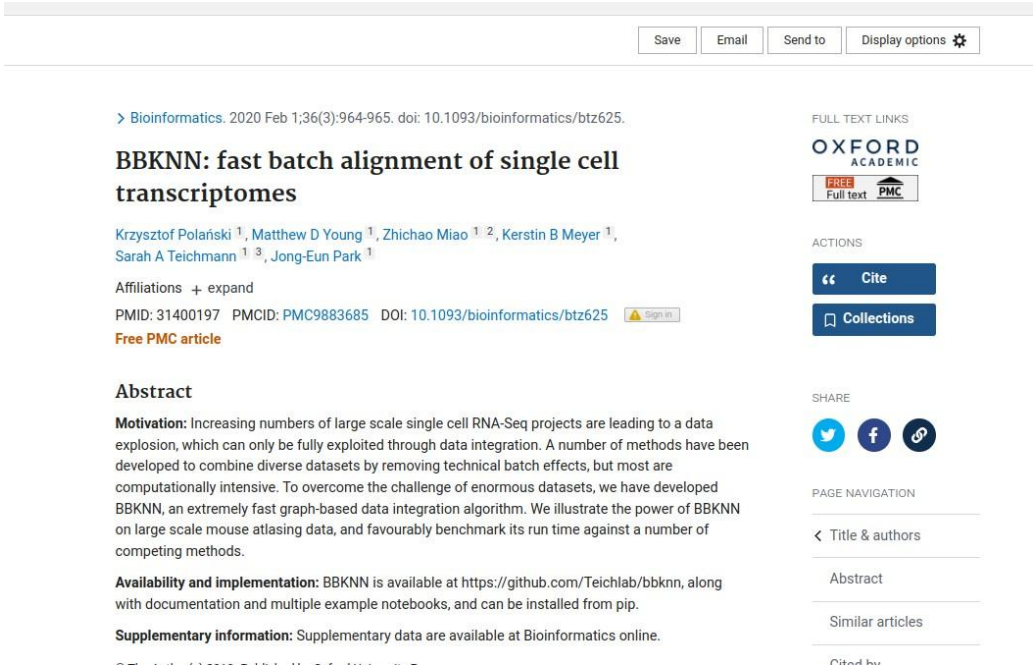
Jialin Liu, Chao Gao, Joshua Sodicoff, Velina Kozareva, Evan Z. Macosko & Joshua D. Welch

Nature Protocols **15**, 3632–3662 (2020) | Cite this article

10k Accesses | **38** Citations | **32** Altmetric | Metrics

Abstract

High-throughput single-cell sequencing technologies hold tremendous potential for defining cell types in an unbiased fashion using gene expression and epigenomic state. A key challenge in realizing this potential is integrating single-cell datasets from multiple protocols, biological contexts, and data modalities into a joint definition of cellular identity. We previously developed an approach, called linked inference of genomic experimental relationships (LIGER), that uses integrative nonnegative matrix factorization to address this challenge. Here, we provide a step-by-step protocol for using LIGER to jointly define cell types from multiple single-cell datasets. The main stages of the protocol are data preprocessing and normalization, joint factorization, quantile normalization and joint clustering, and visualization. We describe how to jointly define cell types from single-cell RNA-seq (scRNA-seq) and single-nucleus ATAC-seq (snATAC-seq) data, but similar steps apply across a wide range of other settings and data types, including cross-species analysis, single-nucleus DNA



Save | Email | Send to | Display options

Bioinformatics. 2020 Feb 1;36(3):964-965. doi: 10.1093/bioinformatics/btz625.

BBKNN: fast batch alignment of single cell transcriptomes

Krzysztof Polański [1], Matthew D Young [1], Zhichao Miao [1 2], Kerstin B Meyer [1], Sarah A Teichmann [1 3], Jong-Eun Park [1]

Affiliations + expand

PMID: 31400197 PMCID: PMC9883685 DOI: 10.1093/bioinformatics/btz625

Free PMC article

Abstract

Motivation: Increasing numbers of large scale single cell RNA-Seq projects are leading to a data explosion, which can only be fully exploited through data integration. A number of methods have been developed to combine diverse datasets by removing technical batch effects, but most are computationally intensive. To overcome the challenge of enormous datasets, we have developed BBKNN, an extremely fast graph-based data integration algorithm. We illustrate the power of BBKNN on large scale mouse atlasing data, and favourably benchmark its run time against a number of competing methods.

Availability and implementation: BBKNN is available at https://github.com/Teichlab/bbknn, along with documentation and multiple example notebooks, and can be installed from pip.

Supplementary information: Supplementary data are available at Bioinformatics online.

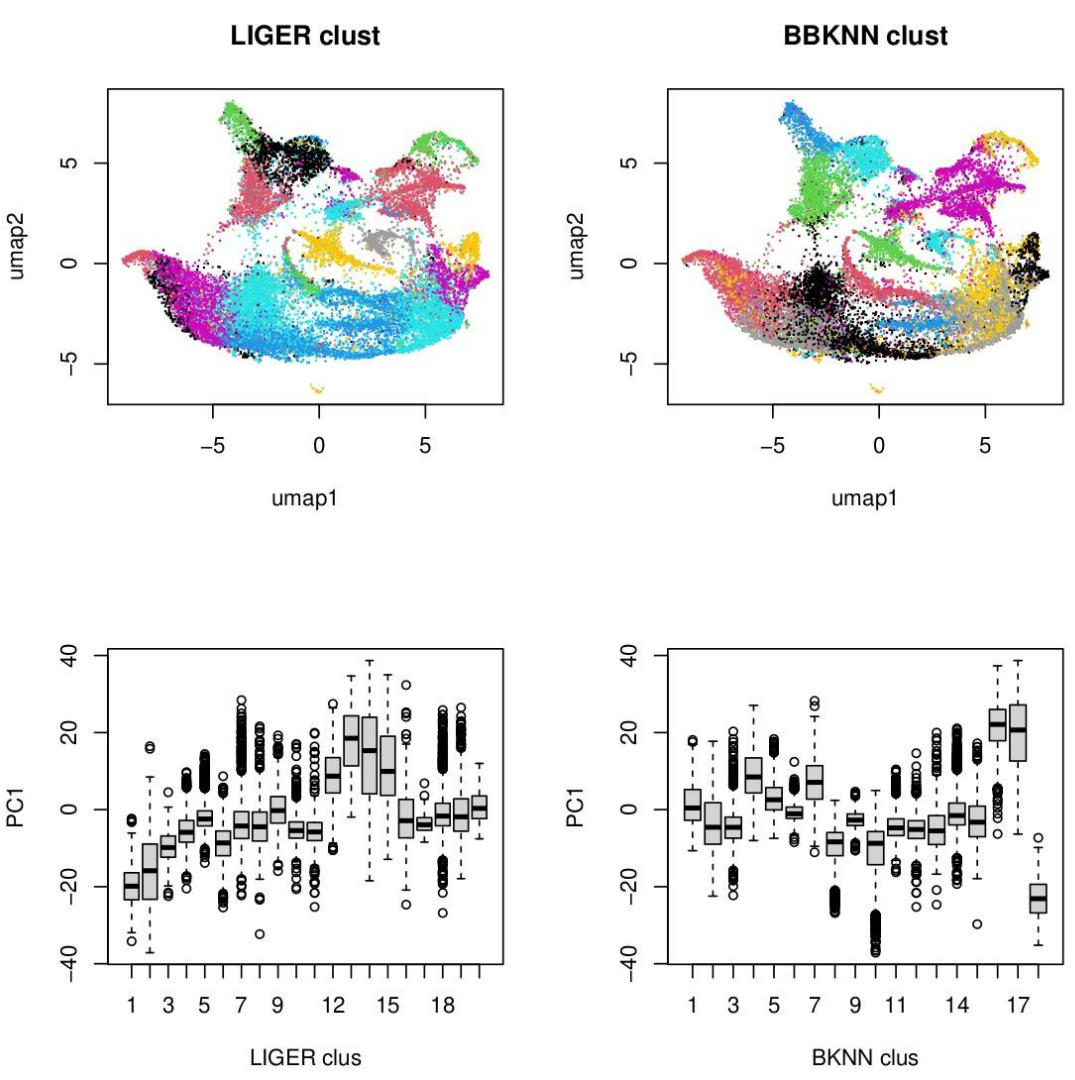
A dataset: Sebastian?

```
> srt_combined

An object of class Seurat

66743 features across 20000 samples within 5 assays

Active assay: RNA (37141 features, 2000 variable features)

 4 other assays present: SCT, Seurat, fastMNNcorrected, ComBatcorrected

 29 dimensional reductions calculated: Seuratpca, SeuratUMAP2D, SeuratUMAP3D,
Uncorrectedpca, UncorrectedUMAP2D, UncorrectedUMAP3D, CSSpca, CSS, CSSUMAP2D, CSSUMAP3D,
LIGER, LIGERUMAP2D, LIGERUMAP3D, Scanorama, ScanoramaUMAP2D, ScanoramaUMAP3D, BBKNNpca,
BBKNNUMAP2D, BBKNNUMAP3D, Harmonypca, Harmony, HarmonyUMAP2D, HarmonyUMAP3D, fastMNN,
fastMNNUMAP2D, fastMNNUMAP3D, ComBatpca, ComBatUMAP2D, ComBatUMAP3D
```

Cumbersome operations

plot(srt_combined@reductions$SeuratUMAP2D@misc$model$embedding, pch=".", col=bkcl, main="BBKNN clust", xlab="umap1",

ylab="umap2")



PC regression for BBKNN and LIGER

```
Call:
lm(formula = srt_combined@reductions$Seuratpca@cell.embeddings[,
    "SeuratPC_1"] ~ bkcl)

Residuals:
     Min       1Q   Median       3Q      Max
-28.0427  -3.0608  -0.0488   3.2483  24.7361

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3921     0.1818   7.657 1.98e-14 ***
bkcl2        -4.8235     0.2504 -19.267  < 2e-16 ***
bkcl3        -5.6520     0.2749 -20.560  < 2e-16 ***
bkcl4         7.4510     0.2790  26.711  < 2e-16 ***
bkcl5         1.6635     0.2508   6.632 3.39e-11 ***
bkcl6        -2.1275     0.2751  -7.733 1.10e-14 ***
bkcl7         5.6735     0.2245  25.271  < 2e-16 ***
bkcl8       -10.7323     0.2155 -49.807  < 2e-16 ***
bkcl9        -4.0534     0.2197 -18.447  < 2e-16 ***
bkcl10      -12.0465     0.2209 -54.540  < 2e-16 ***
bkcl11       -5.8098     0.2356 -24.662  < 2e-16 ***
bkcl12       -6.5344     0.2746 -23.797  < 2e-16 ***
bkcl13       -6.1467     0.3379 -18.193  < 2e-16 ***
bkcl14       -2.1949     0.2236  -9.815  < 2e-16 ***
bkcl15       -3.7388     0.3638 -10.278  < 2e-16 ***
bkcl16       20.3362     0.2269  89.629  < 2e-16 ***
bkcl17       18.2575     0.2927  62.376  < 2e-16 ***
bkcl18      -24.2387     0.6075 -39.901  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.529 on 19982 degrees of freedom
Multiple R-squared:  0.7384,  Adjusted R-squared:  0.7382
F-statistic:  3318 on 17 and 19982 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = srt_combined@reductions$Seuratpca@cell.embeddings[,
    "SeuratPC_1"] ~ ligcl)

Residuals:
    Min      1Q  Median      3Q     Max
-32.659  -3.469  -0.397   3.280  33.572

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.8213     0.2930 -67.652  < 2e-16 ***
ligcl2        3.5800     0.4638   7.718 1.24e-14 ***
ligcl3       10.1786     0.4583  22.211  < 2e-16 ***
ligcl4       14.3637     0.3335  43.067  < 2e-16 ***
ligcl5       17.8882     0.3118  57.363  < 2e-16 ***
ligcl6       10.9846     0.3253  33.765  < 2e-16 ***
ligcl7       16.8481     0.3639  46.302  < 2e-16 ***
ligcl8       15.8816     0.4047  39.243  < 2e-16 ***
ligcl9       19.9424     0.3391  58.809  < 2e-16 ***
ligcl10      14.7702     0.3561  41.483  < 2e-16 ***
ligcl11      14.4113     0.4135  34.854  < 2e-16 ***
ligcl12      28.5405     0.3344  85.356  < 2e-16 ***
ligcl13      37.6173     0.3317 113.422  < 2e-16 ***
ligcl14      34.0304     0.3499  97.260  < 2e-16 ***
ligcl15      31.4944     0.4025  78.243  < 2e-16 ***
ligcl16      18.5877     0.6408  29.005  < 2e-16 ***
ligcl17      16.4688     0.8934  18.434  < 2e-16 ***
ligcl18      18.8936     0.3306  57.149  < 2e-16 ***
ligcl19      19.2920     0.3951  48.830  < 2e-16 ***
ligcl20      20.5970     0.8669  23.760  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.372 on 19980 degrees of freedom
Multiple R-squared:  0.6526,  Adjusted R-squared:  0.6523
F-statistic:  1976 on 19 and 19980 DF,  p-value: < 2.2e-16
```

Upshots

- SCP seems to achieve great convenience of exploration of competing methods
  - How does it keep up with all of them?
- Seurat object as common data structure is cumbersome, can be improved
- Data integration solution – Sebastian?
- Simple display and diagnostic of clusters against PC1 … other PCs and other methods are easily investigated