Bioconductor CZI / HCA Seed Networks Symposium: Work in Progress

July 20, 2020

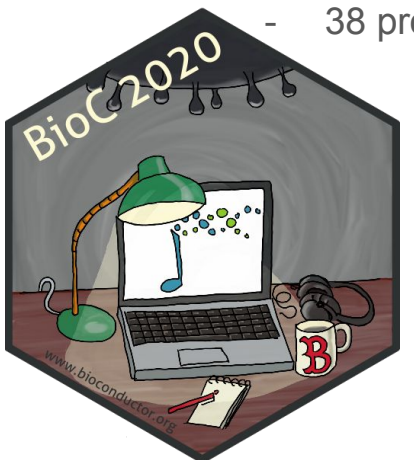
Introduction to the symposium

Bioconductor https://bioconductor.org

- Statistical analysis and comprehension of high-throughput genomic data
- 1900+ R packages contributed by our global user base
- Widely used & well respected

Single cell resources

- 100+ existing packages
- Orchestrating single-cell analysis with Bioconductor
- Annual conference next week!



Seed networks for the human cell atlas see

- "...bring together experimental scientists, computational biologists, software engineers, and physicians to support the continued development of the Human Cell Atlas (HCA)"
- 38 projects, of which we are one

Bioconductor's contributions

- Access and represent HCA data
- Methods and benchmarks for emerging and integrative data
- Methods for scalable, performant analysis

Today

- Recent updates from across our collaboration
- Short talks, with question & answer opportunities

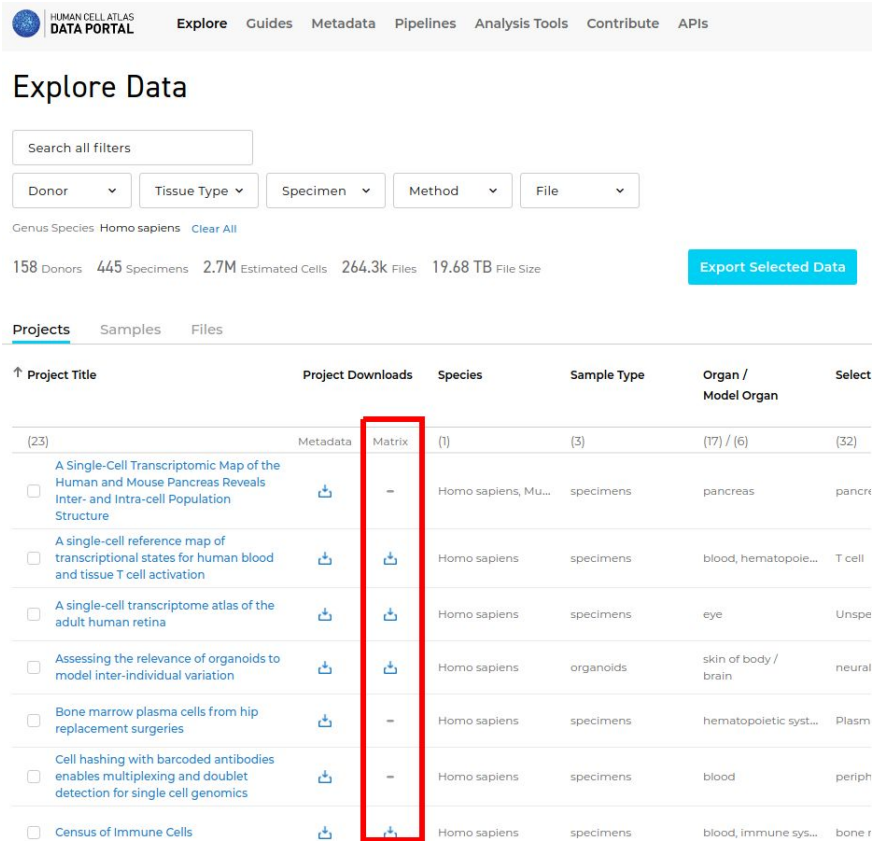
Project leaders

- Aedin Culhane
- Greg Finak
- Kasper Hansen
- Stephanie Hicks
- Wolfgang Huber
- Martin Morgan
- Davide Risso
- Matt Ritchie

Data access and representation

Programmatic access to the HCA using **HCAExplorer**

- Download experiment metadata and pre-computed expression matrices
- Mirrors the functionality of the HCA Data Explorer https://data.humancellatlas.org/explore/projects
- https://bioconductor.org/packages/HCAExplorerBrowser



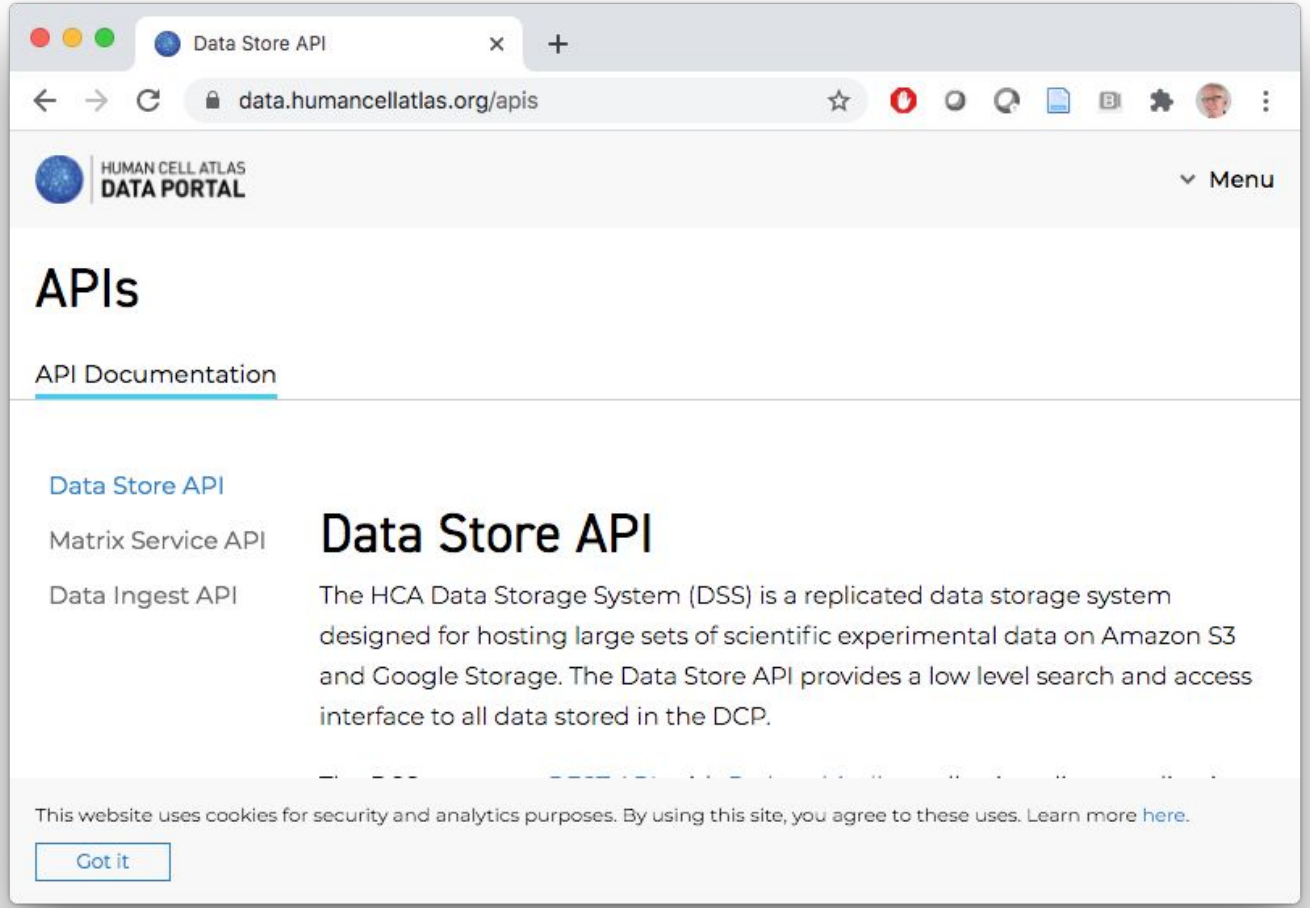
Downloading an expression matrix as a LoomExperiment File using HCAExplorer

```
## Create HCAExplorer object
hca <- HCAExplorer()

## Obtain the first project by subsetting
hca <- hca[1]

## Download project's expression matrix file as a LoomExperiment object
le <- getExpressionMatrix(hca, format = "loom")
```

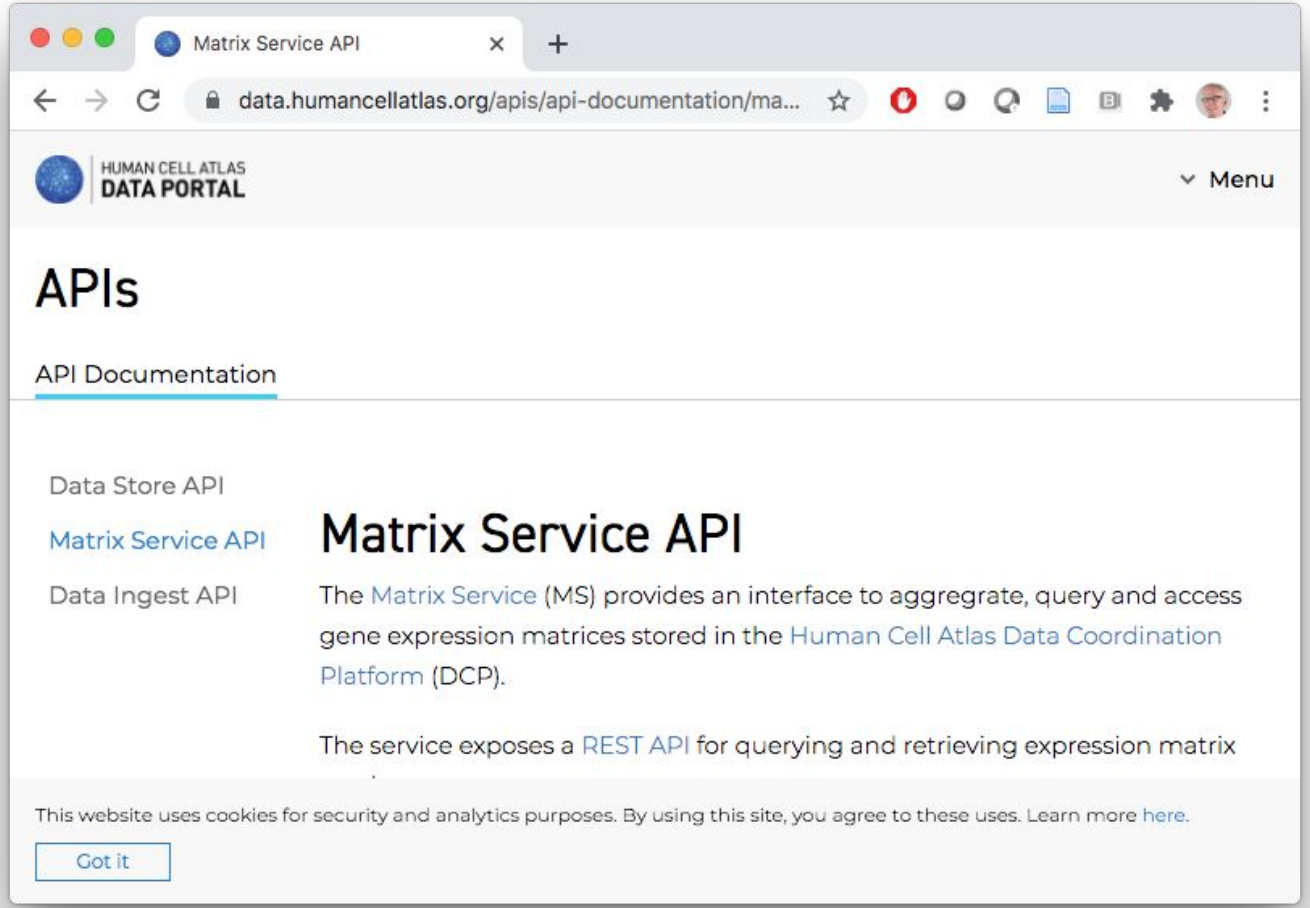
Select first project in the HCAExplorer object and download its matrices as a LoomExperiment object.


Data Store API
data.humancellatlas.org/apis
HUMAN CELL ATLAS
DATA PORTAL
Menu
APIs
API Documentation
Data Store API
Matrix Service API
Data Ingest API
Data Store API
The HCA Data Storage System (DSS) is a replicated data storage system designed for hosting large sets of scientific experimental data on Amazon S3 and Google Storage. The Data Store API provides a low level search and access interface to all data stored in the DCP.
This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. Learn more here.
Got it

Programmatic access to the HCA DSS API using **HCABrowser**

- A more complex interface that implements the HCA DSS's api methods https://dss.data.humancellatlas.org/
- Mirrors the functionality of the python dcp-cli package https://github.com/HumanCellAtlas/dcp-cli
- https://bioconductor.org/packages/HCABrowser

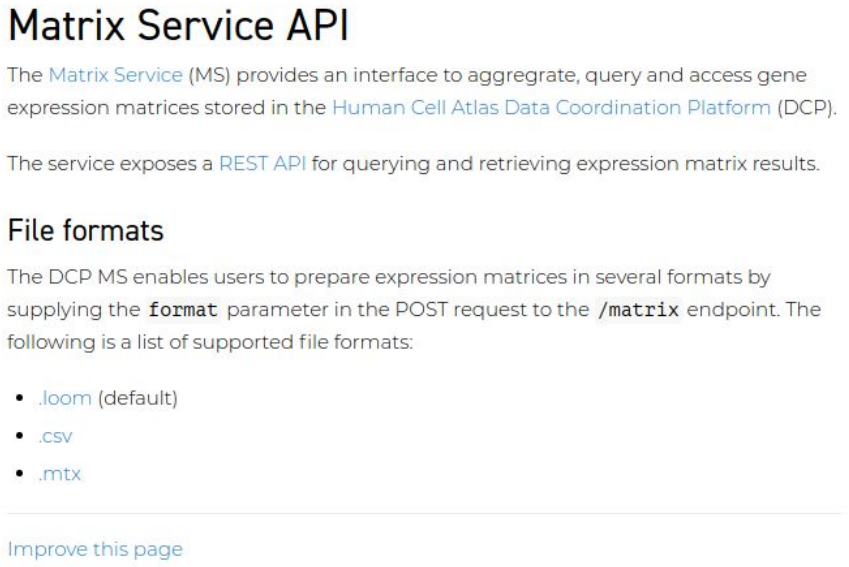
```
(!organ.text %in% c('Brain', 'blood')) &
(files.specimen_from_organism_json.genus_species.text == "Homo sapiens" |
library_preparation_protocol_json.library_construction_approach.text == 'Smart-seq2'))
...
```


Matrix Service API
data.humancellatlas.org/apis/api-documentation/ma...
HUMAN CELL ATLAS
DATA PORTAL
Menu
APIs
API Documentation
Data Store API
Matrix Service API
Data Ingest API
Matrix Service API
The Matrix Service (MS) provides an interface to aggregrate, query and access gene expression matrices stored in the Human Cell Atlas Data Coordination Platform (DCP).
The service exposes a REST API for querying and retrieving expression matrix
This website uses cookies for security and analytics purposes. By using this site, you agree to these uses. Learn more here.
Got it

Programmatic access to the HCA Matrix API using BiocManager::install("**HCAMatrixBrowser**")

- Available in different formats (and representations)
  - .loom (LoomExperiment)
  - .mtx (SingleCellExperiment)
  - .csv (tibble list)

- Easy-to-use R interface
  - Main function: **loadHCAMatrix**
  - Primary input bundle_fqids -- vector of cell bundle identifiers



Matrix Service API

The Matrix Service (MS) provides an interface to aggregrate, query and access gene expression matrices stored in the Human Cell Atlas Data Coordination Platform (DCP).

The service exposes a REST API for querying and retrieving expression matrix results.

File formats

The DCP MS enables users to prepare expression matrices in several formats by supplying the `format` parameter in the POST request to the `/matrix` endpoint. The following is a list of supported file formats:

- .loom (default)
- .csv
- .mtx

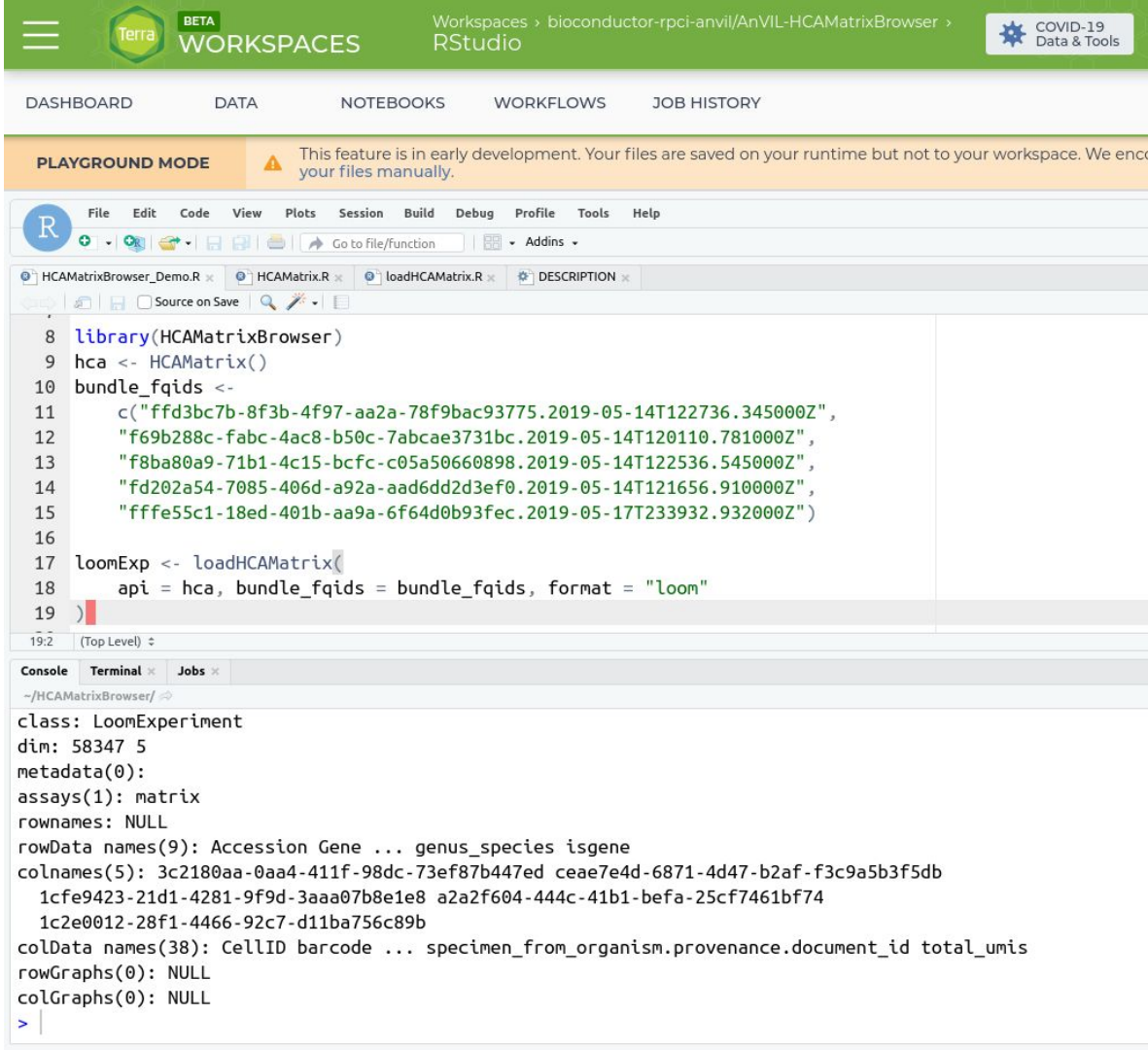
Improve this page

https://bioconductor.org/packages/HCAMatrixBrowser

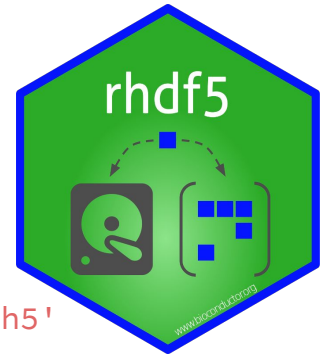
HCA data on Terra

In four commands, we can obtain data from the HCA

- API endpoint: https://matrix.data.humancellatlas.org/



rhdf5 can read files in S3 buckets



- Latest version of rhdf5 distributed with support for reading directly from S3 e.g.

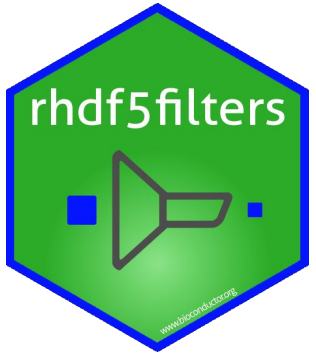
```
public_S3_url <-
'https://rhdf5-public.s3.eu-central-1.amazonaws.com/rhdf5ex_t_float_3d.h5'

## EXPLORE FILE ##
h5ls(file = public_S3_url, s3 = TRUE)
##   group name       otype dclass        dim
## 0     /   a1 H5I_DATASET  FLOAT 5 x 10 x 2

## READ SUBSET ##
h5read(public_S3_url, name = "a1", index = list(1:2, 3, 1), s3 = TRUE)
##  , , 1
##
##             [,1]
##  [1,] 0.2444485
##  [2,] 0.3873723
```

- Works with public and private buckets

rhdf5filters provides additional compression filters in R



- Currently seven filters:
  - BLOSC meta compressor (6 filters)
  - BZIP2
- Compiles C code on all platforms (inc Windows) no pre-built binary required
- Integrated with **rhdf5**
  - Writing: Supply filter argument to functions
  - Reading: Used automatically if needed
- https://bioconductor.org/packages/rhdf5filters/
- Future plan: integrate all plugins distributed by HDF5 Group (link)