

# FAIRifying single cell RNA-seq data from Single Cell Portal

*Hackathon Report*

Eric Weitz  
Broad Institute

Bio-IT World Conference  
Wednesday, May 16, 2018  
Boston, MA, USA

# Single Cell Portal

Portal for single cell RNA-seq data

- **Summarize:** Abstracts and figures from study publications
- **Explore:** visualize gene expression with t-SNE, violin plots, heatmaps, more
- **Analyze:** Run workflows like 10X CellRanger, leveraging a web UI and FireCloud
- **Download:** Get raw data, as individual files or in bulk

[https://portals.broadinstitute.org/single\\_cell](https://portals.broadinstitute.org/single_cell)

Single Cell Portal X Eric

Secure | https://portals.broadinstitute.org/single\_cell

Single Cell Portal BETA ? Help ▾ Sign In

# Single Cell Portal BETA

Visualization portal for single cell RNA-seq data.

Now featuring 34 studies with 432,801 cells.

## Browse Studies ?

Search Studies...



Most Recent

Most Popular

Reset Filters

### Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq) ▾

[View Study](#)

Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus. Habib N, Li Y, Heidenreich M, Swiech L, Avraham-David I, Trombetta J, Hession C, Zhang F, Regev A. *Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons*. *Science* 28 Jul 2016 DOI: 10.1126/science.aad7038 Contact: naomi@broadinstitute.org Single cell RNA-Seq provides rich information about cell types and states. However, it is difficult to capture rare dynamic processes, such as adult neurogenesis, because isolation of rare neurons from adult tissue is challenging and markers for each phase are limited. Here, we develop Div-Seq, which combines scalable single-nucleus RNA-Seq (sNuc-Seq) with pulse labeling of proliferating cells by EdU...  
(continued)

### Retinal Bipolar Neuron Drop-seq ▾

[View Study](#)

Retinal Bipolar Neuron Drop-Seq Karthik Shekhar, Sylvain W. Lapan, Irene E. Whitney, Nicholas M. Tran, Evan Z. Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z. Levin, James Nemesh, Melissa Goldman, Steven A. McCarroll, Constance L. Cepko, Aviv Regev, Joshua R. Sanes. *Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics*. *Cell*. Volume 166, Issue 5, p1308–1323.e30, 25 August 2016. DOI: http://dx.doi.org/10.1016/j.cell.2016.07.054 Contact: Karthik Shekhar at karthik@broadinstitute.org Patterns of gene expression can be used to characterize and classify neuronal types. It is challenging, however, to generate taxonomies that fulfill the essential criteria of being comprehensive, harmonizing with conventional classification schemes, and lacking superfluous subdivisions of genuine types. To address these challenges, we used massively parallel single-cell RNA profiling and optimized computational methods on a heterogeneous class of... (continued)

### 10X LucOS ▾

Single Cell Portal Eric

Secure | [https://portals.broadinstitute.org/single\\_cell/study/1-single-nucleus-rna-seq-of-cell-diversity-in-the-adult-mouse-hippocampus-snuc-seq#study-summary](https://portals.broadinstitute.org/single_cell/study/1-single-nucleus-rna-seq-of-cell-diversity-in-the-adult-mouse-hippocampus-snuc-seq#study-summary)

Single Cell Portal BETA Study Overview Help eweitz

## Study: Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus (sNuc-Seq) 1402 cells

Summary Explore Download Settings

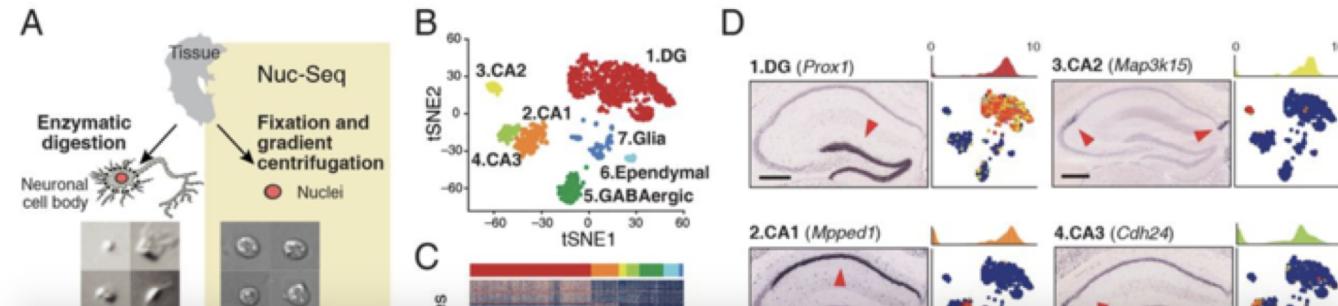
### Single nucleus RNA-seq of cell diversity in the adult mouse hippocampus.

Habib N, Li Y, Heidenreich M, Swiech L, Avraham-David I, Trombetta J, Hession C, Zhang F, Regev A. **Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.** *Science* 28 Jul 2016 DOI: [10.1126/science.aad7038](https://doi.org/10.1126/science.aad7038)

Contact: naomi@broadinstitute.org

Single cell RNA-Seq provides rich information about cell types and states. However, it is difficult to capture rare dynamic processes, such as adult neurogenesis, because isolation of rare neurons from adult tissue is challenging and markers for each phase are limited. Here, we develop Div-Seq, which combines scalable single-nucleus RNA-Seq (sNuc-Seq) with pulse labeling of proliferating cells by EdU to profile individual dividing cells. sNuc-Seq and Div-Seq can sensitively identify closely related hippocampal cell types and track transcriptional dynamics of newborn neurons within the adult hippocampal neurogenic niche, respectively. This study contains the sNuc-Seq analysis performed as a part of the Div-Seq method development.

Using sNuc-Seq, we analyzed 1,367 single nuclei from hippocampal anatomical sub-regions (DG, CA1, CA2, and CA3) from adult mice, including enrichment of genetically-tagged lowly abundant GABAergic neurons (9). sNuc-Seq robustly generated high quality data across animal age groups (including 2 years old mice), detecting 5,100 expressed genes per nucleus on average, with comparable complexity to single neuron RNA-Seq from young mice (1, 2, 3). Analysis of sNuc-Seq data revealed distinct nuclei clusters (Fig. 1B-D shown below) corresponding to known cell types and anatomical distinctions in the hippocampus.



Gene Expression for *Gad1*[Summary](#)[Explore](#)[Download](#)[Settings](#)

Gad1



Distribution

Scatter

View Options

**Q Advanced Gene Search**

Upload gene list

 Choose File No file chosen

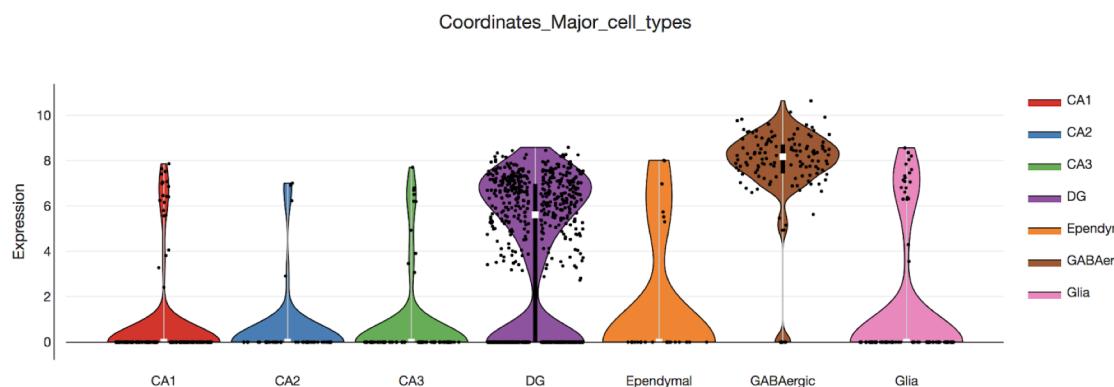
Collapse genes by

None

Back

View Gene Lists

Create Annotations

**Load cluster**

Coordinates\_Major\_cell\_types

**Select annotation**

CLUSTER

**Subsampling threshold**

All Cells

**▼ Distribution****Plot Type**

Violin Plot

**Kernel Type**

Gaussian

**Bandwidth Selector**

nrd0

Toggle Annotations

**Scatter****Heatmap**

Gene Expression for **GAD1**[Summary](#) [Explore](#) [Download](#) [Settings](#) [Bulk Download](#)Study Files (12)

Filename	Description	Download
CLUSTER_AND_SUBCLUSTER_INDEX.txt	Cluster and sub-cluster assignments for all single cells	52.9 KB
Coordinates_CA1.txt (Coordinates_CA1.txt)	Cluster coordinates for sub clusters in CA1	6.46 KB
Coordinates_CA3.txt (Coordinates_CA3.txt)	Cluster coordinates for sub-clusters in CA3	3 KB
Coordinates_DG.txt (Coordinates_DG.txt)	Cluster coordinates for sub-clusters in DG	27.6 KB
Coordinates_GABAergic.txt (Coordinates_GABAergic.txt)	Cluster coordinates for sub-clusters in GABAergic	5.67 KB
Coordinates_Glia.txt (Coordinates_Glia.txt)	Cluster coordinates for sub-clusters in Glia	4.53 KB
Coordinates_Major_cell_types.txt (Coordinates_Major_cell_types.txt)	Major cell types coordinates (top-level clusters)	49.2 KB
DATA_MATRIX_LOG TPM.txt	Expression Matrix, log(TPM) cells X genes	214 MB
GABAergic_edit_subcluster_marker_gene.txt	GABAergic cluster marker genes	26 KB
Glia_subcluster_marker_genes.txt	Glia cluster marker genes	20.7 KB
Major_cell_types_marker_genes.txt	Major cell types marker gene list	32.5 KB
README.txt	Experimental design & file format information	861 Bytes

Primary Data (1)

Show 10 entries

Search:

# Hackathon: Day 1

- Onboard team members
  - Expected 4 Broadies, 2-3 external
  - Actual 4 Broadies, ~**18** external
  - The more the merrier!
- Evaluate FAIRness of Single Cell Portal using Purple Polar Bears web survey app
- Brainstorm tasks for large number of new developers
- Get everyone integrated, unblocked, and coding productively

# Hackathon: Day 2

Implement features!

# Study metadata: Use cases and tasks

**User story:** provide FAIR metadata using Human Cell Atlas (HCA) community standards about studies

**Use case:** SCP has prose abstracts, needs to detect organism, tissue, etc. and map to ontology URL

**NLP analysis:** Soheil Danesh, Asya Lushnikova

**Use case:** Ontology mapping from NLP is list of key-value pairs, needs structure defined in community metadata standard

**TSV-JSON transform:** Tom Madden, David Managadze, Frank O

**Use case:** Community metadata standard uses text values, not URLs

**FAIRify HCA metadata:** Kenny Knecht, Adelaide Rhodes, Tom Madden

**Use case:** Community metadata standard uses JSON Schema, but ontologies often use OWL

**JSON-OWL transform:** Morgan Wahl

**Use case:** Tie together all these tasks in a pipeline

**Pipeline:** Alex Baumann, Tim Lee

**Use case:** Provide adapters to convert from HCA metadata standard to GEO metadata standard

**Converters:** Etienne Gnimpieba, Tayler Hoekstra, Isaac Hanson

**Use case:** NLP can't provide all the metadata we need

**Augment study creation UI/UX:** Anthony Dubois, Maya Bobrovitch, Kate Voss, Morgan Wahl

# Study metadata: Explicit classification and licensing

## Classification

Genus species

Homo sapiens

Organ

Blood

Organ part

Blood dendritic cells and monocytes

Disease

Disease

Please provide a valid disease.

## Project information

Visibility

Private  Public

Workspace

Create a new workspace

All uploaded data files for studies are stored in FireCloud workspaces. If you do not already have a FireCloud account, you will receive an invitation email after you create your study.

Billing Project

Default project (no analysis possible)

Licensing

CC BY ND 4.0

Get more information about licensing on [CreativeCommons.org](http://CreativeCommons.org)

Confirm

An NLP algorithm fills the form and the user can edit it using a standard dictionary (auto complete feature).

Study owner specifies an accessible usage license

# Analysis metadata: Use cases and tasks

- **Use case:** Analysis metadata conforms to community standards, but audience is restricted
  - **Enable public analysis.json:** Jon Bistline
- **Use case:** Analysis metadata is not easily machine-findable
  - **Enable analysis.json search:** Jon Bistline

# Analysis.json ‘FAIR’ payload

```
⚠ Not Secure | https://localhost/single_cell/analysis/download/d2d739e0-10a9-42a9-bec1-2e585394d203

{
  "url": "https://localhost/single_cell/analysis/download/d2d739e0-10a9-42a9-bec1-2e585394d203",
  "study_url": "https://localhost/single_cell/study/cellranger-demo",
  "license": {
    "name": "Creative Commons Attribution-NoDerivatives 4.0 International Public License",
    "url": "https://creativecommons.org/licenses/by-nd/4.0/legalcode"
  },
  "payload": {
    "inputs": [
      {
        "name": "cellranger.secondary",
        "value": true
      },
      {
        "name": "cellranger.expectCells",
        "value": 3000
      },
      {
        "name": "cellranger.fastq",
        "value": [
          "gs://fc-fd009b61-6524-4708-8480-8a4de7e49c88/cell_ranger_fastqs/test_sample_S1_L001_R1_001.fastq.gz",
          "gs://fc-fd009b61-6524-4708-8480-8a4de7e49c88/cell_ranger_fastqs/test_sample_S1_L001_R2_001.fastq.gz",
          "gs://fc-fd009b61-6524-4708-8480-8a4de7e49c88/cell_ranger_fastqs/test_sample_S1_L001_R1_001.fastq.gz"
        ]
      },
      {
        "name": "cellranger.sampleId",
        "value": "test_sample_S1_L001"
      },
      {
        "name": "cellranger.transcriptomeTarGz",
        "value": "gs://fc-bcc55e6c-bec3-4b2e-9fb2-5e1526ddfc2/reference_data/mouse/mm10/refdata-cellranger-mm10-1.2.0.tar.gz"
      },
      {
        "name": "cellranger.diskSpace",
        "value": "250"
      },
      {
        "name": "cellranger.referenceName",
        "value": "mm10"
      }
    ],
    "reference_bundle": "gs://fc-bcc55e6c-bec3-4b2e-9fb2-5e1526ddfc2/reference_data/mouse/mm10/refdata-cellranger-mm10-1.2.0.tar.gz",
    "tasks": [
      {
        "name": "cellranger.CellRanger",
        "disk_size": "local-disk 250 HDD",
        "zone": "us-central1-b,us-central1-c,us-central1-f",
        "log_err": "gs://fc-fd009b61-6524-4708-8480-8a4de7e49c88/d2d739e0-10a9-42a9-bec1-2e585394d203/cellranger/8401124f-c5fd-4c20-ab94-32a158b1aab/call-CellRanger/CellRanger-stderr.log",
        "start_time": "2018-03-22T15:36:30.321Z",
        "cpus": 64,
        "log_out": "gs://fc-fd009b61-6524-4708-8480-8a4de7e49c88/d2d739e0-10a9-42a9-bec1-2e585394d203/cellranger/8401124f-c5fd-4c20-ab94-32a158b1aab/call-CellRanger/CellRanger-stdout.log",
        "stop_time": "2018-03-22T15:36:40.339Z",
        "memory": "416 GB",
        "docker_image": "regevlab/cellranger-2.0.2"
      }
    ]
  }
}
```

# Analysis search API

```
← → C ▲ Not Secure | https://localhost/single_cell/analysis/search?study=SS2&name=cell&query_type=or

{
  "query": {
    "study": "SS2",
    "name": "cell",
    "query_type": "OR"
  },
  "results": [
    {
      "url": "https://localhost/single_cell/analysis/download/b7b61fd1-ebe4-4e92-9bcf-50d3221a79aa",
      "name": "SS2_scRNA_pipeline_3",
      "study_url": "https://localhost/single_cell/study/ss2-integration"
    },
    {
      "url": "https://localhost/single_cell/analysis/download/d2d739e0-10a9-42a9-becl-2e585394d203",
      "name": "cell-ranger-2-0-2_3_test_sample_S1_L001",
      "study_url": "https://localhost/single_cell/study/cellranger-demo"
    }
  ]
}
```

# Relation to FAIR Metrics

- FAIR Evaluator survey is just a start
- Single Cell Portal Resources has multiple resource types
- “Study” resource type increased in Interoperability, Reusability
- “Analysis” resource type increased in Findability, Accessibility

# FAIR score: 29% -> 52%

## Your FAIR score

Average Score

How FAIR is your dataset?

**52%**

Findable

**91%**

Accessible

**50%**

Interoperable

**33%**

Reusable

**33%**

# Collaborative coding

[https://github.com/BioITHackathons/single cell portal core](https://github.com/BioITHackathons/single_cell_portal_core)

45 commits from 12 people in < 2 days

Rapid onboarding to completely new project for 20 of 22 team members

Several people learned how to use Git and GitHub!

# Broad Institute hackathon team



Eric Weitz  
[eweitz@broadinstitute.org](mailto:eweitz@broadinstitute.org)  
<https://github.com/eweitz>



Jonathan Bistline  
[bistline@broadinstitute.org](mailto:bistline@broadinstitute.org)



Kate Voss  
[kvoss@broadinstitute.org](mailto:kvoss@broadinstitute.org)



Alexander Baumann  
[abaumann@broadinstitute.org](mailto:abaumann@broadinstitute.org)

# Thank you!

Eric Weitz

Jon Bistline

Kate Voss

Alex Baumann

Soheil Danesh

Austin Chow

Adelaide Rhodes

Maya Bobrovitch

Tim Lee

Tom Madden

Brian Kang

Brad Nissenbaum

Frank O

Morgan Wahl

Kenny Knecht

Julia Ivashina

David Managadze

Etienne Gnimpieba

Tayler Hoekstra

Isaac Hanson

Asya Lushnikova

Anthony Dubois