BRAIN Initiative Cell Census Network (BICCN) Omics Workshop

Find, export, and analyze 10x single-nuclei RNA-seq data using NeMO and Terra

Table of Contents

Step 0. Clone the BICCN Workshop Terra workspace

Background

This featured workspace is “Read only”. In order to run a workflow, you will need to make your own workspace copy and set up your own billing project (even though running these exercises will cost less than $2.00!).

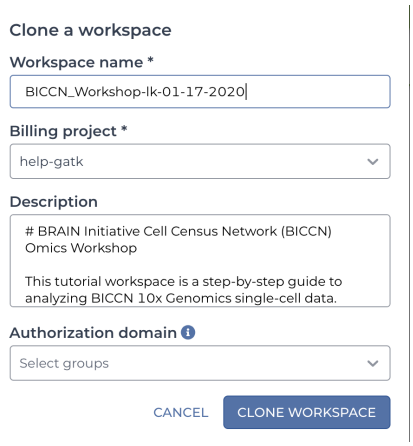
Estimated time and cost: 2 mins and < $0.01

Step-by-step instructions

0.1. Select the three vertical dot icon at the upper right corner of the [BICCN workspace].

0.2. Select “**Clone**” from the dropdown menu.

0.3. Rename your copy with something memorable.



0.4. Choose your billing project.

0.5. Do not select an Authorization Domain. These are only required when using restricted-access data.

0.6. Select the “**Clone Workspace**” button to make your own copy.

Take a few minutes to explore the **Data**, **Notebooks**, and **Workflows** tabs of the workspace. On the Data page, notice there are two preloaded data tables: **file** and

file_set. These tables contain a downsampled data set to be used in **Step 2**.

Select the **Workspace** data table. Here you find all the metadata for the genomic reference files necessary to run the tutorial workflows.

Go to the **Notebooks** page; it contains a Jupyter Notebook featuring the Seurat tutorial which will be used in Step 4.

Go to the **Workflows** page; it contains two workflows: **1-Optimus-mouse-v2** and **2-Cumulus** to be used in Steps 2 and 3 respectively. These workflows are set up to use mouse references that are listed in the Workspace data table.

Step 1. Explore and export NeMO data

Learning objectives

The Neuroscience Muti-omic Data Archive (NeMO) is a data repository for the BRAIN Initiative and related brain research initiatives. This section will teach you how to:

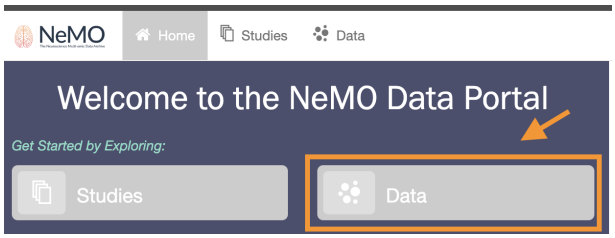
1. Identify data with faceted search in the NeMO Data Portal
2. Export NeMO data to an existing Terra workspace

Estimated time and cost: 10 mins and < $0.01

Step-by-step instructions

1.1. Go to the NeMO Data Portal.

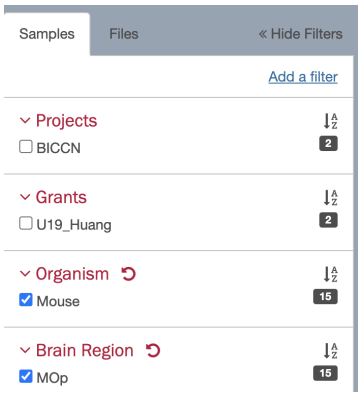
1.2. Under “Get Started by Exploring”, select the Data option.



1.3. Search for 10x data using the faceted search; you search by “Sample” and “File.”

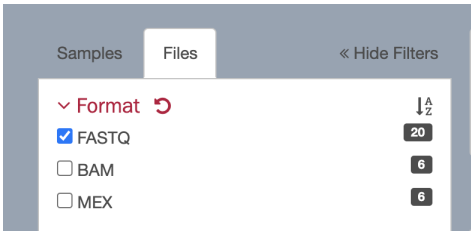
For example, to search for a mouse single-nuclei dataset from isocortex, use the following facets on the “**Sample**” tab:

- Organism: **Mouse**
- Brain Region: **MOp**
- Subspecimen Types: **Nucleus**
- Techniques: **10x Chromium 3' v2**

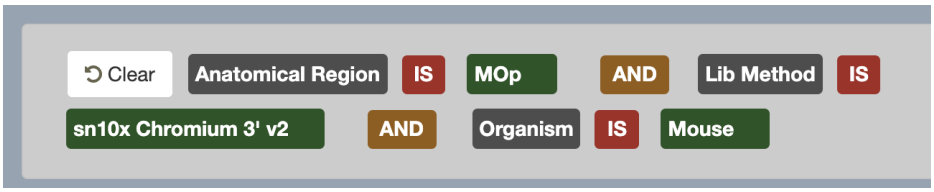


To narrow the search by a file format (like FASTQ) use the following facets on the "**File**" tab:

- Format: **FASTQ**



The selected facets will be listed at the top of the portal.



The graphical summary will display information about your cohort selection.

You can also narrow down cohorts by adding additional filters, like Sample ID. To find the exact files used for this tutorial, add a Sample ID filter to the NeMO search.

1.3.1. At the top of the **Samples** tab, select **add a filter**.

1.3.2. Type "Sample ID" in the search.

1.3.3. Select the Sample ID option.

1.3.4. Paste the sample ID for this tutorial into the search bar:

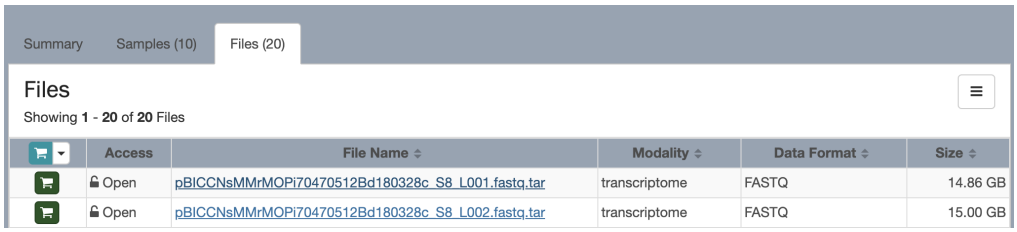
PBICCNSMMRMOPI70470512BD180328C

1.3.5. Select “Go!”

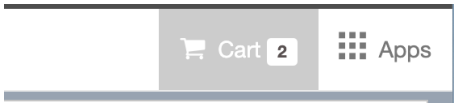
1.4. To export the cohort, select **Add all files to the cart** or go to the **Files** tab to add individual files to the cart by clicking the shopping cart icon left of the file name.

For this workspace example, select the following two files (as shown in the image below):

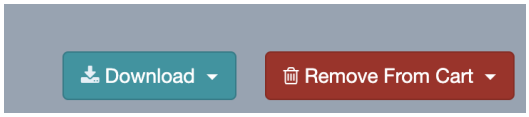
pBICCNsMMrMOPi70470512Bd180328c_S8_L001.fastq.tar
pBICCNsMMrMOPi70470512Bd180328c_S8_L002.fastq.tar



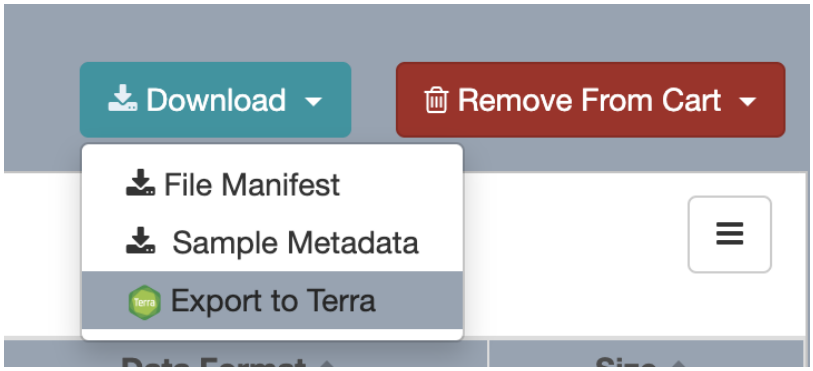
1.5. When ready to export, go to the **Cart** in the upper right of the portal.



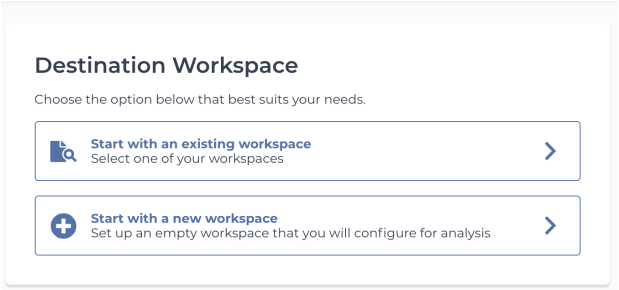
1.6. Select the **Download** icon.



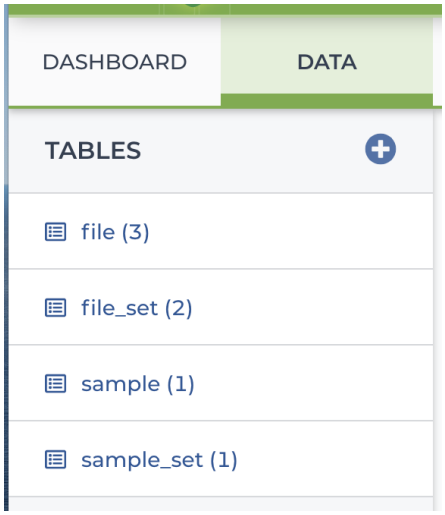
1.7. Select **Export to Terra.** You will be redirected to the Terra Platform.



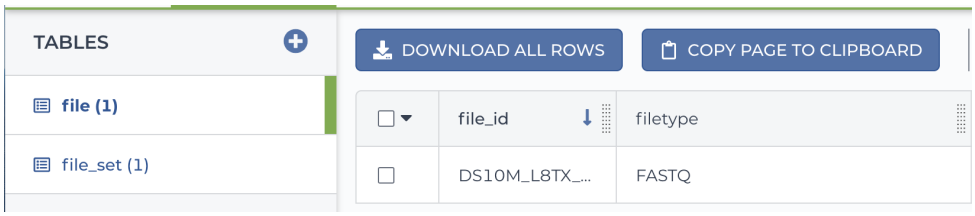
1.8. Select the **“Start with an existing workspace”** option and search for your cloned workspace.



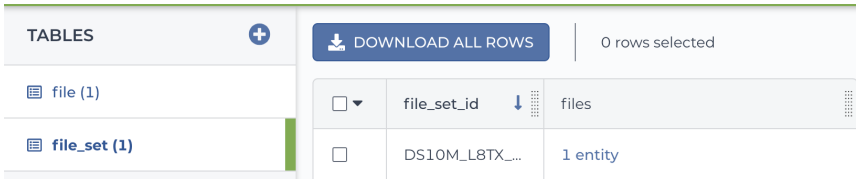
Data will import to the Terra Data page. You might need to refresh your browser to see the imported data tables which now include: **file**, **file_set**, **sample**, and **sample_set**.



The **file** table now contains files for each lane of 10x sequencing (rows of the table); each file has links to the raw data, including the three 10x sequencing FASTQ files (R1, R2, and I1 FASTQs).



The **file_set** table automatically groups lanes of sequencing that belong to the same library preparation.



The **sample** table contains sample metadata, such as organism, library method, and the project from which the sample is derived.

The **sample_set** table groups samples that come from the same project.

Step 2. Process 10x data with the Optimus workflow

Learning objectives

Optimus is an alignment, preprocessing, and quality control pipeline for 10x single-cell and single-nuclei RNA-seq data; it produces a cell by gene matrix containing raw gene counts as well as quality metrics. This section will teach you how to:

1. Set up the Optimus workflow using Terra data tables
2. Run the workflow and confirm outputs in the Terra workspace

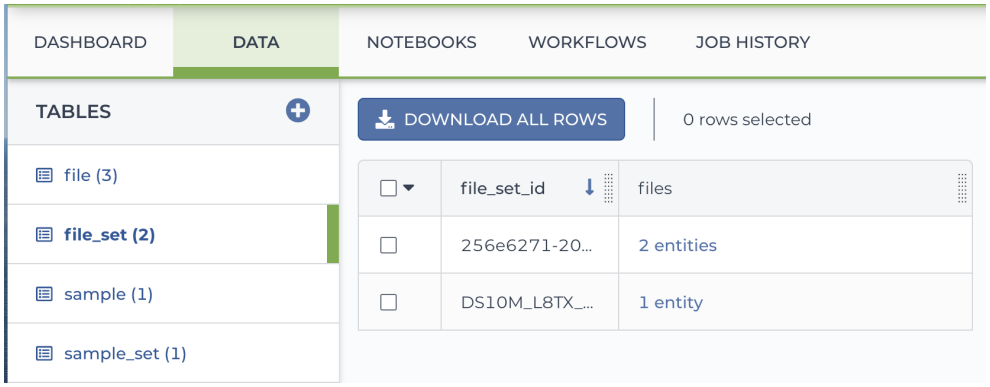
To save time and cost, this section uses a downsampled 10x NeMO dataset derived from primary motor cortex (MOp).

Estimated time and cost: 2 hrs for the workflow to run and ~ $0.26

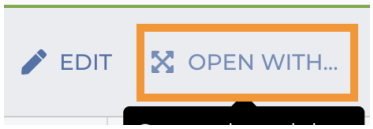
Step-by-step instructions

2.1. Go to to the **file_set** data table.

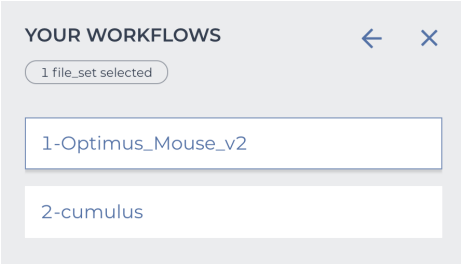
2.2. Check the box left of the **DS10M_L8TX_171026_01_A04** set.



2.3. Select "**Open with**" icon and then **Workflow.**



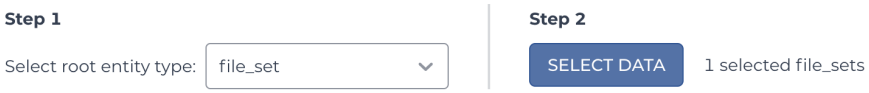
2.4. Select the **1-Optimus-mouse-v2 workflow.**



You will be redirected to the workflow configuration page.

2.5. Go to “Step 1.” on the configuration page.

Notice that the root entity type is **file_set.** Optimus requires an array of FASTQ files as input; Terra defaults to use the file_set data table because it knows the Optimus workflow requires a set and we selected the **DS10M_L8TX_171026_01_A04** file_set in section 2.2.



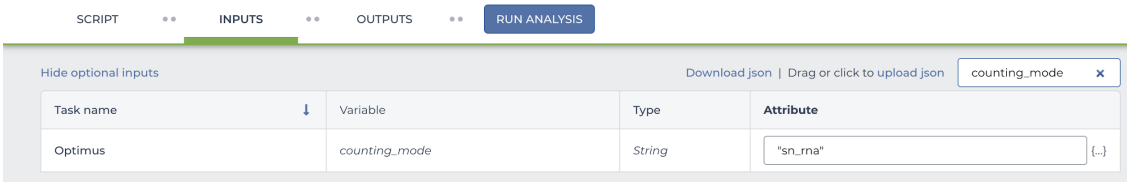
2.6. Go to the **Inputs** tab at the bottom of the configuration page and explore the different input variables and their preset attributes.

Thought Questions
When setting up the Optimus workflow inputs for your own data, think about the following:

1. What species is my data? Adjust reference files to point to those for the correct species (Optimus is validated for human and mouse).

2. What 10x chemistry does my data use? 10x chemistry is important because it affects the length of the cell barcodes in the **whitelist**. Optimus works for both 10x v2 and v3 chemistry which you can specify using the “chemistry” string. Make sure to adjust the whitelist attribute accordingly.

3. Is my data single-cell or single-nuclei? Optimus has a counting_mode variable; it runs sc_rna (single-cell) by default, but you can specify sn_rna mode for single-nuclei.



This workflow in this tutorial workspace is set up to run on mouse single-nuclei v2 data by default, but you should always double-check your setup!

2.7. In the **Inputs** tab, specify the attribute for input **r1_fastq** variable.

This attribute has been left blank to demonstrate how to specify a file from a Terra data table. We want to analyze a group of FASTQ files belonging to the same 10x library preparation; we specified this group of files in the file_set table which is why the file_set table is used as the root entity type. Terra locates files in the root_entity table using the “this.” syntax.

The R1 FASTQ files for our data set are listed in the **r1_fastq** column of the **file** data table.

To specify the r1 FASTQ files belonging to the data set, we use the following attribute:

this.files.r1_fastq

2.8. Select **Save.**

2.9. In the Outputs tab, specify an attribute for the **loom_output_file.**

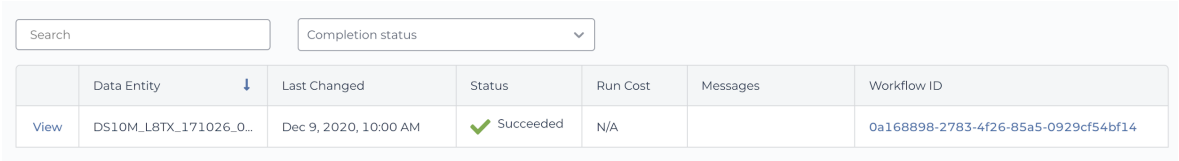
Even if we have multiple sequencing lanes, Optimus produces only one output Loom for the set of files. We want to write this output to the **file_set** data table, which means we will use the “this.” syntax, followed by whatever name we want to give the new data table column.

For example, you could change the attribute to **this.my_loom** to write a new column to the file_set data table called “my_loom”.

2.10. Select **Save.**

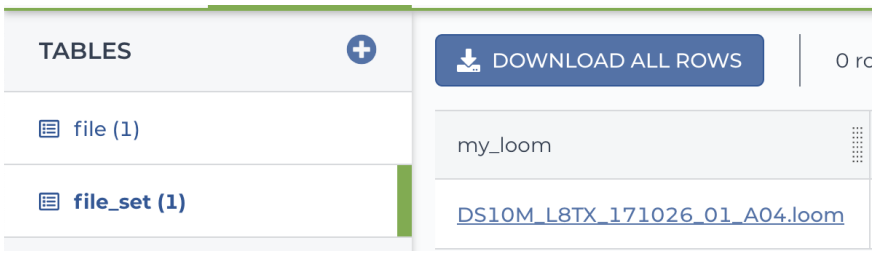
2.11. Select **Run Analysis** and then **Launch.**

You will be redirected to the **Job History** page. When the workflow successfully completes, you will see a green checkmark and the word “Succeeded”.



2.12. Go to the **file_set** data table.

Confirm that the count matrix (Loom format) file is in the **my_loom** column. You can use this file in Step 3 .



Step 3. Filter, normalize, and cluster Optimus output using Cumulus

Learning objectives

Cumulus is a cloud-based analysis workflow for large-scale single-cell and single-nucleus RNA-seq data (Li et al. 2020). The workflow takes in a raw count matrix and performs filtering, normalization, and clustering. This section will teach you how to:

1. Set up the Cumulus workflow using the raw count matrix (Loom) produced by Optimus
2. Run the Cumulus workflow and confirm outputs in the Terra workspace

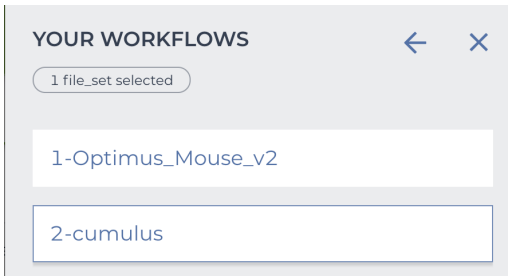
Estimated time and cost: 30 mins (11 mins for the workflow to run) and ~$0.12

Step-by-step instructions

3.1. Go to the **file_set** data table on the Data page and select the checkbox next to the **DS10M_L8TX_171026_01_A04.**

3.2. Select the "**Open with**" and then **Workflow.**

3.3. Select the **2-Cumulus workflow.**



You will be redirected to the workflow configuration page.

3.4. Explore the Input and Output tabs.

The workflow is set up to run Cumulus on the Loom file in the **output_loom** column of the **file_set** table.

Additionally, it produces an optional Seurat-compatible h5ad file and visualizations/embeddings that are compatible with Single Cell Portal, a platform for finding, exploring, and sharing single-cell data.

3.5. Go to the workflow **Inputs** tab.

3.6. Make sure the input_file variable is specified as the column with the Optimus loom file.

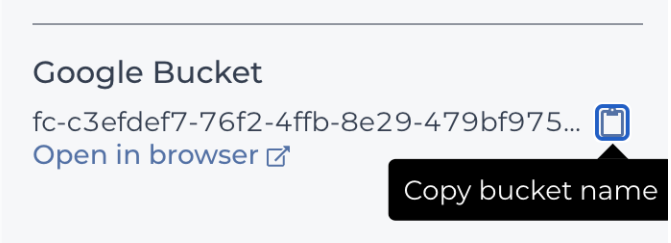
For this tutorial, the input_file attribute should be "**this.output_loom**", which points to a pre-generated Loom file that is identical to the one in the my_loom column you created in Step 2. This is so that you don't have to wait for Optimus to complete prior to trying Cumulus.

Alternatively, if your Optimus run is complete, you can set the attribute to "this.my_loom" to try the output *you* generated with Optimus.

3.7. Change the input attribute for the **output_directory** variable to a Cumulus folder in your Workspace Google Bucket.

The attribute is preset to "**gs://PASTE_BUCKET_ID_HERE/Cumulus/**". Just paste your Google Bucket ID over the PASTE_BUCKET_ID_HERE section.

You can find the Google Bucket location on your **Dashboard**.



For example, if your Google Bucket ID is **fc-c3efdef7-76f2-4ffb-8e29-479bf9758df9**, you would use the following string for the attribute:

"gs://fc-c3efdef7-76f2-4ffb-8e29-479bf9758df9/Cumulus/"

cumulus	output_directory	String	"gs://fc-c3efdef7-76f2-4ffb-8e29-479bf9758df9/Cumulus/" {...}

3.8. Select **Save**.

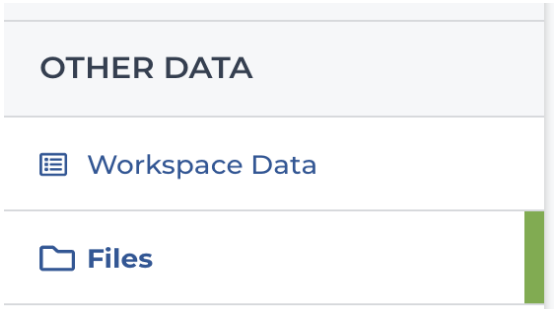
3.9. **Do not specify any output attributes in the Workflow Outputs tab**.

Cumulus will place outputs in the directory specified on the inputs tab, which in this example will write to the Files subsection of the **Other Data** section on the workspace Data page. Cumulus is not currently configured to write outputs to a Terra data table.

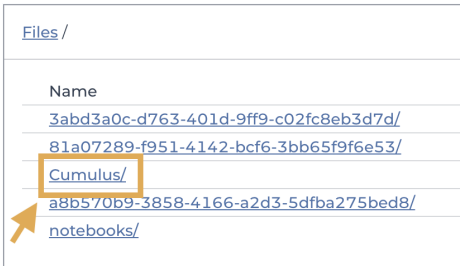
3.10. Select **Run Analysis** and then **Launch**.

You will be redirected to the **Job History** page. When the workflow successfully completes, you will see a green checkmark and the word "Succeeded".

3.11. Go to the Data page and select **Files** in the **Other Data** section.



3.12. Select the "**Cumulus**" folder and then the "**Downsample Brain**" folder to see the outputs.



Take a look at the different file types Cumulus produces. You can check the Cumulus documentation for a full description.

Some of the outputs will have a "**.scp**" suffix; these are outputs that are compatible with Single Cell Portal, a platform allowing you to visualize cell clusters identified by Cumulus.

Step 4. Explore single-cell data using Seurat

Learning objectives

Seurat is an R package designed for single-cell RNA-seq data analysis and exploration (Butler et al. 2018; Stuart, Butler et al. 2019). This section will teach you how to:

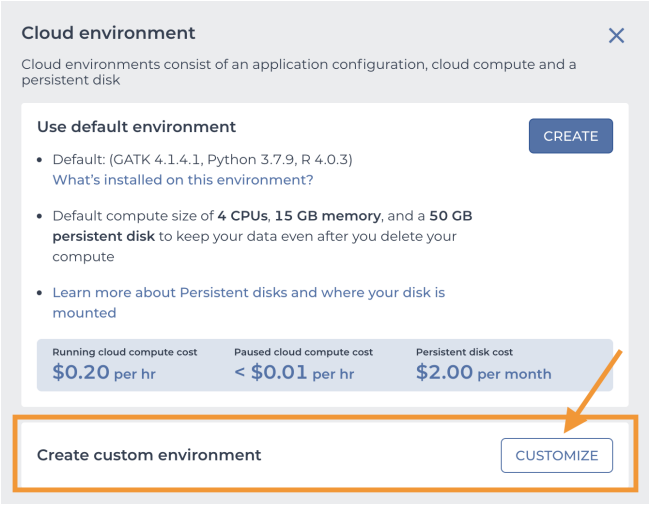
1. Set up a custom Cloud Environment for a Seurat Jupyter Notebook
2. Run a Seurat tutorial in a Jupyter Notebook

Estimated time and cost: 10 mins and < $0.20

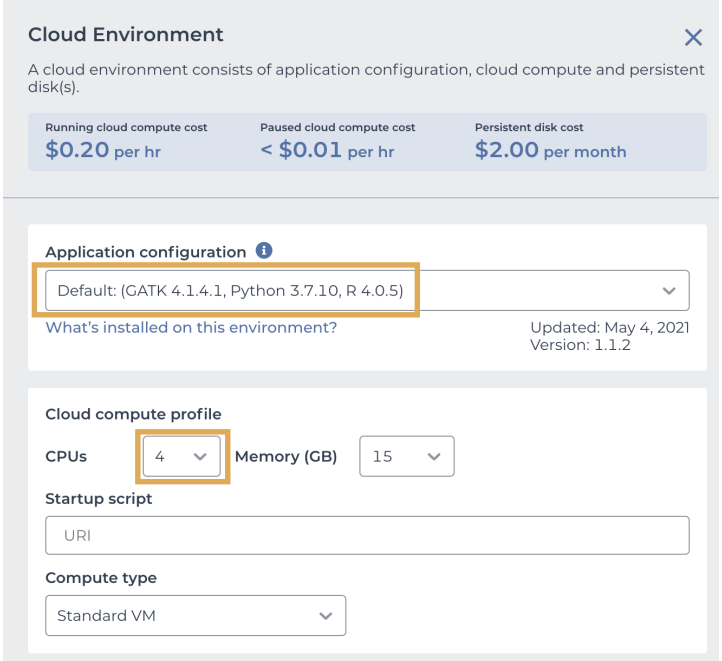
Step-by-step instructions

4.1. Select the **Cloud Environment** widget at the top right corner of the workspace.

4.2. Select the **Create custom environment**.



4.3. Select the **Default Environment** from the **Application configuration** drop-down.



4.4. Choose **4 CPUs** from the Cloud compute profile.

4.5. From the **Notebooks** page, open the **Seurat Notebook** in edit mode.

This Jupyter Notebook contains the Seurat Guided Clustering tutorial which is modified to use the Optimus output Loom matrix. For timing purposes, we've converted the Loom matrix to a Seurat object and stored the RDS file (brain.rds) in a public Google bucket, but the instructions will demonstrate how to convert and import into the Cloud Environment.

4.6. Follow the instructions in the Notebook to filter, normalize, and visualize single-cell data.

Congratulations! You've completed the BICCN Omics Terra Tutorial.

To learn more about how to use Terra or the different tools showcased in this tutorial, see the Resources section below.

Resources

Terra support documentation, videos, and workspaces

Get started on Terra with Terra support materials. These resources will teach you the basics of creating and using data tables, importing and running workflows, and using Jupyter Notebooks.

1. Terra support documentation
2. Terra support videos
3. Terra Data Tables Quickstart Featured Workspace
4. Terra Workflows Quickstart Featured Workspace
5. Terra Notebooks Quickstart Featured Workspace

Optimus resources

Read detailed information about the Optimus workflow in the Optimus overview or try running the workflow on different use cases (i.e. single cell, single nuclei, v2 or v3 chemistry) in the Featured Workspace.

1. Optimus Overview in the WARP (WDL Analysis Research Pipelines) GitHub repository
2. Optimus code and changelog in WARP
3. Optimus Featured Workspace in Terra

Cumulus resources

Learn more about the Cumulus workflow and parameters in the Cumulus documentation or try Cumulus in the Cumulus Featured Workspace.

1. Cumulus documentation
2. Cumulus Featured Workspace in Terra

Seurat resources

Try additional Seurat tutorials or learn more about Seurat tools in the Seurat documentation. You can also find more information about the Guided Clustering tutorial by visiting the original vignette below.

1. Seurat documentation
2. Seurat Guided Clustering Tutorial

References

1. BRAIN Initiative Cell Census Network (BICCN) et al. A multimodal cell census and atlas of the mammalian primary motor cortex. bioRxiv 2020.10.19.343129; doi: https://doi.org/10.1101/2020.10.19.343129

2. Butler, A., Hoffman, P., Smibert, P. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411–420 (2018). https://doi.org/10.1038/nbt.4096

3. Li, B., Gould, J., Yang, Y. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat Methods 17, 793–798 (2020). https://doi.org/10.1038/s41592-020-0905-x

4. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of Single-Cell Data. Cell. 2019 Jun 13;177(7):1888-1902.e21. doi: 10.1016/j.cell.2019.05.031. Epub 2019 Jun 6. PMID: 31178118; PMCID: PMC6687398.