

---

# Cumulus Documentation

*Release 2.3.0*

**Bo Li, Joshua Gould, and et al.**

**Jan 29, 2023**



---

## Contents

---

<b>1</b>	<b>Release Highlights in Current Stable</b>	<b>3</b>
----------	---	----------



All of our docker images are publicly available on [Quay](#) and [Docker Hub](#). Our workflows use Quay as the default Docker registry. Users can use Docker Hub as the Docker registry by entering `cumulusprod` for the workflow input **“docker\_registry”**, or enter a custom registry name of their own choice.

If you use Cumulus in your research, please consider citing:

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>



---

## Release Highlights in Current Stable

---

### 1.1 2.4.0 January 28, 2023

#### Updates:

- In Cellranger workflow:
  - Upgrade *cellranger\_version* default to 7.1.0.
  - Add Mouse probe set v1.0 for [Fixed RNA Profiling](#) analysis.
  - Add probe set v1.0.1 of both Human and Mouse for [Fixed RNA Profiling](#) analysis.
  - Upgrade *cumulus\_feature\_barcoding* version default to 0.11.1.
- In Cellranger\_create\_reference workflow: upgrade *cellranger\_version* default to 7.1.0.
- In Cellranger\_vdj\_create\_reference workflow: upgrade *cellranger\_version* default to 7.1.0.
- In Spaceranger workflow: upgrade *spaceranger\_version* default to 2.0.1.

#### 1.1.1 First Time Running on Terra

##### Authenticate with Google

If you've done this before you can skip this step - you only need to do this once.

1. Ensure the [gcloud CLI](#) is installed on your computer.

Note: Broad users do not have to install this-they can type:

```
reuse Google-Cloud-SDK
```

to make the Google Cloud tools available.

2. Execute the following command to login to Google Cloud.:

```
gcloud auth login
```

3. Copy and paste the link in your unix terminal into your web browser.
  4. Enter authorization code in unix terminal.
- 

### Create a Terra workspace

1. Create a new [Terra](#) workspace by clicking Create New Workspace in Terra

Further reading: [Terra tutorials](#).

### 1.1.2 Cumulus workflows overview

Cumulus workflows are written in [WDL](#) language, and published on [Dockstore](#). Below is an overview of them:



Workflow	First Version	Date Added	Function
Cellranger	0.1.0	2018-07-27	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generating count matrix using cellranger count or cellranger-atac count, running cellranger vdj or feature-barcode extraction.
Spaceranger	1.2.0	2021-01-19	Run Space Ranger tools to process spatial transcriptomics data, which includes extracting sequence reads using spaceranger mkfastq, and generating count matrix using spaceranger count.
STARsolo	1.2.0	2021-01-19	Run STARsolo to generate gene-count matrices from FASTQ files.
GeoMx_fastq_to_dcc	2.2.0	2022-10-04	Run Nanostring GeoMx Digital Spatial NGS Pipeline, and convert FASTQ files into DCC files.
GeoMx_dcc_to_count_matrix	2.2.0	2022-10-04	Take the DCC zip file from <i>GeoMxFastqToDCC</i> workflow, as well as other output of GeoMx DSP machine as the input, and generate an Area Of Interest (AOI) by probe count matrix with pathologists' annotation.
Demultiplexing	0.3.0	2018-10-24	Run tools (demuxEM, souporecell, or popsicle) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
Cumulus	0.1.0	2018-07-27	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
Cellbender	2.1.0	2022-07-13	Run CellBender tool to remove technical artifacts from high-throughput single-cell/single-nucleus RNA sequencing data.
Cellranger_create_reference	0.12.0	2019-12-14	Run Cell Ranger tools to build sc/snRNA-seq references.
Cellranger_atac_create_reference	0.12.0	2019-12-14	Run Cell Ranger tools to build scATAC-seq references.
cellranger_vdj_create_reference	0.12.0	2019-12-14	Run Cell Ranger tools to build single-cell immune profiling references.
STARsolo_create_reference	2.0.0	2022-03-14	Run STAR to build sc/snRNA-seq references for STARsolo count.
Cellranger_atac_aggr	0.13.0	2020-02-07	Run Cell Ranger tools to aggregate scATAC-seq samples.
Smart-Seq2	0.5.0	2018-11-18	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files.
Smart-Seq2_create_reference	0.12.0	2019-12-14	Generate user-customized genome references for SMART-Seq2 data.

## Legacy versions on Broad Method Registry

As Cumulus is now switched to [Dockstore](#) for release, we no longer maintain the Cumulus workflows published on Broad Method Registry.

But Terra users can still check out the legacy snapshots listed below for usage.

**Stable version - v1.5.1**

WDL	Snapshot	Function
cumulus/cellranger_workflow	28	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generating count matrix using cellranger count or cellranger-atac count, running cellranger vdj or feature-barcode extraction.
cumulus/spaceranger_workflow	3	Run Space Ranger tools to process spatial transcriptomics data, which includes extracting sequence reads using spaceranger mkfastq, and generating count matrix using spaceranger count.
cumulus/star_solo	7	Run STARsolo to generate gene-count matrices from FASTQ files.
cumulus/count	18	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	32	Run tools (demuxEM, soupcell, or popsicle) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	40	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	5	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	4	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	10	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files.
cumulus/smartseq2_create_reference	6	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	43	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.

## Stable version - v1.5.0

WDL	Snapshot	Function
cumulus/cellranger_workflow	26	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generating count matrix using cellranger count or cellranger-atac count, running cellranger vdj or feature-barcode extraction.
cumulus/spaceranger_workflow	3	Run Space Ranger tools to process spatial transcriptomics data, which includes extracting sequence reads using spaceranger mkfastq, and generating count matrix using spaceranger count.
cumulus/star_solo	7	Run STARsolo to generate gene-count matrices from FASTQ files.
cumulus/count	18	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	31	Run tools (demuxEM, soupfilter, or popscle) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	40	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	5	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	4	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	10	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files.
cumulus/smartseq2_create_reference	6	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	43	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.

**Stable version - v1.4.0**

WDL	Snapshot	Function
cumulus/cellranger_workflow	26	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generating count matrix using cellranger count or cellranger-atac count, running cellranger vdj or feature-barcode extraction.
cumulus/spaceranger_workflow	3	Run Space Ranger tools to process spatial transcriptomics data, which includes extracting sequence reads using spaceranger mkfastq, and generating count matrix using spaceranger count.
cumulus/star_solo	6	Run STARsolo to generate gene-count matrices from FASTQ files.
cumulus/count	18	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	30	Run tools (demuxEM, souporcell, or popsicle) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	10	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	5	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	4	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	10	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files.
cumulus/smartseq2_create_reference	6	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	41	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.

## Stable version - v1.3.0

WDL	Snapshot	Function
cumulus/cellranger_workflow	15	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generating count matrix using cellranger count or cellranger-atac count, running cellranger vdj or feature-barcode extraction.
cumulus/spaceranger_workflow	1	Run Space Ranger tools to process spatial transcriptomics data, which includes extracting sequence reads using spaceranger mkfastq, and generating count matrix using spaceranger count.
cumulus/star_solo	3	Run STARsolo to generate gene-count matrices from FASTQ files.
cumulus/count	18	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	22	Run tools (demuxEM, souporecell, or demuxlet) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	2	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	2	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	2	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	7	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files.
cumulus/smartseq2_create_reference	2	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	36	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.

## Stable version - v1.2.0

WDL	Snapshot	Function
cumulus/cellranger_workflow	15	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generating count matrix using cellranger count or cellranger-atac count, running cellranger vdj or feature-barcode extraction.
cumulus/spaceranger_workflow	1	Run Space Ranger tools to process spatial transcriptomics data, which includes extracting sequence reads using spaceranger mkfastq, and generating count matrix using spaceranger count.
cumulus/star_solo	3	Run STARsolo to generate gene-count matrices from FASTQ files.
cumulus/count	18	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	22	Run tools (demuxEM, souporecell, or demuxlet) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	2	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	2	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	2	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	7	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files.
cumulus/smartseq2_create_reference	2	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	35	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.

## Stable version - v1.1.0

WDL	Snapshot	Function
cumulus/cellranger_workflow	14	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/star_solo	3	Run STARsolo to generate gene-count matrices from FASTQ files.
cumulus/count	16	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	21	Run tools (demuxEM, souporecell, or demuxlet) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	0	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	2	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	3	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	3	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	7	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/smartseq2_create_reference	0	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	34	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.

## Stable version - v1.0.0

WDL	Snapshot	Function
cumulus/cellranger_workflow	12	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/count	14	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/demultiplexing	20	Run tools (demuxEM, souporecell, or demuxlet) for cell-hashing/nucleus-hashing/genetic-pooling analysis.
cumulus/cellranger_create_reference	2	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	2	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	2	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	7	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/smartseq2_create_reference	8	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	31	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_hashing_cite_seq	10	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis

## Stable version - v0.15.0

WDL	Snapshot	Function
cumulus/cellranger_workflow	10	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/count	14	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/cellranger_create_reference	2	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	2	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference	2	Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference	2	Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	7	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/smartseq2_create_reference	8	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	24	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_subcluster	16	Run subcluster analysis using cumulus
cumulus/cumulus_hashing_cite_seq	10	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis



**Stable version - v0.14.0**

WDL	Snapshot	Function
cumulus/cellranger_workflow	8	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/count	11	Run alternative tools (STARsolo, Optimus, Salmon alevin, or Kallisto BUStools) to generate gene-count matrices from FASTQ files.
cumulus/cellranger_create_reference	6	Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	1	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference		Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference		Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	7	Run HISAT2/STAR/Bowtie2-RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/smartseq2_create_reference	6	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	16	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_subcluster	10	Run subcluster analysis using cumulus
cumulus/cumulus_hashing_cite_seq	4	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis

**Stable version - v0.13.0**

WDL	Snapshot	Function
cumulus/cellranger_workflow	7	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/cellranger_create_reference		Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_aggr	1	Run Cell Ranger tools to aggregate scATAC-seq samples.
cumulus/cellranger_atac_create_reference		Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference		Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	5	Run Bowtie2 and RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/smartseq2_create_reference	4	Generate user-customized genome references for SMART-Seq2 data.
cumulus/cumulus	14	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_subcluster	9	Run subcluster analysis using cumulus
cumulus/cumulus_hashing_cite_seq	7	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis

**Stable version - v0.12.0**

WDL	Snapshot	Function
cumulus/cellranger_workflow	6	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/cellranger_create_reference		Run Cell Ranger tools to build sc/snRNA-seq references.
cumulus/cellranger_atac_create_reference		Run Cell Ranger tools to build scATAC-seq references.
cumulus/cellranger_vdj_create_reference		Run Cell Ranger tools to build single-cell immune profiling references.
cumulus/smartseq2	5	Run Bowtie2 and RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/smartseq2_create_reference		Generate user-customized genome references for SMART-Seq2 workflow.
cumulus/cumulus	11	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_subcluster	8	Run subcluster analysis using cumulus
cumulus/cumulus_hashing_cite_seq	4	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis

**Stable version - v0.11.0**

WDL	Snapshot	Function
cumulus/cellranger_workflow	4	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/smartseq2	3	Run Bowtie2 and RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/cumulus	8	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_subcluster	5	Run subcluster analysis using cumulus
cumulus/cumulus_hashing_cite_seq	4	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis

**Stable version - v0.10.0**

WDL	Snapshot	Function
cumulus/cellranger_workflow	3	Run Cell Ranger tools, which include extracting sequence reads using cellranger mkfastq or cellranger-atac mkfastq, generate count matrix using cellranger count or cellranger-atac count, run cellranger vdj or feature-barcode extraction
cumulus/smartseq2	3	Run Bowtie2 and RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
cumulus/cumulus	7	Run cumulus analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering, visualization, differential expression analysis, cell type annotation, etc.
cumulus/cumulus_subcluster	4	Run subcluster analysis using cumulus
cumulus/cumulus_hashing_cite_seq	4	Run cumulus for cell-hashing/nucleus-hashing/CITE-Seq analysis

**Stable version - HTAPP v2**

WDL	Snapshot	Function
regev/cellranger_mkfastq_count	45	Run Cell Ranger to extract FASTQ files and generate gene-count matrices for 10x genomics data
scCloud/smartseq2	5	Run Bowtie2 and RSEM to generate gene-count matrices for SMART-Seq2 data from FASTQ files
scCloud/scCloud	14	Run scCloud analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering and more
scCloud/scCloud_subcluster	9	Run subcluster analysis using scCloud
scCloud/scCloud_hashing_cite_seq	9	Run scCloud for cell-hashing/nucleus-hashing/CITE-Seq analysis

**Stable version - HTAPP v1**

WDL	Snapshot	Function
regev/cellranger_mkfastq_count	39	Run Cell Ranger to extract FASTQ files and generate gene-count matrices for 10x genomics data
scCloud/scCloud	3	Run scCloud analysis module for variable gene selection, batch correction, PCA, diffusion map, clustering and more

**1.1.3 Import workflows to Terra**

Cumulus workflows are hosted on [Dockstore](#) under the organization of *Li Lab*. For illustration, we'll use *Cellranger* workflow to show how to import Cumulus workflows to your [Terra](#) workspace.

1. Select *Cellranger* workflow from [Cumulus workflow collection](#) by clicking its “View” button:

Notice that all Cumulus workflows have `github.com/lilab-bcb/cumulus/` prefix, which indicates they are imported from Cumulus GitHub repo to Dockstore.

2. In the workflow page, by switching to “Versions” tab, you can view all the available versions of *Cellranger* workflow, where the default version is on the top:

Git Reference	Date Modified	Valid	Verified Platforms	Snapshot	Actions
master	2022-03-14 21:33	✓			Actions
Default 1.5.0	2021-07-20 01:10	✓			Actions
1.4.0	2021-05-17 16:05	✓			Actions
1.3.0	2021-02-02 23:31	✓			Actions

To change to a non-default version, simply clicking the version name in “Git Reference” column. After that, click “Terra” button on the right panel.

**Note:** The **master** version refers to the development branch of Cumulus workflows, which is always under rapid change.

For stable usage, please always refer to a [released version](#).

3. You’ll be asked to log in to Terra if not. Then you can see the following page:

**Importing from Dockstore**

github.com/lilab-bcb/cumulus/Cellranger  
V. master

**⚠ Please note:** Dockstore cannot guarantee that the WDL and Docker image referenced by this Workflow will not change. We advise you to review the WDL before future runs.

```

1 version 1.0
2
3 import "cellranger_mkfastq.wdl" as crm
4 import "cellranger_count.wdl" as crc
5 import "cellranger_multi.wdl" as crmulti
6 import "cellranger_vdj.wdl" as crv
7 import "../cumulus/cumulus_adt.wdl" as ca
8 import "cellranger_atac_mkfastq.wdl" as cram
9 import "cellranger_atac_count.wdl" as crac
10 import "cellranger_arc_mkfastq.wdl" as crarm
11 import "cellranger_arc_count.wdl" as crarc
12
13 workflow cellranger_workflow {
14   input {
15     # 5 - 9 columns (Sample, Reference, Flowcell, Lane, Index, [Chemis
16     File input_csv_file
17     # Output directory, gs URL
18     String output_directory
19
20     # If run mkfastq
21     Boolean run_mkfastq = true
22     # If run count
23     Boolean run_count = true
  
```

**Workflow Name**

Cellranger

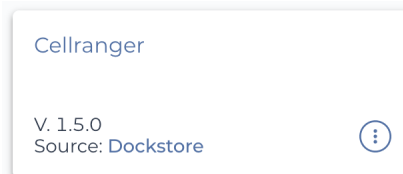
**Destination Workspace**

Select a workspace ▼

IMPORT Or create a new workspace

In “Destination Workspace” drop-down menu on the right panel, you can select the target Terra workspace to import *CellRanger* workflow. Optionally, you can even rename the workflow in “Workflow Name” field. When everything is done, click “IMPORT” button below to finish.

4. When finished, you can see *Cellranger* workflow appearing in “WORKFLOWS” tab of your Terra workspace:



Moreover, in its workflow page (as below)

The screenshot shows the Cumulus Workflows interface. At the top, there are tabs: DASHBOARD, DATA, NOTEBOOKS, WORKFLOWS (selected), and JOB HISTORY. Below the tabs, there is a "Back to list" link. The main section is titled "Cellranger" with an information icon. Below the title, there is a "Version" dropdown menu set to "1.5.0" and a "Source" field with the link "github.com/lilab-bcb/cumulus/Cellranger:1.5.0". Below the source, there is a "Synopsis" section with the text "No documentation provided". There are two radio buttons for running the workflow: "Run workflow with inputs defined by file paths" and "Run workflow(s) with inputs defined by data table" (selected). Below the radio buttons, there are two steps: "Step 1" and "Step 2". "Step 1" has a "Select root entity type:" dropdown menu set to "pair\_set". "Step 2" has a "SELECT DATA" button and the text "No data selected". Below the steps, there are four checkboxes: "Use call caching" (checked), "Delete intermediate outputs" (unchecked), "Use reference disks" (unchecked), and "Retry with more memory" (unchecked). Below the checkboxes, there are four tabs: SCRIPT, INPUTS (selected), OUTPUTS, and a "RUN ANALYSIS" button. Below the tabs, there is a table with the following data:

Task name ↓	Variable	Type	Attribute
cellranger_workflow	input_csv_file	File	Required

you can even switch the workflow's version in "Version" drop-down menu, and click the link in "Source" field to view the workflow's WDL source code.

## 1.1.4 Release notes

### Version 2.4

#### 2.4.0 January 28, 2023

##### Updates:

- In Cellranger workflow:
  - Upgrade *cellranger\_version* default to 7.1.0.
  - Add Mouse probe set v1.0 for **Fixed RNA Profiling** analysis.
  - Add probe set v1.0.1 of both Human and Mouse for **Fixed RNA Profiling** analysis.
  - Upgrade *cumulus\_feature\_barcoding* version default to 0.11.1.
- In Cellranger\_create\_reference workflow: upgrade *cellranger\_version* default to 7.1.0.
- In Cellranger\_vdj\_create\_reference workflow: upgrade *cellranger\_version* default to 7.1.0.
- In Spaceranger workflow: upgrade *spaceranger\_version* default to 2.0.1.

### Version 2.3

#### 2.3.0 October 30, 2022

##### New Features:

- Add support for [Fixed RNA Profiling](#) to Cellranger workflow.

**Updates:**

- In Cellranger workflow:
  - Upgrade *cellranger\_version* default to 7.0.1.
  - Upgrade *cellranger\_arc\_version* default to 2.0.2.
- In Cellranger\_create\_reference workflow: upgrade *cellranger\_version* default to 7.0.1.
- In Cellranger\_vdj\_create\_reference workflow: upgrade *cellranger\_version* default to 7.0.1.

**Version 2.2****2.2.0 October 4, 2022****New Features:**

- Add [Nanostring GeoMx DSP](#) workflows. It consists of two steps:
  - **GeoMx\_fastq\_to\_dcc** workflow to convert FASTQ files into DCC files by wrapping Nanostring GeoMx Digital Spatial NGS Pipeline.
  - **GeoMx\_dcc\_to\_count\_matrix** workflow to generate probe count matrix from DCC files with pathologists' annotation.

**Updates:**

- Spaceranger workflow:
  - Add support on 10x Space Ranger v2.0.0.
  - Add *human\_probe\_v2* Probe Set for FFPE samples, which is compatible with CytAssist FFPE samples.
- Upgrade *cumulus\_feature\_barcoding\_version* default to [v0.11.0](#) for Feature Barcoding in Cellranger workflow.

**API Changes:**

- Across all workflows, for AWS backend:
  - All workflows now have *awsQueueArn* input, which is used for explicitly specifying the Arn string of an AWS Compute Environment.
  - Remove *awsMaxRetries* input for all workflows. Namely, use Cromwell's default value 0.

**Bug Fix:**

- Fix the issue on localizing GCP folders in Cellranger workflow for ATAC-Seq and 10x Multiome data.
- 

**Version 2.1****2.1.1 July 18, 2022**

- Make *cumulus* workflow work with Cromwell v81+.

## 2.1.0 July 13, 2022

### New Features:

- Add [CellBender](#) workflow for ambient RNA removal.
- *CellRanger*:
  - For **ATAC-Seq** data, add `ARC-v1` chemistry keyword for analyzing only the ATAC part of 10x multiome data. See [CellRanger scATAC-seq sample sheet](#) section for details.
  - For **antibody/hashing/citeseq/crispr** data, add `multiome` chemistry keyword for the feature barcoding on 10x multiome data.
- *STARsolo*:
  - In workflow output, besides `mtx` format gene-count matrices, the workflow also generates matrices in 10x-compatible `hdf5` format.

### Improvements:

- *CellRanger*: For **antibody/hashing/citeseq/crispr** data,
  - `cumulus_feature_barcoding` v0.9.0+ now supports multi-threading and faster gzip file I/O.
- Workflows check if `output_directory` is a valid Cloud URI based on the given backend value before execution. (Feature request #322 )

### Updates:

- Genome Reference:
  - Add Cellranger VDJ v7.0.0 genome references: `GRCh38_vdj_v7.0.0` and `GRCm38_vdj_v7.0.0` in [CellRanger scIR-seq sample sheet](#) section.
- Default version upgrade:
  - Update *cellranger* default to v7.0.0.
  - Update *cellranger-atac* default to v2.1.0.
  - Update *cellranger-arc* default to v2.0.1.
  - Update *cumulus\_feature\_barcoding* default to v0.10.0.
  - *STARsolo* workflow uses STAR v2.7.10a by default.

---

## Version 2.0

### 2.0.0 March 14, 2022

#### Overall:

- Cumulus workflows are now released on [Dockstore](#):
  - Add the tutorial on [importing Cumulus workflows to Terra](#).
  - Archive the [legacy versions](#) on Broad Method Registry.
- Add support on multiple platforms via **backend** input: `gcp` for Google Cloud, `aws` for Amazon AWS, `local` for local machine. Enable Google Cloud support by default.



- For Amazon AWS backend, add **awsMaxRetries** input to set the maximum retries allowed for job execution at runtime. By default, use 5.
- Update the [command-line job submission](#) tutorial to work with [Altocumulus v2.0.0](#) or later.
- On Examples:
  - Update [gene expression, hashing and CITE-Seq](#) example tutorial.
  - Add tutorial on [10x CellPlex analysis](#) using Cumulus workflows on Cloud.

#### Workflow-specific:

- Add *STARsolo\_create\_reference* workflow to build genome references for STARsolo counting. See its [documentation](#) for details.
- On *Cellranger* workflow:
  - Add support for 10x Cell Ranger version 6.1.1 and 6.1.2, and use 6.1.2 by default. See [Cell Ranger v6.1 release notes](#).
  - Add support for 10x Cell Ranger ARC version 2.0.1, and use it by default. See [Cell Ranger ARC v2.0 release notes](#) for the release notes.
  - Upgrade [cumulus\\_feature\\_barcoding](#) to version 0.7.0 to allow manually set barcode starting position (via input **crispr\_barcode\_pos**).
  - Add support for non 10x CRISPR assays. See the description of **crispr** **DataType** value in [this section](#) for details.
  - For input data consisting of fastq files, it's able to handle folder structure of both flat (all fastq files in one folder) and nested (one subfolder per sample listed in the input sample sheet) forms.
  - Add **fastq\_outputs** to workflow output, which contains *mkfastq* step output folders for samples listed in the input sample sheet.
  - Add **count\_outputs** to workflow output, which contains *count* step output folders for samples listed in the input sample sheet.
- On *Spaceranger* workflow:
  - Add support for 10x Space Ranger version 1.3.0 and 1.3.1, and use 1.3.1 by default. See [Space Ranger v1.3 release notes](#) for the release notes.
  - For input data consisting of fastq files, it's able to handle folder structure of both flat (all fastq files in one folder) and nested (one subfolder per library) forms.
  - Add output section for the workflow. See [here](#) for details.
  - Retire old genome references:
    - \* Keep GRCh38-2020-A and mm10-2020-A.
    - \* Retire GRCh38, mm10, GRCh38-2020-A-premrna and mm10-2020-A-premrna. Users can still reach out to [Cumulus team](#) to ask for URIs to these old references, but they are not provided by default.
  - In the description of **ReorientImages** field of input sample sheet, add the information on its valid values.
- On *STARsolo* workflow:
  - Add support for STAR version 2.7.9a, and use it by default. See [STAR v2.7.9a release notes](#) for the release notes.
  - Reorganize the workflow by exposing more inputs to users.

- Add support on more protocols: 10x multiome, 10x 5' (both SC5P-R2 and SC5P-PE), Slide-Seq and Share-Seq. See [here](#) <./starsolo.html#prepare-a-sample-sheet> for details.
  - Use input **read1\_fastq\_pattern** and **read2\_fastq\_pattern** to support fastq files generated by Cell Ranger or SeqWell, as well as Sequence Read Archive (SRA) data.
  - For input data consisting of fastq files, it's able to handle folder structure of both flat (all fastq files in one folder) and nested (one subfolder per library) forms.
  - Do not attach filename prefix to output files to avoid the incorrect SJ raw *feature.tsv* symlink error, which would cause the folder delocalization fail. (see [discussion](#) with [STAR](#) team)
  - Add STAR log file to workflow output. This is the *Log.out* file if running STAR locally, which can be used for tracking the process and sharing with [STAR](#) team when opening an issue there.
  - Retire old genome references:
    - \* Keep GRCh38-2020-A, mm10-2020-A, and GRCh38-and-mm10-2020-A.
    - \* Retire old references listed [here](#). Users can still reach out to [Cumulus team](#) to ask for URIs to them, but they are not provided by default.
  - On *Demultiplexing* workflow:
    - Upgrade [demuxEM](#) to version 0.1.7 for bug fix.
  - On *Cellranger\_create\_reference* workflow:
    - Add the generated reference file to the workflow output.
    - Bug fix in using input **memory**.
    - Update documentation to suggest only using Cell Ranger version 6.1.1 or later for building reference, as v6.0.1 has issues which leave the job running without terminating.
  - On *Cellranger\_atac\_create\_reference* workflow:
    - Add the generated reference file to the workflow output.
  - On *Cellranger\_vdj\_create\_reference* workflow:
    - Add the generated reference file to the workflow output.
- 

## Version 1.x

### Version 1.5.1 September 15, 2021

- Fix the issue of WDLs after Terra platform updates the Cromwell engine.

### Version 1.5.0 July 20, 2021

- On *demultiplexing* workflow
  - Update *demuxEM* to v0.1.6.
- On *cumulus* workflow
  - Add Nonnegative Matrix Factorization (NMF) feature: *run\_nmf* and *nmf\_n* inputs.
  - Add integrative NMF (iNMF) data integration method: *inmf* option in *correction\_method* input; the number of expected factors is also specified by *nmf\_n* input.

- When NMF or iNMF is enabled, word cloud plots and gene program UMAP plots of NMF/iNMF results will be generated.
- Update *Pegasus* to v1.4.2.

### Version 1.4.0 May 17, 2021

- **On *cellranger* workflow**

- Add support for multiomics analysis using linked samples, *cellranger-arc count*, *cellranger multi* and *cellranger count* will be automatically triggered based on the sample sheet
- Add support for cellranger version 6.0.1 and 6.0.0
- Add support for cellranger-arc version 2.0.0, 1.0.1, 1.0.0
- Add support for cellranger-atac version 2.0.0
- Add support for cumulus\_feature\_barcoding version 0.6.0, which handles CellPlex CMO tags
- Add *GRCh38-2020-A\_arc\_v2.0.0*, *mm10-2020-A\_arc\_v2.0.0*, *GRCh38-2020-A\_arc\_v1.0.0* and *mm10-2020-A\_arc\_v1.0.0* references for *cellranger-arc*.
- Fixed bugs in *cellranger\_atac\_create\_reference*
- Add delete undetermined FASTQs option for *mkfastq*

- **On *demultiplexing* workflow**

- Replace *demuxlet* with *popsicle*, which includes both *demuxlet* and *freemuxlet*

- **On *cumulus* workflow**

- Fixed bug that *remap\_singlets* and *subset\_singlets* don't work when input is in sample sheet format.

- Modified workflows to remove trailing spaces and support spaces within *output\_directory*

### Version 1.3.0 February 2, 2021

- **On *cumulus* workflow:**

- Change *cumulus\_version* to *pegasus\_version* to avoid confusion.
- Update to use Pegasus v1.3.0 for analysis.

### Version 1.2.0 January 19, 2021

- **Add *spaceranger* workflow:**

- Wrap up *spaceranger* version 1.2.1

- **On *cellranger* workflow:**

- Fix workflow WDL to support both single index and dual index
- Add support for cellranger version 5.0.1 and 5.0.0
- Add support for targeted gene expression analysis
- Add support for *--include-introns* and *--no-bam* options for *cellranger count*

- Remove `--force-cells` option for cellranger vdj as noted in cellranger 5.0.0 release note
- Add `GRCh38_vdj_v5.0.0` and `GRCm38_vdj_v5.0.0` references
- Bug fix on *cumulus* workflow.
- Reorganize the sidebar of Cumulus documentation website.

### Version 1.1.0 December 28, 2020

- **On *cumulus* workflow:**
  - Add CITE-Seq data analysis back. (See section [Run CITE-Seq analysis](#) for details)
  - Add doublet detection. (See `infer_doublets`, `expected_doublet_rate`, and `doublet_cluster_attribute` input fields)
  - For tSNE visualization, only support FIt-SNE algorithm. (see `run_tsne` and `plot_tsne` input fields)
  - Improve efficiency on log-normalization and DE tests.
  - Support multiple marker JSON files used in cell type annotation. (see `organism` input field)
  - More preset gene sets provided in gene score calculation. (see `calc_signature_scores` input field)
- **Add *star\_solo* workflow (see [STARsolo section](#) for details):**
  - Use [STARsolo](#) to generate count matrices from FASTQ files.
  - Support chemistry protocols such as 10X-V3, 10X-V2, DropSeq, and SeqWell.
- Update the example of analyzing hashing and CITE-Seq data (see [Example section](#)) with the new workflows.
- Bug fix.

### Version 1.0.0 September 23, 2020

- Add *demultiplexing* workflow for cell-hashing/nucleus-hashing/genetic-pooling analysis.
- Add support on CellRanger version 4.0.0.
- **Update *cumulus* workflow with Pegasus version 1.0.0:**
  - Use `zarr` file format to handle data, which has a better I/O performance in general.
  - Support focus analysis on Unimodal data, and appending other Unimodal data to it. (`focus` and `append` inputs in *cluster* step).
  - Quality-Control: Change `percent_mito` default from 10.0 to 20.0; by default remove bounds on UMIs (`min_umis` and `max_umis` inputs in *cluster* step).
  - Quality-Control: Automatically figure out name prefix of mitochondrial genes for GRCh38 and mm10 genome reference data.
  - Support signature / gene module score calculation. (`calc_signature_scores` input in *cluster* step)
  - Add *Scanorama* method to batch correction. (`correction_method` input in *cluster* step).
  - Calculate UMAP embedding by default, instead of FIt-SNE.

- Differential Expression (DE) analysis: remove inputs `mwu` and `auc` as they are calculated by default. And cell-type annotation uses MWU test result by default.
  - Remove `cumulus_subcluster` workflow.
- 

## **Version 0.x**

### **Version 0.15.0 May 6, 2020**

- Update all workflows to OpenWDL version 1.0.
- Cumulus now supports multi-job execution from Terra data table input.
- Cumulus generates CirroCumulus input in `.cirro` folder, instead of a huge `.parquet` file.

### **Version 0.14.0 February 28, 2020**

- Added support for gene-count matrices generation using alternative tools (STARsolo, Optimus, Salmon alevin, Kallisto BUSTools).
- Cumulus can process demultiplexed data with remapped singlets names and subset of singlets.
- Update VDJ related inputs in Cellranger workflow.
- SMART-Seq2 and Count workflows are in OpenWDL version 1.0.

### **Version 0.13.0 February 7, 2020**

- Added support for aggregating scATAC-seq samples.
- Cumulus now accepts mtx format input.

### **Version 0.12.0 December 14, 2019**

- Added support for building references for sc/snRNA-seq, scATAC-seq, single-cell immune profiling, and SMART-Seq2 data.

### **Version 0.11.0 December 4, 2019**

- Reorganized Cumulus documentation.

### **Version 0.10.0 October 2, 2019**

- scCloud is renamed to Cumulus.
- Cumulus can accept either a sample sheet or a single file.

### Version 0.7.0 February 14, 2019

- Added support for 10x genomics scATAC assays.
- scCloud runs FIt-SNE as default.

### Version 0.6.0 January 31, 2019

- Added support for 10x genomics V3 chemistry.
- Added support for extracting feature matrix for Perturb-Seq data.
- Added R script to convert output\_name.seurat.h5ad to Seurat object. Now the raw.data slot stores filtered raw counts.
- Added min\_umis and max\_umis to filter cells based on UMI counts.
- Added QC plots and improved filtration spreadsheet.
- Added support for plotting UMAP and FLE.
- Now users can upload their JSON file to annotate cell types.
- Improved documentation.
- Added lightGBM based marker detection.

### Version 0.5.0 November 18, 2018

- Added support for plated-based SMART-Seq2 scRNA-Seq data.

### Version 0.4.0 October 26, 2018

- Added CITE-Seq module for analyzing CITE-Seq data.

### Version 0.3.0 October 24, 2018

- Added the demuxEM module for demultiplexing cell-hashing/nuclei-hashing data.

### Version 0.2.0 October 19, 2018

- Added support for V(D)J and CITE-Seq/cell-hashing/nuclei-hashing.

### Version 0.1.0 July 27, 2018

- KCO tools released!

## 1.1.5 Run Cell Ranger tools using cellranger\_workflow

**cellranger\_workflow** wraps Cell Ranger to process single-cell/nucleus RNA-seq, single-cell ATAC-seq and single-cell immune profiling data, and supports feature barcoding (cell/nucleus hashing, CITE-seq, Perturb-seq). It also provide routines to build cellranger references.

## A general step-by-step instruction

This section mainly considers jobs starting from BCL files. If your job starts with FASTQ files, and only need to run `cellranger count` part, please refer to [this subsection](#).

### 1. Import `cellranger_workflow`

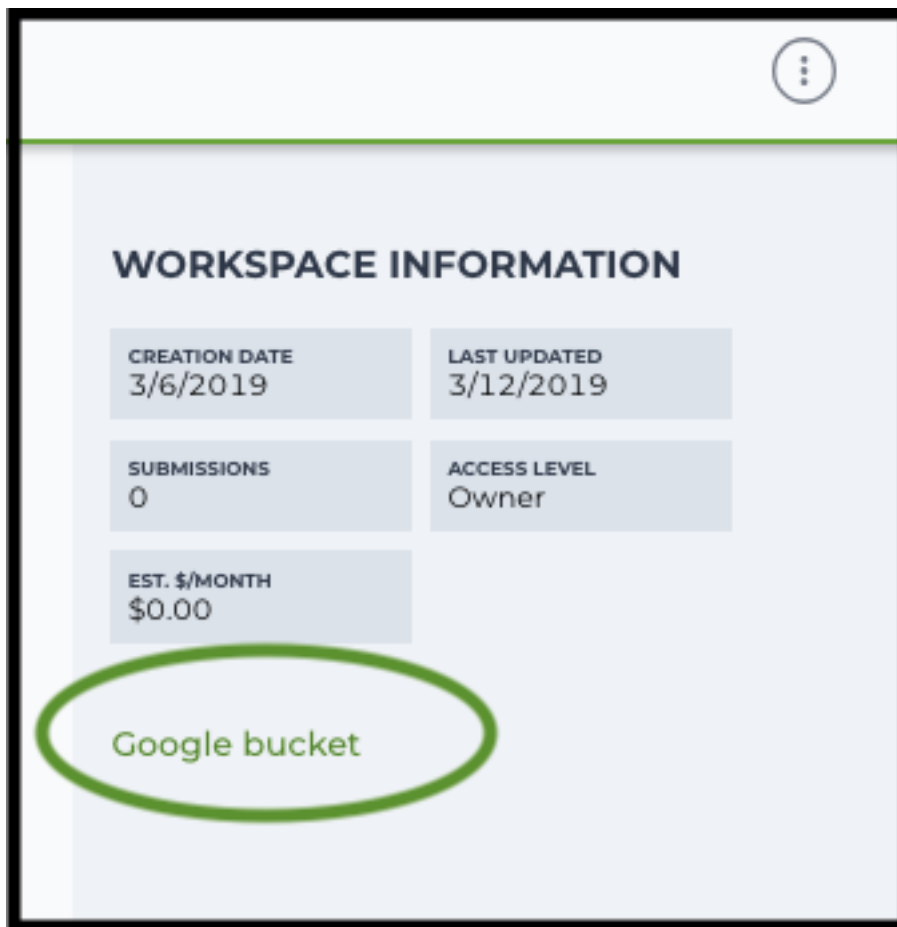
Import `cellranger_workflow` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose workflow [github.com/lilab-bcb/cumulus/CellRanger](https://github.com/lilab-bcb/cumulus/CellRanger) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `cellranger_workflow` workflow in the drop-down menu.

### 2. Upload sequencing data to Google bucket

Copy your sequencing output to your workspace bucket using `gsutil` (you already have it if you've installed Google cloud SDK) in your unix terminal.

You can obtain your bucket URL in the dashboard tab of your Terra workspace under the information panel.



Use `gsutil cp [OPTION]... src_url dst_url` to copy data to your workspace bucket. For example, the following command copies the directory at `/foo/bar/nextseq/Data/VK18WBC6Z4` to a Google bucket:

```
gsutil -m cp -r /foo/bar/nextseq/Data/VK18WBC6Z4 gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4
```

`-m` means copy in parallel, `-r` means copy the directory recursively, and `gs://fc-e0000000-0000-0000-0000-000000000000` should be replaced by your own workspace Google bucket URL.

---

**Note:** If input is a folder of BCL files, users do not need to upload the whole folder to the Google bucket. Instead, they only need to upload the following files:

```
RunInfo.xml
RTAComplete.txt
runParameters.xml
Data/Intensities/s.locs
Data/Intensities/BaseCalls
```

If data are generated using MiSeq or NextSeq, the location files are inside lane subfolders L001 under `Data/Intensities/`. In addition, if users' data only come from a subset of lanes (e.g. L001 and L002), users only need to upload lane subfolders from the subset (e.g. `Data/Intensities/BaseCalls/L001`, `Data/Intensities/BaseCalls/L002` and `Data/Intensities/L001`, `Data/Intensities/L002` if sequencer is MiSeq or NextSeq).

---

Alternatively, users can submit jobs through command line interface (CLI) using [altocumulus](#), which will smartly upload BCL folders according to the above rules.

### 3. Prepare a sample sheet

#### 3.1 Sample sheet format:

Please note that the columns in the CSV can be in any order, but that the column names must match the recognized headings.

The sample sheet describes how to demultiplex flowcells and generate channel-specific count matrices. Note that *Sample*, *Lane*, and *Index* columns are defined exactly the same as in 10x's simple CSV layout file.

A brief description of the sample sheet format is listed below (**required column headers are shown in bold**).



Column	Description
<b>Sample</b>	Contains sample names. Each 10x channel should have a unique sample name. Sample name can only contain characters from [a-zA-Z0-9_-].
<b>Reference</b>	Provides the reference genome used by Cell Ranger for each 10x channel. The elements in the <i>reference</i> column can be either Google bucket URLs to reference tarballs or keywords such as <i>GRCh38-2020-A</i> . A full list of available keywords is included in each of the following data type sections (e.g. sc/snRNA-seq) below.
<b>Flowcell</b>	Indicates the Google bucket URLs of uploaded BCL folders. If starts with FASTQ files, this should be Google bucket URLs of uploaded FASTQ folders. The FASTQ folders should contain one subfolder for each sample in the flowcell with the sample name as the subfolder name. Each subfolder contains FASTQ files for that sample.
<b>Lane</b>	Tells which lanes the sample was pooled into. Can be either single lane (e.g. 8) or a range (e.g. 7-8) or all (e.g. *).
<b>Index</b>	Sample index (e.g. SI-GA-A12).
<b>Chemistry</b>	Describes the 10x chemistry used for the sample. This column is optional.
<b>DataType</b>	Describes the data type of the sample — <i>rna</i> , <i>vdj</i> , <i>citeseq</i> , <i>hashing</i> , <i>cmo</i> , <i>crispr</i> , <i>atac</i> . <b>rna</b> refers to gene expression data ( <i>cellranger count</i> ), <b>vdj</b> refers to V(D)J data ( <i>cellranger vdj</i> ), <b>citeseq</b> refers to CITE-Seq tag data, <b>hashing</b> refers to cell-hashing or nucleus-hashing tag data, <b>adt</b> , which refers to the case where <i>hashing</i> and <i>citeseq</i> reads are in a sample library. <b>cmo</b> refers to cell multiplexing oligos used in 10x Genomics' CellPlex assay, <b>crispr</b> refers to Perturb-seq guide tag data, <b>atac</b> refers to scATAC-Seq data ( <i>cellranger-atac count</i> ), <b>frp</b> refers to Fixed RNA Profiling (FRP) gene expression data, This column is optional and the default data type is <i>rna</i> .
<b>ProbeSet</b>	Probe set reference for FRP samples. Currently <i>FRP_human_probe_v1</i> is the only available and thus the default reference. Only works for samples of <i>DataType frp</i> .
<b>FeatureBarcodeFile</b>	Google bucket urls pointing to feature barcode files for <i>rna</i> , <i>citeseq</i> , <i>hashing</i> , <i>cmo</i> and <i>crispr</i> data. Features can be either targeted genes for targeted gene expression analysis, antibody for CITE-Seq, cell-hashing, nucleus-hashing or gRNA for Perturb-seq. If <i>cmo</i> data is analyzed separately using <i>cumulus_feature_barcoding</i> , file format should follow the guide in Feature barcoding assays section, otherwise follow the guide in Single-cell multiomics section. This column is only required for targeted gene expression analysis ( <i>rna</i> ), CITE-Seq ( <i>citeseq</i> ), cell-hashing or nucleus-hashing ( <i>hashing</i> ), CellPlex ( <i>cmo</i> ) and Perturb-seq

<b>Link</b>	Designed for Single Cell Multiome ATAC + Gene Expression, Feature Barcoding,
-------------	--

The sample sheet supports sequencing the same 10x channels across multiple flowcells. If a sample is sequenced across multiple flowcells, simply list it in multiple rows, with one flowcell per row. In the following example, we have 4 samples sequenced in two flowcells.

Example:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry,DataType
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK18WBC6Z4,1-2,SI-GA-A8,threeprime,rna
sample_2,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK18WBC6Z4,3-4,SI-GA-B8,SC3Pv3,rna
sample_3,mm10-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4,
↪5-6,SI-GA-C8,fiveprime,rna
sample_4,mm10-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4,
↪7-8,SI-GA-D8,fiveprime,rna
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK10WBC9Z2,1-2,SI-GA-A8,threeprime,rna
sample_2,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK10WBC9Z2,3-4,SI-GA-B8,SC3Pv3,rna
sample_3,mm10-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9Z2,
↪5-6,SI-GA-C8,fiveprime,rna
sample_4,mm10-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9Z2,
↪7-8,SI-GA-D8,fiveprime,rna
```

### 3.2 Upload your sample sheet to the workspace bucket:

Example:

```
gsutil cp /foo/bar/projects/sample_sheet.csv gs://fc-e0000000-0000-
↪0000-0000-000000000000/
```

## 4. Launch analysis

In your workspace, open `cellranger_workflow` in **WORKFLOWS** tab. Select the desired snapshot version (e.g. latest). Select Run workflow with inputs defined by file paths as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click **SAVE** button. Select `Use call caching` and click **INPUTS**. Then fill in appropriate values in the **Attribute** column. Alternative, you can upload a JSON file to configure input by clicking **Drag** or click to upload json.

Once **INPUTS** are appropriated filled, click **RUN ANALYSIS** and then click **LAUNCH**.

## 5. Notice: run `cellranger mkfastq` if you are non Broad Institute users

Non Broad Institute users that wish to run `cellranger mkfastq` must create a custom docker image that contains `bcl2fastq`.

See [bcl2fastq](#) instructions.

## 6. Run `cellranger` count only

Sometimes, users might want to perform demultiplexing locally and only run the count part on the cloud. This section describes how to only run the count part via `cellranger_workflow`.

1. Copy your FASTQ files to the workspace using `gsutil` in your unix terminal. There are two cases:

- **Case 1:** All the FASTQ files are in one top-level folder. Then you can simply upload this folder to Cloud, and in your sample sheet, make sure **Sample** names are consistent with the filename prefix of their corresponding FASTQ files.
- **Case 2:** In the top-level folder, each sample has a dedicated subfolder containing its FASTQ files. In this case, you need to upload the whole top-level folder, and in your sample sheet, make sure **Sample** names and their corresponding subfolder names are identical.

Notice that if your FASTQ files are downloaded from the Sequence Read Archive (SRA) from NCBI, you must rename your FASTQs to follow the `bcl2fastq` [file naming conventions](#).

Example:

```
gsutil -m cp -r /foo/bar/fastq_path/K18WBC6Z4 gs://fc-e0000000-0000-0000-0000-000000000000/K18WBC6Z4_fastq
```

2. Create a sample sheet following the similar structure as [above](#), except the following differences:

- **Flowcell** column should list Google bucket URLs of the FASTQ folders for flowcells.
- **Lane** and **Index** columns are NOT required in this case.

Example:

```
Sample,Reference,Flowcell
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/K18WBC6Z4_fastq
```

3. Set optional input `run_mkfastq` to `false`.

## 7. Workflow outputs

See the table below for workflow level outputs.

Name	Type	Description
fastq_outputs	Array[Array[String]?]	The top-level array contains results (as arrays) for different data modalities. The inner-level array contains cloud locations of FASTQ files, one url per flowcell.
count_outputs	Array[Array[String]?]	The top-level array contains results (as arrays) for different data modalities. The inner-level array contains cloud locations of count matrices, one url per sample.
count_matrix	String	Cloud url for a template count_matrix.csv to run Cumulus. It only contains sc/snRNA-Seq samples.

## Single-cell and single-nucleus RNA-seq

To process sc/snRNA-seq data, follow the specific instructions below.

### Sample sheet

#### 1. Reference column.

Pre-built scRNA-seq references are summarized below.

Keyword	Description
<b>GRCh38-2020-A</b>	Human GRCh38 (GENCODE v32/Ensembl 98)
<b>mm10-2020-A</b>	Mouse mm10 (GENCODE vM23/Ensembl 98)
<b>GRCh38_and_mm10-2020-A</b>	Human GRCh38 (GENCODE v32/Ensembl 98) and mouse mm10 (GENCODE vM23/Ensembl 98)
<b>GRCh38_v3.0.0</b>	Human GRCh38, cellranger reference 3.0.0, Ensembl v93 gene annotation
<b>hg19_v3.0.0</b>	Human hg19, cellranger reference 3.0.0, Ensembl v87 gene annotation
<b>mm10_v3.0.0</b>	Mouse mm10, cellranger reference 3.0.0, Ensembl v93 gene annotation
<b>GRCh38_and_mm10_v3.1.0</b>	Human (GRCh38) and mouse (mm10), cellranger references 3.1.0, Ensembl v93 gene annotations for both human and mouse
<b>hg19_and_mm10_v3.0.0</b>	Human (hg19) and mouse (mm10), cellranger reference 3.0.0, Ensembl v93 gene annotations for both human and mouse
<b>GRCh38_v1.2.0 or GRCh38</b>	Human GRCh38, cellranger reference 1.2.0, Ensembl v84 gene annotation
<b>hg19_v1.2.0 or hg19</b>	Human hg19, cellranger reference 1.2.0, Ensembl v82 gene annotation
<b>mm10_v1.2.0 or mm10</b>	Mouse mm10, cellranger reference 1.2.0, Ensembl v84 gene annotation
<b>GRCh38_and_mm10_v1.2.0 or GRCh38_and_mm10</b>	Human and mouse, built from GRCh38 and mm10 cellranger references, Ensembl v84 gene annotations are used
<b>GRCh38_and_SARSCoV2</b>	Human GRCh38 and SARS-COV-2 RNA genome, cellranger reference 3.0.0, generated by Carly Ziegler. The SARS-COV-2 viral sequence and gtf are as described in [Kim et al. Cell 2020] ( <a href="https://github.com/hyeshik/sars-cov-2-transcriptome">https://github.com/hyeshik/sars-cov-2-transcriptome</a> , BetaCov/South Korea/KCDC03/2020 based on NC_045512.2). The GTF was edited to include only CDS regions, and regions were added to describe the 5' UTR ("SARSCoV2_5prime"), the 3' UTR ("SARSCoV2_3prime"), and reads aligning to anywhere within the Negative Strand("SARSCoV2_NegStrand"). Additionally, trailing A's at the 3' end of the virus were excluded from the SARSCoV2 fasta, as these were found to drive spurious viral alignment in pre-COVID19 samples.

Pre-built snRNA-seq references are summarized below.

Keyword	Description
<b>GRCh38_premrna_v3.0.0</b>	GRCh38, introns included, built from GRCh38 cellranger reference 3.0.0, Ensembl v93 gene annotation, treating annotated transcripts as exons
<b>GRCh38_premrna_v1.2.0</b> or <b>GRCh38_premrna</b>	GRCh38, introns included, built from GRCh38 cellranger reference 1.2.0, Ensembl v84 gene annotation, treating annotated transcripts as exons
<b>mm10_premrna_v1.2.0</b> or <b>mm10_premrna</b>	mm10, introns included, built from mm10 cellranger reference 1.2.0, Ensembl v84 gene annotation, treating annotated transcripts as exons
<b>GRCh38_premrna_human_and_mm10_premrna_v1.2.0</b> or <b>GRCh38_premrna_and_mm10_premrna</b>	Human and mm10, introns included, built from GRCh38_premrna_v1.2.0 and mm10_premrna_v1.2.0
<b>GRCh38_premrna_human_and_SARSCoV2</b>	Human and SARS-CoV-2, introns included, built from GRCh38_premrna_v3.0.0, and SARS-CoV-2 RNA genome. This reference was generated by <a href="#">Carly Ziegler</a> . The SARS-CoV-2 RNA genome is from <a href="#">[Kim et al. Cell 2020]</a> ( <a href="https://github.com/hyeshik/sars-cov-2-transcriptome">https://github.com/hyeshik/sars-cov-2-transcriptome</a> , BetaCov/South Korea/KCDC03/2020 based on NC_045512.2). Please see the description of <i>GRCh38_and_SARSCoV2</i> above for details.

## 2. Index column.

Put 10x single cell RNA-seq sample index set names (e.g. SI-GA-A12) here.

## 3. Chemistry column.

According to *cellranger count*'s documentation, chemistry can be

Chemistry	Explanation
<b>auto</b>	autodetection (default). If the index read has extra bases besides cell barcode and UMI, autodetection might fail. In this case, please specify the chemistry
<b>threeprime</b>	Single Cell 3
<b>fiveprime</b>	Single Cell 5
<b>SC3Pv1</b>	Single Cell 3 v1
<b>SC3Pv2</b>	Single Cell 3 v2
<b>SC3Pv3</b>	Single Cell 3 v3. You should set cellranger version input parameter to >= 3.0.2
<b>SC5P-PE</b>	Single Cell 5 paired-end (both R1 and R2 are used for alignment)
<b>SC5P-R2</b>	Single Cell 5 R2-only (where only R2 is used for alignment)

## 4. DataType column.

This column is optional with a default **rna**. If you want to put a value, put **rna** here.

## 5. FeatureBarcodeFile column.

Put target panel CSV file here for targeted expression data. Note that if a target panel CSV is present, cell ranger version must be >= 4.0.0.

## 6. Example:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry,DataType,FeatureBarcodeFile
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4,1-
↪2,SI-GA-A8,threeprime,rna
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9Z2,1-
↪2,SI-GA-A8,threeprime,rna
```

(continues on next page)

(continued from previous page)

```
sample_2,mm10-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4,5-6,  
↪SI-GA-C8,fiveprime,rna  
sample_2,mm10-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9Z2,5-6,  
↪SI-GA-C8,fiveprime,rna  
sample_3,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4,3,  
↪SI-TT-A1,auto,rna,gs://fc-e0000000-0000-0000-0000-000000000000/immunology_v1.0_  
↪GRCh38-2020-A.target_panel.csv
```

## Workflow input

For sc/snRNA-seq data, `cellranger_workflow` takes Illumina outputs as input and runs `cellranger mkfastq` and `cellranger count`. Relevant workflow inputs are described below, with required inputs highlighted in bold.

Name	Description	Example	Default
<code>input_sample_sheet</code>	Sample Sheet (contains Sample, Reference, Flowcell, Lane, Index as required and Chemistry, DataType, FeatureBarcodeFile as optional)	"gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv"	
<code>output_directory</code>	Output directory	"gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_output"	Results are written under directory <i>output_directory</i> and will overwrite any existing files at this location.
<code>run_mkfastq</code>	Do you want to run cellranger mkfastq	true	true
<code>run_cellranger_count</code>	Do you want to run cellranger count	true	true
<code>delete_input_bcl_directories</code>	Delete BCL directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	false	false
<code>mkfastq_number_of_mismatches</code>	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	
<code>mkfastq_force_use_index_i7</code>	If a flowcell has paired indices are specified, but the flow-cell was run with only one sample index, allow the demultiplex to proceed using the i7 half of the sample index pair	false	false
<code>mkfastq_only_use_index_i7</code>	Only use single index samples identified by an i7-only sample index, ignoring dual-indexed samples. Dual-indexed samples will not be demultiplexed	false	false
<code>mkfastq_use_database_lengths</code>	Use database read lengths as specified in <i>RunInfo.xml</i>	"Y28n*,I8n*,N10,Y90n*"	
<code>mkfastq_delete_unidentified_FASTQ_files</code>	Delete unidentified FASTQ files generated by bcl2fastq2	true	false
<code>force_cells</code>	Force pipeline to use this number of cells, bypassing the cell detection algorithm, mutually exclusive with <code>expect_cells</code>	6000	
<code>expect_cells</code>	Expected number of recovered cells. Mutually exclusive with <code>force_cells</code>	3000	
<code>include_introns</code>	Turn this option on to also count reads mapping to intronic regions. With this option, users do not need to use pre-mRNA references. Note that if this option is set, <code>cellranger_version</code> must be $\geq 5.0.0$ .	true	true
<code>no_bam</code>	Turn this option on to disable BAM file generation. This option is only available if <code>cellranger_version</code> $\geq 5.0.0$ .	false	false
<code>secondary_analysis</code>	Perform Cell Ranger secondary analysis (dimensionality reduction, clustering, etc.)	false	false

## Workflow output

See the table below for important sc/snRNA-seq outputs.

Name	Type	Description
cellranger_mkfastq.output_fastq_files	Array[String]? Array[String]	Subworkflow output. A list of cloud urls containing FASTQ files, one url per flowcell.
cellranger_count.output_count_matrices	Array[String]? Array[String]	Subworkflow output. A list of cloud urls containing gene count matrices, one url per sample.
cellranger_count.output_web_summary_files	Array[String]? Array[String]	Subworkflow output. A list of htmls visualizing QCs for each sample (cellranger count output).
collect_summaries.metrics_summary_files	File?	Task output. A excel spreadsheet containing QCs for each sample.
count_matrix	String	Workflow output. Cloud url for a template count_matrix.csv to run Cumulus.

## Feature barcoding assays (cell & nucleus hashing, CITE-seq and Perturb-seq)

cellranger\_workflow can extract feature-barcode count matrices in CSV format for feature barcoding assays such as *cell and nucleus hashing*, *CellPlex*, *CITE-seq*, and *Perturb-seq*. For cell and nucleus hashing as well as CITE-seq, the feature refers to antibody. For Perturb-seq, the feature refers to guide RNA. Please follow the instructions below to configure cellranger\_workflow.

### Prepare feature barcode files

Prepare a CSV file with the following format: feature\_barcode,feature\_name. See below for an example:

```
TTCCTGCCATTACTA, sample_1
CCGTACCTCATTGTT, sample_2
GGTAGATGTCCTCAG, sample_3
TGGTGCATTCTTGA, sample_4
```

The above file describes a cell hashing application with 4 samples.

If cell hashing and CITE-seq data share a same sample index, you should concatenate hashing and CITE-seq barcodes together and add a third column indicating the feature type. See below for an example:

```
TTCCTGCCATTACTA, sample_1, hashing
CCGTACCTCATTGTT, sample_2, hashing
GGTAGATGTCCTCAG, sample_3, hashing
TGGTGCATTCTTGA, sample_4, hashing
CTCATTGTAACCT, CD3, citeseq
GCGCAACTTGATGAT, CD8, citeseq
```

Then upload it to your google bucket:

```
gsutil antibody_index.csv gs://fc-e0000000-0000-0000-0000-000000000000/
↪antibody_index.csv
```



## Sample sheet

### 1. Reference column.

This column is not used for extracting feature-barcode count matrix. To be consistent, please put the reference for the associated scRNA-seq assay here.

### 2. Index column.

The ADT/HTO index can be either Illumina index primer sequence (e.g. ATTACTCG, also known as D701), or 10x single cell RNA-seq sample index set names (e.g. SI-GA-A12).

**Note 1:** All ADT/HTO index sequences (including 10x's) should have the same length (8 bases). If one index sequence is shorter (e.g. ATCACG), pad it with P7 sequence (e.g. ATCACGAT).

**Note 2:** It is users' responsibility to avoid index collision between 10x genomics' RNA indexes (e.g. SI-GA-A8) and Illumina index sequences for used here (e.g. ATTACTCG).

**Note 3:** For NextSeq runs, please reverse complement the ADT/HTO index primer sequence (e.g. use reverse complement CGAGTAAT instead of ATTACTCG).

### 3. Chemistry column.

The following keywords are accepted for *Chemistry* column:

Chemistry	Explanation
<b>auto</b>	Default. This is an alias for Single Cell 3' v3 (SC3Pv3)
<b>threeprime</b>	This is another alias for Single Cell 3' v3
<b>SC3Pv3</b>	Single Cell 3 v3
<b>SC3Pv2</b>	Single Cell 3 v2
<b>fiveprime</b>	Single Cell 5
<b>SC5P-PE</b>	Single Cell 5 paired-end (both R1 and R2 are used for alignment)
<b>SC5P-R2</b>	Single Cell 5 R2-only (where only R2 is used for alignment)
<b>multiome</b>	10x Multiome barcodes

### 4. DataType column.

The following keywords are accepted for *DataType* column:

DataType	Explanation
<b>citeseq</b>	CITE-seq
<b>hashing</b>	Cell or nucleus hashing
<b>cmo</b>	CellPlex
<b>adt</b>	Hashing and CITE-seq are in the same library
<b>crispr</b>	<p>Perturb-seq/CROP-seq</p> <p>If neither <i>crispr_barcode_pos</i> nor <i>scaffold_sequence</i> (see Workflow input) is set, <b>crispr</b> refers to 10x CRISPR assays. If in addition <i>Chemistry</i> is set to be <b>SC3Pv3</b> or its aliases, Cumulus automatically complement the middle two bases to convert 10x feature barcoding cell barcodes back to 10x RNA cell barcodes.</p> <p>Otherwise, <b>crispr</b> refers to non 10x CRISPR assays, such as CROP-Seq. In this case, we assume feature barcoding cell barcodes are the same as the RNA cell barcodes and no cell barcode conversion will be conducted.</p>

5. *FeatureBarcodeFile* column.

Put Google Bucket URL of the feature barcode file here.

## 6. Example:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry,DataType,FeatureBarcodeFile
sample_1_rna,GRCh38_v3.0.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK18WBC6Z4,1-2,SI-GA-A8,threeprime,rna
sample_1_adt,GRCh38_v3.0.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK18WBC6Z4,1-2,ATTACTCG,SC3Pv3,adt,gs://fc-e0000000-0000-0000-0000-000000000000/
↪antibody_index.csv
sample_2_adt,GRCh38_v3.0.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK18WBC6Z4,3-4,TCCGAGGA,SC3Pv3,adt,gs://fc-e0000000-0000-0000-0000-000000000000/
↪antibody_index.csv
sample_3_crispr,GRCh38_v3.0.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK18WBC6Z4,5-6,CGCTCATT,SC3Pv3,crispr,gs://fc-e0000000-0000-0000-0000-
↪000000000000/crispr_index.csv
```

In the sample sheet above, despite the header row,

- First row describes the normal 3' RNA assay;
- Second row describes its associated antibody tag data, which can from either a CITE-seq, cell hashing, or nucleus hashing experiment.
- Third row describes another tag data, which is in 10x genomics' V3 chemistry. For tag and crispr data, it is important to explicitly state the chemistry (e.g. SC3Pv3).
- Last row describes one gRNA guide data for Perturb-seq (see *crispr* in *DataType* field).

## Workflow input

For feature barcoding data, *cellranger\_workflow* takes Illumina outputs as input and runs *cellranger mkfastq* and *cumulus adt*. Relevant workflow inputs are described below, with required inputs highlighted in bold.

Name	Description	Example	Default
<code>input_sample_sheet</code>	File Sheet (contains Sample, Reference, Flowcell, Lane, Index as required and Chemistry, DataType, FeatureBarcodeFile as optional)	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv”	
<code>output_directory</code>	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_output”	
<code>run_cellranger_mkfastq</code>	Do you want to run cellranger mkfastq	true	true
<code>run_cdf</code>	Do you want to run cumulus adt	true	true
<code>delete_index_bcl_directories</code>	Delete BCL directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	false	false
<code>mkfastq_number_of_mismatches</code>	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	
<code>mkfastq_flowcell_index</code>	If flowcell index is specified, but the flow-cell was run with only one sample index, allow the demultiplex to proceed using the i7 half of the sample index pair	false	false
<code>mkfastq_only_single_index</code>	Only filter single index samples identified by an i7-only sample index, ignoring dual-indexed samples. Dual-indexed samples will not be demultiplexed	false	false
<code>mkfastq_override_read_lengths</code>	Override read lengths as specified in <i>RunInfo.xml</i>	“Y28n*,I8n*,N10,Y90n*”	
<code>mkfastq_delete_unidentified</code>	Delete unidentified FASTQ files generated by bcl2fastq2	true	false
<code>crispr_barcode_start</code>	Barcode start position at Read 2 (0-based coordinate) for CRISPR	19	0
<code>scaffold_sequence</code>	Reference sequence in sgRNA for Perturb-seq, only used for crispr data type.	“GTTTAAGAGCTAAGCTGGAA”	“”
<code>max_mismatch</code>	Maximum hamming distance in feature barcodes for the adt task (changed to 2 as default)	2	2
<code>min_read_ratio</code>	Minimum read count ratio (non-inclusive) to justify a feature given a cell barcode and feature combination, only used for the adt task and crispr data type	0.1	0.1
<code>cellranger_version</code>	cellranger version, could be: 7.1.0, 7.0.1, 7.0.0, 6.1.2, 6.1.1, 6.0.2, 6.0.1, 6.0.0, 5.0.1, 5.0.0	“7.1.0”	“7.1.0”
<code>cumulus_version</code>	Cumulus version, could be: 0.11.1, 0.11.0, 0.10.0, 0.9.0, 0.8.0, 0.7.0, 0.6.0, 0.5.0, 0.4.0, 0.3.0, 0.2.0.	“0.11.1”	“0.11.1”
<code>docker_registry</code>	Docker registry to use for cellranger_workflow. Options: • “quay.io/cumulus” for im-	“quay.io/cumulus”	“quay.io/cumulus”

## Parameters used for feature count matrix extraction

If the chemistry is V2, [10x genomics v2 cell barcode white list](#) will be used, a hamming distance of 1 is allowed for matching cell barcodes, and the UMI length is 10. If the chemistry is V3, [10x genomics v3 cell barcode white list](#) will be used, a hamming distance of 0 is allowed for matching cell barcodes, and the UMI length is 12.

For Perturb-seq data, a small number of sgRNA protospace sequences will be sequenced ultra-deeply and we may have PCR chimeric reads. Therefore, we generate filtered feature count matrices as well in a data driven manner:

1. First, plot the histogram of UMIs with certain number of read counts. The number of UMIs with  $x$  supporting reads decreases when  $x$  increases. We start from  $x = 1$ , and a valley between two peaks is detected if we find  $\text{count}[x] < \text{count}[x + 1] < \text{count}[x + 2]$ . We filter out all UMIs with  $< x$  supporting reads since they are likely formed due to chimeric reads.
2. In addition, we also filter out barcode-feature-UMI combinations that have their read count ratio, which is defined as total reads supporting barcode-feature-UMI over total reads supporting barcode-UMI, no larger than `min_read_ratio` parameter set above.

## Workflow outputs

See the table below for important outputs.

Name	Type	Description
<code>cellranger_mkfastq.output_fastq[Array[String]]?</code>	<code>Array[String]</code>	Subworkflow output. A list of cloud urls containing FASTQ files, one url per flowcell.
<code>cumulus_adt.output_count_matrices[Array[String]]?</code>	<code>Array[String]</code>	Subworkflow output. A list of cloud urls containing feature-barcode count matrices, one url per sample.

In addition, For each antibody tag or crispr tag sample, a folder with the sample ID is generated under `output_directory`. In the folder, two files — `sample_id.csv` and `sample_id.stat.csv.gz` — are generated.

`sample_id.csv` is the feature count matrix. It has the following format. The first line describes the column names: `Antibody/CRISPR, cell_barcode_1, cell_barcode_2, ..., cell_barcode_n`. The following lines describe UMI counts for each feature barcode, with the following format: `feature_name, umi_count_1, umi_count_2, ..., umi_count_n`.

`sample_id.stat.csv.gz` stores the gzipped sufficient statistics. It has the following format. The first line describes the column names: `Barcode, UMI, Feature, Count`. The following lines describe the read counts for every barcode-umi-feature combination.

If the feature barcode file has a third column, there will be two files for each feature type in the third column. For example, if hashing presents, `sample_id.hashing.csv` and `sample_id.hashing.stat.csv.gz` will be generated.

`sample_id.report.txt` is a summary report in TXT format. The first lines describe the total number of reads parsed, the number of reads with valid cell barcodes (and percentage over all parsed reads), the number of reads with valid feature barcodes (and percentage over all parsed reads) and the number of reads with both valid cell and feature barcodes (and percentage over all parsed reads). It is then followed by sections describing each feature type. In each section, 7 lines are shown: section title, number of valid cell barcodes (with matching cell barcode and feature barcode) in this section, number of reads for these cell barcodes, mean number of reads per cell barcode, number of UMIs for these cell barcodes, mean number of UMIs per cell barcode and sequencing saturation.

If data type is `crispr`, three additional files, `sample_id.umi_count.pdf`, `sample_id.filt.csv` and `sample_id.filt.stat.csv.gz`, are generated.

`sample_id.umi_count.pdf` plots number of UMIs against UMI with certain number of reads and colors UMIs with high likelihood of being chimeric in blue and other UMIs in red. This plot is generated purely based on number of reads each UMI has. For better visualization, we do not show UMIs with > 50 read counts (rare in data).

`sample_id.filt.csv` is the filtered feature count matrix. It has the same format as `sample_id.csv`.

`sample_id.filt.stat.csv.gz` is the filtered sufficient statistics. It has the same format as `sample_id.stat.csv.gz`.

## Single-cell ATAC-seq

To process scATAC-seq data, follow the specific instructions below.

### Sample sheet

#### 1. Reference column.

Pre-built scATAC-seq references are summarized below.

Keyword	Description
<b>GRCh38-2020-A_arc_v2.0.0</b>	Human GRCh38, cellranger-arc/atac reference 2.0.0
<b>mm10-2020-A_arc_v2.0.0</b>	Mouse mm10, cellranger-arc/atac reference 2.0.0
<b>GRCh38_and_mm10-2020-A_atac_v2.0.0</b>	Human GRCh38 and mouse mm10, cellranger-atac reference 2.0.0
<b>GRCh38_atac_v1.2.0</b>	Human GRCh38, cellranger-atac reference 1.2.0
<b>mm10_atac_v1.2.0</b>	Mouse mm10, cellranger-atac reference 1.2.0
<b>hg19_atac_v1.2.0</b>	Human hg19, cellranger-atac reference 1.2.0
<b>b37_atac_v1.2.0</b>	Human b37 build, cellranger-atac reference 1.2.0
<b>GRCh38_and_mm10_atac_v1.2.0</b>	Human GRCh38 and mouse mm10, cellranger-atac reference 1.2.0
<b>hg19_and_mm10_atac_v1.2.0</b>	Human hg19 and mouse mm10, cellranger-atac reference 1.2.0
<b>GRCh38_atac_v1.1.0</b>	Human GRCh38, cellranger-atac reference 1.1.0
<b>mm10_atac_v1.1.0</b>	Mouse mm10, cellranger-atac reference 1.1.0
<b>hg19_atac_v1.1.0</b>	Human hg19, cellranger-atac reference 1.1.0
<b>b37_atac_v1.1.0</b>	Human b37 build, cellranger-atac reference 1.1.0
<b>GRCh38_and_mm10_atac_v1.1.0</b>	Human GRCh38 and mouse mm10, cellranger-atac reference 1.1.0
<b>hg19_and_mm10_atac_v1.1.0</b>	Human hg19 and mouse mm10, cellranger-atac reference 1.1.0

#### 2. Index column.

Put 10x single cell ATAC sample index set names (e.g. SI-NA-B1) here.

#### 3. Chemistry column.

By default is **auto**, which will not specify a given chemistry. To analyze just the individual ATAC library from a 10x multiome assay using cellranger-atac count, use **ARC-v1** in the Chemistry column.

#### 4. DataType column.

Set it to **atac**.

#### 5. FeatureBarcodeFile column.

Leave it blank for scATAC-seq.

6. Example:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry,DataType
sample_atac,GRCh38_atac_v1.1.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK10WBC9YB,*,SI-NA-A1,auto,atac
```

## Workflow input

`cellranger_workflow` takes Illumina outputs as input and runs `cellranger-atac mkfastq` and `cellranger-atac count`. Please see the description of inputs below. Note that required inputs are shown in bold.

Name	Description	Example	Default
<code>input_sample_sheet</code>	Sample Sheet (contains Sample, Reference, Flowcell, Lane, Index as required and Chemistry, DataType, FeatureBarcodeFile as optional)	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv”	
<code>output_directory</code>	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_atac_output”	
<code>run_mkfastq</code>	you want to run cellranger-atac mkfastq	true	true
<code>run_count</code>	you want to run cellranger-atac count	true	true
<code>delete_index_dirs</code>	Delete index directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	false	false
<code>mkfastq_num_mismatches</code>	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	
<code>mkfastq_force_suplex_i7</code>	If force-suplex i7/15 paired indices are specified, but the flowcell was run with only one sample index, allow the demultiplex to proceed using the i7 half of the sample index pair	false	false
<code>mkfastq_only_dual_index</code>	Only dual-index samples identified by an i7-only sample index, ignoring dual-indexed samples. Dual-indexed samples will not be demultiplexed	false	false
<code>mkfastq_override_mask</code>	Override the mask lengths as specified in <i>RunInfo.xml</i>	“Y28n*,I8n*,N10,Y90n*”	
<code>mkfastq_delete_unidentified</code>	Delete unidentified FASTQ files generated by bcl2fastq2	true	false
<code>force_cells</code>	Force pipeline to use this number of cells, bypassing the cell detection algorithm	6000	
<code>atac_dimensionality</code>	Choose the algorithm for dimensionality reduction prior to clustering and tsne: “lsa”, “pls”, or “pca”	“lsa”	“lsa”
<code>peaks</code>	A 3-column BED file of peaks to override cellranger atac peak caller. Peaks must be sorted by position and not contain overlapping peaks; comment lines beginning with # are allowed	“gs://fc-e0000000-0000-0000-0000-000000000000/common_peaks.bed”	
<code>cellranger_version</code>	cellranger version. Available options: 2.1.0, 2.0.0, 1.2.0, 1.1.0	“2.1.0”	“2.1.0”
<code>docker_registry</code>	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>“quay.io/cumulus” for images on Red Hat registry;</li> <li>“cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<code>mkfastq_docker_registry</code>	Docker registry to use for cellranger-atac mkfastq. Default is the registry to which only Broad users have access. See <i>bcl2fastq</i> for making your own registry.	“gcr.io/broad-cumulus”	“gcr.io/broad-cumulus”
<b>1.1. 2.4.0 January 28, 2023</b>			<b>43</b>
<code>acronym_file</code>	The path to the acronym file TSV	“s3://xxxx/index.tsv”	“gs://regev-lab/resources/cellranger/index.tsv”

## Workflow output

See the table below for important scATAC-seq outputs.

Name	Type	Description
cellranger_atac_mkfastq.output	Array[String]	Subworkflow output. A list of cloud urls containing FASTQ files, one url per flowcell.
cellranger_atac_count.output	Array[String]	Subworkflow output. A list of cloud urls containing cellranger-atac count outputs, one url per sample.
cellranger_atac_count.output	Array[File]	Subworkflow output. A list of htmls visualizing QCs for each sample (cellranger-atac count output).
collect_summaries_atac.metrics	File	Task output. A excel spreadsheet containing QCs for each sample.

## Aggregate scATAC-Seq Samples

To aggregate multiple scATAC-Seq samples, follow the instructions below:

1. Import `cellranger_atac_aggr` workflow. Please see Step 1 [here](#), and the name of workflow is “**cumulus/cellranger\_atac\_aggr**”.
2. Set the inputs of workflow. Please see the description of inputs below. Notice that required inputs are shown in bold:



Name	Description	Example	Default
<b>aggr_id</b>	Aggregate ID.	"aggr_sample"	
<b>input_dirs</b>	Comma-separated URLs to directories of samples to be aggregated.	"gs://fc-e0000000-0000-0000-0000-000000000000/data/sample1,gs://fc-e0000000-0000-0000-0000-000000000000/data/sample2"	
<b>output_dir</b>	Output directory	"gs://fc-e0000000-0000-0000-0000-000000000000/aggregate_result"	
<b>genome</b>	The reference genome name used by Cell Ranger, can be either a keyword of pre-built genome, or a Google Bucket URL. See <a href="#">this table</a> for the list of keywords of pre-built genomes.	"GRCh38_atac_v1.2.0"	
<b>normalization</b>	Sample normalization mode. Options are: none, depth, or signal.	"none"	"none"
<b>secondary</b>	Perform secondary analysis (dimensionality reduction, clustering and visualization).	false	false
<b>dim_reduction</b>	Choose the algorithm for dimensionality reduction prior to clustering and tsne. Options are: lsa, plsa, or pca.	"lsa"	"lsa"
<b>peaks</b>	A 3-column BED file of peaks to override cellranger atac peak caller. Peaks must be sorted by position and not contain overlapping peaks; comment lines beginning with # are allowed	"gs://fc-e0000000-0000-0000-0000-000000000000/common_peaks.bed"	
<b>cellranger_atac</b>	Cell Ranger ATAC version to use. Options: 2.1.0, 2.0.0, 1.2.0, 1.1.0.	"2.1.0"	"2.1.0"
<b>zones</b>	Google cloud zones	"us-central1-a us-west1-a"	"us-central1-b"
<b>num_cpus</b>	Number of cpus to request for cellranger atac aggr.	64	64
<b>backend</b>	Cloud backend for file transfer. Available options: <ul style="list-style-type: none"> <li>• "gcp" for Google Cloud;</li> <li>• "aws" for Amazon AWS;</li> <li>• "local" for local machine.</li> </ul>	"gcp"	"gcp"
<b>memory</b>	Memory size string for cellranger atac aggr.	"57.6G"	"57.6G"
<b>disk_space</b>	Disk space in GB needed for cellranger atac aggr.	500	500
<b>preemptible</b>	Number of preemptible tries.	2	2
<b>aws_queue</b>	The AWS ARN string of the job queue to be used. This only works for aws backend.	"arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf"	""
<b>docker_registry</b>	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>• "quay.io/cumulus" for images on Red Hat registry;</li> <li>• "cumulusprod" for backup images on Docker Hub.</li> </ul>	"quay.io/cumulus"	"quay.io/cumulus"

3. Check out the output in `output_directory/aggr_id` folder, where `output_directory` and `aggr_id` are the inputs you set in Step 2.

## Single-cell immune profiling

To process single-cell immune profiling (scIR-seq) data, follow the specific instructions below.

### Sample sheet

1. **Reference** column.

Pre-built scIR-seq references are summarized below.

Keyword	Description
<b>GRCh38_vdj_v7.0.0</b>	Human GRCh38 V(D)J sequences, cellranger reference 7.0.0, annotation built from Ensembl <i>Homo_sapiens.GRCh38.94.chr_patch_hapl_scaff.gtf</i>
<b>GRCm38_vdj_v7.0.0</b>	Mouse GRCm38 V(D)J sequences, cellranger reference 7.0.0, annotation built from Ensembl <i>Mus_musculus.GRCm38.94.gtf</i>
<b>GRCh38_vdj_v5.0.0</b>	Human GRCh38 V(D)J sequences, cellranger reference 5.0.0, annotation built from Ensembl <i>Homo_sapiens.GRCh38.94.chr_patch_hapl_scaff.gtf</i>
<b>GRCm38_vdj_v5.0.0</b>	Mouse GRCm38 V(D)J sequences, cellranger reference 5.0.0, annotation built from Ensembl <i>Mus_musculus.GRCm38.94.gtf</i>
<b>GRCh38_vdj_v4.0.0</b>	Human GRCh38 V(D)J sequences, cellranger reference 4.0.0, annotation built from Ensembl <i>Homo_sapiens.GRCh38.94.chr_patch_hapl_scaff.gtf</i>
<b>GRCm38_vdj_v4.0.0</b>	Mouse GRCm38 V(D)J sequences, cellranger reference 4.0.0, annotation built from Ensembl <i>Mus_musculus.GRCm38.94.gtf</i>
<b>GRCh38_vdj_v3.1.0</b>	Human GRCh38 V(D)J sequences, cellranger reference 3.1.0, annotation built from Ensembl <i>Homo_sapiens.GRCh38.94.chr_patch_hapl_scaff.gtf</i>
<b>GRCm38_vdj_v3.1.0</b>	Mouse GRCm38 V(D)J sequences, cellranger reference 3.1.0, annotation built from Ensembl <i>Mus_musculus.GRCm38.94.gtf</i>
<b>GRCh38_vdj_v2.0.0</b> or <b>GRCh38_vdj</b>	Human GRCh38 V(D)J sequences, cellranger reference 2.0.0, annotation built from Ensembl <i>Homo_sapiens.GRCh38.87.chr_patch_hapl_scaff.gtf</i> and <i>vdj_GRCh38_alts_ensembl_10x_genes-2.0.0.gtf</i>
<b>GRCm38_vdj_v2.2.0</b> or <b>GRCm38_vdj</b>	Mouse GRCm38 V(D)J sequences, cellranger reference 2.2.0, annotation built from Ensembl <i>Mus_musculus.GRCm38.90.chr_patch_hapl_scaff.gtf</i>

2. **Index** column.

Put 10x single cell V(D)J sample index set names (e.g. SI-GA-A3) here.

3. **Chemistry** column.

This column is not used for scIR-seq data. Put **fiveprime** here as a placeholder if you decide to include the Chemistry column.

4. **DataType** column.

Set it to **vdj**.

5. **FeatureBarcodeFile** column.

Leave it blank for scIR-seq.

## 6. Example:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry,DataType
sample_vdj,GRCh38_vdj_v3.1.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK10WBC9ZZ,1,SI-GA-A1,fiveprime,vdj
```

## Workflow input

For scIR-seq data, `cellranger_workflow` takes Illumina outputs as input and runs `cellranger mkfastq` and `cellranger vdj`. Relevant workflow inputs are described below, with required inputs highlighted in bold.

Name	Description	Example	Default
input_sample_sheet_file	Sample Sheet (contains Sample, Reference, Flowcell, Lane, Index as required and Chemistry, DataType, FeatureBarcodeFile as optional)	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv”	
output_directory	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_output”	
run_mkfastq	Do you want to run cellranger mkfastq	true	true
run_cellranger_vdj	Do you want to run cellranger vdj	true	true
delete_bcl2fastq_directories	Delete bcl2fastq directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	false	false
mkfastq_number_of_mismatches	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	
mkfastq_force_suspect_i7_index	If force suspected i7 index paired indices are specified, but the flowcell was run with only one sample index, allow the demultiplex to proceed using the i7 half of the sample index pair	false	false
mkfastq_only_demultiplex_index	Only demultiplex samples identified by an i7-only sample index, ignoring dual-indexed samples. Dual-indexed samples will not be demultiplexed	false	false
mkfastq_override_read_lengths	Override the read lengths as specified in <i>RunInfo.xml</i>	“Y28n*,I8n*,N10,Y90n*”	
mkfastq_delete_undetermined	Delete undetermined FASTQ files generated by bcl2fastq2	true	false
vdj_de novo	Do not align reads to reference V(D)J sequences before de novo assembly	false	false
vdj_chain	Force the analysis to be carried out for a particular chain type. The accepted values are: <ul style="list-style-type: none"> <li>• “auto” for auto detection based on TR vs IG representation;</li> <li>• “TR” for T cell receptors;</li> <li>• “IG” for B cell receptors.</li> </ul>	“auto”	“auto”
cellranger_version	cellranger version, could be 7.1.0, 7.0.1, 7.0.0, 6.1.2, 6.1.1, 6.0.2, 6.0.1, 6.0.0, 5.0.1, 5.0.0	“7.1.0”	“7.1.0”
docker_registry	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
mkfastq_docker_registry	Docker registry to use for cellranger mkfastq. Default is the registry to which only Broad users have access. See <i>bcl2fastq</i> for making your own registry.	“gcr.io/broad-cumulus”	“gcr.io/broad-cumulus”
48		<b>Chapter 1. Release Highlights in Current Stable</b>	
acronym_file	The link/path of an index file in TSV format for a table	“s3://xxxx/index.tsv”	“gs://regev-lab/resources/cellranger/index.tsv”

## Workflow output

See the table below for important scIR-seq outputs.

Name	Type	Description
cellranger_mkfastq.output_fastq_files	Array[String]?	Subworkflow output. A list of cloud urls containing FASTQ files, one url per flowcell.
cellranger_vdj.output_vdj_dir	Array[String]?	Subworkflow output. A list of cloud urls containing vdj results, one url per sample.
cellranger_vdj.output_web_summary	Array[File]?	Subworkflow output. A list of htmls visualizing QCs for each sample (cellranger vdj output).
collect_summaries_vdj.metrics	FileSummaries	Task output. A excel spreadsheet containing QCs for each sample.

## Single-cell multiomics

To utilize cellranger arc/cellranger multi/cellranger count for single-cell multiomics, follow the specific instructions below. In particular, we put each single modality in one separate line in the sample sheet as described above. We then use the *Link* column to link multiple modalities together. Depending on the modalities included, *cellranger arc* (Multiome ATAC + Gene Expression), *cellranger multi* (CellPlex), or *cellranger count* (Feature Barcode) will be triggered. Note that cumulus\_feature\_barcoding/demuxEM would not be triggered for hashing/citeseq in this setting.

## Sample sheet

### 1. Reference column.

Pre-built Multiome ATAC + Gene Expression references are summarized below. CellPlex and Feature Barcode use the same reference as in Single-cell and single-nucleus RNA-seq.

Keyword	Description
<b>GRCh38-2020-A_arc_v2.0.0</b>	Human GRCh38 sequences (GENCODE v32/Ensembl 98), cellranger arc reference 2.0.0
<b>mm10-2020-A_arc_v2.0.0</b>	Mouse GRCm38 sequences (GENCODE vM23/Ensembl 98), cellranger arc reference 2.0.0
<b>GRCh38-2020-A_arc_v1.0.0</b>	Human GRCh38 sequences (GENCODE v32/Ensembl 98), cellranger arc reference 1.0.0
<b>mm10-2020-A_arc_v1.0.0</b>	Mouse GRCm38 sequences (GENCODE vM23/Ensembl 98), cellranger arc reference 1.0.0

### 2. DataType column.

For each modality, set it to the corresponding data type.

### 3. FeatureBarcodeFile column.

For RNA-seq modality, only set this if a target panel is provided. For CMO (CellPlex), provide sample name - CMO tag association as follows:

```
sample1, CMO301|CMO302
sample2, CMO303
```

For CITESeq, Perturb-seq and hashing, provide one CSV file as defined in [Feature Barcode Reference](#). Note that one feature barcode reference should be provided for all feature-barcode related modalities (e.g. *citeseq*, *hashing*, *crispr*) and all these modalities should put the same reference file in *FeatureBarcodeFile* column.

#### 4. *Link* column.

Put a sample unique link name for all modalities that are linked.

#### 5. Example:

```
Sample,Reference,Flowcell,Lane,Index,DataType,FeatureBarcodeFile,Link
sample1_rna,GRCh38-2020-A_arc_v2.0.0,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK10WBC9ZZ,*,SI-TT-A1,rna,,sample1
sample1_atac,GRCh38-2020-A_arc_v2.0.0,gs://fc-e0000000-0000-0000-0000-
↪000000000000/VK10WBC9ZZ,*,SI-TT-N1,atac,,sample1
sample2_rna,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9ZX,
↪*,SI-TT-A2,rna,,sample2
sample2_cmo,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9ZX,
↪*,SI-TT-N2,cmo,gs://fc-e0000000-0000-0000-0000-000000000000/cmo.csv,sample2
sample3_rna,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9ZY,
↪*,SI-TT-A3,rna,,sample3
sample3_citeseq,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↪VK10WBC9ZY,*,SI-TT-N3,citeseq,gs://fc-e0000000-0000-0000-0000-000000000000/
↪feature_ref.csv,sample3
```

In the above example, three linked samples are provided. *cellranger arc*, *cellranger multi* and *cellranger count* will be triggered respectively.

## Workflow input

For single-cell multiomics data, *cellranger\_workflow* takes Illumina outputs as input and runs *cellranger-arc mkfastq/cellranger mkfastq* and *cellranger-arc count/cellranger multi/cellranger count*. Relevant workflow inputs are described below, with required inputs highlighted in bold.

Name	Description	Example	Default
<b>input_sample_sheet</b>	Sample Sheet (contains Sample, Reference, Flowcell, Lane, Index as required and Chemistry, DataType, FeatureBarcodeFile, Link as optional)	"gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv"	
<b>output_directory</b>	Output directory	"gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_output"	
<b>run_mkfastq</b>	you want to run cellranger-arc mkfastq/cellranger mkfastq	true	true
<b>run_count</b>	you want to run cellranger-arc count/cellranger multi/cellranger count	true	true
<b>delete_intermediate_directories</b>	Delete intermediate directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	false	false

Continued on next page

Table 1 – continued from previous page

Name	Description	Example	Default
mkfastq_num_bcl2fastq2_matches	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	
mkfastq_force_use_i7	If force_use_i7 is specified, but the flowcell was run with only one sample index, allow the demultiplex to proceed using the i7 half of the sample index pair	false	false
mkfastq_only_demultiplex_i7	Only demultiplex samples identified by an i7-only sample index, ignoring dual-indexed samples. Dual-indexed samples will not be demultiplexed	false	false
mkfastq_override_read_lengths	Override the read lengths as specified in <i>RunInfo.xml</i>	“Y28n*,I8n*,N10,Y90n*”	
mkfastq_delete_undetermined	Delete undetermined FASTQ files generated by bcl2fastq2	true	false
force_cells	Force pipeline to use this number of cells, bypassing the cell detection algorithm, mutually exclusive with expect_cells. This option is used by <i>cellranger multi</i> and <i>cellranger count</i> .	6000	
expect_cells	Expected number of recovered cells. Mutually exclusive with force_cells. This option is used by <i>cellranger multi</i> and <i>cellranger count</i> .	3000	
include_introns	Turn this option on to also count reads mapping to intronic regions. With this option, users do not need to use pre-mRNA references. Note that if this option is set, cellranger_version must be >= 5.0.0. This option is used by <i>cellranger multi</i> and <i>cellranger count</i> .	true	true
arc_gex_exclude_introns	Disable counting of intronic reads. In this mode, only reads that are exonic and compatible with annotated splice junctions in the reference are counted. <b>Note:</b> using this mode will reduce the UMI counts in the feature-barcode matrix.	false	false
no_bam	Turn this option on to disable BAM file generation. This option is only available if cellranger_version >= 5.0.0. This option is used by <i>cellranger-arc count</i> , <i>cellranger multi</i> and <i>cellranger count</i> .	false	false

Continued on next page

Table 1 – continued from previous page

Name	Description	Example	Default
arc_min_atac_count	Cell caller override to define the minimum number of ATAC transposition events in peaks (ATAC counts) for a cell barcode. <b>Note:</b> this input must be specified in conjunction with <code>arc_min_gex_count</code> input. With both inputs set, a barcode is defined as a cell if it contains at least <code>arc_min_atac_count</code> ATAC counts AND at least <code>arc_min_gex_count</code> GEX UMI counts.	100	
arc_min_gex_count	Cell caller override to define the minimum number of GEX UMI counts for a cell barcode. <b>Note:</b> this input must be specified in conjunction with <code>arc_min_atac_count</code> . See the description of <code>arc_min_atac_count</code> input for details.	200	
peaks	A 3-column BED file of peaks to override cellranger arc peak caller. Peaks must be sorted by position and not contain overlapping peaks; comment lines beginning with # are allowed	“gs://fc-e0000000-0000-0000-0000-000000000000/common_peaks.bed”	
secondary	Perform Cell Ranger secondary analysis (dimensionality reduction, clustering, etc.). This option is used by <i>cellranger multi</i> and <i>cellranger count</i> .	false	false
cmo_set	CMO set CSV file, declaring CMO constructs and associated barcodes. See <a href="#">CMO reference</a> for details. Used only for <i>cellranger multi</i> .	“gs://fc-e0000000-0000-0000-0000-000000000000/cmo_set.csv”	
cellranger_version	cellranger version, could be 7.1.0, 7.0.1, 7.0.0, 6.1.2, 6.1.1, 6.0.2, 6.0.1, 6.0.0, 5.0.1, 5.0.0	“7.1.0”	“7.1.0”
cellranger_umi_version	cellranger umi version, could be 2.0.2, 2.0.1, 2.0.0, 1.0.1, 1.0.0	“2.0.2”	“2.0.2”

Continued on next page



Table 1 – continued from previous page

Name	Description	Example	Default
docker_registry	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>“quay.io/cumulus” for images on Red Hat registry;</li> <li>“cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
mkfastq_docker_registry	Docker registry to use for cellranger-arc mkfastq/cellranger mkfastq. Default is the registry to which only Broad users have access. See <a href="#">bcl2fastq</a> for making your own registry.	“gcr.io/broad-cumulus”	“gcr.io/broad-cumulus”
acronym_file	The link/path of an index file in TSV format for fetching preset genome references, chemistry whitelists, etc. by their names. Set an GS URI if <i>backend</i> is <i>gcp</i> ; an S3 URI for <i>aws</i> backend; an absolute file path for <i>local</i> backend.	“s3://xxxx/index.tsv”	“gs://regev-lab/resources/cellranger/index.tsv”
zones	Google cloud zones	“us-central1-a us-west1-a”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
num_cpus	Number of cpus to request for one node for cellranger mkfastq and cellranger vdj	32	32
memory	Memory size string for cellranger/cellranger-arc mkfastq and cellranger vdj	“120G”	“120G”
mkfastq_disk_space	On-disk space in GB for mkfastq	1500	1500
count_disk_space	Disk space in GB needed for cellranger count	500	500
arc_num_cpus	Number of cpus to request for one node for cellranger-arc count	64	64
arc_memory	Memory size string for cellranger-arc count	“160G”	“160G”
arc_disk_space	Disk space in GB needed for cellranger-arc count	700	700
backend	Cloud backend for file transfer. Available options: <ul style="list-style-type: none"> <li>“gcp” for Google Cloud;</li> <li>“aws” for Amazon AWS;</li> <li>“local” for local machine.</li> </ul>	“gcp”	“gcp”
preemptible	Number of preemptible tries	2	2

Continued on next page

Table 1 – continued from previous page

Name	Description	Example	Default
awsQueueArn	The AWS ARN string of the job queue to be used. This only works for <code>aws</code> backend.	<code>"arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf"</code>	<code>""</code>

## Workflow output

See the table below for important sc/snRNA-seq outputs.

Name	Type	Description
<code>cellranger_arc_mkfastq.output_fastqs_directory</code> / <code>cell-ranger_mkfastq.output_fastqs_directory</code>	Array[String]	Subworkflow output. A list of cloud urls containing FASTQ files, one url per flowcell.
<code>cellranger_arc_count.output_multi_directory</code> / <code>cell-ranger_multi.output_multi_directory</code> / <code>cell-ranger_count_fbc.output_count_directory</code>	Array[String]	Subworkflow output. A list of cloud urls containing <i>cellranger-arc count</i> , <i>cellranger multi</i> or <i>cellranger count</i> outputs, one url per sample.
<code>cellranger_arc_count.output_web_summary</code> / <code>cell-ranger_count_fbc.output_web_summary</code>	Array[String]	A list of htmls visualizing QCs for each sample ( <i>cellranger-arc count</i> / <i>cellranger count</i> output).
<code>collect_summaries_arc.metrics_summaries</code> / <code>collect_summaries_fbc.metrics_summaries</code>	File	A excel spreadsheet containing QCs for each sample.

## Fixed RNA Profiling

Cellranger multi supports [Fixed RNA Profiling](#) since version 7.0.0.

## Sample Sheet

### 1. Reference column.

Prebuilt scRNA-seq references for FRP data processing are summarized below.

Keyword	Description
<b>GRCh38-2020-A</b>	Human GRCh38 (GENCODE v32/Ensembl 98)

### 2. DataType column.

Set `frp` for RNA-Seq modalities of your FRP samples. For other modalities (e.g. citeseq or anti-body), set to their corresponding data types.

### 3. ProbeSet column.

Preset probe set references for FRP samples:

Keyword	Description
<b>FRP_human_probe_v1.0.1</b>	FRP probe set v1.0.1 for human. <i>Only work with Cell Ranger v7.1+.</i>
<b>FRP_mouse_probe_v1.0.1</b>	FRP probe set v1.0.1 for mouse. <i>Only work with Cell Ranger v7.1+.</i>
<b>FRP_human_probe_v1.0</b>	FRP probe set v1.0 for human. Work with Cell Ranger v7.0+.
<b>FRP_mouse_probe_v1.0</b>	FRP probe set v1.0 for mouse. Work with Cell Ranger v7.0+.

If *ProbeSet* column is not set, use **FRP\_human\_probe\_v1.0.1** by default.

#### 4. *FeatureBarcodeFile* column.

Provide sample name - Probe Barcode association as follows:

```
sample1,BC001|BC002,Control
sample2,BC003|BC004,Treated
```

where the third column (i.e. *Control* and *Treated* above) is optional, which specifies the description of the samples.

#### 5. *Link* column.

Put a sample unique link name for all modalities that are linked.

If *Link* column is not set, only consider RNA-seq modalities (i.e. samples of *DataType* *frp*) and use their *Sample* names as the *Link* names.

#### 6. Example:

```
Sample,Reference,ProbeSet,Flowcell,DataType,FeatureBarcodeFile,Link
sample1,GRCh38-2020-A,FRP_human_probe_v1,/path/to/sample1/fastq/folder,frp,/path/
↳to/sample1/fbf/file,
sample2_rna,GRCh38-2020-A,FRP_human_probe_v1,/path/to/sample2/rna/fastq/folder,
↳frp,/path/to/sample2/rna/fbf/file,sample2
sample2_citeseq,GRCh38-2020-A,,/path/to/sample2/citeseq/fastq/folder,citeseq,/
↳path/to/sample2/citeseq/fbf/file,sample2
```

In the example above, two linked samples are provided.

## Workflow Input

For FRP data, `cellranger_workflow` takes Illumina outputs as input and runs `cellranger mkfastq` and `cellranger multi`. Relevant workflow inputs are described below, with required inputs highlighted in **bold**:

Name	Description	Example	Default
<b>input_sample_sheet</b>	Sample Sheet (contains Sample, Reference, DataType, Flowcell as required; Lane and Index are required if <code>run_mkfastq</code> is <code>true</code> ; ProbeSet, FeatureBarcodeFile and Link are optional)	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv”	
<b>output_directory</b>	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_output”	
<b>run_mkfastq</b>	Do you want to run cellranger mkfastq	<code>true</code>	<code>true</code>
<b>run_cellranger_multi</b>	Do you want to run cellranger multi	<code>true</code>	<code>true</code>
<b>delete_bcl_directories</b>	Delete BCL directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	<code>false</code>	<code>false</code>
<b>mkfastq_num_mismatches</b>	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	1
<b>mkfastq_force_suplex_i7_i5</b>	If force-suplex i7/i5 paired indices are specified, but the flowcell was run with only one sample index, allow the demultiplex to proceed using the i7 half of the sample index pair	<code>false</code>	<code>false</code>
<b>mkfastq_only_demultiplex</b>	Only demultiplex samples identified by an i7-only sample index, ignoring dual-indexed samples. Dual-indexed samples will not be demultiplexed	<code>false</code>	<code>false</code>
<b>mkfastq_override_read_lengths</b>	Override the read lengths as specified in RunInfo.xml	““Y28n*,I8n*,N10,Y90n*””	
<b>mkfastq_delete_unidentified</b>	Delete unidentified FASTQ files generated by bcl2fastq2	<code>false</code>	<code>false</code>
<b>force_cells</b>	Force pipeline to use this number of cells, bypassing the cell detection algorithm, mutually exclusive with <code>expect_cells</code> . This option is used by <code>cellranger multi</code> .	6000	
<b>expect_cells</b>	Expected number of recovered cells. Mutually exclusive with <code>force_cells</code> . This option is used by <code>cellranger multi</code> .	3000	
<b>include_introns</b>	Turn this option on to also count reads mapping to intronic regions. With this option, users do not need to use pre-mRNA references. Note that if this option is set, <code>cellranger_version</code> must be <code>&gt;= 5.0.0</code> . This option is used by <code>cellranger multi</code> .	<code>true</code>	<code>true</code>
<b>no_bam</b>	Turn this option on to disable BAM file generation. This option is only available if <code>cellranger_version &gt;= 5.0.0</code> . This option is used by <code>cellranger multi</code> .	<code>false</code>	<code>false</code>
<b>secondary_analysis</b>	Perform Cell Ranger secondary analysis (dimensionality reduction, clustering, etc.). This option is used by <code>cellranger multi</code> .	<code>false</code>	<code>false</code>
<b>cellranger_version</b>	Cell Ranger version to use. Available	“7.1.0”	“7.1.0”

## Workflow Output

See the table below for important outputs:

Name	Type	Description
fastq_outputs	Array[Array[String]]	<code>fastq_outputs[0]</code> gives the list of cloud urls containing FASTQ files for RNA-Seq modalities of FRP data, one url per flowcell.
count_outputs	Map[String, Array[String]]	<code>count_outputs["multi"]</code> gives the list of cloud urls containing <i>cellranger multi</i> outputs, one url per sample.

## Build Cell Ranger References

We provide routines wrapping Cell Ranger tools to build references for sc/snRNA-seq, scATAC-seq and single-cell immune profiling data.

### Build references for sc/snRNA-seq

We provide a wrapper of `cellranger mkref` to build sc/snRNA-seq references. Please follow the instructions below.

#### 1. Import `cellranger_create_reference`

Import `cellranger_create_reference` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose [github.com/kalarman-cell-observatory/cumulus/Cellranger\\_create\\_reference](https://github.com/kalarman-cell-observatory/cumulus/Cellranger_create_reference) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `cellranger_create_reference` workflow in the drop-down menu.

#### 2. Upload required data to Google Bucket

Required data may include input sample sheet, genome FASTA files and gene annotation GTF files.

#### 3. Input sample sheet

If multiple species are specified, a sample sheet in CSV format is required. We describe the sample sheet format below, with required columns highlighted in bold:

Column	Description
<b>Genome</b>	Genome name
<b>Fasta</b>	Location to the genome assembly in FASTA/FASTA.gz format
<b>Genes</b>	Location to the gene annotation file in GTF/GTF.gz format
Attributes	Optional, A list of <code>key:value</code> pairs separated by <code>;</code> . If set, <code>cellranger mkgtf</code> will be called to filter the user-provided GTF file. See <a href="#">10x filter with mkgtf</a> for more details

Please note that the columns in the CSV can be in any order, but that the column names must match the recognized headings.

See below for an example for building Example:

```
Genome,Fasta,Genes,Attributes
GRCh38,gs://fc-e0000000-0000-0000-0000-000000000000/GRCh38.fa.gz,gs://fc-
↪e0000000-0000-0000-0000-000000000000/GRCh38.gtf.gz,gene_biotype:protein_
↪coding;gene_biotype:lincRNA;gene_biotype:antisense
mm10,gs://fc-e0000000-0000-0000-0000-000000000000/mm10.fa.gz,gs://fc-
↪e0000000-0000-0000-0000-000000000000/mm10.gtf.gz
```

If multiple species are specified, the reference will built under **Genome** names concatenated by ‘\_and\_’s. In the above example, the reference is stored under ‘GRCh38\_and\_mm10’.

#### 4. Workflow input

Required inputs are highlighted in bold. Note that **input\_sample\_sheet** and **input\_fasta**, **input\_gtf** , **genome** and attributes are mutually exclusive.

Name	Description	Example	Default
<b>input_sample_sheet</b>	<b>input_sample_sheet</b> in CSV format allows users to specify more than 1 genomes to build references (e.g. human and mouse). If a sample sheet is provided, <b>input_fasta</b> , <b>input_gtf</b> , and attributes will be ignored.	“gs://fc-e0000000-0000-0000-0000-000000000000/input_sample_sheet.csv”	
<b>input_fasta</b>	<b>input_fasta</b> genome reference in either FASTA or FASTA.gz format	“gs://fc-e0000000-0000-0000-0000-000000000000/Homo_sapiens.GRCh38.dna.toplevel.fa.gz”	
<b>input_gtf</b>	<b>input_gtf</b> gene annotation file in either GTF or GTF.gz format	“gs://fc-e0000000-0000-0000-0000-000000000000/Homo_sapiens.GRCh38.94.chr_patch_hapl_scaff.gtf.gz”	
<b>genome</b>	genome reference name. New reference will be stored in a folder named <b>genome</b>	refdata-cellranger-vdj-GRCh38-alts-ensembl-3.1.0	
<b>output_directory</b>	<b>output_directory</b>	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_reference”	
<b>attributes</b>	Attributes list of key:value pairs separated by ;. If this option is not None, cellranger mkgtf will be called to filter the user-provided GTF file. See <a href="#">10x filter with mkgtf</a> for more details	“gene_biotype:protein_coding;gene_biotype:lincRNA;gene_biotype:antisense”	
<b>pre_mRNA</b>	Do you want to build pre-mRNA references, in which we use full length transcripts as exons in the annotation file. We follow <a href="#">10x build Cell Ranger compatible pre-mRNA Reference Package</a> to build pre-mRNA references	true	false
<b>ref_version</b>	reference version string	Ensembl v94	
<b>cellranger_version</b>	cellranger version, could be: 7.1.0, 7.0.1, 7.0.0, 6.1.2, 6.1.1	“7.1.0”	“7.1.0”
<b>docker_registry</b>	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>“quay.io/cumulus” for images on Red Hat registry;</li> <li>“cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>zones</b>	Google cloud zones	“us-central1-a us-west1-a”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
<b>num_cpus</b>	Number of cpus to request for one node for building indices	1	1
<b>memory</b>	Memory size string for cellranger mkref	“32G”	“32G”
<b>disk_space</b>	Optional disk space in GB	100	100
<b>backend</b>	Optional cloud backend for file transfer. Available options: <ul style="list-style-type: none"> <li>“gcp” for Google Cloud;</li> <li>“aws” for Amazon AWS;</li> </ul>	“gcp”	“gcp”

## 5. Workflow output

Name	Type	Description
output_reference	File	Gzipped reference folder with name <i>genome.tar.gz</i> . We will also store a copy of the gzipped tarball under <b>output_directory</b> specified in the input.

---

### Build references for scATAC-seq

We provide a wrapper of `cellranger-atac mkref` to build scATAC-seq references. Please follow the instructions below.

#### 1. Import `cellranger_atac_create_reference`

Import `cellranger_atac_create_reference` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus/Cellranger\\_atac\\_create\\_reference](https://github.com/lilab-bcb/cumulus/Cellranger_atac_create_reference) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `cellranger_atac_create_reference` workflow in the drop-down menu.

#### 2. Upload required data to Google Bucket

Required data include config JSON file, genome FASTA file, gene annotation file (GTF or GFF3 format) and motif input file (JASPAR format).

#### 3. Workflow input

Required inputs are highlighted in bold.



Name	Description	Example	Default
<b>genome</b>	Genome reference name. New reference will be stored in a folder named <b>genome</b>	refdata-cellranger-atac-mm10-1.1.0	
<b>input_fasta</b>	URL for input fasta file	“gs://fc-e0000000-0000-0000-0000-000000000000/GRCh38.fa”	
<b>input_gtf</b>	URL for input GTF file	“gs://fc-e0000000-0000-0000-0000-000000000000/annotation.gtf”	
<b>organism</b>	Name of the organism	“human”	
<b>non_nuclear_contigs</b>	Comma-separated list of names of contigs that are not in nucleus	“chrM”	“chrM”
<b>input_motifs</b>	Optional file containing transcription factor motifs in JASPAR format	“gs://fc-e0000000-0000-0000-0000-000000000000/motifs.pfm”	
<b>output_directory</b>	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_atac_reference”	
<b>cellranger-atac version</b>	cellranger-atac version, could be: 2.1.0, 2.0.0, 1.2.0, 1.1.0	“2.1.0”	“2.1.0”
<b>docker_registry</b>	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>zones</b>	Google cloud zones	“us-central1-a us-west1-a”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
<b>memory</b>	Memory size string for cellranger-atac mkref	“32G”	“32G”
<b>disk_space</b>	Optional disk space in GB	100	100
<b>backend</b>	Cloud backend for file transfer. Available options: <ul style="list-style-type: none"> <li>• “gcp” for Google Cloud;</li> <li>• “aws” for Amazon AWS;</li> <li>• “local” for local machine.</li> </ul>	“gcp”	“gcp”
<b>preemptible</b>	Number of preemptible tries	2	2
<b>aws_queue</b>	The AWS ARN string of the job queue to be used. This only works for aws backend.	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf”	“”

## 4. Workflow output

Name	Type	Description
output_reference	File	Gzipped reference folder with name <i>genome.tar.gz</i> . We will also store a copy of the gzipped tarball under <b>output_directory</b> specified in the input.

---

### Build references for single-cell immune profiling data

We provide a wrapper of `cellranger mkvdjref` to build single-cell immune profiling references. Please follow the instructions below.

#### 1. Import `cellranger_vdj_create_reference`

Import `cellranger_vdj_create_reference` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus/Cellranger\\_vdj\\_create\\_reference](https://github.com/lilab-bcb/cumulus/Cellranger_vdj_create_reference) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `cellranger_vdj_create_reference` workflow in the drop-down menu.

#### 2. Upload required data to Google Bucket

Required data include genome FASTA file and gene annotation file (GTF format).

#### 3. Workflow input

Required inputs are highlighted in bold.

Name	Description	Example	Default
<b>input_fasta</b>	Genome reference in either FASTA or FASTA.gz format	“gs://fc-e0000000-0000-0000-0000-000000000000/Homo_sapiens.GRCh38.dna.toplevel.fa.gz”	
<b>input_gtf</b>	Gene annotation file in either GTF or GTF.gz format	“gs://fc-e0000000-0000-0000-0000-000000000000/Homo_sapiens.GRCh38.94.chr_patch_hapl_scaff.gtf.gz”	
<b>genome</b>	Genome reference name. New reference will be stored in a folder named <b>genome</b>	refdata-cellranger-vdj-GRCh38-alts-ensembl-3.1.0	
<b>output_directory</b>	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/cellranger_vdj_reference”	
<b>ref_version</b>	Reference version string	Ensembl v94	
<b>cellranger_version</b>	Cell Ranger version, could be: 7.1.0, 7.0.1, 7.0.0, 6.1.2, 6.1.1	“7.1.0”	“7.1.0”
<b>docker_registry</b>	Docker registry to use for cellranger_workflow. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>zones</b>	Google cloud zones	“us-central1-a us-west1-a”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
<b>memory</b>	Memory size string for cellranger mkvdjref	“32G”	“32G”
<b>disk_space</b>	Optional disk space in GB	100	100
<b>backend</b>	Cloud backend for file transfer. Available options: <ul style="list-style-type: none"> <li>• “gcp” for Google Cloud;</li> <li>• “aws” for Amazon AWS;</li> <li>• “local” for local machine.</li> </ul>	“gcp”	“gcp”
<b>preemptible</b>	Number of preemptible tries	2	2
<b>aws_queue</b>	The AWS ARN string of the job queue to be used. This only works for aws backend.	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf”	“”

#### 4. Workflow output

Name	Type	Description
output_reference	File	Gzipped reference folder with name <i>genome.tar.gz</i> . We will also store a copy of the gzipped tarball under <b>output_directory</b> specified in the input.

## 1.1.6 Run STARsolo to generate gene-count matrices from FASTQ files

This `starsolo_workflow` workflow generates gene-count matrices from FASTQ data using STARsolo.

### Prepare input data and import workflow

#### 1. Run `cellranger_workflow` to generate FASTQ data

You can skip this step if your data are already in FASTQ format.

Otherwise, for 10X data, you need to first run `cellranger_workflow` to generate FASTQ files from BCL raw data for each sample. Please follow [cellranger\\_workflow manual](#).

Notice that you should set `run_mkfastq` to `true` to get FASTQ output. You can also set `run_count` to `false` to skip Cell Ranger count step.

For Non-Broad users, you'll need to build your own docker for `bcl2fastq` step. Instructions are [here](#).

#### 2. Import `starsolo_workflow`

Import `starsolo_workflow` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose workflow [github.com/lilab-bcb/cumulus/STARsolo](https://github.com/lilab-bcb/cumulus/STARsolo) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `starsolo_workflow` in the drop-down menu.

### 3. Prepare a sample sheet

#### 3.1 Sample sheet format:

Please note that the columns in the CSV can be in any order, but that the column names must match the recognized headings.

The sample sheet describes how to identify flowcells and generate sample/channel-specific count matrices.

A brief description of the sample sheet format is listed below (**required column headers are shown in bold**).

Column	Description
<b>Sample</b>	Contains the sample name. Each sample should have a unique sample name.
<b>Reference</b>	<p>Provides the reference genome used by STARSolo for each sample.</p> <p>The elements in this column can be either Cloud bucket URIs to reference tarballs or keywords such as <i>GRCh38-2020-A</i>.</p> <p>A full list of available keywords is included in <a href="#">genome reference</a> section below.</p>
<b>Location</b>	Indicates the Cloud bucket URI of the folder holding FASTQ files of each sample.
<b>Assay</b>	<p>Indicates the assay type of each sample. Available options:</p> <ul style="list-style-type: none"> <li>• <code>tenX_v3</code> for 10x 3' v3</li> <li>• <code>tenX_multiome</code> for 10x multiome</li> <li>• <code>tenX_v2</code> for 10x 3' v2</li> <li>• <code>tenX_5p</code> for 10x 5' (only use R2 for alignment; equivalent to 10x chemistry <i>SC5P-R2</i>)</li> <li>• <code>tenX_5p_pe</code> for 10x 5' (use both R1 and R2 for alignment, and R1 has length longer than 39 nt; equivalent to 10x chemistry <i>SC5P-PE</i>)</li> <li>• DropSeq</li> <li>• SeqWell</li> <li>• SlideSeq</li> <li>• ShareSeq</li> <li>• None</li> </ul> <p>If not specified, use the default <code>tenX_v3</code>.</p>

### 3.2 Assay-specific preset STARSolo options

If **tenX\_v3**, The following STARSolo options would be applied (could be overwritten by user-specified options):

```
--soloType CB_UMI_Simple --soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --
↪soloUMIlen 12 --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts --soloUMIfiltering_
↪MultiGeneUMI_CR --soloUMIdedup 1MM_CR --clipAdapterType CellRanger4 --
↪outFilterScoreMin 30 --outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR_
↪CY UY CB UB
```

If **tenX\_multiome**, use the same STARSolo options as for *tenX\_v3* assay, but with the 10X ARC Multiome Gene Expression whitelist.

If **tenX\_v2**, the following STARSolo options would be applied (could be overwritten by user-specified options):

```
--soloType CB_UMI_Simple --soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --
↪soloUMIlen 10 --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts --soloUMIfiltering_
↪MultiGeneUMI_CR --soloUMIdedup 1MM_CR --clipAdapterType CellRanger4 --
↪outFilterScoreMin 30 --outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR_
↪CY UY CB UB
```

If **tenX\_5p**, the following STARSolo options would be applied (could be overwritten by user-specified options):

```
--soloType CB_UMI_Simple --soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --
↪soloUMIlen 10 --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts --soloUMIfiltering_
↪MultiGeneUMI_CR --soloStrand Reverse --soloUMIdedup 1MM_CR --outFilterScoreMin 30 --
↪outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR CY UY CB UB
```

If **tenX\_5p\_pe**, the following STARSolo options would be applied (could be overwritten by user-specified options):

```
--soloType CB_UMI_Simple --soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --
↪soloUMIlen 10 --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts --soloUMIfiltering_
↪MultiGeneUMI_CR --soloBarcodeMate 1 --clip5pNbases 39 0 --soloUMIdedup 1MM_CR
↪outFilterScoreMin 30 --outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR_
↪CY UY CB UB
```

(continues on next page)

(continued from previous page)

If **ShareSeq**, the following STARsolo options would be applied (could be overwritten by user-specific options):

```
--soloType CB_UMI_Simple --soloCBstart 1 --soloCBlen 24 --soloUMIstart 25 --
↪soloUMIlen 10 --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts --soloUMIfiltering_
↪MultiGeneUMI_CR --soloUMIdedup 1MM_CR --clipAdapterType CellRanger4 --
↪outFilterScoreMin 30 --outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR_
↪CY UY CB UB
```

If **SeqWell** or **DropSeq**, the following STARsolo options would be applied (could be overwritten by user-specified options):

```
--soloType CB_UMI_Simple --soloCBstart 1 --soloCBlen 12 --soloUMIstart 13 --
↪soloUMIlen 8 --outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR CY UY CB_
↪UB
```

If **None**, no preset option would be applied.

The sample sheet supports sequencing the same sample across multiple flowcells. In case of multiple flowcells, you should specify one line for each flowcell using the same sample name. In the following example, we have 2 samples and sample\_1 is sequenced in two flowcells.

Example:

```
Sample,Reference,Location,Assay
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4/sample_
↪1_fastqs,tenX_v3
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9Z2/sample_
↪1_fastqs,tenX_v3
sample_2,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4/sample_
↪2_fastqs,tenX_v2
```

### 3.2 Upload your sample sheet to the workspace bucket:

Example:

```
gsutil cp /foo/bar/projects/sample_sheet.csv gs://fc-e0000000-0000-0000-0000-
↪000000000000/
```

## 1. Launch analysis

In your workspace, open `starsolo_workflow` in **WORKFLOWS** tab. Select the desired snapshot version (e.g. latest). Select **Process single workflow** from files as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click **SAVE** button. Select **Use call caching** and click **INPUTS**. Then fill in appropriate values in the **Attribute** column. Alternative, you can upload a JSON file to configure input by clicking **Drag** or click to **upload json**.

Once **INPUTS** are appropriated filled, click **RUN ANALYSIS** and then click **LAUNCH**.

## Workflow inputs

Below are inputs for *count* workflow. Notice that required inputs are in bold.

Name	Description	Example	Default
<b>input_csv_file</b>	Input CSV sample sheet describing metadata of each sample.	"gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.tsv"	
<b>output_directory</b>	GCloud bucket URI of output directory.	"gs://fc-e0000000-0000-0000-0000-000000000000/count_result"	
read1_fastq_pattern	<p>Filename suffix pattern in wildcards for Read 1. This is used for looking for Read 1 fastq files.</p> <p>If fastq files are generated by CellRanger count, use <code>_S*_L*_R1_001.fastq.gz</code>, which means Read 1 files must have names such as "<code>&lt;Sample&gt;_S1_L1_R1_001.fastq.gz</code>", where <code>&lt;Sample&gt;</code> is specified in <b>input_csv_file</b>.</p> <p>If fastq files are Sequence Read Archive (SRA) data, use something like <code>_1.fastq.gz</code>, where <code>_1</code> refers to the first reads, so that Read 1 files must have names such as "<code>&lt;Sample&gt;_1.fastq.gz</code>" where <code>&lt;Sample&gt;</code> is specified in <b>input_csv_file</b>.</p> <p>If fastq files are not zipped, substitute <code>.fastq</code> for <code>.fastq.gz</code> in the corresponding pattern above.</p>	"_S*_L*_R1_001.fastq.gz"	"_S*_L*_R1_001.fastq.gz"
read2_fastq_pattern	<p>Filename suffix pattern in wildcards for Read 2. This is used for looking for Read 2 fastq files.</p> <p>If fastq files are generated by CellRanger count, use <code>_S*_L*_R2_001.fastq.gz</code>, which means Read 2 files must have names such as "<code>&lt;Sample&gt;_S1_L1_R2_001.fastq.gz</code>", where <code>&lt;Sample&gt;</code> is specified in <b>input_csv_file</b>.</p> <p>If fastq files are Sequence Read Archive (SRA) data, use something like <code>_2.fastq.gz</code>, where <code>_2</code> refers to the second reads, so that Read 2 files must have names such as "<code>&lt;Sample&gt;_2.fastq.gz</code>" where <code>&lt;Sample&gt;</code> is specified in <b>input_csv_file</b>.</p> <p>If fastq files are not zipped, substitute <code>.fastq</code> for <code>.fastq.gz</code> in the corresponding pattern above.</p>	"_S*_L*_R2_001.fastq.gz"	"_S*_L*_R2_001.fastq.gz"

Continued on next page

Table 2 – continued from previous page

Name	Description	Example	Default
barcode_read	Specify which read contains cell barcodes and UMIs: either <code>read1</code> or <code>read2</code> . This only applies to samples with <i>Assay</i> <code>None</code> in <b>input_csv_file</b> . Otherwise, samples with <i>Assay</i> type <code>ShareSeq</code> automatically specify <code>read2</code> for cell barcodes and UMIs, while <code>read1</code> for cDNAs; samples of all the other know <i>Assay</i> types automatically specify <code>read1</code> for cell barcodes and UMIs, while <code>read2</code> for cDNAs.	“read1”	“read1”
soloType	[STARsolo option] Type of single-cell RNA-seq, choosing from <i>CB_UMI_Simple</i> , <i>CB_UMI_Complex</i> , <i>CB_samTagOut</i> , <i>SmartSeq</i> .	“CB_UMI_Simple”	None
soloCBwhitelist	[STARsolo option] Cell barcode white list in either plain text or gzipped format. <b>Notice:</b> If specified, it will overwrite the white lists for <b>ALL</b> the samples in your sample sheet.	gs://my_bucket/my_white_list.txt	None
soloFeatures	[STARsolo option] Genomic features for which the UMI counts per Cell Barcode are collected (can choose multiple items): <ul style="list-style-type: none"><li>• <i>Gene</i>: reads match the gene transcript</li><li>• <i>SJ</i>: splice junctions reported in SJ.out.tab</li><li>• <i>GeneFull</i>: count all reads overlapping genes’ exons and introns</li><li>• <i>Velocity</i>: calculate Spliced, Unspliced, and Ambiguous counts per cell per gene similar to the velocity.py tool developed by LaManno et al. Note that <i>Velocity</i> requires <i>Gene</i>.</li></ul>	“Gene GeneFull SJ Velocity”	“Gene”
soloMultiMapping	[STARsolo option] Counting method for reads mapping to multiple genes (can choose multiple items): <ul style="list-style-type: none"><li>• <i>Unique</i>: count only reads that map to unique genes</li><li>• <i>Uniform</i>: uniformly distribute multi-genic UMIs to all genes</li><li>• <i>Rescue</i>: distribute UMIs proportionally to unique+uniform counts (first iteration of EM)</li><li>• <i>PropUnique</i>: distribute UMIs proportionally to unique mappers, if present, and uniformly if not</li><li>• <i>EM</i>: use Maximum Likelihood Estimation (MLE) to distribute multi-gene UMIs among their genes</li></ul>	“Unique”	“Unique”
soloCBstart	[STARsolo option] Cell barcode start position (1-based coordinate).	1	1
soloCBlen	[STARsolo option] Cell barcode length.	16	16
soloUMIstart	[STARsolo option] UMI start position (1-based coordinate).	17	17

Continued on next page



Table 2 – continued from previous page

Name	Description	Example	Default
soloUMIlen	[STARsolo option] UMI length.	10	10
soloBarcodeReadLength	[STARsolo option] Length of the barcode read - 1: equals to sum of <i>soloCBlen</i> and <i>soloUMIlen</i> . - 0: not defined, do not check. <b>Notice:</b> 0 is set to be default, which is different from STAR. This is in case users have barcode read sequenced of length 28 nt (standard for 10x 3'), but assay is 5' (CB+UMI length is 26 nt).	0	0
soloBarcodeMate	[STARsolo option] Identifies which read mate contains the barcode (CB+UMI) sequence: <ul style="list-style-type: none"> <li>• 0: barcode sequence is on separate read, which should always be the last file in the input Read1 file list</li> <li>• 1: barcode sequence is a part of mate 1</li> <li>• 2: barcode sequence is a part of mate 2</li> </ul>	0	0
soloCBposition	[STARsolo option] Position of Cell Barcode(s) on the barcode read. Presently only works when <i>solo_type</i> is CB_UMI_Complex, and barcodes are assumed to be on Read2. Format for each barcode: “startAnchor_startPosition_endAnchor_endPosition” start(end)Anchor defines the Anchor Base for the CB: 0: read start; 1: read end; 2: adapter start; 3: adapter end start(end)Position is the 0-based position with of the CB start(end) with respect to the Anchor Base String for different barcodes are separated by space.	“0_0_2_-1 3_1_3_8”	
soloUMIposition	[STARsolo option] Position of the UMI on the barcode read, same as soloCBposition	“3_9_3_14”	
soloAdapterSequence	[STARsolo option] Adapter sequence to anchor barcodes.		
soloAdapterMismatch	[STARsolo option] Maximum number of mismatches allowed in adapter sequence.	1	1

Continued on next page

Table 2 – continued from previous page

Name	Description	Example	Default
soloCBmatchWLType	[STARsolo option] Matching the Cell Barcodes to the WhiteList, choosing from <ul style="list-style-type: none"> <li><i>Exact</i>: only exact matches allowed</li> <li><i>IMM</i>: only one match in whitelist with 1 mismatched base allowed. Allowed CBs have to have at least one read with exact match</li> <li><i>IMM_multi</i>: multiple matches in whitelist with 1 mismatched base allowed, posterior probability calculation is used choose one of the matches. Allowed CBs have to have at least one read with exact match. This option matches best with CellRanger 2.2.0</li> <li><i>IMM_multi_pseudocounts</i>: same as <i>IMM_multi</i>, but pseudocounts of 1 are added to all whitelist barcodes</li> <li><i>IMM_multi_Nbase_pseudocounts</i>: same as <i>IMM_multi_pseudocounts</i>, multimatching to WL is allowed for CBs with N-bases. This option matches best with CellRanger &gt;= 3.0.0</li> </ul>	“IMM_multi”	“IMM_multi”
soloInputSAMattrBarcodeSeq	[STARsolo option] When inputting reads from a SAM file ( <code>--readsFileType SAM SE/PE</code> ), these SAM attributes mark the barcode qualities (in proper order). For instance, for 10X CellRanger or STARsolo BAMs, use <code>--soloInputSAMattrBarcodeSeq CR UR</code> . This parameter is required when running STARsolo with input from SAM.	“CR UR”	
soloInputSAMattrBarcodeQual	[STARsolo option] When inputting reads from a SAM file ( <code>--readsFileType SAM SE/PE</code> ), these SAM attributes mark the barcode sequence (in proper order). For instance, for 10X CellRanger or STARsolo BAMs, use <code>--soloInputSAMattrBarcodeQual CY UY</code> . If this parameter is – (default), the quality ‘H’ will be assigned to all bases.	“CY UY”	
soloStrand	[STARsolo option] Strandedness of the solo libraries: <ul style="list-style-type: none"> <li><i>Unstranded</i>: no strand information</li> <li><i>Forward</i>: read strand same as the original RNA molecule</li> <li><i>Reverse</i>: read strand opposite to the original RNA molecule</li> </ul>	“Forward”	“Forward”

Continued on next page

Table 2 – continued from previous page

Name	Description	Example	Default
soloUMIdedup	<p>[STARsolo option] Type of UMI deduplication (collapsing) algorithm:</p> <ul style="list-style-type: none"> <li>• <i>1MM_All</i>: all UMIs with 1 mismatch distance to each other are collapsed (i.e. counted once)</li> <li>• <i>1MM Directional UMIttools</i>: follows the “directional” method from the UMI-tools by Smith, Heger and Sudbery (Genome Research 2017)</li> <li>• <i>1MM Directional</i>: same as 1MM Directional UMIttools, but with more stringent criteria for duplicate UMIs</li> <li>• <i>Exact</i>: only exactly matching UMIs are collapsed</li> <li>• <i>NoDedup</i>: no deduplication of UMIs, count all reads</li> <li>• <i>1MM CR</i>: CellRanger2-4 algorithm for 1MM UMI collapsing</li> </ul>	“1MM_All”	“1MM_All”
soloUMIfiltering	<p>[STARsolo option] Type of UMI filtering (for reads uniquely mapping to genes):</p> <ul style="list-style-type: none"> <li>• -: basic filtering: remove UMIs with N and homopolymers (similar to CellRanger 2.2.0)</li> <li>• <i>MultiGeneUMI</i>: basic + remove lower-count UMIs that map to more than one gene</li> <li>• <i>MultiGeneUMI_All</i>: basic + remove all UMIs that map to more than one gene</li> <li>• <i>MultiGeneUMI_CR</i>: basic + remove lower-count UMIs that map to more than one gene, matching CellRanger &gt; 3.0.0. Only works with <code>--soloUMIdedup 1MM CR</code></li> </ul>	“MultiGeneUMI”	“-“

Continued on next page

Table 2 – continued from previous page

Name	Description	Example	Default
soloCellFilter	[STARsolo option] Cell filtering type and parameters: <ul style="list-style-type: none"> <li><i>None</i>: do not output filtered cells</li> <li><i>TopCells</i>: only report top cells by UMI count, followed by the exact number of cells</li> <li><i>CellRanger2.2</i>: simple filtering of CellRanger 2.2. Can be followed by numbers: number of expected cells, robust maximum percentile for UMI count, maximum to minimum ratio for UMI count. The hardcoded values are from CellRanger: nExpectedCells=3000; maxPercentile=0.99; maxMinRatio=10</li> <li><i>EmptyDrops_CR</i>: EmptyDrops filtering in CellRanger flavor. Please cite the original EmptyDrops paper: A.T.L Lun et al, Genome Biology, 20, 63 (2019): <a href="https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y">https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y</a>. Can be followed by 10 numeric parameters: nExpectedCells maxPercentile maxMinRatio indMin indMax umiMin umiMinFracMedian candMaxN FDR simN. The hardcoded values are from CellRanger: 3000 0.99 10 45000 90000 500 0.01 20000 0.01 10000</li> </ul>	“CellRanger2.2 3000 0.99 10”	“CellRanger2.2 3000 0.99 10”
soloOutFormat	[STARsolo option] Field 3 in the Gene features.tsv file. If “-“, then no 3rd field is output.	“Gene Expression”	“Gene Expression”
outSAMtype	[STAR option] Type of SAM/BAM output.	“BAM SortedByCoordinate”	“BAM SortedByCoordinate” for <i>tenX_v3</i> , <i>tenX_v2</i> , <i>SeqWell</i> and <i>DropSeq</i> assay types, “BAM Unsorted” otherwise.
star_version	STAR version to use. Currently support: 2.7.9a, 2.7.10a (2.7.10a_alpha_220601).	“2.7.10a”	“2.7.10a”
docker_registry	Docker registry to use: <ul style="list-style-type: none"> <li><a href="https://quay.io/cumulus">quay.io/cumulus</a> for images on Red Hat registry;</li> <li><a href="https://cumulusprod.docker.com">cumulusprod</a> for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”

Continued on next page

Table 2 – continued from previous page

Name	Description	Example	Default
zones	Google cloud zones to consider for execution.	“us-east1-d us-west1-a us-west1-b”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
num_cpu	Number of CPUs to request for count per sample.	32	32
memory	Memory size string for count per sample.	“120G”	“120G”
disk_space	Disk space in GB needed for count per sample.	500	500
backend	Cloud infrastructure backend to use. Available options: <ul style="list-style-type: none"> <li>• <code>gcp</code> for Google Cloud;</li> <li>• <code>aws</code> for Amazon AWS;</li> <li>• <code>local</code> for local machine.</li> </ul>	“gcp”	“gcp”
preemptible	Number of maximum preemptible tries allowed. This works only when <i>backend</i> is <code>gcp</code> .	2	2
awsQueueArn	The AWS ARN string of the job queue to be used. This only works for <code>aws</code> backend.	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf”	“”

## Workflow outputs

See the table below for *star\_solo* workflow outputs.

Name	Type	Description
count_outputs	Array[String]	<p>Google Bucket URI of output directories of all samples. Each folder is for one sample in the input sample sheet.</p> <p>For the count matrices generated, taking Gene solo feature for example, they are in  <code>&lt;output_folder&gt;/&lt;sample_id&gt;/Solo.out/Gene/raw/</code> and  <code>&lt;output_folder&gt;/&lt;sample_id&gt;/Solo.out/Gene/filtered/</code>  subfolders.</p> <p>Inside each subfolder, there are 2 formats: <code>mtx</code>, and <code>h5</code> following <a href="#">10x HDF5 format</a>.</p>
starsoloLogs	Array[File]	Google Bucket URIs of STAR logs for each sample, respectively. This is the <code>Log.out</code> if running STAR locally, which is important for debugging.

## Prebuilt genome references

We've built the following scRNA-seq references for users' convenience:

Keyword	Description
<b>GRCh38-2020-A</b>	Human GRCh38, comparable to cellranger reference 2020-A (GENCODE v32/Ensembl 98)
<b>mm10-2020-A</b>	Mouse mm10, comparable to cellranger reference 2020-A (GENCODE vM23/Ensembl 98)
<b>GRCh38-and-mm10-2020-A</b>	Human GRCh38 (GENCODE v32/Ensembl 98) and mouse mm10 (GENCODE vM23/Ensembl 98)

**Note:** For **snRNA-seq** data, please choose the corresponding scRNA-seq reference above, and add `GeneFull` in the `soloFeatures` input.

## Build STARSolo References

We provide a wrapper of STAR to build sc/snRNA-seq references. Please follow the instructions below.

### 1. Import `starsolo_create_reference`

Import `starsolo_create_reference` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/STARsolo\\_create\\_reference](https://github.com/lilab-bcb/STARsolo_create_reference) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `starsolo_create_reference` workflow in the drop-down menu.

## 2. Upload required data to Cloud bucket

Required data include the genome FASTA file and gene annotation GTF file of the target genome reference.

## 3. Workflow input

Required inputs are highlighted **in bold**.

Name	Description	Example	Default
<b>input_fasta</b>	Input genome reference in FASTA format.	“gs://fc-e0000000-0000-0000-0000-000000000000/mm-10/genome.fa”	
<b>input_gtf</b>	Input gene annotation file in GTF format.	“gs://fc-e0000000-0000-0000-0000-000000000000/mm-10/genes.gtf”	
<b>genome</b>	Genome reference name. This is used for specifying the name of the genome index generated.	“mm-10”	
<b>output_directory</b>	Cloud bucket URI of the output directory.	“gs://fc-e0000000-0000-0000-0000-000000000000/starsolo-reference”	
<b>docker_registry</b>	Docker registry to use: <ul style="list-style-type: none"> <li>quay.io/cumulus for images on Red Hat registry;</li> <li>cumulusprod for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>star_version</b>	STAR version to use. Currently support: 2.7.9a and 2.7.10a (2.7.10a_alpha_220601).	“2.7.10a”	“2.7.10a”
<b>num_cpu</b>	Number of CPUs to request for count per sample.	32	32
<b>memory</b>	Memory size string for count per sample.	“80G”	“80G”
<b>disk_space</b>	Disk space in GB needed for count per sample.	100	100
<b>zones</b>	Google cloud zones to consider for execution.	“us-east1-d us-west1-a us-west1-b”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
<b>backend</b>	Cloud infrastructure backend to use. Available options: <ul style="list-style-type: none"> <li>gcp for Google Cloud;</li> <li>aws for Amazon AWS;</li> <li>local for local machine.</li> </ul>	“gcp”	“gcp”
<b>preemptible</b>	Number of maximum preemptible tries allowed. This works only when <i>backend</i> is <i>gcp</i> .	2	2
<b>awsQueueArn</b>	The AWS ARN string of the job queue to be used. This only works for <i>aws</i> backend.	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf”	“”



## 4. Workflow Output

Name	Type	Description
output_reference	File	Gzipped reference folder with name “<genome>-starsolo.tar.gz”, where <genome> is specified by workflow input <b>genome</b> above. The workflow will save a copy of it under <b>output_directory</b> specified in workflow input above.

### 1.1.7 Demultiplex genetic-pooling/cell-hashing/nucleus-hashing sc/snRNA-Seq data

This demultiplexing workflow generates gene-count matrices from cell-hashing/nucleus-hashing/genetic-pooling data by demultiplexing.

In the workflow, demuxEM is used for analyzing cell-hashing/nucleus-hashing data, while souporcell and popscle (including *demuxlet* and *freemuxlet*) are for genetic-pooling data.

## Prepare input data and import workflow

### 1. Run cellranger\_workflow

To demultiplex, you’ll need raw gene count and hashtag matrices for cell-hashing/nucleus-hashing data, or raw gene count matrices and genome BAM files for genetic-pooling data. You can generate these data by running the `cellranger_workflow`.

Please refer to the [cellranger\\_workflow tutorial](#) for details.

When finished, you should be able to find the raw gene count matrix (e.g. `raw_gene_bc_matrices_h5.h5`), hashtag matrix (e.g. `sample_1_ADT.csv`) / genome BAM file (e.g. `possorted_genome_bam.bam`) for each sample.

### 2. Import demultiplexing

Import *demultiplexing* workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus/Demultiplexing](https://github.com/lilab-bcb/cumulus/Demultiplexing) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export *demultiplexing* workflow in the drop-down menu.

## 3. Prepare a sample sheet

### 3.1 Sample sheet format:

Create a sample sheet, **sample\_sheet\_demux.csv**, which describes the metadata for each pair of RNA and hashtag data. A brief description of the sample sheet format is listed below (**required column headers are shown in bold**).

Column	Description
<b>OUTNAME</b>	Output name for one pair of RNA and hashtag data. Must be unique per pair.
<b>RNA</b>	Google bucket url to the raw gene count matrix generated in Step 1.
<b>TagFile/ADT</b>	Google bucket url to the hashtag file generated in Step 1. The column name can be either <i>TagFile</i> or <i>ADT</i> , where <i>ADT</i> is for backward compatibility with older snapshots.
<b>TYPE</b>	Assay type, which can be cell-hashing, nucleus-hashing, or genetic-pooling.
Genotype	<p>Google bucket url to the reference genotypes in <i>vcf.gz</i> format. This column is <b>required</b> in the following cases:</p> <ul style="list-style-type: none"> <li>Run genetic-pooling assay with <i>souporcell</i> algorithm (i.e. <i>TYPE</i> is genetic-pooling, <i>demultiplexing_algorithm</i> input is <i>souporcell</i>): <ul style="list-style-type: none"> <li>Run with reference genotypes, i.e. <i>souporcell_de_novo_mode</i> is false.</li> <li>Run in <i>de novo</i> mode (i.e. <i>souporcell_de_novo_mode</i> is true), but need to match the resulting cluster names by information from reference genotypes (see description of <i>souporcell_rename_donors</i> input below).</li> </ul> </li> <li>Run genetic-pooling assay with <i>popscle</i> algorithm (i.e. <i>TYPE</i> is genetic-pooling, <i>demultiplexing_algorithm</i> input is <i>popscle</i>): <ul style="list-style-type: none"> <li><i>popscle_num_samples</i> input is 0. In this case, <i>demuxlet</i> will be run with reference genotypes.</li> <li><i>popscle_num_samples</i> input is larger than 0. In this case, reference genotypes will be only used to generate pileups, then <i>freemuxlet</i> will be used for demultiplexing <b>without</b> reference genotypes.</li> </ul> </li> </ul>

Example:

```
OUTNAME, RNA, TagFile, TYPE, Genotype
sample_1, gs://exp/data_1/raw_gene_bc_matrices_h5.h5, gs://exp/data_1/sample_1_
↪ADT.csv, cell-hashing
sample_2, gs://exp/data_2/raw_gene_bc_matrices_h5.h5, gs://exp/data_2/sample_2_
↪ADT.csv, nucleus-hashing
sample_3, gs://exp/data_3/raw_gene_bc_matrices_h5.h5, gs://exp/data_3/
↪possorted_genome_bam.bam, genetic-pooling
sample_4, gs://exp/data_4/raw_gene_bc_matrices_h5.h5, gs://exp/data_4/
↪possorted_genome_bam.bam, genetic-pooling, gs://exp/variants/ref_genotypes.
↪vcf.gz
```

### 3.2 Upload your sample sheet to the workspace bucket:

Use *gsutil* (you already have it if you've installed *gcloud* CLI) in your unix terminal to upload your sample sheet to workspace bucket.

Example:

```
gsutil cp /foo/bar/projects/sample_sheet_demux.csv gs://fc-e0000000-0000-
↪0000-0000-000000000000/
```

## Workflow inputs

Below are inputs for *demultiplexing* workflow. We'll first introduce global inputs, and then inputs for each of the demultiplexing tools. Notice that required inputs are in bold.



## global inputs

Name	Description	Example	Default
<b>input_sample_sheet</b>	Input CSV file describing metadata of RNA and hashtag data pairing.	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet_demux.csv”	
<b>output_directory</b>	This is the output directory (gs url + path) for all results. There will be one folder per RNA-hashtag data pair under this directory.	“gs://fc-e0000000-0000-0000-0000-000000000000/demux_output”	
genome	Reference genome name. Its usage depends on the assay type: <ul style="list-style-type: none"> <li>For <i>cell-hashing</i> or <i>nucleus-hashing</i>, only write this name as an annotation into the resulting count matrix file.</li> <li>For <i>genetic-pooling</i>, if <i>demultiplexing_algorithm</i> input is <code>souporcell</code>, you should choose one name from this <a href="#">genome reference</a> list.</li> <li>For <i>genetic-pooling</i>, if <i>demultiplexing_algorithm</i> input is <code>popsicle</code>, reference genome name is not needed.</li> </ul>	“GRCh38”	
demultiplexing_algorithm	Demultiplexing algorithm to use for <i>genetic-pooling</i> data. Options: <ul style="list-style-type: none"> <li>“souporcell”: Use <code>souporcell</code>, a reference-genotypes-free algorithm for demultiplexing droplet scRNA-Seq data.</li> <li>“popsicle”: Use <code>popsicle</code>, a canonical algorithm for demultiplexing droplet scRNA-Seq data, including <i>demuxlet</i> (with reference genotypes) and <i>freemuxlet</i> (reference-genotype-free) components.</li> </ul>	“souporcell”	“souporcell”
min_num_genes	Only demultiplex cells/nuclei with at least <min_num_genes> expressed genes	100	100
zones	Google cloud zones to consider for execution.	“us-east1-d us-west1-a us-west1-b”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
docker_registry	Docker registry to use. <ul style="list-style-type: none"> <li>“quay.io/cumulus” for images on Red Hat registry;</li> <li>“cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
config_version	Version of config docker image to use. This docker is used for parsing the input sample sheet for downstream execution. Available options: 0.2, 0.1.	0.2	0.2
backend	Cloud infrastructure backend to use. Available options:	“gcp”	“gcp”

**demuxEM inputs**

Name	Description	Example	Default
demuxEM_alpha	demuxEM parameter. The Dirichlet prior concentration parameter (alpha) on samples. An alpha value < 1.0 will make the prior sparse.	0.0	0.0
demuxEM_min_num_umis	demuxEM parameter. Only demultiplex cells/nuclei with at least <demuxEM_min_num_umis> of UMIs.	100	100
demuxEM_min_signal_hashtag	demuxEM parameter. Any cell/nucleus with less than <demuxEM_min_signal_hashtag> hashtags from the signal will be marked as unknown.	10.0	10.0
demuxEM_random_seed	demuxEM parameter. The random seed used in the KMeans algorithm to separate empty ADT droplets from others.	0	0
demuxEM_generate_plots	demuxEM parameter. If generate a series of diagnostic plots, including the background/signal between HTO counts, estimated background probabilities, HTO distributions of cells and non-cells, etc.	true	true
demuxEM_generate_gender_plot	demuxEM parameter. If generate violin plots using gender-specific genes (e.g. Xist). <demuxEM_generate_gender_plot> is a comma-separated list of gene names	"XIST"	
demuxEM_version	demuxEM version to use. Choose from "0.1.7", "0.1.6" and "0.1.5".	"0.1.7"	"0.1.7"
demuxEM_num_cpus	demuxEM parameter. Number of CPUs to request for demuxEM per pair.	8	8
demuxEM_memory	demuxEM parameter. Memory size string for demuxEM per pair.	"10G"	"10G"
demuxEM_disk_space	demuxEM parameter. Disk space (integer) in GB needed for demuxEM per pair.	20	20

## souporcell inputs

Name	Description	Example	Default
souporcell_version	souporcell version to use. Available versions: <ul style="list-style-type: none"> <li>2021.03: Based on commitment <a href="#">1bd9f1</a> on 2021/03/07.</li> <li>2020.07: Based on commitment <a href="#">0d09fb</a> on 2020/07/27.</li> <li>2020.03: Based on commitment <a href="#">eeddcd</a> on 2020/03/31.</li> </ul>	“2021.03”	“2021.03”
souporcell_num_clusters	souporcell parameter. Number of expected clusters when doing clustering. <b>This needs to be set when running souporcell.</b>	8	1
souporcell_de_novo_mode	souporcell parameter. <ul style="list-style-type: none"> <li>If <code>true</code>, run souporcell in de novo mode without reference genotypes: <ul style="list-style-type: none"> <li>If input <code>souporcell_common_variants</code> is further provided, use this common variants list instead of calling SNPs de novo.</li> <li>If a reference genotype vcf file is provided in the sample sheet, use it <b>only</b> for matching the cluster labels computed by souporcell.</li> </ul> </li> <li>If <code>false</code>, run souporcell with <code>--known-genotypes</code> option using the reference genotype vcf file specified in sample sheet.</li> </ul>	true	true
souporcell_num_clusters	souporcell parameter. Number of expected clusters when doing clustering. <b>This needs to be set when running souporcell.</b>	8	1
souporcell_common_variants	souporcell parameter. Users can provide a common variants list in VCF format for Souporcell to use, instead of calling SNPs de novo. <b>Notice:</b> This input is enabled only when <code>souporcell_de_novo_mode</code> is <code>false</code> .	“1000genome.common.variants.vcf.gz”	
souporcell_skip_remap	souporcell parameter. Skip remap step. Only recommended in non denovo mode or common variants are provided.	true	false
souporcell_rename_clusters	souporcell parameter. A comma-separated list of donor names for matching clusters achieved by souporcell. Must be consistent with <code>souporcell_num_clusters</code> input. <ul style="list-style-type: none"> <li>If this input is empty, use cluster labels from the reference genotype vcf file if provided in the sample sheet; if this vcf file is not provided, simply name clusters as <i>Donor1</i>, <i>Donor2</i>, ...</li> </ul>	“CB1,CB2,CB3,CB4”	
82	<ul style="list-style-type: none"> <li>If this input is not empty, and a reference genotype vcf file is provided in the sample sheet, first match the cluster labels using those from this vcf file, then rename to donor names specified in this input.</li> </ul>	Chapter 1. Release Highlights in Current Stable	

## Popsicle inputs

Name	Description	Example	Default
popsicle_num_samples	popsicle parameter. Number of samples to be multiplexed together: <ul style="list-style-type: none"> <li>If 0, run with <i>demuxlet</i> using reference genotypes.</li> <li>Otherwise, run with <i>freemuxlet</i> in de novo mode without reference genotypes.</li> </ul>	4	0
popsicle_min_MQ	popsicle parameter. Minimum mapping quality to consider (lower MQ will be ignored).	20	20
popsicle_min_TD	popsicle parameter. Minimum distance to the tail (lower will be ignored).	0	0
popsicle_tag_group	popsicle parameter. Tag representing readgroup or cell barcodes, in the case to partition the BAM file into multiple groups. For 10x genomics, use CB.	“CB”	“CB”
popsicle_tag_umi	popsicle parameter. Tag representing UMIs. For 10x genomics, use UB.	“UB”	“UB”
popsicle_field	popsicle parameter. FORMAT field to extract from: genotype (GT), genotype likelihood (GL), or posterior probability (GP).	“GT”	“GT”
popsicle_alpha	popsicle parameter. Grid of alpha to search for, in a comma separated list format of all alpha values to be considered.	“0.1,0.2,0.3,0.4,0.5”	“0.1,0.2,0.3,0.4,0.5”
popsicle_rename_donors	popsicle parameter. A comma-separated list of donor names for renaming clusters achieved by popsicle. Must be consistent with <i>popsicle_num_samples</i> input. By default, the resulting donors are <i>Donor1</i> , <i>Donor2</i> , ...	“CB1,CB2,CB3,CB4”	
popsicle_version	popsicle parameter. popsicle version to use. Available options: <ul style="list-style-type: none"> <li>2021.05: Based on commitment <a href="#">da70fc7</a> on 2021/05/05.</li> <li>0.1b: Based on version <a href="#">0.1-beta</a> released on 2019/10/03.</li> </ul>	“2021.05”	“2021.05”
popsicle_num_cpu	popsicle parameter. Number of CPU used by popsicle per pair.	1	1
popsicle_memory	popsicle parameter. Memory size string per pair.	“120G”	“120G”
popsicle_extra_disk_space	popsicle parameter. Extra disk space size (integer) in GB needed for popsicle per pair, besides the disk size required to hold input files specified in the sample sheet.	100	100

## Workflow outputs

See the table below for *demultiplexing* workflow outputs.

Name	Type	Description
output_folders	Array[String]	A list of Google Bucket URLs of the output folders. Each folder is associated with one RNA-hashtag pair in the given sample sheet.
output_zarr_files	Array[File]	A list of demultiplexed RNA count matrices in zarr format. Each zarr file is associated with one RNA-hashtag pair in the given sample sheet. Please refer to section <a href="#">load demultiplexing results into Python and R</a> for its structure.

In the output subfolder of each cell-hashing/nuclei-hashing RNA-hashtag data pair, you can find the following files:



Name	Description
output_name_demux.zarr.zip	Demultiplexed RNA raw count matrix in zarr format. Please refer to section <a href="#">load demultiplexing results into Python and R</a> for its structure.
output_name.out.demuxEM.zarr.zip	<p>This file contains intermediate results for both RNA and hashing count matrices.</p> <p>To load this file into Python, you need to first install <a href="#">Pegasusio</a> on your local machine. Then use <code>import pegasusio as io; data = io.read_input("output_name.out.demuxEM.zarr.zip")</code> in Python environment.</p> <p>It contains 2 UnimodalData objects: one with key name suffix <code>-hashing</code> is the hashtag count matrix, the other one with key name suffix <code>-rna</code> is the demultiplexed RNA count matrix.</p> <p>To load the hashtag count matrix, type <code>hash_data = data.get_data('&lt;genome&gt;-hashing')</code>, where <code>&lt;genome&gt;</code> is the genome name of the data. The count matrix is <code>hash_data.X</code>; cell barcode attributes are stored in <code>hash_data.obs</code>; sample names are in <code>hash_data.var_names</code>. Moreover, the estimated background probability regarding hashtags is in <code>hash_data.uns['background_probs']</code>.</p> <p>To load the RNA matrix, type <code>rna_data = data.get_data('&lt;genome&gt;-rna')</code>, where <code>&lt;genome&gt;</code> is the genome name of the data. It only contains cells which have estimated sample assignments. The count matrix is <code>rna_data.X</code>. Cell barcode attributes are stored in <code>rna_data.obs</code>:  <code>rna_data.obs['demux_type']</code> stores the estimated droplet types (singlet/doublet/unknown) of cells; <code>rna_data.obs['assignment']</code> stores the estimated hashtag(s) that each cell belongs to. Moreover, for cell-hashing/nucleus-hashing data, you can find estimated sample fractions (sample1, sample2, ..., sampleN, background) for each droplet in <code>rna_data.obsm['raw_probs']</code>.</p>
output_name.ambient_hashtag.hist.pdf	Optional output. A histogram plot depicting hashtag distributions of empty droplets and non-empty droplets.
output_name.background_probabilities.pdf	Optional output. A bar plot visualizing the estimated hashtag background probability distribution.
output_name.real_content.hist.pdf	Optional output. A histogram plot depicting hashtag distributions of not-real-cells and real-cells as defined by total number of expressed genes in the RNA assay.
output_name.rna_demux.hist.pdf	Optional output. A histogram plot depicting RNA UMI distribution for singlets, doublets and unknown cells.
output_name.gene_name.violin.pdf	Optional outputs. Violin plots depicting gender-specific gene expression across samples. We can have multiple plots if a gene list is provided in <code>demuxEM_generate_gender_plot</code> field of <code>cumulus_hashing_cite_seq</code> inputs.

In the output subfolder of each genetic-pooling RNA-hashtag data pair generated by *souporcell*, you can find the following files:

Name	Description
output_name_demux.zarr.zip	Demultiplexed RNA count matrix in zarr format. Please refer to section <a href="#">load demultiplexing results into Python and R</a> for its structure.
clusters.tsv	Inferred droplet type and cluster assignment for each cell barcode.
cluster_genotypes.vcf	Inferred genotypes for each cluster.
match_donors.log	Log of matching donors step, with information of donor matching included.

In the output subfolder of each genetic-pooling RNA-hashtag data pair generated by *demuxlet*, you can find the following files:

Name	Description
output_name_demux.zarr.zip	Demultiplexed RNA count matrix in zarr format. Please refer to section <a href="#">load demultiplexing results into Python and R</a> for its structure.
output_name.best (demuxlet) or output_name.clust1.samples.gz (freemuxlet)	Inferred droplet type and cluster assignment for each cell barcode.

## Load demultiplexing results into Python and R

To load demultiplexed RNA count matrix into Python, you need to install Python package *pegasusio* first. Then follow the codes below:

```
import pegasusio as io
data = io.read_input('output_name_demux.zarr.zip')
```

Once you load the data object, you can find estimated droplet types (singlet/doublet/unknown) in `data.obs['demux_type']`. Notices that there are cell barcodes with no sample associated, and therefore have no droplet type.

You can also find estimated sample assignments in `data.obs['assignment']`.

For cell-hashing/nucleus-hashing data, if one sample name can correspond to multiple feature barcodes, each feature barcode is assigned to a unique sample name, and this deduplicated sample assignment results are in `data.obs['assignment.dedup']`.

To load the results into R, you need to install R package *reticulate* in addition to Python package *pegasusio*. Then follow the codes below:

```
library(reticulate)
ad <- import("pegasusio", convert = FALSE)
data <- ad$read_input("output_name_demux.zarr.zip")
```

Results are in `data$obs['demux_type']`, `data$obs['assignment']`, and similarly as above, for cell-hashing/nucleus-hashing data, you'll find an additional field `data$obs['assignment.dedup']` for deduplicated sample assignment in the case that one sample name can correspond to multiple feature barcodes.

### 1.1.8 Run CellBender for ambient RNA removal

**cellbender** workflow wraps *CellBender* tool for removing technical artifacts from high-throughput single-cell/single-nucleus RNA sequencing data.

This workflow is modified from the official BSD-3-Clause licensed [CellBender WDL workflow](#), with adding the support on scattering over multiple samples simultaneously.

## Prepare input data and import workflow

### 1. Run `cellranger_workflow`

To demultiplex, you'll need raw gene count and hashtag matrices for cell-hashing/nucleus-hashing data, or raw gene count matrices and genome BAM files for genetic-pooling data. You can generate these data by running the `cellranger_workflow`.

Please refer to the [cellranger\\_workflow tutorial](#) for details.

When finished, you should be able to find the raw gene count matrix (e.g. `raw_feature_bc_matrix.h5` or `raw_gene_bc_matrices_h5.h5`) for each sample.

### 2. Import `cellbender`

Import `cellbender` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus/CellBender](#) to import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export `cellbender` workflow in the drop-down menu.

### 3. Prepare a sample sheet

Create a TSV-format sample sheet (say `cellbender_sheet.tsv`), which describes the metadata for each scRNA-Seq sample. Notice that the first column specifies sample names, and the second column specifies the Cloud URI of the **raw** gene-count matrices generated in Step 1.

An example sample sheet is the follow:

sample_1	gs://exp/data_1/raw_feature_bc_matrix.h5
sample_2	gs://exp/data_2/raw_feature_bc_matrix.h5
sample_3	gs://exp/data_3/raw_feature_bc_matrix.h5

Then upload your sample sheet to your Terra workspace's bucket using `gsutil`. For example:

```
gsutil cp /foo/bar/projects/cellbender_sheet.tsv gs://fc-e0000000-0000-0000-
↪0000-000000000000/my-project/
```

## Workflow inputs

Below are inputs for `CellBender` workflow. Notice that required inputs are **in bold**:

Name	Description	Example	Default
<b>input_tsv_file</b>	Input TSV file describing metadata of scRNA-Seq samples.	“gs://fc-e0000000-0000-0000-0000-000000000000/my-project/cellbender_sheet.tsv”	
<b>output_directory</b>	This is the output directory URI for all results. There will be one subfolder per sample under this directory.	“gs://fc-e0000000-0000-0000-0000-000000000000/my-project/cellbender_output”	
<b>expected_cells</b>	Number of cells expected in the dataset (a rough estimate within a factor of 2 is sufficient).	2048	None
<b>total_droplets</b>	The number of droplets from the rank-ordered UMI plot that will be analyzed. The largest <i>total_droplets_included</i> droplets will have their cell probabilities inferred as an output.	25000	25000
<b>model</b>	Which model is being used for count data: <ul style="list-style-type: none"> <li>“simple” does not model either ambient RNA or random barcode swapping (for debugging purposes – not recommended).</li> <li>“ambient” assumes background RNA is incorporated into droplets.</li> <li>“swapping” assumes background RNA comes from random barcode swapping.</li> <li>“full” uses a combined ambient and swapping model.</li> </ul>	“full”	“full”
<b>low_count_threshold</b>	Droplets with UMI counts below this number are completely excluded from the analysis. This can help identify the correct prior for empty droplet counts in the rare case where empty counts are extremely high (over 200).	15	15
<b>fpr</b>	Target false positive rate in (0, 1). A false positive is a true signal count that is erroneously removed. More background removal is accompanied by more signal removal at high values of FPR. You can specify multiple values by giving a space-separated string, which will create multiple output files.	“0.01 0.05 0.1”	“0.01”
<b>epochs</b>	Number of epochs to train.	150	150
<b>z_dim</b>	Dimension of latent variable $z$ .	100	100
<b>z_layers</b>	Dimension of hidden layers in the encoder for $z$ . For multiple layers, specify them in space-separated string format.	“500 100 300”	“500”
<b>empty_drop_training_fraction</b>	the fraction of the training data each epoch that is drawn (randomly sampled) from surely empty droplets.	0.5	0.5
<b>blacklist_genes</b>	Integer indices of genes to ignore entirely. In the output count matrix, the counts for these genes will be set to zero. For multiple genes, specify them in space-separated string format.	“0 1 2”	“”
<b>learning_rate</b>	Training detail: lower learning rate for inference. A OneCycle learning rate schedule is used, where the upper learning rate is ten times this value. (For this value, probably do not exceed 1e-3).	1e-4	1e-4
<b>exclude_antibody_features</b>	If this flag will cause remove-background to operate on gene counts only, ignoring other features.	false	false
<b>docker_registry</b>	Docker registry to use. <ul style="list-style-type: none"> <li>“quay.io/cumulus” for images on Red Hat registry;</li> <li>“cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”

## Workflow outputs

See the table below for *cellbender* workflow outputs:

Name	Type	Description
cellbender_outputs	Array[String]	A list of Cloud URIs of the output folders. Each folder is associated with one scRNA-seq sample in the given sample sheet.

## 1.1.9 Run Cumulus for sc/snRNA-Seq data analysis

### Run Cumulus analysis

#### Prepare Input Data

#### Case One: Sample Sheet

Follow the steps below to run **cumulus** on [Terra](#).

1. Create a sample sheet, **count\_matrix.csv**, which describes the metadata for each sample count matrix. The sample sheet should at least contain 2 columns — *Sample* and *Location*. *Sample* refers to sample names and *Location* refers to the location of the channel-specific count matrix in either of
  - 10x format with v2 chemistry. For example, `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/sample_1/raw_gene_bc_matrices_h5.h5`.
  - 10x format with v3 chemistry. For example, `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/sample_1/raw_feature_bc_matrices.h5`.
  - Drop-seq format. For example, `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/sample_2/sample_2.umi.dge.txt.gz`.
  - Matrix Market format (mtx). If the input is mtx format, location should point to a mtx file with a file suffix of either ‘.mtx’ or ‘.mtx.gz’. In addition, the associated barcode and gene tsv/txt files should be located in the same folder as the mtx file. For example, if we generate mtx file using BUSTools, we set *Location* to `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/mm10/cells_x_genes.mtx`. We expect to see `cells_x_genes.barcodes.txt` and `cells_x_genes.genes.txt` under folder `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/mm10/`. We support loading mtx files in HCA DCP, 10x Genomics V2/V3, SCUMI, dropEST and BUSTools format. Users can also set *Location* to a folder `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/hg19_and_mm10/`. This folder must be generated by Cell Ranger for multi-species samples and we expect one subfolder per species (e.g. ‘hg19’) and each subfolder should contain mtx file, barcode file, gene name file as generated by Cell Ranger.
  - csv format. If it is HCA DCP csv format, we expect the expression file has the name of `expression.csv`. In addition, we expect that `cells.csv` and `genes.csv` files are located under the same folder as the `expression.csv`. For example, `gs://fc-e0000000-0000-0000-0000-000000000000/my_dir/sample_3/`.
  - tsv or loom format.

An optional Reference column can be used to select samples generated from a same reference (e.g. mm10). If the count matrix is in either DGE, mtx, csv, tsv, or loom format, the value in this column will be used as the reference since the count matrix file does not contain reference name information. The only exception is mtx format. If users do not provide a Reference column, we will use the basename of the folder containing the mtx file as its reference. In addition, the Reference column can be used to aggregate count matrices generated from different genome versions or gene annotations together under a unified reference. For example, if we have one matrix generated from mm9 and the other one generated from mm10, we can write mm9\_10 for these two matrices in their Reference column. Pegasus will change their references to mm9\_10 and use the union of gene symbols from the two matrices as the gene symbols of the aggregated matrix. For HDF5 files (e.g. 10x v2/v3), the reference name contained in the file does not need to match the value in this column. In fact, we use this column to rename references in HDF5 files. For example, if we have two HDF files, one generated from mm9 and the other generated from mm10. We can set these two files' Reference column value to mm9\_10, which will rename their reference names into mm9\_10 and the aggregated matrix will contain all genes from either mm9 or mm10. This renaming feature does not work if one HDF5 file contain multiple references (e.g. mm10 and GRCh38).

The sample sheet can optionally contain two columns - nUMI and nGene. These two columns define minimum number of UMIs and genes for cell selection for each sample in the sample sheet. nGene column overwrites minimum\_number\_of\_genes parameter.

You are free to add any other columns and these columns will be used in selecting channels for further analysis. In the example below, we have *Source*, which refers to the tissue of origin, *Platform*, which refers to the sequencing platform, *Donor*, which refers to the donor ID, and *Reference*, which refers to the reference genome.

Example:

```
Sample,Source,Platform,Donor,Reference,Location
sample_1,bone_marrow,NextSeq,1,GRCh38,gs://fc-e0000000-0000-0000-0000-
↪000000000000/my_dir/sample_1/raw_gene_bc_matrices_h5.h5
sample_2,bone_marrow,NextSeq,2,GRCh38,gs://fc-e0000000-0000-0000-0000-
↪000000000000/my_dir/sample_2/raw_gene_bc_matrices_h5.h5
sample_3,pbmc,NextSeq,1,GRCh38,gs://fc-e0000000-0000-0000-0000-000000000000/
↪my_dir/sample_3/raw_feature_bc_matrices.h5
sample_4,pbmc,NextSeq,2,GRCh38,gs://fc-e0000000-0000-0000-0000-000000000000/
↪my_dir/sample_4/raw_feature_bc_matrices.h5
```

If you ran **cellranger\_workflow** previously, you should already have a template **count\_matrix.csv** file that you can modify from **generate\_count\_config**'s outputs.

1. Upload your sample sheet to the workspace.

Example:

```
gsutil cp /foo/bar/projects/my_count_matrix.csv gs://fc-e0000000-0000-
↪0000-0000-000000000000/
```

where `/foo/bar/projects/my_count_matrix.csv` is the path to your sample sheet in local machine, and `gs://fc-e0000000-0000-0000-0000-000000000000/` is the location on Google bucket to hold it.

2. Import *cumulus* workflow to your workspace.

Import by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus](https://github.com/lilab-bcb/cumulus) for import.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export *cumulus* workflow in the drop-down menu.

3. In your workspace, open *cumulus* in WORKFLOWS tab. Select Run workflow with inputs defined by file paths as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click the **SAVE** button.

### Case Two: Single File

Alternatively, if you only have one single count matrix for analysis, you can go without sample sheets. **Cumulus** currently supports the following formats:

- 10x genomics v2/v3 format (hdf5);
- Drop-seq dge format;
- csv (no HCA DCP format), tsv or loom formats.

Simply upload your data to the Google Bucket of your workspace, and specify its URL in `input_file` field of Cumulus' global inputs (see below). For hdf5 files, there is no need to specify genome names. For other formats, you can specify genome name in `considered_refs` field in cluster inputs; otherwise, default name ' ' will be used.

In this case, the **aggregate\_matrices** step will be skipped.

### Case Three: Multiple samples without aggregation

Sometimes, you may want to run Cumulus on multiple samples simultaneously. This is different from Case one, because samples are analyzed separately without aggregation.

1. To do it, you need to first [create a data table](#) on Terra. An example TSV file is the following:

```
entity:cumulus_test_id input_h5
5k_pbmc_v3 gs://fc-e0000000-0000-0000-0000-000000000000/5k_pbmc_v3/raw_feature_
↪bc_matrix.h5
1k_pbmc_v3 gs://fc-e0000000-0000-0000-0000-000000000000/1k_pbmc_v3/raw_feature_
↪bc_matrix.h5
```

You are free to add more columns, but sample ids and URLs to RNA count matrix files are required. I'll use this example TSV file for the rest of steps in this case.

1. Upload your TSV file to your workspace. Open the **DATA** tab on your workspace. Then click the upload button on left **TABLE** panel, and select the TSV file above. When uploading is done, you'll see a new data table with name "cumulus\_test":

	cumulus_test_id ↓	input_h5
<input type="checkbox"/>	1k_pbmc_v3	<a href="#">raw_feature_bc_matrix.h5</a>
<input type="checkbox"/>	5k_pbmc_v3	<a href="#">raw_feature_bc_matrix.h5</a>

- Import *cumulus* workflow to your workspace as in Case one. Then open *cumulus* in WORKFLOW tab. Select Run workflow(s) with inputs defined by data table, and choose *cumulus\_test* from the drop-down menu.

- ☐ Run workflow with inputs defined by file paths  
☒ Run workflow(s) with inputs defined by data table

### Step 1

Select root entity type: cumulus\_test

- In the input field, specify:

- input\_file:** Type `this.input_h5`, where `this` refers to the data table selected, and `input_h5` is the column name in this data table for RNA count matrices.
- output\_directory:** Type Google bucket URL for the main output folder. For example, `gs://fc-e0000000-0000-0000-0000-000000000000/cumulus_results`.
- output\_name:** Type `this.cumulus_test_id`, where `cumulus_test_id` is the column name in data table for sample ids.

An example is in the screen shot below:

Task name	Variable	Type	Attribute
cumulus	input_file	File	this.input_h5
cumulus	output_directory	String	"gs://fc-e0000000-0000-0000-0000-000000000000/cumulus_results"
cumulus	output_name	String	this.cumulus_test_id

Then finish setting up other inputs following the description in sections below. When you are done, click **SAVE**, and then **RUN ANALYSIS**.

When all the jobs are done, you'll find output for the 2 samples in subfolders `gs://fc-e0000000-0000-0000-0000-000000000000/cumulus_results/5k_pbmc_v3` and `gs://fc-e0000000-0000-0000-0000-000000000000/cumulus_results/1k_pbmc_v3`, respectively.

## Cumulus steps:

**Cumulus** processes single cell data in the following steps:



1. **aggregate\_matrices** (optional). When given a CSV format sample sheet, this step aggregates channel-specific count matrices into one big count matrix. Users can specify which channels they want to analyze and which sample attributes they want to import to the count matrix in this step. Otherwise, if a single count matrix file is given, skip this step.
2. **cluster**. This is the main analysis step. In this step, **Cumulus** performs low quality cell filtration, highly variable gene selection, batch correction, dimension reduction, diffusion map calculation, graph-based clustering and 2D visualization calculation (e.g. t-SNE/UMAP/FLE).
3. **de\_analysis**. This step is optional. In this step, **Cumulus** can calculate potential markers for each cluster by performing a variety of differential expression (DE) analysis. The available DE tests include Welch's t test, Fisher's exact test, and Mann-Whitney U test. **Cumulus** can also calculate the area under ROC (AUROC) curve values for putative markers. If `find_markers_lightgbm` is on, **Cumulus** will try to identify cluster-specific markers by training a LightGBM classifier. If the samples are human or mouse immune cells, **Cumulus** can also optionally annotate putative cell types for each cluster based on known markers.
4. **plot**. This step is optional. In this step, **Cumulus** can generate 6 types of figures based on the **cluster** step results:
  - **composition** plots which are bar plots showing the cell compositions (from different conditions) for each cluster. This type of plots is useful to fast assess library quality and batch effects.
  - **umap** and **net\_umap**: UMAP like plots based on different algorithms, respectively. Users can specify cell attributes (e.g. cluster labels, conditions) for coloring side-by-side.
  - **tsne**: FIt-SNE plots. Users can specify cell attributes (e.g. cluster labels, conditions) for coloring side-by-side.
  - **fle** and **net\_fle**: FLE (Force-directed Layout Embedding) like plots based on different algorithms, respectively. Users can specify cell attributes (e.g. cluster labels, conditions) for coloring side-by-side.
  - If input is CITE-Seq data, there will be **citeseq\_umap** plots which are UMAP plots based on epitope expression.
5. **cirro\_output**. This step is optional. Generate [CirroCumulus](#) inputs for visualization using [CirroCumulus](#).
6. **scp\_output**. This step is optional. Generate analysis result in [Single Cell Portal](#) (SCP) compatible format.

In the following sections, we will first introduce global inputs and then introduce the WDL inputs and outputs for each step separately. But please note that you need to set inputs from all steps simultaneously in the Terra WDL.

Note that we will make the required inputs/outputs bold and all other inputs/outputs are optional.

---

## global inputs

Name	Description	Example	Default
<b>input_file</b>	Input CSV sample sheet describing metadata of each 10x channel, or a single input count matrix file	“gs://fc-e0000000-0000-0000-0000-000000000000/my_count_matrix.csv”	
<b>output_directory</b>	Google bucket URL of the output directory.	“gs://fc-e0000000-0000-0000-0000-000000000000/my_results_dir”	
<b>output_name</b>	This is the name of subdirectory for the current sample; and all output files within the subdirectory will have this string as the common filename prefix.	“my_sample”	
<b>default_reference</b>	If sample count matrix is in either DGE, mtx, csv, tsv or loom format and there is no Reference column in the csv_file, use default_reference as the reference string.	“GRCh38”	
<b>pegasus_version</b>	Pegasus version to use for analysis. Versions available: 1.4.3, 1.4.2, 1.4.0, 1.3.0.	“1.4.3”	“1.4.3”
<b>docker_registry</b>	Docker registry to use. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>zones</b>	Google cloud zones to consider for execution.	“us-east1-d us-west1-a us-west1-b”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
<b>num_cpu</b>	Number of CPUs per Cumulus job	32	64
<b>memory</b>	Memory size string	“200G”	“200G”
<b>disk_space</b>	Total disk space in GB	100	100
<b>backend</b>	Cloud infrastructure backend to use. Available options: <ul style="list-style-type: none"> <li>• “gcp” for Google Cloud;</li> <li>• “aws” for Amazon AWS;</li> <li>• “local” for local machine.</li> </ul>	“gcp”	“gcp”
<b>preemptible</b>	Number of preemptible tries. This works only when <i>backend</i> is <i>gcp</i> .	2	2
<b>awsQueueArn</b>	The AWS ARN string of the job queue to be used. This only works for <i>aws</i> backend.	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwl”	“”

## aggregate\_matrices

### aggregate\_matrices inputs

Name	Description	Example	Default
restrictions	Select channels that satisfy all restrictions. Each restriction takes the format of name:value,...,value. Multiple restrictions are separated by ‘;’	“Source:bone_marrow;Platform:NextSeq”	
attributes	Specify a comma-separated list of outputted attributes. These attributes should be column names in the count_matrix.csv file	“Source,Platform,Donor”	
select_only_singlets	If we have demultiplexed data, turning on this option will make cumulus only include barcodes that are predicted as singlets.	true	false
remap_singlets	<p>For demultiplexed data, user can remap singlet names using assignment in String in this input. This string assignment takes the format “new_name_i:old_name_1,old_name_2;new_name_ii:old_name_3;...”.</p> <p>For example, if we hashed 5 libraries from 3 samples: sample1_lib1, sample1_lib2; sample2_lib1, sample2_lib2; sample3, we can remap them to 3 samples using this string: "sample1:sample1_lib1,sample1_lib2;sample2:sample2_lib1,sample2_lib2".</p> <p>In this way, the new singlet names will be in metadata field with key assignment, while the old names are kept in metadata with key assignment.orig.</p> <p><b>Notice:</b> This input is enabled only when <i>select_only_singlets</i> input is true.</p>	“Group1:CB1,CB2;Group2:CB3,CB4,CB5”	
subset_singlets	<p>For demultiplexed data, user can use this input to choose a subset of singlets based on their names. This string takes the format “name1,name2,...”.</p> <p>Note that if <i>remap_singlets</i> input is specified, subsetting happens after remapping, i.e. you should use the new singlet names for choosing subset.</p> <p><b>Notice:</b> This input is enabled only when <i>select_only_singlets</i> input is true.</p>	“Group2,CB6,CB7”	
minimum_number_of_genes	Only keep barcodes with at least this number of expressed genes	100	100
is_dropseq	If inputs are DropSeq data.	false	false

**aggregate\_matrices output**

Name	Type	Description
output_aggr_zarr	File	Aggregated count matrix in Zarr format

**cluster****cluster inputs**

Name	Description	Example	Default
focus	Focus analysis on Unimodal data with <keys>. <keys> is a comma-separated list of keys. If None, the <code>self._selected</code> will be the focused one. Focus key consists of two parts: reference genome name, and data type, connected with a hyphen marker “-”. Reference genome name depends on the reference you used when running Cellranger workflow. See details in <a href="#">reference list</a> .	“GRCh38-rna”	
append	Append Unimodal data <key> to any <keys> in <i>focus</i> . Similarly as focus keys, append key also consists of two parts: reference genome name, and data type, connected with a hyphen marker “-”. See <a href="#">reference list</a> for details.	“SARSCoV2-rna”	
channel	Specify the cell barcode attribute to represent different samples.	“Donor”	
black_list	Cell barcode attributes in black list will be popped out. Format is “attr1,attr2,...,attrn”.	“attr1,attr2,attr3”	
min_genes_before_filtration	If a filtered matrix is input, empty barcodes will dominate pre-filtration statistics. To avoid this, for raw data matrix, only consider barcodes with at least <min_genes_before_filtration> genes for pre-filtration condition.	100	100
select_only_singlets	If we have demultiplexed data, turning on this option will make cumulus only include barcodes that are predicted as singlets	false	false

Continued on next page

Table 3 – continued from previous page

Name	Description	Example	Default
remap_singlets	<p>For demultiplexed data, user can remap singlet names using assignment in String in this input. This string assignment takes the format “new_name_i:old_name_1,old_name_2;new_name_ii:old_name_3;...”.</p> <p>For example, if we hashed 5 libraries from 3 samples: sample1_lib1, sample1_lib2; sample2_lib1, sample2_lib2; sample3, we can remap them to 3 samples using this string: "sample1:sample1_lib1,sample1_lib2;sample2:sample2_lib1,sample2_lib2".</p> <p>In this way, the new singlet names will be in metadata field with key assignment, while the old names are kept in metadata with key assignment.orig.</p> <p><b>Notice:</b> This input is enabled only when <i>select_only_singlets</i> input is true.</p>	“Group1:CB1,CB2;Group2:CB3,CB4,CB5”	
subset_singlets	<p>For demultiplexed data, user can use this input to choose a subset of singlets based on their names. This string takes the format “name1,name2,...”.</p> <p>Note that if <i>remap_singlets</i> is specified, subsetting happens after remapping, i.e. you should use the new singlet names for choosing subset.</p> <p><b>Notice:</b> This input is enabled only when <i>select_only_singlets</i> input is true.</p>	“Group2,CB6,CB7”	
output_filtration_results	If results of cell and gene filtration results to a spreadsheet	true	true
plot_filtration_results	If plots of filtration results as PDF files	true	true
plot_filtration_figsize	Figure size for filtration plots. <figsize> is a comma-separated list of two numbers, the width and height of the figure (e.g. 6,4)	6,4	
output_h5ad	Generate Seurat-compatible h5ad file. Must set to true if performing DE analysis, cell type annotation, or plotting.	true	true
output_loom	If generate loom-formatted file	false	false
min_genes	Only keep cells with at least <min_genes> of genes	500	500
max_genes	Only keep cells with less than <max_genes> of genes	6000	6000
min_umis	Only keep cells with at least <min_umis> of UMIs. By default, don’t filter cells due to UMI lower bound.	100	
max_umis	Only keep cells with less than <max_umis> of UMIs. By default, don’t filter cells due to UMI upper bound.	600000	

Continued on next page

Table 3 – continued from previous page

Name	Description	Example	Default
mito_prefix	Prefix of mitochondrial gene names. This is to identify mitochondrial genes.	“mt-“	“MT-” for <i>GRCh38</i> reference genome data; “mt-” for <i>mm10</i> reference genome data; for other reference genome data, must specify this prefix manually.
percent_mito	Only keep cells with mitochondrial ratio less than <percent_mito>% of total counts	50	20.0
gene_percent_cells	Only use genes that are expressed in at <gene_percent_cells>% of cells to select variable genes	50	0.05
counts_per_cell	Total counts per cell after normalization, before transforming the count matrix into Log space.	1e5	1e5
select_hvf_flavor	Highly variable feature selection method. Options: <ul style="list-style-type: none"> <li>• “pegasus”: New selection method proposed in Pegasus, the analysis module of Cumulus workflow.</li> <li>• “Seurat”: Conventional selection method used by Seurat and SCANPY.</li> </ul>	“pegasus”	“pegasus”
select_hvf_ngenes	Select top <select_hvf_ngenes> highly variable features. If <select_hvf_flavor> is “Seurat” and <select_hvf_ngenes> is “None”, select HVGs with z-score cutoff at 0.5.	2000	2000
no_select_hvf	Do not select highly variable features.	false	false
plot_hvf	Plot highly variable feature selection. Will not work if no_select_hvf is true.	false	false
correct_batch_effects	Correct batch effects	false	false
correction_method	Batch correction method. Options: <ul style="list-style-type: none"> <li>• “harmony”: Harmony algorithm (Korsunsky et al. Nature Methods 2019).</li> <li>• “L/S”: Location/Scale adjustment algorithm (Li and Wong. The analysis of Gene Expression Data, 2003).</li> <li>• “scanorama”: Scanorama algorithm (Hie et al. Nature Biotechnology 2019).</li> </ul>	“harmony”	“harmony”

Continued on next page

Table 3 – continued from previous page

Name	Description	Example	Default
batch_group_by	<p>Batch correction assumes the differences in gene expression between channels are due to batch effects. However, in many cases, we know that channels can be partitioned into several groups and each group is biologically different from others.</p> <p>In this case, we will only perform batch correction for channels within each group. This option defines the groups.</p> <p>If &lt;expression&gt; is None, we assume all channels are from one group. Otherwise, groups are defined according to &lt;expression&gt;.</p> <p>&lt;expression&gt; takes the form of either 'attr', or 'attr1+attr2+...+attrn', or 'attr=value11,...,value1n_1;value21,...,value2n_2;...;valuem1,...,valuemn_m'.</p> <p>In the first form, 'attr' should be an existing sample attribute, and groups are defined by 'attr'.</p> <p>In the second form, 'attr1',..., 'attrn' are n existing sample attributes and groups are defined by the Cartesian product of these n attributes.</p> <p>In the last form, there will be m + 1 groups.</p> <p>A cell belongs to group i (i &gt; 0) if and only if its sample attribute 'attr' has a value among valuei1,...,valuein_i.</p> <p>A cell belongs to group 0 if it does not belong to any other groups</p>	"Donor"	None
random_state	Random number generator seed	0	0
calc_signature_genes	<p>Geneset for calculating signature scores. It can be either of the following forms:</p> <ul style="list-style-type: none"> <li>String chosen from: cell_cycle_human, cell_cycle_mouse, gender_human, gender_mouse, mitochondrial_genes_human, mitochondrial_genes_mouse, ribosomal_genes_human, ribosomal_genes_mouse, apoptosis_human, and apoptosis_mouse.</li> <li>Google bucket URL of a GMT format file. For example: gs://fc-e0000000-0000-0000-0000-000000000000/cell_cycle_sig.gmt.</li> </ul>	"cell_cycle_human"	
nPC	Number of principal components	50	50
knn_K	Number of nearest neighbors used for constructing affinity matrix.	50	100

Continued on next page

Table 3 – continued from previous page

Name	Description	Example	Default
knn_full_speed	For the sake of reproducibility, we only run one thread for building kNN indices. Turn on this option will allow multiple threads to be used for index building. However, it will also reduce reproducibility due to the racing between multiple threads.	false	false
run_diffmap	Whether to calculate diffusion map or not. It will be automatically set to <code>true</code> when input <code>run_fle</code> or <code>run_net_fle</code> is set.	false	false
diffmap_ndc	Number of diffusion components	100	100
diffmap_maxt	Maximum time stamp in diffusion map computation to search for the knee point.	5000	5000
run_louvain	Run Louvain clustering algorithm	true	true
louvain_resolution	Resolution parameter for the Louvain clustering algorithm	1.3	1.3
louvain_class_label	Louvain cluster label name in analysis result.	“louvain_labels”	“louvain_labels”
run_leiden	Run Leiden clustering algorithm.	false	false
leiden_resolution	Resolution parameter for the Leiden clustering algorithm.	1.3	1.3
leiden_niter	Number of iterations of running the Leiden algorithm. If negative, run Leiden iteratively until no improvement.	2	-1
leiden_class_label	Leiden cluster label name in analysis result.	“leiden_labels”	“leiden_labels”
run_spectral_louvain	Run Spectral Louvain clustering algorithm	false	false
spectral_louvain_basis	Basis used for KMeans clustering. Use diffusion map by default. If diffusion map is not calculated, use PCA coordinates. Users can also specify “pca” to directly use PCA coordinates.	“diffmap”	“diffmap”
spectral_louvain_resolution	Resolution parameter for louvain.	1.3	1.3
spectral_louvain_class_label	Spectral Louvain label name in analysis result.	“spectral_louvain_labels”	“spectral_louvain_labels”
run_spectral_leiden	Run Spectral Leiden clustering algorithm.	false	false
spectral_leiden_basis	Basis used for KMeans clustering. Use diffusion map by default. If diffusion map is not calculated, use PCA coordinates. Users can also specify “pca” to directly use PCA coordinates.	“diffmap”	“diffmap”
spectral_leiden_resolution	Resolution parameter for leiden.	1.3	1.3
spectral_leiden_class_label	Spectral Leiden label name in analysis result.	“spectral_leiden_labels”	“spectral_leiden_labels”
run_tsne	Run FIt-SNE for visualization.	false	false
tsne_perplexity	SNE’s perplexity parameter.	30	30
tsne_initialization	Initialization method for FIt-SNE. It can be either: ‘random’ refers to random initialization; ‘pca’ refers to PCA initialization as described in [Kobak et al. 2019].	“pca”	“pca”
run_umap	Run UMAP for visualization	true	true
umap_K	K neighbors for UMAP.	15	15
umap_min_dis	UMAP parameter.	0.5	0.5
umap_spread	UMAP parameter.	1.0	1.0
run_fle	Run force-directed layout embedding (FLE) for visualization	false	false
fle_K	Number of neighbors for building graph for FLE	50	50
fle_target_change_per_node	Target change per node to stop FLE.	2.0	2.0
fle_target_steps	Maximum number of iterations before stopping the algorithm	5000	5000

Continued on next page



Table 3 – continued from previous page

Name	Description	Example	Default
net_down_sampling_fraction	Downsampling fraction for net-related visualization	0.1	0.1
run_net_umap	Run Net UMAP for visualization	false	false
net_umap_out_basis	Basis name for Net UMAP coordinates in analysis result	“net_umap”	“net_umap”
run_net_fle	Run Net FLE for visualization	false	false
net_fle_out_basis	Basis name for Net FLE coordinates in analysis result.	“net_fle”	“net_fle”
infer_doublets	Infer doublets using the <a href="#">Pegasus method</a> . When finished, Scrublet-like doublet scores are in cell attribute <code>doublet_score</code> , and “doublet/singlet” assignment on cells are stored in cell attribute <code>demux_type</code> .	false	false
expected_doublet_rate	The expected doublet rate per sample. If not specified, calculate the expected rate based on number of cells from the 10x multipler rate table.	0.05	
doublet_cluster_attribute	Specify which cluster attribute (e.g. “louvain_labels”) should be used for doublet inference. Then doublet clusters will be marked with the following criteria: passing the Fisher’s exact test and having $\geq 50\%$ of cells identified as doublets.  If not specified, the first computed cluster attribute in the list of “leiden”, “louvain”, “spectral_leiden” and “spectral_louvain” will be used.	“louvain_labels”	
citeseq	Perform CITE-Seq data analysis. Set to <code>true</code> if input data contain both RNA and CITE-Seq modalities. This will set <i>focus</i> to be the RNA modality and <i>append</i> to be the CITE-Seq modality. In addition, “ADT-” will be added in front of each antibody name to avoid name conflict with genes in the RNA modality.	false	false
citeseq_umap	For high quality cells kept in the RNA modality, calculate a distinct UMAP embedding based on their antibody expression.	false	false
citeseq_umap_exclude	Comma-separated list of antibodies to be excluded from the CITE-Seq UMAP calculation (e.g. Mouse-IgG1,Mouse-IgG2a).	“Mouse-IgG1,Mouse-IgG2a”	

## cluster outputs

Name	Type	Description
<b>output_zarr</b>	File	<p>Output file in zarr format (output_name.zarr.zip).</p> <p>To load this file in Python, you need to first install <a href="#">PegasusIO</a> on your local machine. Then use <code>import pegasusio as io; data = io.read_input('output_name.zarr.zip')</code> in Python environment.</p> <p><code>data</code> is a <i>MultimodalData</i> object, and points to its default <i>UnimodalData</i> element. You can set its default <i>UnimodalData</i> to others by <code>data.set_data(focus_key)</code> where <code>focus_key</code> is the key string to the wanted <i>UnimodalData</i> element.</p> <p>For its default <i>UnimodalData</i> element, the log-normalized expression matrix is stored in <code>data.X</code> as a Scipy CSR-format sparse matrix, with cell-by-gene shape.</p> <p>Alternatively, to get the raw count matrix, first run <code>data.select_matrix('raw.X')</code>, then <code>data.X</code> will be switched to point to the raw matrix.</p> <p>The <code>obs</code> field contains cell related attributes, including clustering results.</p> <p>For example, <code>data.obs_names</code> records cell barcodes;  <code>data.obs['Channel']</code> records the channel each cell comes from;  <code>data.obs['n_genes']</code>, <code>data.obs['n_counts']</code>, and  <code>data.obs['percent_mito']</code> record the number of expressed genes, total UMI count, and mitochondrial rate for each cell respectively;  <code>data.obs['louvain_labels']</code>,  <code>data.obs['leiden_labels']</code>,  <code>data.obs['spectral_louvain_labels']</code>, and  <code>data.obs['spectral_leiden_labels']</code> record each cell's cluster labels using different clustering algorithms;</p> <p>The <code>var</code> field contains gene related attributes.</p> <p>For example, <code>data.var_names</code> records gene symbols,  <code>data.var['gene_ids']</code> records Ensembl gene IDs, and  <code>data.var['highly_variable_features']</code> records selected variable genes.</p> <p>The <code>obsm</code> field records embedding coordinates.</p> <p>For example, <code>data.obsm['X_pca']</code> records PCA coordinates,  <code>data.obsm['X_tsne']</code> records t-SNE coordinates,  <code>data.obsm['X_umap']</code> records UMAP coordinates,  <code>data.obsm['X_diffmap']</code> records diffusion map coordinates,  and <code>data.obsm['X_flow']</code> records the force-directed layout coordinates.</p> <p>The <code>uns</code> field stores other related information, such as reference genome (<code>data.uns['genome']</code>), kNN on PCA coordinates (<code>data.uns['pca_knn_indices']</code> and <code>data.uns['pca_knn_distances']</code>), etc.</p>
<b>output_log</b>	File	This is a copy of the logging module output, containing important intermediate messages
output_h5ad	Array[File]	List of output file(s) in Seurat-compatible h5ad format
102		<p><b>Chapter 1. Release Highlights in Current Stable</b></p> <p>(output_name.focus_key.h5ad), in which each file is associated with a focus of the input data.</p> <p>To load this file in Python, first install <a href="#">PegasusIO</a> on your local machine. Then use <code>import pegasusio as io; data = io.read_input(output_name.focus_key.h5ad!)</code> in Python</p>

## de\_analysis

### de\_analysis inputs

Name	Description	Example	Default
perform_de_analysis	Whether perform differential expression (DE) analysis. If performing, by default calculate AUROC scores and Mann-Whitney U test.	true	true
cluster_labels	Specify the cluster label used for DE analysis	"louvain_labels"	"louvain_labels"
alpha	Control false discovery rate at <alpha>	0.05	0.05
fisher	Calculate Fisher's exact test	false	false
t_test	Calculate Welch's t-test.	false	false
find_markers_lightgbm	Whether detect markers using LightGBM	false	false
remove_ribosomal	Remove ribosomal genes with either RPL or RPS as prefixes. Currently only works for human data	false	false
min_gain	Only report genes with a feature importance score (in gain) of at least <gain>	1.0	1.0
annotate_clusters	if also annotate cell types for clusters based on DE results	false	false
annotate_de_test	Differential Expression test to use for inference on cell types. Options: mwu, t, or fisher	"mwu"	"mwu"
organism	Organism, could either of the follow: <ul style="list-style-type: none"> <li>Preset markers: human_immune, mouse_immune, human_brain, mouse_brain, human_lung, or a combination of them as a string separated by comma.</li> <li>User-defined marker file: A Google bucket link to a user-specified JSON file describing the markers. For example: gs://fc-e00000000/my_markers.json.</li> </ul>	"mouse_immune,mouse_brain,human_immune"	"human_immune"
minimum_report_score	Minimum cell type score to report a potential cell type	0.5	0.5

## de\_analysis outputs

Name	Type	Description
output_de_h5ad	Array[File]	<p>List of h5ad-formatted results with DE results updated (output_name.focus_key.h5ad), in which each file is associated with a focus of the input data.</p> <p>To load this file in Python, you need to first install <a href="#">PegasusIO</a> on your local machine. Then type <code>import pegasusio as io; data = io.read_input('output_name.focus_key.h5ad')</code> in Python environment.</p> <p>After loading, data has the similar structure as <i>UnimodalData</i> object in Description of <b>output_zarr</b> in <a href="#">cluster outputs</a> section.</p> <p>Besides, there is one additional field <code>varm</code> which records DE analysis results in <code>data.varm['de_res']</code>. You can use Pandas DataFrame to convert it into a reader-friendly structure: <code>import pandas as pd; df = pd.DataFrame(data.varm['de_res'], index=data.var_names)</code>. Then in the resulting data frame, genes are rows, and those DE test statistics are columns.</p> <p>DE analysis in cumulus is performed on each cluster against cells in all the other clusters. For instance, in the data frame, column <code>1:log2Mean</code> refers to the mean expression of genes in log-scale for cells in Cluster 1. The number before colon refers to the cluster label to which this statistic belongs.</p>
output_de_xlsx	Array[File]	<p>List of spreadsheets reporting DE results (output_name.focus_key.de.xlsx), in which each file is associated with a focus of the input data.</p> <p>Each cluster has two tabs: one for up-regulated genes for this cluster, one for down-regulated ones. In each tab, genes are ranked by AUROC scores. Genes which are not significant in terms of q-values in any of the DE test are not included (at false discovery rate specified in <b>alpha</b> field of <a href="#">de_analysis inputs</a>).</p>
output_markers_xlsx	Array[File]	<p>List of Excel spreadsheets containing detected markers (output_name.focus_key.markers.xlsx), in which each file is associated with a focus of the input data. Each cluster has one tab in the spreadsheet and each tab has three columns, listing markers that are strongly up-regulated, weakly up-regulated and down-regulated.</p>
output_anno_file	Array[File]	<p>List of cluster-based cell type annotation files (output_name.focus_key.anno.txt), in which each file is associated with a focus of the input data.</p>

## How cell type annotation works

In this subsection, we will describe the format of input JSON cell type marker file, the *ad hoc* cell type inference algorithm, and the format of the output putative cell type file.

## JSON file

The top level of the JSON file is an object with two name/value pairs:

- **title:** A string to describe what this JSON file is for (e.g. “Mouse brain cell markers”).
- **cell\_types:** List of all cell types this JSON file defines. In this list, each cell type is described using a separate object with 2 to 3 name/value pairs:
  - **name:** Cell type name (e.g. “GABAergic neuron”).
  - **markers:** List of gene-marker describing objects, each of which has 2 name/value pairs:
    - \* **genes:** List of positive and negative gene markers (e.g. [“Rbfox3+”, “Flt1-”]).
    - \* **weight:** A real number between 0.0 and 1.0 to describe how much we trust the markers in **genes**.

All markers in **genes** share the weight evenly. For instance, if we have 4 markers and the weight is 0.1, each marker has a weight of  $0.1 / 4 = 0.025$ .

The weights from all gene-marker describing objects of the same cell type should sum up to 1.0.

- **subtypes:** Description on cell subtypes for the cell type. It has the same structure as the top level JSON object.

See below for an example JSON snippet:

```
{
  "title" : "Mouse brain cell markers",
  "cell_types" : [
    {
      "name" : "Glutamatergic neuron",
      "markers" : [
        {
          "genes" : ["Rbfox3+", "Reln+", "Slc17a6+", "Slc17a7+"],
          "weight" : 1.0
        }
      ]
    },
    {
      "name" : "GABAergic neuron",
      "markers" : [
        {
          "genes" : ["Gad67+", "Gad67-"],
          "weight" : 1.0
        }
      ]
    }
  ],
  "subtypes" : {
    "title" : "Glutamatergic neuron subtype markers",
    "cell_types" : [
      {
        "name" : "Glutamatergic layer 4",
        "markers" : [
          {
            "genes" : ["Rorb+", "Paqr8+"],
            "weight" : 1.0
          }
        ]
      }
    ]
  }
}
```

## Inference Algorithm

We have already calculated the up-regulated and down-regulated genes for each cluster in the differential expression analysis step.

First, load gene markers for each cell type from the JSON file specified, and exclude marker genes, along with their associated weights, that are not expressed in the data.

Then scan each cluster to determine its putative cell types. For each cluster and putative cell type, we calculate a score between 0 and 1, which describes how likely cells from the cluster are of this cell type. The higher the score is, the more likely cells are from the cell type.

To calculate the score, each marker is initialized with a maximum impact value (which is 2). Then do case analysis as follows:

- For a positive marker:
  - If it is not up-regulated, its impact value is set to 0.
  - Otherwise, if it is up-regulated:
    - \* If it additionally has a fold change in percentage of cells expressing this marker (within cluster vs. out of cluster) no less than 1.5, it has an impact value of 2 and is recorded as a **strong supporting marker**.
    - \* If its fold change ( $fc$ ) is less than 1.5, this marker has an impact value of  $1 + (fc - 1) / 0.5$  and is recorded as a **weak supporting marker**.
- For a negative marker:
  - If it is up-regulated, its impact value is set to 0.
  - If it is neither up-regulated nor down-regulated, its impact value is set to 1.
  - Otherwise, if it is down-regulated:
    - \* If it additionally has  $1 / fc$  (where  $fc$  is its fold change) no less than 1.5, it has an impact value of 2 and is recorded as a **strong supporting marker**.
    - \* If  $1 / fc$  is less than 1.5, it has an impact value of  $1 + (1 / fc - 1) / 0.5$  and is recorded as a **weak supporting marker**.

The score is calculated as the weighted sum of impact values weighted over the sum of weights multiplied by 2 from all expressed markers. If the score is larger than 0.5 and the cell type has cell subtypes, each cell subtype will also be evaluated.

## Output annotation file

For each cluster, putative cell types with scores larger than `minimum_report_score` will be reported in descending order with respect to their scores. The report of each putative cell type contains the following fields:

- **name:** Cell type name.
- **score:** Score of cell type.
- **average marker percentage:** Average percentage of cells expressing marker within the cluster between all positive supporting markers.
- **strong support:** List of strong supporting markers. Each marker is represented by a tuple of its name and percentage of cells expressing it within the cluster.
- **weak support:** List of week supporting markers. It has the same structure as **strong support**.

## plot

The h5ad file contains a default cell attribute `Channel`, which records which channel each that single cell comes from. If the input is a CSV format sample sheet, `Channel` attribute matches the `Sample` column in the sample sheet. Otherwise, it's specified in `channel` field of the cluster inputs.

Other cell attributes used in plot must be added via `attributes` field in the `aggregate_matrices` inputs.

## plot inputs

Name	Description	Example	Default
plot_composition	Takes the format of “label:attr,label:attr,...,label:attr”. If non-empty, generate composition plot for each “label:attr” pair. “label” refers to cluster labels and “attr” refers to sample conditions	“louvain_labels:Donor”	None
plot_tsne	Takes the format of “attr,attr,...,attr”. If non-empty, plot attr colored FIt-SNEs side by side	“louvain_labels,Donor”	None
plot_umap	Takes the format of “attr,attr,...,attr”. If non-empty, plot attr colored UMAP side by side	“louvain_labels,Donor”	None
plot_fle	Takes the format of “attr,attr,...,attr”. If non-empty, plot attr colored FLE (force-directed layout embedding) side by side	“louvain_labels,Donor”	None
plot_net_umap	Takes the format of “attr,attr,...,attr”. If non-empty, plot attr colored UMAP side by side based on net UMAP result.	“leiden_labels,Donor”	None
plot_net_fle	Takes the format of “attr,attr,...,attr”. If non-empty, plot attr colored FLE (force-directed layout embedding) side by side based on net FLE result.	“leiden_labels,Donor”	None
plot_citeseq_umap	Takes the format of “attr,attr,...,attr”. If non-empty, plot attr colored UMAP side by side based on CITE-Seq UMAP result.	“louvain_labels,Donor”	None

## plot outputs

Name	Type	Description
output_pdfs	Array[File]	Outputted pdf files
output_htmls	Array[File]	Outputted html files



## Generate input files for CirroCumulus

Generate [CirroCumulus](#) inputs for visualization using [CirroCumulus](#).

### cirro\_output inputs

Name	Description	Example	Default
generate_cirro_inputs	Whether to generate input files for CirroCumulus	false	false

### cirro\_output outputs

Name	Type	Description
output_cirro_path	Google Bucket URL	Path to CirroCumulus inputs

## Generate SCP-compatible output files

Generate analysis result in [Single Cell Portal](#) (SCP) compatible format.

### scp\_output inputs

Name	Description	Example	Default
generate_scp_outputs	Whether to generate SCP format output or not.	false	false
output_dense	Output dense expression matrix, instead of the default sparse matrix format.	false	false

### scp\_output outputs

Name	Type	Description
output_scp_files	Array[File]	Outputted SCP format files.

## Run CITE-Seq analysis

Users now can use *cumulus/cumulus* workflow solely to run CITE-Seq analysis.

1. Prepare a sample sheet in the following format:

```
Sample,Location,Modality
sample_1,gs://your-bucket/rna_raw_counts.h5,rna
sample_1,gs://your-bucket/citeseq_cell_barcodes.csv,citeseq
```

Each row stands for one modality:

- **Sample:** Sample name, which *must* be the same in the two rows to let Cumulus aggregate RNA and CITE-Seq matrices.
- **Location:** Google bucket URL of the corresponding count matrix file.
- **Modality:** Modality type. `rna` for RNA count matrix; `citeseq` for CITE-Seq antibody count matrix.

2. Run `cumulus/cumulus` workflow using this sample sheet as the input file, and specify the following input fields:

- **`citeseq`:** Set this to `true` to enable CITE-Seq analysis.
  - **`citeseq_umap`:** Set this to `true` to calculate the CITE-Seq UMAP embedding on cells.
  - **`citeseq_umap_exclude`:** A list of CITE-Seq antibodies to be excluded from UMAP calculation. This list should be written in a string format with each antibody name separated by comma.
  - **`plot_citeseq_umap`:** A list of cell barcode attributes to be plotted based on CITE-Seq UMAP embedding. This list should be written in a string format with each attribute separated by comma.
- 

## Load Cumulus results into Pegasus

[Pegasus](#) is a Python package for large-scale single-cell/single-nucleus data analysis, and it uses [PegasusIO](#) for read/write. To load Cumulus results into Pegasus, we provide instructions based on file format:

- **`zarr`:** Annotated Zarr file in zip format. This is the standard output format of Cumulus. You can load it by:

```
import pegasusio as io
data = io.read_input("output_name.zarr.zip")
```

- **`h5ad`:** When setting “`output_h5ad`” field in *Cumulus cluster* to `true`, a list of annotated H5AD file(s) will be generated besides Zarr result. If the input data have multiple foci, Cumulus will generate one H5AD file per focus. You can load it by:

```
import pegasusio as io
adata = io.read_input("output_name.focus_key.h5ad")
```

Sometimes you may also want to specify how the result is loaded into memory. In this case, `read_input` has argument `mode`. Please see [its documentation](#) for details.

- **`loom`:** When setting “`output_loom`” field in *Cumulus cluster* to `true`, a list of loom format file(s) will be generated besides Zarr result. Similarly as H5AD output, Cumulus generates multiple loom files if the input data have more than one foci. To load loom file, you can optionally set its genome name in the following way as this information is not contained by loom file:

```
import pegasusio as io
data = pg.read_input("output_name.focus_key.loom", genome = "GRCh38")
```

After loading, Pegasus manipulate the data matrix in *PegasusIO MultimodalData* structure.

---

## Load Cumulus results into Seurat

Seurat is a single-cell data analysis package written in R.

### Load H5AD File into Seurat

First, you need to set “**output\_h5ad**” field to **true** in cumulus cluster inputs to generate Seurat-compatible output files `output_name.focus_key.h5ad`, in addition to the standard result `output_name.zarr.zip`. If the input data have multiple foci, Cumulus will generate one H5AD file per focus.

Notice that Python, and Python package `anndata` with version at least `0.6.22.post1`, and R package `reticulate` are required to load the result into Seurat.

Execute the R code below to load the h5ad result into Seurat (working with both Seurat v2 and v3):

```
source("https://raw.githubusercontent.com/lilab-bcb/cumulus/master/workflows/cumulus/
↪h5ad2seurat.R")
ad <- import("anndata", convert = FALSE)
test_ad <- ad$read_h5ad("output_name.focus_key.h5ad")
result <- convert_h5ad_to_seurat(test_ad)
```

The resulting Seurat object `result` has three data slots:

- **raw.data** records filtered raw count matrix.
- **data** records filtered and log-normalized expression matrix.
- **scale.data** records variable-gene-selected, standardized expression matrix that are ready to perform PCA.

### Load loom File into Seurat

First, you need to set “**output\_loom**” field to **true** in cumulus cluster inputs to generate a loom format output file, say `output_name.focus_key.loom`, in addition to the standard result `output_name.zarr.zip`. If the input data have multiple foci, Cumulus will generate one loom file per focus.

You also need to install *loomR* package in your R environment:

```
install.packages("devtools")
devtools::install_github("mojaveazure/loomR", ref = "develop")
```

Execute the R code below to load the loom file result into Seurat (working with Seurat v3 only):

```
source("https://raw.githubusercontent.com/lilab-bcb/cumulus/master/workflows/cumulus/
↪loom2seurat.R")
result <- convert_loom_to_seurat("output_name.focus_key.loom")
```

In addition, if you want to set an active cluster label field for the resulting Seurat object, do the following:

```
Idents(result) <- result@meta.data$louvain_labels
```

where `louvain_labels` is the key to the Louvain clustering result in Cumulus, which is stored in cell attributes `result@meta.data`.

## Load Cumulus results into SCANPY

**SCANPY** is another Python package for single-cell data analysis. We provide instructions on loading Cumulus output into SCANPY based on file format:

- **h5ad**: Annotated H5AD file. This is the standard output format of Cumulus:

```
import scanpy as sc
adata = sc.read_h5ad("output_name.h5ad")
```

Sometimes you may also want to specify how the result is loaded into memory. In this case, `read_h5ad` has argument `backed`. Please see [SCANPY documentation](#) for details.

- **loom**: This format is generated when setting “`output_loom`” field in Cumulus cluster to **true**:

```
import scanpy as sc
adata = sc.read_loom("output_name.loom")
```

Besides, `read_loom` has a boolean `sparse` argument to decide whether to read the data matrix as sparse, with default value `True`. If you want to load it as a dense matrix, simply type:

```
adata = sc.read_loom("output_name.loom", sparse = False)
```

After loading, SCANPY manipulates the data matrix in `anndata` structure.

---

## Visualize Cumulus results in Python

Ensure you have [Pegasus](#) installed.

Download your analysis result data, say `output_name.zarr.zip`, from Google bucket to your local machine.

Follow [Pegasus plotting tutorial](#) for visualizing your data in Python.

### 1.1.10 Run Terra pipelines via command line

You can run Terra pipelines via the command line by installing the [Altocumulus](#) package (version 2.0.0 or later is required).

#### Install Altocumulus

1. Make sure you have `conda` installed. If you haven't installed `conda`, use the following commands to install it on Linux:

```
wget https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh .
bash Miniconda3-latest-Linux-x86_64.sh -p /home/foo/miniconda3
mv Miniconda3-latest-Linux-x86_64.sh /home/foo/miniconda3
```

where `/home/foo/miniconda3` should be replaced by your own folder holding Miniconda3.

Or use the following commands for MacOS installation:

```
curl -O curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-MacOSX-x86_64.sh
bash Miniconda3-latest-MacOSX-x86_64.sh -p /Users/foo/miniconda3
mv Miniconda3-latest-MacOSX-x86_64.sh /Users/foo/miniconda3

where ``/Users/foo/miniconda3`` should be replaced by your own folder holding
↳Miniconda3.
```

1. Create a conda environment named “alto” and install Altocumulus:

```
conda create -n alto -y pip
source activate alto
pip install altocumulus
```

When the installation is done, type `alto -h` in terminal to see if you can see the help information.

## Set up Google Cloud Account

Install [gcloud CLI](#) on your local machine.

Then type the following command in your terminal

```
gcloud auth application-default login
```

and follow the pop-up instructions to set up your Google cloud account.

## Run workflows on Terra

**alto terra run** submits workflows to Terra for execution. Features:

- Uploads local files/directories in your inputs to a Google Cloud bucket updates the file paths to point to the Google Cloud bucket.  
Your sample sheet can point to local file paths. In this case, `alto terra run` will take care of uploading directories smartly (e.g. only upload necessary files in BCL folders) and modifying the sample sheet to point to a Google Cloud bucket.
- Creates or uses an existing workspace.
- Uses the latest version of a method unless the method version is specified.

## Options

Required options are in bold.

Name	Description
<b>-m &lt;METHOD&gt;</b> <b>-method</b> <b>&lt;METHOD&gt;</b>	<p>Specify a Terra workflow <i>&lt;METHOD&gt;</i> to use.</p> <p><i>&lt;METHOD&gt;</i> is of format <i>Namespace/Name</i> (e.g. cumulus/cellranger_workflow).</p> <p>Workflow name. The workflow can come from either Dockstore or Broad Methods Repository. If it comes from Dockstore, specify the name as organization:collection:name:version (e.g. broadinstitute:cumulus:cumulus:1.5.0) and the default version would be used if version is omitted. If it comes from Broad Methods Repository, specify the name as namespace/name/version (e.g. cumulus/cumulus/43) and the latest snapshot would be used if version is omitted.</p>
<b>-w</b> <b>&lt;WORKSPACE&gt;</b> <b>-workspace</b> <b>&lt;WORKSPACE&gt;</b>	<p>Specify which Terra workspace <i>&lt;WORKSPACE&gt;</i> to use.</p> <p><i>&lt;WORKSPACE&gt;</i> is also of format <i>Namespace/Name</i> (e.g. foo/bar). The workspace will be created if it does not exist.</p>
<b>-i</b> <b>&lt;WDL_INPUTS&gt;</b> <b>-inputs</b> <b>&lt;WDL_INPUTS&gt;</b>	<p>Specify the WDL input JSON file to use.</p> <p>It can be a local file, a JSON string, or a Google bucket URL directing to a remote JSON file.</p>
<b>-bucket-folder</b> <b>&lt;folder&gt;</b>	<p>Store inputs to <i>&lt;folder&gt;</i> under workspace's google bucket.</p>
<b>-o &lt;updated_json&gt;</b> <b>-upload</b> <b>&lt;updated_json&gt;</b>	<p>Upload files/directories to Google bucket of the workspace, and generate an updated input JSON file (with local paths replaced by Google bucket URLs) to <i>&lt;updated_json&gt;</i> on local machine.</p>
<b>-no-cache</b>	<p>Disable Terra cache calling</p>

## Example run on Terra

This example shows how to use `alto terra run` to run `cellranger_workflow` to extract gene-count matrices from sequencing output.

1. Prepare your sample sheet `example_sample_sheet.csv` as the following:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry
sample_1,GRCh38,/my-local-path/flowcell11,1-2,SI-GA-A8,threeprime
sample_2,GRCh38,/my-local-path/flowcell11,3-4,SI-GA-B8,threeprime
sample_3,mm10,/my-local-path/flowcell11,5-6,SI-GA-C8,fiveprime
```

(continues on next page)

(continued from previous page)

```
sample_4,mm10,/my-local-path/flowcell1,7-8,SI-GA-D8,fiveprime
sample_1,GRCh38,/my-local-path/flowcell2,1-2,SI-GA-A8,threeprime
sample_2,GRCh38,/my-local-path/flowcell2,3-4,SI-GA-B8,threeprime
sample_3,mm10,/my-local-path/flowcell2,5-6,SI-GA-C8,fiveprime
sample_4,mm10,/my-local-path/flowcell2,7-8,SI-GA-D8,fiveprime
```

where `/my-local-path` is the top-level directory of your BCL files on your local machine.

Note that `sample_1`, `sample_2`, `sample_3`, and `sample_4` are sequenced on 2 flowcells.

2. Prepare your JSON input file `inputs.json` for `cellranger_workflow`:

```
{
  "cellranger_workflow.input_csv_file" : "/my-local-path/sample_sheet.csv",
  "cellranger_workflow.output_directory" : "gs://url/outputs",
  "cellranger_workflow.delete_input_bcl_directory": true
}
```

where `gs://url/outputs` is the folder on Google bucket of your workspace to hold output.

3. Run the following command to kick off your Terra workflow:

```
alto terra run -m cumulus/cellranger_workflow -i inputs.json -w myworkspace_
↪namespace/myworkspace_name -o inputs_updated.json
```

where `myworkspace_namespace/myworkspace_name` should be replaced by your workspace namespace and name.

Upon success, `alto terra run` returns a URL pointing to the submitted Terra job for you to monitor.

If for any reason, your job failed. You could rerun it without uploading files again via the following command:

```
alto terra run -m cumulus/cellranger_workflow -i inputs_updated.json -w myworkspace_
↪namespace/myworkspace_name
```

because `inputs_updated.json` is the updated version of `inputs.json` with all local paths being replaced by their corresponding Google bucket URLs after uploading.

## Run workflows on a Cromwell server

**alto cromwell run** submits WDL jobs to a Cromwell server for execution. Features:

- Uploads local files/directories in your inputs to an appropriate location depending on backend chosen and updates the file paths to point to the bucket information.
- Uses the `method` parameter to pull in appropriate workflow to import and run.

## Options

Required options are in bold.

Name	Description
<b>-s &lt;SERVER&gt;</b> <b>-server</b> <b>&lt;SERVER&gt;</b>	Server hostname or IP address.
<b>-p &lt;PORT&gt;</b> <b>-port &lt;PORT&gt;</b>	Port number for Cromwell service. The default port is 8000.
<b>-m</b> <b>&lt;METHOD_STR&gt;</b> <b>-method</b> <b>&lt;METHOD_STR&gt;</b>	Workflow name from Dockstore, with name specified as organization:collection:name:version (eg. broadinstitute:cumulus:cumulus:1.5.0). The default version would be used if version is omitted.
<b>-i &lt;INPUT&gt;</b> <b>-input &lt;INPUT&gt;</b>	Path to a local JSON file specifying workflow inputs.
<b>-o &lt;updated_json&gt;</b> <b>-upload &lt;INPUT&gt;</b>	Upload files/directories to the workspace cloud bucket and output updated input json (with local path replaced by cloud bucket urls) to <updated_json>.
<b>-b</b> <b>&lt;[s3 gs]://&lt;bucket-name&gt;/&lt;bucket-folder&gt;&gt;</b> <b>-bucket</b> <b>&lt;[s3 gs]://&lt;bucket-name&gt;/&lt;bucket-folder&gt;&gt;</b>	Cloud bucket folder for uploading local input data. Start with 's3://' if an AWS S3 bucket is used, 'gs://' for a Google bucket. Must be specified when '-o' option is used.
<b>-no-ssl-verify</b>	Disable SSL verification for web requests. Not recommended for general usage, but can be useful for intra-networks which don't support SSL verification.

### Example import of any Cumulus workflow

This example shows how to use `alto cromwell run` to run demultiplexing workflow on any backend.

1. Prepare your sample sheet `demux_sample_sheet.csv` as the following:

```
OUTNAME, RNA, TagFile, TYPE
sample_1, gs://exp/data_1/raw_feature_bc_matrix.h5, gs://exp/data_1/sample_1_ADT.
--csv, cell-hashing
```

(continues on next page)



(continued from previous page)

```
sample_2,gs://exp/data_2/raw_feature_bc_matrix.h5,gs://exp/data_3/possorted_
↪genome_bam.bam,genetic-pooling
```

2. Prepare your JSON input file `cumulus_inputs.json` for `cellranger_workflow`:

```
{
  "demultiplexing.input_sample_sheet" : "demux_sample_sheet.csv",
  "demultiplexing.output_directory" : "gs://url/outputs",
  "demultiplexing.zones" : "us-west1-a us-west1-b us-west1-c",
  "demultiplexing.backend" : "gcp",
  "demultiplexing.genome" : "GRCh38-2020-A"
}
```

where `gs://url/outputs` is the folder on Google bucket of your workspace to hold output.

3. Run the following command to kick off your run on a chosen backend:

```
alto cromwell run -s 10.10.10.10 -p 3000 -m_
↪broadinstitute:cumulus:Demultiplexing:master \
-i cumulus_inputs.json
```

## 1.1.11 Examples

### Example of Gene expression, Hashing and CITE-Seq Analysis on Cloud

In this example, you'll learn how to perform Gene expression, Hashing and CITE-Seq data analysis on Cloud.

This example covers the cases of both [Terra](#) platform and a custom cloud server running [Cromwell](#). When reading through the tutorial, you may check out the corresponding part based on your working situation.

## 0. Prerequisite

### 0-a. Cromwell server

If you use a Cromwell server on Cloud, on your local machine, you need to install the corresponding Cloud SDK tool if not:

- [gcloud CLI](#) if your Cloud bucket is on Google Cloud.
- [AWS CLI v2](#) if your Cloud bucket is on Amazon AWS Cloud.

And then install [Altocumulus](#) in your Python environment. This is the tool for data transfer between local machine and cloud bucket, as well as communication with the Cromwell server on cloud.

### 0-b. Terra Platform

If you use Terra, after registering on Terra and creating a workspace there, you'll need the following information:

- **Terra workspace name.** This is shown on your Terra workspace webpage, with format "`<workspace-namespace>/<workspace-name>`". For example, if your Terra workspace has full name `ws-lab/ws-01`, then **ws-lab** is the namespace and **ws-01** is the workspace name within that namespace.

- The corresponding **Google Cloud Bucket** of your Terra workspace. You can check it under “*Google Bucket*” title on the right panel of your Terra workspace’s *Dashboard* tab. The bucket name associated with your workspace starts with `fc-` followed by a sequence of heximal numbers. For example, `gs://fc-e0000000`, where “`gs://`” is the header of GS URI.

Besides, install [gcloud CLI](#) and [Altocumulus](#) on your local machine for data uploading. These tools will be used for data transfer between local machine and Cloud bucket.

Alternatively, you can also use Terra web UI for job submission instead of command-line submission. This will be discussed in [Section Run Analysis with Terra Web UI](#) below.

---

## 1. Extract Gene-Count Matrices

This phase is to extract gene-count matrices from sequencing output.

There are two cases: (1) from BCL data, which includes *mkfastq* step to generate FASTQ files and *count* step to generate gene-count matrices; (2) from FASTQ files, which only runs the *count* step.

### 1-a. Extract Genen-Count Matrices from BCL data

This section covers the case starting from BCL data.

#### Step 1. Sample Sheet Preparation

First, prepare a feature index file for your dataset. Say its filename is `antibody_index.csv`, which has format “*feature\_barcode,feature\_name,feature\_type*”. See an example below:

```
TTCCTGCCATTACTA,HTO_1,hashing
CCGTACCTCATTGTT,HTO_2,hashing
GGTAGATGTCCTCAG,HTO_3,hashing
TGGTGTCTTCTTGA,Ab1,citeseq
CTCATGTAACTCCT,Ab2,citeseq
GCGCAACTTGATGAT,Ab3,citeseq
... ..
```

where each line contains the barcode and the name of a Hashing/CITE-Seq index: `hashing` indicates a Cell/Nucleus-Hashing index, while `citeseq` indicates a CITE-Seq index.

Next, create a sample sheet `cellranger_sample_sheet.csv` for Cell Ranger processing on your local machine. Below is an example:

```
Sample,Reference,Flowcell,Lane,Index,Chemistry,DataType,FeatureBarcodeFile
sample_control,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-TT-A1,fiveprime,rna
sample_gex,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-TT-A2,fiveprime,rna
sample_cell_hashing,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-NN-A1,fiveprime,
↪hashing,/path/to/antibody_index.csv
sample_cite_seq,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-NN-A2,fiveprime,citeseq,/
↪path/to/antibody_index.csv
```

where

- `GRCh38-2020-A` is the is the Human GRCh38 (GENCODE v32/Ensembl 98) genome reference prebuilt by Cumulus. See [Cumulus single-cell genome reference list](#) for a complete list of genome references.

- `/path/to/flowcell/folder` should be replaced by the actual local path to the BCL folder of your sequencing data.
- `/path/to/antibody_index.csv` should be replaced by the actual local path to `antibody_index.csv` file we just created above.
- `rna`, `hashing` and `citeseq` refer to gene expression data, cell/nucleus-hashing data, and CITE-Seq data, respectively.
- Samples of type `rna` do not need any feature barcode file for indexing.

For the details on how to prepare this sample sheet, please refer to Step 3 of [Cell Ranger sample sheet instruction](#).

## Step 2. Workflow Input Preparation

Now prepare a JSON file for **cellranger\_workflow** WDL workflow input on your local machine (say named `cellranger_inputs.json`):

```
{
  "cellranger_workflow.input_csv_file": "/path/to/cellranger_sample_sheet.csv",
  "cellranger_workflow.output_directory": "gs://my-bucket/cellranger_output"
}
```

where

- `/path/to/cellranger_sample_sheet.csv` should be replaced by the actual local path to your sample sheet created above.
- `gs://my-bucket/cellranger_output` is the target folder on Google bucket to store your result when the workflow job is finished, where `my-bucket` should be replaced by your own Google bucket name.

For details on the all the workflow inputs of *cellranger\_workflow*, please refer to [Cell Ranger workflow inputs](#).

## Step 3. Job Submission

Now we are ready to submit a job to cloud for computing:

- If you use a Cromwell server on cloud, run the following Altocumulus command:

```
alto cromwell run -s <server-address> -p <port-number> -m
↪broadinstitute:cumulus:cellranger -i /path/to/cellranger_inputs.json -o
↪cellranger_inputs_updated.json -b gs://my-bucket/data_source
```

where

- `-s` specifies the server's IP address (or hostname), where `<server-address>` should be replaced by the actual IP address (or hostname).
- `-p` specifies the server's port number for Cromwell, where `<port-number>` should be replaced by the actual port number.
- `-m` specifies which WDL workflow to use. You should use the Dockstore name of Cumulus [cellranger\\_workflow](#). Here, the version is omitted, so that the default version will be used. Alternatively, you can explicitly specify which version to use, e.g. `broadinstitute:cumulus:cellranger:master` to use its development version in *master* branch.
- `-i` specifies the workflow input JSON file.
- `-o` and `-b` are used when the input data (which are specified in the workflow input JSON file and sample sheet CSV file) are local and need to be uploaded to Cloud bucket first.

- `-o` specifies the updated workflow input JSON file after uploading the input data, with all the local paths updated to Cloud bucket URIs. This is useful when resubmitting jobs running the same input data, without uploading the same input data again.
- `-b` specifies which folder on Cloud bucket to upload the local input data, where `my-bucket` should be replaced by your own Google bucket name. Feel free to choose the folder name other than `data_source`.

Notice that `-o` and `-b` options can be dropped if all of your input data are already on Cloud bucket.

After submission, you'll get the job's ID for tracking its status:

```
alto cromwell check_status -s <server-address> -p <port-number> --id <your-job-ID>
```

where `<your-job-ID>` should be replaced by the actual Cromwell job ID.

- If you use Terra, run the following Altocumulus command:

```
alto terra run -m broadinstitute:cumulus:cellranger -w ws-lab/ws-01 --bucket-  
→folder data_source -i /path/to/cellranger_inputs.json -o cellranger_inputs_  
→updated.json
```

where

- `-m` specifies which WDL workflow to use. You should use the Dockstore name of Cumulus `cellranger_workflow`. Here, the version is omitted, so that the default version will be used. Alternatively, you can explicitly specify which version to use, e.g. `broadinstitute:cumulus:cellranger:master` to use its development version in *master* branch.
- `-w` specifies the Terra workspace full name to use, where `ws-lab/ws-01` should be replaced by your own Terra workspace full name.
- `--bucket-folder` specifies the folder name on the Google bucket associated with the Terra workspace to store the uploaded data. Feel free to choose folder name other than `data_source`.
- `-i` specifies the workflow input JSON file, where `/path/to/cellranger_inputs.json` should be replaced by the actual local path to `cellranger_inputs.json` file.
- `-o` specifies the updated workflow input JSON file after uploading the input data, with all the local paths updated to Cloud bucket URIs. This is useful when resubmitting jobs running the same input data, without uploading the same input data again.

Notice that `--bucket-folder` and `-o` options can be dropped if all of your input data are already on Cloud bucket.

After submission, you can check the job's status in the *Job History* tab of your Terra workspace page.

When the job is done, you'll get results in `gs://my-bucket/cellranger_output`, which is specified in `cellranger_inputs.json` above. It should contain 4 subfolders, each of which is associated with one sample specified in `cellranger_sample_sheet.csv` above.

For the next phases, you'll need 3 files from the output:

- RNA count matrix of the sample group of interest: `gs://my-bucket/cellranger_output/sample_gex/raw_feature_bc_matrix.h5`;
- Cell-Hashing Antibody count matrix: `gs://my-bucket/cellranger_output/sample_cell_hashing/sample_cell_hashing.csv`;
- CITE-Seq Antibody count matrix: `gs://my-bucket/cellranger_output/sample_cite_seq/sample_cite_seq.csv`.

## 1-b. Extract Gene-Count Matrices from FASTQ files

This section covers the case starting from FASTQ files.

Similarly as above, First, prepare a feature index file for your dataset. Say its filename is `antibody_index.csv`, which has format “*feature\_barcode, feature\_name, feature\_type*”. See an example below:

```
TTCCTGCCATTACTA, HTO_1, hashing
CCGTACCTCATTGTT, HTO_2, hashing
GGTAGATGTCCTCAG, HTO_3, hashing
TGGTGTCACTTCTGA, Ab1, citeseq
CTCATTGTAACTCCT, Ab2, citeseq
GCGCAACTTGATGAT, Ab3, citeseq
... ..
```

where each line contains the barcode and the name of a Hashing/CITE-Seq index: `hashing` indicates a Cell/Nucleus-Hashing index, while `citeseq` indicates a CITE-Seq index.

Next, create a sample sheet `cellranger_sample_sheet.csv` for Cell Ranger processing on your local machine. Below is an example:

```
Sample,Reference,Flowcell,Chemistry,DataType,FeatureBarcodeFile
sample_1_rna,GRCh38-2020-A,/path/to/fastq/gex,fiveprime,rna
sample_2_rna,GRCh38-2020-A,/path/to/fastq/gex,fiveprime,rna
sample_3_rna,GRCh38-2020-A,/path/to/fastq/gex,fiveprime,rna
sample_1_adt,GRCh38-2020-A,/path/to/fastq/hashing_citeseq,fiveprime,adt,/path/to/
↪antibody_index.csv
sample_2_adt,GRCh38-2020-A,/path/to/fastq/hashing_citeseq,fiveprime,adt,/path/to/
↪antibody_index.csv
sample_3_adt,GRCh38-2020-A,/path/to/fastq/hashing_citeseq,fiveprime,adt,/path/to/
↪antibody_index.csv
```

where

- `GRCh38-2020-A` is the is the Human GRCh38 (GENCODE v32/Ensembl 98) genome reference prebuilt by Cumulus. See [Cumulus single-cell genome reference list](#) for a complete list of genome references.
- `/path/to/fastq/gex` should be replaced by the actual local path to the folder containing FASTQ files of RNA samples.
- `/path/to/fastq/hashing_citeseq` should be replaced by the actual local path to the folder containing FASTQ files of Cell/Nucleus-Hashing and CITE-Seq samples.
- `/path/to/antibody_index.csv` should be replaced by the actual local path to `antibody_index.csv` file we just created above.
- `rna` and `adt` refer to gene expression data and antibody data, respectively. In specific, `adt` covers both `citeseq` and `hashing` types, i.e. it includes both Hashing and CITE-Seq data types.
- Samples of type `rna` do not need any feature barcode file for indexing.
- Columns *Lane* and *Index* are not needed if starting from FASTQ files, as *mkfastq* step will be skipped.

For the details on how to prepare this sample sheet, please refer to Step 3 of [Cell Ranger sample sheet instruction](#).

Now prepare a JSON file for **cellranger\_workflow** WDL workflow input on your local machine (say named `cellranger_inputs.json`):

```
{
    "cellranger_workflow.input_csv_file": "/path/to/cellranger_sample_sheet.csv",
```

(continues on next page)

(continued from previous page)

```
"cellranger_workflow.output_directory": "gs://my-bucket/cellranger_output",
"cellranger_workflow.run_mkfastq": false
}
```

where

- `/path/to/cellranger_sample_sheet.csv` should be replaced by the actual local path to your sample sheet created above.
- `gs://my-bucket/cellranger_output` is the target folder on Google bucket to store your result when the workflow job is finished, where `my-bucket` should be replaced by your own Google bucket name.
- Set `run_mkfastq` to `false` to skip the `mkfastq` step, as we start from FASTQ files.

For details on the all the workflow inputs of `cellranger_workflow`, please refer to [Cell Ranger workflow inputs](#).

Now we are ready to submit a job to cloud for computing. Follow instructions in [Section 1-a](#) above.

When finished, you'll get results in `gs://my-bucket/cellranger_output`, which is specified in `cellranger_inputs.json` above. It should contain 6 subfolders, each of which is associated with one sample specified in `cellranger_sample_sheet.csv` above.

In specific, for each `adt` type sample, there are both count matrix of Hashing data and that of CITE-Seq data generated inside its corresponding subfolder, with filename suffix `.hashing.csv` and `.citeseq.csv`, respectively.

---

## 2. Demultiplex Cell-Hashing Data using DemuxEM

### Run Workflow on Cloud

Next, we need to demultiplex the resulting RNA gene-count matrices. We use [DemuxEM](#) method in this example.

To be brief, we use the output of [Section 1-a](#) for illustration:

1. On your local machine, prepare a CSV-format sample sheet `demux_sample_sheet.csv` with the following content:

```
OUTNAME, RNA, TagFile, TYPE
exp, gs://my-bucket/cellranger_output/sample_gex/raw_feature_bc_matrix.h5, gs://my-
↪ bucket/cellranger_output/sample_cell_hashing/sample_cell_hashing.csv, cell-
↪ hashing
```

where **OUTNAME** specifies the subfolder and file names of output, which is free to be changed, **RNA** and **TagFile** columns specify the RNA and hashing tag meta-data of samples, and **TYPE** is `cell-hashing` for this phase.

2. On your local machine, also prepare an input JSON file `demux_inputs.json` for **demultiplexing** WDL workflow, `demux_inputs.json` with the following content:

```
{
  "demultiplexing.input_sample_sheet" : "/path/to/demux_sample_sheet.csv",
  "demultiplexing.output_directory" : "gs://my-bucket/demux_output"
}
```

where `/path/to/demux_sample_sheet.csv` should be replaced by the actual local path to `demux_sample_sheet.csv` created above.

For the details on these options, please refer to [demultiplexing workflow inputs](#).

3. Submit a *demultiplexing* job with `demux_inputs.json` input above to cloud for execution.

For job submission:

- If you use a Cromwell server on cloud, run the following Altocumulus command on your local machine:

```
alto cromwell run -s <server-address> -p <port-number> -m
↪broadinstitute:cumulus:demultiplexing -i /path/to/demux_inputs.json -o demux_
↪inputs_updated.json -b gs://my-bucket/data_source
```

where

- `broadinstitute:cumulus:demultiplexing` refers to [demultiplexing](#) WDL workflow published on Dockstore. Here, the version is omitted, so that the default version will be used. Alternatively, you can explicitly specify which version to use, e.g. `broadinstitute:cumulus:demultiplexing:master` to use its development version in *master* branch.
- `/path/to/demux_inputs.json` should be replaced by the actual local path to `demux_inputs.json` created above.
- Replace `my-bucket` in `-b` option by your own Google bucket name, and feel free to choose folder name other than `data_source` for uploading.
- We still need `-o` and `-b` options because `demux_sample_sheet.csv` is on the local machine.

Similarly, when the submission succeeds, you'll get another job ID for demultiplexing. You can use it to track the job status.

- If you use Terra, run the following Altocumulus command:

```
alto terra run -m broadinstitute:cumulus:demultiplexing -w ws-lab/ws-01 --bucket-
↪folder data_source -i /path/to/demux_inputs.json -o demux_inputs_updated.json
```

where

- `broadinstitute:cumulus:demultiplexing` refers to [demultiplexing](#) WDL workflow published on Dockstore. Here, the version is omitted, so that the default version will be used. Alternatively, you can explicitly specify which version to use, e.g. `broadinstitute:cumulus:demultiplexing:master` to use its development version in *master* branch.
- `/path/to/demux_inputs.json` should be replaced by the actual local path to `demux_inputs.json` created above.
- `ws-lab/ws-01` should be replaced by your own Terra workspace full name.
- `--bucket-folder`: Feel free to choose folder name other than `data_source` for uploading.
- We still need `-o` and `--bucket-folder` options because `demux_sample_sheet.csv` is on the local machine.

After submission, you can check the job's status in the *Job History* tab of your Terra workspace page.

When finished, demultiplexing results are in `gs://my-bucket/demux_output/exp` folder, with the following important output files:

- `exp_demux.zarr.zip`: Demultiplexed RNA raw count matrix. This will be used for downstream analysis.
- `exp.out.demuxEM.zarr.zip`: This file contains intermediate results for both RNA and hashing count matrices, which is useful for compare with other demultiplexing methods.
- DemuxEM plots in PDF format. They are used for evaluating the performance of DemuxEM on the data.

### (Optional) Extract Demultiplexing results

This is performed on your local machine with demultiplexing results downloaded from cloud to your machine.

To download the demultiplexed count matrix `exp_demux.zarr.zip`, you can either do it in Google cloud console, or using `gsutil` in command line:

```
gsutil -m cp gs://my-bucket/demux_output/exp/exp_demux.zarr.zip .
```

After that, in your Python environment, install `Pegasus` package, and follow the steps below to extract the demultiplexing results:

1. Load Libraries:

```
import numpy as np
import pandas as pd
import pegasus as pg
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Load demuxEM output. For demuxEM, load RNA expression matrix with demultiplexed sample identities in Zarr format. These can be found in Google cloud console. QC 500 <= # of genes < 6000, % mito <= 10%:

```
data = pg.read_input('exp_demux.zarr.zip')
pg.qc_metrics(data, min_genes=500, max_genes=6000, mito_prefix='Mt-', percent_
↪ mito=10)
pg.filter_data(data)
```

3. Demultiplexing results showing singlets, doublets and unknown:

```
data.obs['demux_type'].value_counts()
```

4. Show assignments in singlets:

```
idx = data.obs['demux_type'] == 'singlet'
data.obs.loc[idx, 'assignment'].value_counts()[0:10]
```

5. Write assignment outputs to CSV:

```
data.obs[['demux_type', 'assignment']].to_csv('demux_exp.csv')
```

---

## 3. Data Analysis on CITE-Seq Data

In this phase, we merge RNA and ADT matrices for CITE-Seq data, and perform the downstream analysis.

To be brief, we use the CITE-Seq count matrix generated from Section 1-a and demultiplexing results in Section 2 for illustration here:

1. On your local machine, prepare a CSV-format sample sheet `count_matrix.csv` with the following content:

```
Sample,Location,Modality
exp,gs://my-bucket/demux_output/exp/exp_demux.zarr.zip,rna
exp,gs://my-bucket/cellranger_output/sample_cite_seq/sample_cite_seq.csv,citeseq
```

This sample sheet describes the metadata for each modality (as one row in the sheet):



- **Sample** specifies the name of the modality, and all the modalities of the same sample should have one common name, as otherwise their count matrices won't be aggregated together;
  - **Location** specifies the file location. For RNA data, this is the output of Phase 2; for CITE-Seq antibody data, it's the output of Phase 1.
  - **Modality** specifies the modality type, which is either `rna` for RNA matrix, or `citeseq` for CITE-Seq antibody matrix.
2. On your local machine, also prepare a JSON file `cumulus_inputs.json` for **cumulus** WDL workflow, with the following content:

```
{
    "cumulus.input_file": "/path/to/count_matrix.csv",
    "cumulus.output_directory": "gs://my-bucket/cumulus_output",
    "cumulus.output_name": "exp_merged_out",
    "cumulus.select_only_singlets": true,
    "cumulus.run_louvain": true,
    "cumulus.run_umap": true,
    "cumulus.citeseq": true,
    "cumulus.citeseq_umap": true,
    "cumulus.citeseq_umap_exclude": "Mouse_IgG1,Mouse_IgG2a,Mouse_IgG2b,Rat_
    ↪IgG2b",
    "cumulus.plot_composition": "louvain_labels:assignment",
    "cumulus.plot_umap": "louvain_labels,assignment",
    "cumulus.plot_citeseq_umap": "louvain_labels,assignment",
    "cumulus.cluster_labels": "louvain_labels",
    "cumulus.annotate_cluster": true,
    "cumulus.organism": "human_immune"
}
```

where `/path/to/count_matrix.csv` should be replaced by the actual local path to `count_matrix.csv` created above.

A typical Cumulus WDL pipeline consists of 4 steps, which is given [here](#). For details on Cumulus workflow inputs above, please refer to [cumulus inputs](#).

3. Submit a *demultiplexing* job with `cumulus_inputs.json` input above to cloud for execution.

For job submission:

- If you use a Cromwell server on cloud, run the following Altocumulus command to submit the job:

```
alto cromwell run -s <server-address> -p <port-number> -m_
    ↪broadinstitute:cumulus:cumulus -i /path/to/cumulus_inputs.json -o cumulus_
    ↪inputs_updated.json -b gs://my-bucket/data_source
```

where

- `broadinstitute:cumulus:cumulus` refers to [cumulus](#) WDL workflow published on Dockstore. Here, the version is omitted, so that the default version will be used. Alternatively, you can explicitly specify which version to use, e.g. `broadinstitute:cumulus:cumulus:master` to use its development version in *master* branch.
- `/path/to/cumulus_inputs.json` should be replaced by the actual local path to `cumulus_inputs.json` created above.
- `my-bucket` in `-b` option should be replaced by your own Google bucket name, and feel free to choose folder name other than `data_source` for uploading data.
- We still need `-o` and `-b` options because `count_matrix.csv` is on the local machine.

Similarly, when the submission succeeds, you'll get another job ID for demultiplexing. You can use it to track the job status.

- If you use Terra, run the following Altocumulus command:

```
alto terra run -m broadinstitute:cumulus:cumulus -w ws-lab/ws-01 --bucket-folder_
↪data_source -i /path/to/cumulus_inputs.json -o cumulus_inputs_updated.json
```

where

- `broadinstitute:cumulus:cumulus` refers to [cumulus](#) WDL workflow published on Dockstore. Here, the version is omitted, so that the default version will be used. Alternatively, you can explicitly specify which version to use, e.g. `broadinstitute:cumulus:cumulus:master` to use its development version in *master* branch.
- `ws-lab/ws-01` should be replaced by your own Terra workspace full name.
- `--bucket-folder`: Feel free to choose folder name other than `data_source` for uploading data.
- `/path/to/cumulus_inputs.json` should be replaced by the actual local path to `cumulus_inputs.json` created above.
- We still need `-o` and `--bucket-folder` options because `count_matrix.csv` is on the local machine.

After submission, you can check the job's status in the *Job History* tab of your Terra workspace page.

When finished, all the output files are in `gs://my-bucket/cumulus_output` folder, with the following important files:

- `exp_merged_out.aggr.zarr.zip`: The *ZARR* format file containing both the aggregated count matrix in `<genome>-rna` modality, as well as CITE-Seq antibody count matrix in `<genome>-citeseq` modality, where `<genome>` is the genome reference name of your count matrices, e.g. *GRCh38-2020-A*.
- `exp_merged_out.zarr.zip`: The *ZARR* format file containing the analysis results in `<genome>-rna` modality, and CITE-Seq antibody count matrix in `<genome>-citeseq` modality.
- `exp_merged_out.<genome>-rna.h5ad`: The processed RNA matrix data in *H5AD* format.
- `exp_merged_out.<genome>-rna.filt.xlsx`: The Quality-Control (QC) summary of the raw data.
- `exp_merged_out.<genome>-rna.filt.{UMI, gene, mito}.pdf`: The QC plots of the raw data.
- `exp_merged_out.<genome>-rna.de.xlsx`: Differential Expression analysis result.
- `exp_merged_out.<genome>-rna.anno.txt`: The putative cell type annotation output.
- `exp_merged_out.<genome>-rna.umap.pdf`: UMAP plot.
- `exp_merged_out.<genome>-rna.citeseq.umap.pdf`: CITE-Seq UMAP plot.
- `exp_merged_out.<genome>-rna.louvain_labels.assignment.composition.pdf`: Composition plot.

---

## Run Analysis with Terra Web UI

For Terra users, instead of using Altocumulus to submit jobs in command line, they can also use the Terra web UI.

First, upload the local BCL data or FASTQ files to the Google bucket associated with your Terra workspace (say `gs://fc-e0000000`) using [gsutil](#):

```
gsutil -m cp -r /path/to/your/data/folder gs://fc-e0000000/data_source/
```

where `/path/to/your/data/folder` should be replaced by the actual local path to your data folder, and `data_source` is the folder on Google bucket to store the uploaded data.

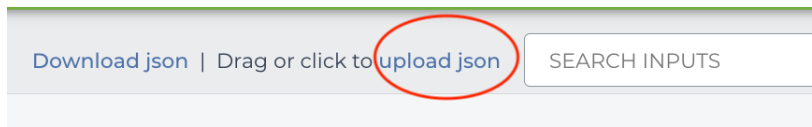
Then for each of the 3 phases above:

1. When preparing the sample sheet, remember to replace all the local paths by the GS URIs of the corresponding folders/files that you uploaded to Google bucket. Then upload it to Google bucket as well:

```
gsutil cp /path/to/sample/sheet gs://fc-e0000000/data_source/
```

where `/path/to/sample/sheet` should be replaced by the actual local path to your sample sheet. Notice that for Phase 1, `antibody_index.csv` file should also be uploaded to Google bucket, and its references in the sample sheet must be replaced by its GS URI.

2. When preparing the workflow input JSON file, change the field of sample sheet to its GS URI on cloud.
3. Import the corresponding WDL workflow to your Terra workspace by following steps in [import workflows](#) tutorial.
4. In the workflow page (Workspace -> Workflows -> your WDL workflow), upload your input JSON file by clicking the “upload json” button:



5. Click “SAVE” button to save the configuration, and click “RUN ANALYSIS” button to submit the job:



You can check the job’s status in the *Job History* tab of your Terra workspace page.

## Example of 10X Genomics CellPlex Analysis on Cloud

In this example, you’ll learn how to perform Cellplex analysis on Cloud using [Cromwell](#).

### 0. Prerequisite

You need to install the corresponding Cloud SDK tool on your local machine if not:

- [gcloud CLI](#) for Google Cloud.
- [AWS CLI v2](#) for Amazon AWS Cloud.

And then install [Altocumulus](#) in your Python environment. This is the tool for data transfer between local machine and Cloud VM instance.

In this example, we assume that your Cromwell server is already deployed on Cloud at IP address `10.0.0.0` with port `8000`, and also assume using Google Cloud with bucket `gs://my-bucket`.

## 1. Extract Genen-Count Matrices

First step is to extract gene-count matrices from sequencer output.

In this example, we have the following experiment setting:

- A sample named `cellplex_gex` by pooling all RNA data together for sequencing, with index `SI-TT-A1`;
- A sample named `cellplex_barcode` for hashing data, with index `SI-NN-A1`;
- Three samples to perform individual control:
  - Sample A with index `SI-TT-A2` and CMO ID `CMO_301`,
  - Sample B with index `SI-TT-A3` and CMO ID `CMO_302`,
  - Sample C with index `SI-TT-A4` and CMO ID `CMO_303`

To extract feature barcodes for the hashing data, we need to create a feature barcoding file (say named `feature_barcode.csv`). Please refer to [10X Multi CMO Reference](#) for the sequence information of these CMO IDs:

```
ATGAGGAATTCCTGC, A
CATGCCAATAGAGCG, B
CCGTCGTCCAAGCAT, C
```

After that, create a sample sheet in CSV format (say named `cellranger_sample_sheet.csv`) as the following:

```
Sample,Reference,Flowcell,Lane,Index,DataType,FeatureBarcodeFile
cellplex_gex,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-TT-A1,rna
cellplex_barcode,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-NN-A1,cmo,/path/to/
↪feature_barcode.csv
A,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-TT-A2,rna
B,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-TT-A3,rna
C,GRCh38-2020-A,/path/to/flowcell/folder,*,SI-TT-A4,rna
```

where

- `GRCh38-2020-A` is the Human GRCh38 (GENCODE v32/Ensembl 98) genome reference prebuilt by Cumulus. See [Cumulus single-cell genome reference list](#) for a complete list of genome references.
- `/path/to/flowcell/folder` should be replaced by the local path to the BCL folder of your sequencer output.
- `/path/to/feature_barcode.csv` should be replaced by the local path to `feature_barcode.csv` file we just created above.
- `rna` and `cmo` refer to gene expression data and cell multiplexing oligos used in 10X Genomics CellPlex assay, respectively.
- Only the sample of `cmo` type needs a feature barcode file for indexing.

For details on preparing this sample sheet, please refer to [Cell Ranger workflow sample sheet format](#).

Now let's prepare an input JSON file for `cellranger_workflow` WDL workflow to execute (say named `cellranger_inputs.json`):

```
{
  "cellranger_workflow.input_csv_file": "/path/to/cellranger_sample_sheet.csv",
  "cellranger_workflow.output_directory": "gs://my-bucket/cellplex/cellranger_output
↪"
}
```

where

- `/path/to/cellranger_sample_sheet.csv` should be replaced by the local path to your sample sheet created above.
- `gs://my-bucket/cellplex/cellranger_output` is the target folder on Google bucket to store your result when the workflow job is finished.

For details on these workflow inputs, please refer to [CellRanger workflow inputs](#).

Now we are ready to submit a job to the Cromwell server on Cloud for computing. On your local machine, run the following command:

```
alto cromwell run -s 10.0.0.0 -p 8000 -m broadinstitute:cumulus:cellranger:master -i /
↳path/to/cellranger_inputs.json -o cellranger_inputs_updated.json -b gs://my-bucket/
↳cellplex
```

where

- `-s` to specify the server's IP address (or hostname), `-p` to specify the server's port number.
- `-m` to specify which WDL workflow to use. You should use the Dockstore name of Cumulus [cellranger\\_workflow](#). Here, the latest version `master` is used. If omit the version info, i.e. `broadinstitute:cumulus:cellranger`, the default version will be used.
- `-i` to specify the workflow input JSON file.
- `-o` and `-b` are used when the input data are local and need to be uploaded to Cloud bucket first. This can be inferred from the workflow input JSON file and sample sheet CSV file.
- `-o` to specify the updated workflow input JSON file after uploading the input data, with all the local paths updated to Cloud bucket URLs.
- `-b` to specify which folder on Cloud bucket to upload the local input data.

Notice that `-o` and `-b` options can be dropped if all of your input data are already on Cloud bucket.

After submission, you'll get the job's ID for tracking its status:

```
alto cromwell check_status -s 10.0.0.0 -p 8000 --id <your-job-ID>
```

where `<your-job-ID>` should be replaced by the actual Cromwell job ID.

When the job is done, you'll get results in `gs://my-bucket/cellplex/cellranger_output`. It should contain 6 subfolders, each of which is associated with one sample in `cellranger_sample_sheet.csv`.

## 2. Demultiplexing

Next, we need to demultiplex the resulting gene-count matrices. In this example, we perform both [DemuxEM](#) and [Souporecell](#) methods, respectively.

For **DemuxEM**, we'll need the RNA raw count matrix in HDF5 format (`gs://my-bucket/cellplex/cellranger_output/cellplex_gex/raw_feature_bc_matrix.h5`) and the hashing count matrix in CSV format (`gs://my-bucket/cellplex/cellranger_output/cellplex_barcode/cellplex_barcode.csv`).

For **Souporecell**, both the RNA raw count matrix above and its corresponding BAM file (`gs://my-bucket/cellplex/cellranger_output/cellplex_gex/possorted_genome_bam.bam`) are needed.

Prepare a sample sheet in CSV format (say named `demux_sample_sheet.csv`) for demultiplexing, one line for DemuxEM, one for Souporecell:

```
OUTNAME, RNA, TagFile, TYPE
cellplex_demux, gs://my-bucket/cellplex/cellranger_output/cellplex_gex/raw_feature_bc_
↪matrix.h5, gs://my-bucket/cellplex/cellranger_output/cellplex_barcode/cellplex_
↪barcode.csv, cell-hashing
cellplex_souporcell, gs://my-bucket/cellplex/cellranger_output/cellplex_gex/raw_
↪feature_bc_matrix.h5, gs://my-bucket/cellplex/cellranger_output/cellplex_gex/
↪possorted_genome_bam.bam, genetic-pooling
```

where

- cell-hashing indicates using DemuxEM for demultiplexing, while genetic-pooling indicates using genetic pooling methods for demultiplexing, with Souporcell being the default.

For details on this sample sheet, please refer to [Demultiplexing workflow sample sheet format](#).

Then prepare a workflow input JSON file (say named `demux_inputs.json`) for demultiplexing:

```
{
  "demultiplexing.input_sample_sheet": "/path/to/demux_sample_sheet.csv",
  "demultiplexing.output_directory": "gs://my-bucket/cellplex/demux_output",
  "demultiplexing.genome": "GRCh38-2020-A",
  "demultiplexing.souporcell_num_clusters": 3
}
```

where

- `/path/to/demux_sample_sheet.csv` should be replaced by the local path to your `demux_sample_sheet.csv` created above.
- `gs://my-bucket/cellplex/demux_output` is the Bucket folder to write the results when the job is finished.
- `GRCh38-2020-A` is the genome reference used by Souporcell, which should be consistent with your settings in Step 1.
- `souporcell_num_clusters` is to set the number of clusters you expect to see for Souporcell clustering. Since we have 3 donors, so set it to 3.

For details, please refer to [Demultiplexing workflow inputs](#).

Now submit the demultiplexing job to Cromwell server on Cloud:

```
alto cromwell run -s 10.0.0.0 -p 8000 -m broadinstitute:cumulus:demultiplexing:master_
↪-i demux_inputs.json -o demux_inputs_updated.json -b gs://my-bucket/cellplex
```

where

- `broadinstitute:cumulus:demultiplexing` refers to [demultiplexing](#) workflow published on Dockerstore.
- We still need `-o` and `-b` options because `demux_sample_sheet.csv` is on the local machine.

Similarly, when the submission succeeds, you'll get another job ID for demultiplexing. You can use it to track the job status.

When finished, below are the important output files:

- DemuxEM output: In folder `gs://my-bucket/cellplex/demux_output/cellplex_demux`,
  - `cellplex_demux_demux.zarr.zip`: Demultiplexed RNA raw count matrix. This will be used for downstream analysis.

- `cellplex_demux.out.demuxEM.zarr.zip`: This file contains intermediate results for both RNA and hashing count matrices, which is useful for compare with other demultiplexing methods.
- DemuxEM plots in PDF format. They are used for estimating the performance of DemuxEM on the data.
- **Souporcell output:** In folder `gs://my-bucket/cellplex/demux_output/cellplex_souporcell`,
  - `cellplex_souporcell_demux.zarr.zip`: Demultiplexed RNA raw count matrix. This will be used for downstream analysis.
  - `clusters.tsv`: Inferred droplet type and cluster assignment for each cell barcode.
  - `cluster_genotypes.vcf`: Inferred genotypes for each cluster.

### 3. Interactive Data Analysis

You may use [Cumulus workflow](#) to perform the downstream analysis in a batch way. Alternatively, you can also download the demultiplexing results from the Cloud bucket to your local machine, and perform the analysis interactively. This section introduces how to use Cumulus' analysis module Pegasus to load demultiplexing results, perform quality control (QC), and compare the performance of the two methods.

You'll need to first install [Pegasus](#) in your local Python environment. Also, download the demultiplexed raw counts in `.zarr.zip` format mentioned above to your local machine.

#### 3.1. Extract Singlet/Doublet Type and Assignment

We can load the DemuxEM result, and perform QC by:

```
import pegasus as pg
data_demuxEM = pg.read_input("cellplex_demux_demux.zarr.zip")
pg.qc_metrics(data_demuxEM, min_genes=500, max_genes=6000, mito_prefix='MT-', percent_
↪mito=20)
pg.filter_data(data_demuxEM)
```

where `qc_metrics` and `filter_data` are Pegasus functions to filter out low quality cells, and keep those with number of genes within range [500, 6000) and having expression of mitochondrial genes < 20%. Please see [Pegasus preprocess tools](#) for details.

There are two columns in `data_demuxEM.obs` field related to demultiplexing results:

- **demux\_type**: This column stores the singlet/doublet type of each cell: `singlet`, `doublet`, or `unknown`.
- **assignment**: This column stores the more detailed assignment of cells regarding samples/donors.

To get the distribution regarding these columns, e.g. `demux_type`:

```
data_demuxEM.obs['demux_type'].value_counts()
```

Besides, you can export the cell barcodes along with their singlet/doublet type and assignment as a CSV file by:

```
data_demuxEM.obs[['demux_type', 'assignment']].to_csv("demuxEM_assignment.csv")
```

We can also do it similarly for the Souporcell result as above, by reading `cellplex_souporcell_demux.zarr.zip` instead.

### 3.2. Compare the Two Demultiplexing Methods

We can compare the performance of DemuxEM and Souporcell by plotting a heatmap showing their singlet/doublet assignment results.

Assume we've already loaded the two results (`data_demuxEM` for DemuxEM result, `data_souporcell` for Souporcell result), and performed QC as in 3.1. The following Python code will generate this heatmap in an interactive Python environment (e.g. in a Jupyter notebook):

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

def extract_assignment(data):
    assign = data.obs['demux_type'].values.astype('object')
    idx_singlet = (data.obs['demux_type'] == 'singlet').values
    assign[idx_singlet] = data.obs.loc[idx_singlet, 'assignment'].values.
    ↪astype(object)
    return assign

assign_demuxEM = extract_assignment(data_demuxEM)
assign_souporcell = extract_assignment(data_souporcell)

df = pd.crosstab(assign_demuxEM, assign_souporcell)
df.columns.name = df.index.name = ""
ax = plt.gca()
ax.xaxis.tick_top()
ax = sns.heatmap(df, annot=True, fmt='d', cmap='inferno', ax=ax)
plt.tight_layout()
plt.gcf().dpi=500
```

### 3.3. Downstream Analysis

To perform further downstream analysis on the singlets, please refer to [Pegasus tutorials](#).

Examples using Terra to perform single-cell sequencing analysis are provided here. Please click the topics on the left panel under title “**Examples**” to explore.

#### 1.1.12 10x Visium

##### Run Space Ranger tools using `spaceranger_workflow`

`spaceranger_workflow` wraps Space Ranger to process 10x Visium data.

##### A general step-by-step instruction

This section mainly considers jobs starting from BCL files. If your job starts with FASTQ files, and only need to run `spaceranger count` part, please refer to [this subsection](#).

##### 1. Import `spaceranger_workflow`

Import `spaceranger_workflow` workflow to your workspace by following instructions in [Import workflows to Terra](#). You should choose workflow [github.com/lilab-bcb/cumulus/Spaceranger](https://github.com/lilab-bcb/cumulus/Spaceranger) to import.

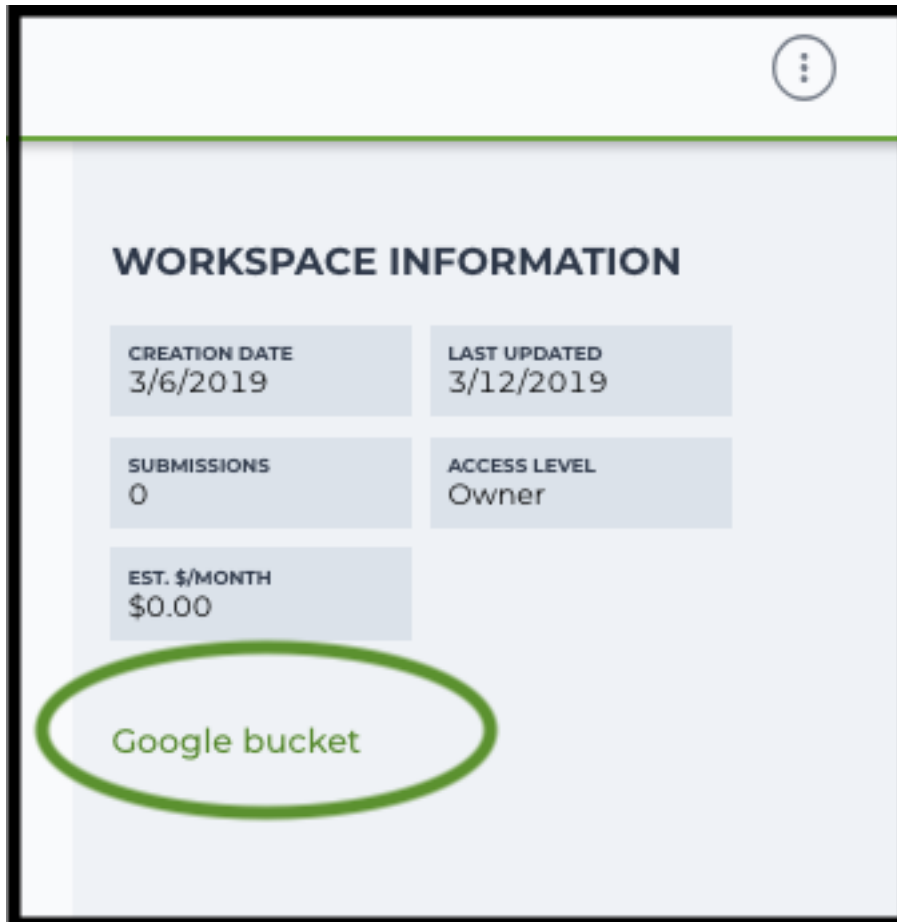


Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export *spaceranger\_workflow* workflow in the drop-down menu.

## 2. Upload sequencing and image data to Google bucket

Copy your sequencing output to your workspace bucket using `gsutil` (you already have it if you've installed Google cloud SDK) in your unix terminal.

You can obtain your bucket URL in the dashboard tab of your Terra workspace under the information panel.



Use `gsutil cp [OPTION]... src_url dst_url` to copy data to your workspace bucket. For example, the following command copies the directory at `/foo/bar/nextseq/Data/VK18WBC6Z4` to a Google bucket:

```
gsutil -m cp -r /foo/bar/nextseq/Data/VK18WBC6Z4 gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4
```

`-m` means copy in parallel, `-r` means copy the directory recursively, and `gs://fc-e0000000-0000-0000-0000-000000000000` should be replaced by your own workspace Google bucket URL.

Similarly, copy all images for spatial data to the same google bucket.

**Note:** If input is a folder of BCL files, users do not need to upload the whole folder to the Google bucket. Instead,

they only need to upload the following files:

```
RunInfo.xml
RTAComplete.txt
runParameters.xml
Data/Intensities/s.locs
Data/Intensities/BaseCalls
```

If data are generated using MiSeq or NextSeq, the location files are inside lane subfolders L001 under Data/Intensities/. In addition, if users' data only come from a subset of lanes (e.g. L001 and L002), users only need to upload lane subfolders from the subset (e.g. Data/Intensities/BaseCalls/L001, Data/Intensities/BaseCalls/L002 and Data/Intensities/L001, Data/Intensities/L002 if sequencer is MiSeq or NextSeq).

---

Alternatively, users can submit jobs through command line interface (CLI) using [altocumulus](#), which will smartly upload BCL folders according to the above rules.

### 3. Prepare a sample sheet

#### 3.1 Sample sheet format:

Please note that the columns in the CSV can be in any order, but that the column names must match the recognized headings.

For **FFPE** data, **ProbeSet** column is mandatory.

The sample sheet describes how to demultiplex flowcells and generate channel-specific count matrices. Note that *Sample*, *Lane*, and *Index* columns are defined exactly the same as in 10x's simple CSV layout file.

A brief description of the sample sheet format is listed below (**required column headers are shown in bold**).

Column	Description
<b>Sample</b>	Contains sample names. Each 10x channel should have a unique sample name.
<b>Reference</b>	<p>Provides the reference genome used by Space Ranger for each 10x channel.</p> <p>The elements in the <i>reference</i> column can be either Google bucket URLs to reference tarballs or keywords such as <i>GRCh38-2020-A</i>.</p> <p>A full list of available keywords is included in each of the following data type sections (e.g. sc/snRNA-seq) below.</p>
<b>Flowcell</b>	<p>Indicates the Google bucket URLs of uploaded BCL folders.</p> <p>If starts with FASTQ files, this should be Google bucket URLs of uploaded FASTQ folders.</p> <p>The FASTQ folders should contain one subfolder for each sample in the flowcell with the sample name as the subfolder name.</p> <p>Each subfolder contains FASTQ files for that sample.</p>
<b>Lane</b>	<p>Tells which lanes the sample was pooled into.</p> <p>Can be either single lane (e.g. 8) or a range (e.g. 7-8) or all (e.g. *).</p>
<b>Index</b>	Sample index (e.g. SI-GA-A12).
ProbeSet	Probe set for FFPE samples. <b>Choosing</b> from <code>human_probe_v1</code> (10x human probe set, CytoAssist-incompatible), <code>human_probe_v2</code> (10x human probe set, CytoAssist-compatible) and <code>mouse_probe_v1</code> (10x mouse probe set). Alternatively, a CSV file describing the probe set can be directly used. Setting ProbeSet to "" for a sample implies the sample is not FFPE.
Image	Cloud bucket url for a brightfield tissue H&E image in .jpg or .tiff format. This column is mutually exclusive with DarkImage and ColorizedImage columns.
DarkImage	Cloud bucket urls for Multi-channel, dark-background fluorescence image as either a single, multi-layer .tiff file, multiple .tiff or .jpg files, or a pre-combined color .tiff or .jpg file. If multiple files are provided, please separate them by ';'. This column is mutually exclusive with Image and ColorizedImage columns.
ColorizedImage	Cloud bucket url for a color composite of one or more fluorescence image channels saved as a single-page, single-file color .tiff or .jpg. This column is mutually exclusive with Image and DarkImage columns.
CytImage	Cloud bucket url for a brightfield image generated by the CytAssist instrument.
Slide	Visium slide serial number. If both Slide and Area are empty, the –unknown-slide option would be set.
Area	Visium capture area identifier. Options for Visium are A1, B1, C1, D1. If both Slide and Area are empty, the –unknown-slide option would be set.
SlideFile	Slide layout file indicating capture spot and fiducial spot positions. Only required if internet access is not available.
LoupeAlignment	Alignment file produced by the manual Loupe alignment step.
TargetPanel	Cloud bucket url for a target panel CSV for targeted gene expression analysis.

The sample sheet supports sequencing the same 10x channels across multiple flowcells. If a sample is sequenced across multiple flowcells, simply list it in multiple rows, with one flowcell per row. In the following example, we have 2 samples sequenced in two flowcells.

Example:

```
Sample,Reference,Flowcell,Lane,Index,Image,Slide,Area
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↳VK18WBC6Z4,1-2,SI-GA-A8,gs://image/image1.tif,V19J25-123,A1
sample_2,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↳VK18WBC6Z4,3-4,SI-GA-B8,gs://image/image2.tif,V19J25-123,B1
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↳VK10WBC9Z2,1-2,SI-GA-A8,gs://image/image1.tif,V19J25-123,A1
sample_2,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/
↳VK10WBC9Z2,3-4,SI-GA-B8,gs://image/image2.tif,V19J25-123,B1
```

### 3.2 Upload your sample sheet to the workspace bucket:

Example:

```
gsutil cp /foo/bar/projects/sample_sheet.csv gs://fc-e0000000-0000-
↳0000-0000-000000000000/
```

## 4. Launch analysis

In your workspace, open `spaceranger_workflow` in **WORKFLOWS** tab. Select the desired snapshot version (e.g. latest). Select **Run workflow with inputs defined by file paths** as below

- ☒ Run workflow with inputs defined by file paths  
☐ Run workflow(s) with inputs defined by data table

and click **SAVE** button. Select **Use call caching** and click **INPUTS**. Then fill in appropriate values in the **Attribute** column. Alternative, you can upload a JSON file to configure input by clicking **Drag** or click to upload json.

Once **INPUTS** are appropriated filled, click **RUN ANALYSIS** and then click **LAUNCH**.

## 5. Notice: run `spaceranger mkfastq` if you are non Broad Institute users

Non Broad Institute users that wish to run `spaceranger mkfastq` must create a custom docker image that contains `bcl2fastq`.

See [bcl2fastq](#) instructions.

## 6. Run `spaceranger count` only

Sometimes, users might want to perform demultiplexing locally and only run the count part on the cloud. This section describes how to only run the count part via `spaceranger_workflow`.

1. Copy your FASTQ files to the workspace using `gsutil` in your unix terminal. There are two cases:
  - **Case 1:** All the FASTQ files are in one top-level folder. Then you can simply upload this folder to Cloud, and in your sample sheet, make sure **Sample** names are consistent with the filename prefix of their corresponding FASTQ files.

- **Case 2:** In the top-level folder, each sample has a dedicated subfolder containing its FASTQ files. In this case, you need to upload the whole top-level folder, and in your sample sheet, make sure **Sample** names and their corresponding subfolder names are identical.

Notice that if your FASTQ files are downloaded from the Sequence Read Archive (SRA) from NCBI, you must rename your FASTQs to follow the [bcl2fastq file naming conventions](#).

Example:

```
gsutil -m cp -r /foo/bar/fastq_path/K18WBC6Z4 gs://fc-e0000000-0000-0000-0000-000000000000/K18WBC6Z4_fastq
```

2. Create a sample sheet following the similar structure as [above](#), except the following differences:

- **Flowcell** column should list Google bucket URLs of the FASTQ folders for flowcells.
- **Lane** and **Index** columns are NOT required in this case.

Example:

```
Sample,Reference,Flowcell,Image,Slide,Area
sample_1,GRCh38-2020-A,gs://fc-e0000000-0000-0000-0000-000000000000/K18WBC6Z4_fastq,gs://image/image1.tif,V19J25-123,A1
```

3. Set optional input `run_mkfastq` to `false`.

## Visium spatial transcriptomics data

To process spatial transcriptomics data, follow the specific instructions below.

### Sample sheet

1. **Reference** column.

Pre-built scRNA-seq references are summarized below.

Keyword	Description
<b>GRCh38-2020-A</b>	Human GRCh38 (GENCODE v32/Ensembl 98)
<b>mm10-2020-A</b>	Mouse mm10 (GENCODE vM23/Ensembl 98)

### Workflow input

For spatial data, `spaceranger_workflow` takes Illumina outputs and related images as input and runs `spaceranger mkfastq` and `spaceranger count`. Relevant workflow inputs are described below, with required inputs highlighted in bold.

Name	Description	Example	Default
<b>input_sample_sheet</b>	File Sheet (contains Sample, Reference, Flowcell, Lane, Index as required and ProbeSet, Image, DarkImage, ColorizedImage, CytaImage, Slide, Area, SlideFile, LoupeAlignment, TargetPanel as optional)	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.csv”	
<b>output_directory</b>	Output directory	“gs://fc-e0000000-0000-0000-0000-000000000000/spaceranger_output”	Results are written under directory <i>output_directory</i> and will overwrite any existing files at this location.
<b>run_mkfastq</b>	you want to run spaceranger mkfastq	true	true
<b>run_count</b>	you want to run spaceranger count	true	true
<b>delete_input_bcl_directories</b>	Delete BCL directories after demux. If false, you should delete this folder yourself so as to not incur storage charges	false	false
<b>mkfastq_number_of_mismatches</b>	Number of mismatches allowed in matching barcode indices (bcl2fastq2 default is 1)	0	
<b>reorient_images</b>	Reorient images with automatic fiducial alignment. This option will apply to all samples in the sample sheet. Spaceranger will attempt to find the best alignment of the fiducial markers by any rotation or mirroring of the image.	true	true
<b>filter_probe_set</b>	Whether to filter the probe set using the “included” column of the probe set CSV.	true	true
<b>dapi_index</b>	Index of DAPI channel (1-indexed) of fluorescence image, only used in the CytaAssist case, with dark background image.	2	
<b>unknown_slide</b>	Use slide option if the slide serial number and area identifier have been lost. Choose from visium-1, visium-2 and visium-2-large.	visium-2	
<b>no_bam</b>	Turn this option on to disable BAM file generation.	false	false
<b>secondary</b>	Perform Space Ranger secondary analysis (dimensionality reduction, clustering, etc.)	false	false
<b>r1_length</b>	Trim the input Read 1 to this length before analysis	28	
<b>r2_length</b>	Trim the input Read 2 to this length before analysis. This value will be set to 50 automatically for FFPE samples if spaceranger version < 2.0.0.	50	
<b>spaceranger_version</b>	Spaceranger version, could be: 2.0.1, 2.0.0, 1.3.1, 1.3.0	“2.0.1”	2.0.1
<b>config_version</b>	Config version used for processing sample sheets, could	“0.3”	“0.3”

## Workflow output

See the table below for important sc/snRNA-seq outputs.

Name	Type	Description
fastq_outputs	Array[String]?	A list of cloud urls containing FASTQ files, one url per flowcell.
count_outputs	Array[String]?	A list of cloud urls containing spaceranger count outputs, one url per sample.
metrics_summaries	File?	A excel spreadsheet containing QCs for each sample.
spaceranger_count.output_urls	Array[String]?	A list of htmls visualizing QCs for each sample (spaceranger count output).

## Build Space Ranger References

Reference built by Cell Ranger for sc/snRNA-seq should be compatible with Space Ranger. For more details on building references using Cell Ranger, please refer to [here](#).

### 1.1.13 Nanostring GeoMx DSP

This section contains two workflows: **geomxngs\_fastq\_to\_dcc** and **geomxngs\_dcc\_to\_count\_matrix**.

**geomxngs\_fastq\_to\_dcc** workflow wraps [Nanostring GeoMx Digital Spatial NGS Pipeline](#) and can convert FASTQ files into DCC files.

**geomxngs\_dcc\_to\_count\_matrix** workflow takes the DCC zip file from **geomxngs\_fastq\_to\_dcc** and other files produced by the GeoMx DSP machine as inputs, and outputs an area of illumination (AOI) by probe count matrix with pathologists' annotation.

## Convert FASTQ files into DCC files by the Nanostring GeoMx Digital Spatial NGS Pipeline

The **geomxngs\_fastq\_to\_dcc** workflow converts FASTQ files to DCC files by wrapping the [Nanostring GeoMx Digital Spatial NGS Pipeline](#). After generating DCC files, use the **geomxngs\_dcc\_to\_count\_matrix** workflow to generate an area of interest by probe count matrix.

## Workflow Input

Relevant workflow inputs are described below (required inputs in bold)

Name	Description	Example	Default
<b>fastq_directory</b>	FASTQ directory URL	“gs://foo/bar/fastqs” or “s3://foo/bar/fastqs”	
<b>ini</b>	Configuration file in INI format, containing pipeline processing parameters	“gs://foo/bar/config.ini”	
<b>output_directory</b>	Directory to write results	“gs://foo/bar/out” or “s3://foo/bar/out”	
<b>fastq_rename</b>	Optional 2 column TSV file with no header used to map original FASTQ names to FASTQ names that GeoMX recognizes.	“gs://foo/bar/fastq_rename.tsv”	
<b>delete_fastq_directory</b>	Whether to delete the input fastqs upon successful completion	true	false
<b>geomxngs_version</b>	Version of the geomx software, currently only “2.3.3.10”.	“2.3.3.10”	“2.3.3.10”
<b>docker_registry</b>	Docker registry to use for this workflow. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>backend</b>	<b>Backend for computation. Available options:</b> <ul style="list-style-type: none"> <li>• “gcp” for Google Cloud</li> <li>• “aws” for Amazon AWS</li> <li>• “local” for local machine</li> </ul>	“aws”	“gcp”
<b>zones</b>	Google cloud zones	“us-central1-a”	“us-central1-a us-central1-b us-central1-c us-central1-f”
<b>preemptible</b>	Number of preemptible tries	2	2
<b>memory</b>	Memory string	“64GB”	“64GB”
<b>cpu</b>	Number of CPUs	4	4
<b>disk_space</b>	Disk space in GB	500	500
<b>aws_queue</b>	The ARN URI of the AWS job queue to be used. Only works when backend is aws.	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf”	“”

## Workflow Output

Name	Description	Type
<b>dcc_zip</b>	URL to the output DCC zip file	String
<b>geomxngs_output</b>	URL to the output of geomxngspipeline; the DCC zip file is part of the output here	String

## Generate probe count matrix with pathologists’ annotation

The **geomxngs\_dcc\_to\_count\_matrix** workflow generates an area of illumination (AOI) by probe count matrix with pathologists’ annotation from the output of the **geomxngs\_fastq\_to\_dcc** workflow and user inputs.



## Workflow Input

Workflow inputs are described below (required inputs in bold).

Name	Description	Example	Default
<b>dcc_zip</b>	DCC zip file from <b>geomxngs_fastq_to_dcc</b> workflow output	“gs://foo/bar/out/DCC-20221001.zip”	
<b>ini</b>	Configuration file in INI format, containing pipeline processing parameters	“gs://foo/bar/config.ini”	
<b>lab_worksheet</b>	Excel file containing library setups	“gs://foo/bar/LabWorksheet.txt”	
<b>dataset</b>	Data QC and annotation file (Excel) downloaded from instrument after uploading DCC zip file; we only use the first tab (SegmentProperties)	“gs://foo/bar/BioprobeQC.xlsx”	
<b>pkc</b>	GeoMx DSP configuration file to associate assay targets with GeoMx HybCode barcodes and Seq Code primers. Options: - CTA_v1.0-4 for Cancer Transcriptome Atlas - COVID-19_v1.0 for COVID-19 Immune Response Atlas - Human_WTA_v1.0 for Human Whole Transcriptome Atlas - Mouse_WTA_v1.0 for Mouse Whole Transcriptome Atlas If your configuration file is not listed, you can provide a URL to a PKC zip file or PKC file instead.	“Human_WTA_v1.0”	
<b>output_directory</b>	Directory to write results	“gs://foo/bar/out” or “s3://foo/bar/out”	
<b>backend</b>	Backend for computation. Available options: - “gcp” for Google Cloud - “aws” for Amazon AWS - “local” for local machine	“aws”	“gcp”
<b>docker_registry</b>	Docker registry to use for this workflow. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>docker_version</b>	Docker image version.	“1.0.0”	“1.0.0”
<b>preemptible</b>	Number of preemptible tries	2	2
<b>memory</b>	Memory string	“8GB”	“8GB”
<b>cpu</b>	Number of CPUs	1	1
<b>extra_disk_space</b>	Extra disk space in GB.	5	5
<b>aws_queue_name</b>	The name URI of the AWS job queue to be used. Only works when backend is aws	“arn:aws:batch:us-east-1:xxx:job-queue/priority-gwf”	“”

## Workflow Output

Name	Description	Type
count_matrix_h5ad	URL to a count matrix in h5ad format. X contains the count matrix, obs contains AOI information, and .var contains probe metadata	String
count_matrix_text	URL to a count matrix in text format. Each row is one probe and each column is one AOI. First column is RTS_ID (Readout Tag Sequence-ID (RTS-ID)). Second column is Gene (if multiple probes map to the same gene, their values are the same). Third columns is Probe (if multiple probes map to the same gene, values are different control_1, control_2). Starting from column 4, we have counts.	String
count_matrix_metadata	URL to a count matrix metadata in text format. All columns from dataset file are included; each row describes one AOI (area of illumination)	String

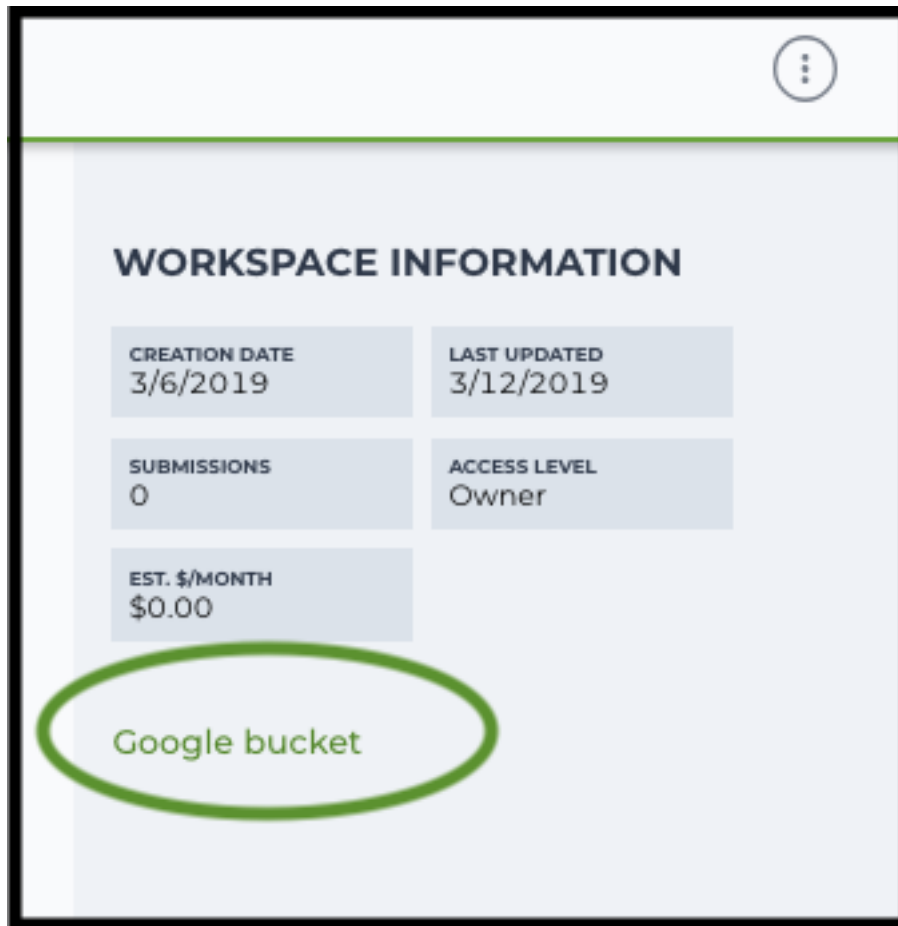
### 1.1.14 Extract gene-count matrices from plated-based SMART-Seq2 data

#### Run SMART-Seq2 Workflow

Follow the steps below to extract gene-count matrices from SMART-Seq2 data on [Terra](#). This WDL aligns reads using *STAR*, *HISAT2*, or *Bowtie 2* and estimates expression levels using *RSEM*.

1. Copy your sequencing output to your workspace bucket using [gsutil](#) in your unix terminal.

You can obtain your bucket URL in the dashboard tab of your Terra workspace under the information panel.



Use `gsutil cp [OPTION]... src_url dst_url` to copy data to your workspace bucket. For example, the following command copies the directory at `/foo/bar/nextseq/Data/VK18WBC6Z4` to a Google bucket:

```
gsutil -m cp -r /foo/bar/nextseq/Data/VK18WBC6Z4 gs://fc-e0000000-0000-
↪0000-0000-000000000000/VK18WBC6Z4
```

`-m` means copy in parallel, `-r` means copy the directory recursively.

## 2. Create a sample sheet.

Please note that the columns in the TSV can be in any order, but that the column names must match the recognized headings.

The sample sheet provides metadata for each cell:

Column	Description
entity:sample	Cell name.
plate	Plate name. Cells with the same plate name are from the same plate.
read1	Location of the FASTQ file for read1 in the cloud (gsurl).
read2	(Optional). Location of the FASTQ file for read2 in the cloud (gsurl). This field can be skipped for single-end reads.

Example:

```
entity:sample  plate  read1  read2
cell-1  plate-1  gs://fc-e0000000-0000-0000-0000-000000000000/smartseq2/
↪cell-1_L001_R1_001.fastq.gz      gs://fc-e0000000-0000-0000-0000-
↪000000000000/smartseq2/cell-1_L001_R2_001.fastq.gz
cell-2  plate-1  gs://fc-e0000000-0000-0000-0000-000000000000/smartseq2/
↪cell-2_L001_R1_001.fastq.gz      gs://fc-e0000000-0000-0000-0000-
↪000000000000/smartseq2/cell-2_L001_R2_001.fastq.gz
cell-3  plate-2  gs://fc-e0000000-0000-0000-0000-000000000000/smartseq2/
↪cell-3_L001_R1_001.fastq.gz
cell-4  plate-2  gs://fc-e0000000-0000-0000-0000-000000000000/smartseq2/
↪cell-4_L001_R1_001.fastq.gz
```

3. Upload your sample sheet to the workspace bucket.

Example:

```
gsutil cp /foo/bar/projects/sample_sheet.csv gs://fc-e0000000-0000-0000-
↪0000-000000000000/
```

4. Import *smartseq2* workflow to your workspace.

Import by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus/Smart-Seq2](https://github.com/lilab-bcb/cumulus/Smart-Seq2) to import.

Moreover, in the workflow page, click **Export to Workspace...** button, and select the workspace to which you want to export *smartseq2* workflow in the drop-down menu.

5. In your workspace, open *smartseq2* in **WORKFLOWS** tab. Select **Run workflow with inputs defined by file paths** as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click **SAVE** button.

### Inputs:

Please see the description of inputs below. Note that required inputs are shown in bold.

Name	Description	Example	Default
input_tsv	TSV file (contains entity:sample, plate, read1, read2)	<code>"gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.tsv"</code>	
output_directory	Output directory	<code>"gs://fc-e0000000-0000-0000-0000-000000000000/smartseq2_output"</code>	
reference	Reference transcriptome to align reads to. Acceptable values: <ul style="list-style-type: none"> <li>Pre-created genome references: <ul style="list-style-type: none"> <li>"GRCh38_ens93filt" for human, genome version is GRCh38, gene annotation is generated using human Ensembl 93 GTF according to <a href="#">cellranger mkgtf</a>;</li> <li>"GRCm38_ens93filt" for mouse, genome version is GRCm38, gene annotation is generated using mouse Ensembl 93 GTF according to <a href="#">cellranger mkgtf</a>;</li> </ul> </li> <li>Create a custom genome reference using <a href="#">smartseq2_create_reference</a> workflow, and specify its Google bucket URL here.</li> </ul>	<code>"GRCh38_ens93filt"</code> , or <code>"gs://fc-e0000000-0000-0000-0000-000000000000/rsem_ref.tar.gz"</code>	
aligner	Which aligner to use for read alignment. Options are "hisat2-hca", "star" and "bowtie"	<code>"star"</code>	<code>"hisat2-hca"</code>
output_genome_bam	Whether to output bam file with alignments mapped to genomic coordinates and annotated with their posterior probabilities.	<code>false</code>	<code>false</code>
smartseq2_version	SMART-seq2 version to use. Versions available: 1.3.0.	<code>"1.3.0"</code>	<code>"1.3.0"</code>
docker_registry	Docker registry to use. Options: <ul style="list-style-type: none"> <li>"quay.io/cumulus" for images on Red Hat registry;</li> <li>"cumulusprod" for backup images on Docker Hub.</li> </ul>	<code>"quay.io/cumulus"</code>	<code>"quay.io/cumulus"</code>
zones	Google cloud zones	<code>"us-east1-d us-west1-a us-west1-b"</code>	<code>"us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c"</code>
<b>1.1. 2.4.0 January 28, 2023</b>			<b>145</b>
num_cpus	Number of cpus to request for one node	<code>4</code>	<code>4</code>
memory	Memory size string	<code>"3.60G"</code>	If

## Outputs:

Name	Type	Description
output_count_matrix	String	Point to a Google bucket URL for count matrix in matrix market format.
rsem_trans_bam	Array[String?]	An array of Google bucket URLs for RSEM transcriptomic BAM files
rsem_genome_bam	Array[String?]	An array of Google bucket URLs for RSEM genomic BAM files if <code>output_genome_bam</code> is <code>true</code> .
rsem_gene	Array[File?]	An array of RSEM gene expression estimation files.
rsem_isoform	Array[File?]	An array of RSEM isoform expression estimation files.
rsem_time	Array[File?]	An array of RSEM execution time log files.
aligner_log	Array[File?]	An array of Aligner log files.
rsem_cnt	Array[File?]	An array of RSEM count files.
rsem_model	Array[File?]	An array of RSEM model files.
rsem_theta	Array[File?]	An array of RSEM generated theta files.

This WDL generates one gene-count matrix in matrix market format:

- `output_count_matrix` is a folder containing three files: `matrix.mtx.gz`, `barcodes.tsv.gz`, and `features.tsv.gz`.
- `matrix.mtx.gz` is a gzipped matrix in matrix market format.
- `barcodes.tsv.gz` is a gzipped TSV file, containing 5 columns. ‘barcodekey’ is cell name. ‘plate’ is the plate name, which can be used for batch correction. ‘total\_reads’ is the total number of reads. ‘alignment\_rate’ is the alignment rate obtained from the aligner. ‘unique\_rate’ is the percentage of reads aligned uniquely to a gene. Cells sequenced with single-end reads appear first in ‘barcodekey’.
- `features.tsv.gz` is a gzipped TSV file, containing 2 columns. ‘featurekey’ is gene symbol. ‘featureid’ is Ensembl ID.

The gene-count matrix can be fed directly into **cumulus** for downstream analysis.

TPM-normalized counts are calculated as follows:

1. Estimate the gene expression levels in TPM using *RSEM*.
2. Suppose  $c$  reads are achieved for one cell, then calculate TPM-normalized count for gene  $i$  as  $\text{TPM}_i / 1e6 * c$ .

TPM-normalized counts reflect both the relative expression levels and the cell sequencing depth.

---

## Custom Genome

We also provide a way of generating user-customized Genome references for SMART-Seq2 workflow.

1. Import `smartseq2_create_reference` workflow to your workspace.

Import by following instructions in [Import workflows to Terra](#). You should choose [github.com/lilab-bcb/cumulus/Smart-Seq2\\_create\\_reference](https://github.com/lilab-bcb/cumulus/Smart-Seq2_create_reference) to import.

Moreover, in the workflow page, click `Export to Workflow...` button, and select the workspace to which you want to export `smartseq2_create_reference` in the drop-down menu.

2. In your workspace, open `smartseq2_create_reference` in **WORKFLOWS** tab. Select `Run workflow` with inputs defined by file paths as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click `SAVE` button.

### Inputs:

Please see the description of inputs below. Note that required inputs are shown in bold.

Name	Description	Type or Example	Default
<b>fasta</b>	Genome fasta file	File. For example, “gs://fc-e0000000-0000-0000-0000-000000000000/Homo_sapiens.GRCh38.dna.primary_assembly.fa”	
<b>gtf</b>	GTF gene annotation file (e.g. Homo_sapiens.GRCh38.83.gtf)	File. For example, “gs://fc-e0000000-0000-0000-0000-000000000000/Homo_sapiens.GRCh38.83.gtf”	
<b>output_directory</b>	S3 bucket url for the output folder	“gs://fc-e0000000-0000-0000-0000-000000000000/output_refs”	
<b>genome</b>	Output reference genome name. Output reference is a gzipped tarball with name genome_aligner.tar.gz	“GRCm38_ens97filt”	
<b>aligner</b>	Build indices for which aligner, choices are hisat2-hca, star, or bowtie2.	“hisat2-hca”	“hisat2-hca”
<b>smartseq2_version</b>	SMART-Seq2 version to use. Versions available: 1.3.0.	“1.3.0”	“1.3.0”
<b>docker_registry</b>	Docker registry to use. Options: <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>zones</b>	Google cloud zones	“us-central1-c”	“us-central1-b”
<b>cpu</b>	Number of CPUs	Integer	If aligner is bowtie2 or hisat2-hca, 8; otherwise 32
<b>memory</b>	Memory size string	String	If aligner is bowtie2 or hisat2-hca, “7.2G”; otherwise “120G”
<b>148</b>		<b>Chapter 1. Release Highlights in Current Stable</b>	



## Outputs

Name	Type	Description
output_reference	File	The custom Genome reference generated. Its default file name is <code>genome_aligner.tar.gz</code> .
monitoring_log	File	CPU and memory profiling log.

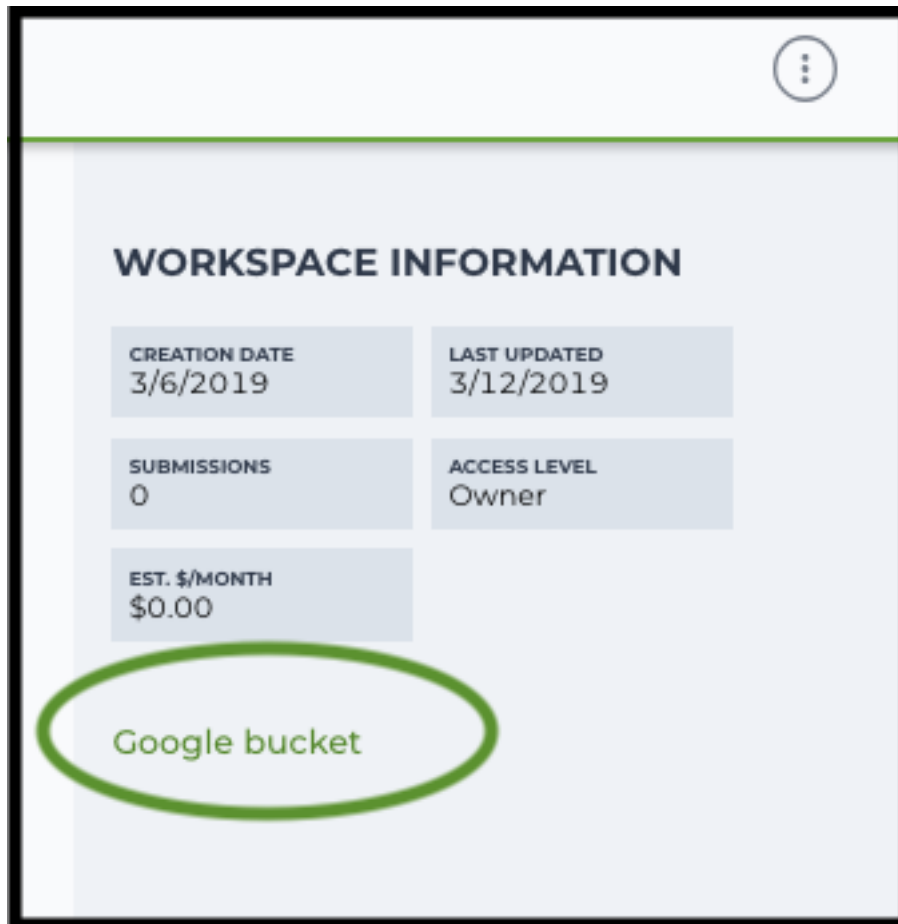
### 1.1.15 Bulk RNA-Seq

#### Run Bulk RNA-Seq Workflow

Follow the steps below to generate count matrices from bulk RNA-Seq data on [Terra](#). This WDL estimates expression levels using *RSEM*.

1. Copy your sequencing output to your workspace bucket using [gsutil](#) in your unix terminal.

You can obtain your bucket URL in the dashboard tab of your Terra workspace under the information panel.



Note: Broad users need to be on an UGER node (not a login node) in order to use the `-m` flag

Request an UGER node:

```
reuse UGER
qrsh -q interactive -l h_vmem=4g -pe smp 8 -binding linear:8 -P regevlab
```

The above command requests an interactive node with 4G memory per thread and 8 threads. Feel free to change the memory, thread, and project parameters.

Once you're connected to an UGER node, you can make [gsutil](#) available by running:

```
reuse Google-Cloud-SDK
```

Use `gsutil cp [OPTION]... src_url dst_url` to copy data to your workspace bucket. For example, the following command copies the directory at `/foo/bar/nextseq/Data/VK18WBC6Z4` to a Google bucket:

```
gsutil -m cp -r /foo/bar/nextseq/Data/VK18WBC6Z4 gs://fc-e0000000-0000-
↪0000-0000-000000000000/VK18WBC6Z4
```

`-m` means copy in parallel, `-r` means copy the directory recursively.

## 2. Create a Terra [data table](#)

Example:

```
entity:sample_id  read1 read2
sample-1  gs://fc-e0000000/data-1/sample1-1_L001_R1_001.fastq.gz    gs://
↪fc-e0000000/data-1/sample-1_L001_R2_001.fastq.gz
sample-2  gs://fc-e0000000/data-1/sample-2_L001_R1_001.fastq.gz    gs://fc-
↪e0000000/data-1/sample-2_L001_R2_001.fastq.gz
```

You are free to add more columns, but sample ids and URLs to fastq files are required.

3. Upload your TSV file to your workspace. Open the **DATA** tab on your workspace. Then click the upload button on left **TABLE** panel, and select the TSV file above. When uploading is done, you'll see a new data table with name "sample":
4. Import *bulk\_rna\_seq* workflow to your workspace. Then open *bulk\_rna\_seq* in the **WORKFLOW** tab. Select Run workflow(s) with inputs defined by data table, and choose *sample* from the drop-down menu.

## Inputs:

Please see the description of important inputs below. Note that required inputs are in bold.

Name	Description	Default
<b>sample_name</b>	Sample name	
<b>read1</b>	Array of URLs to read 1	
<b>read2</b>	Array of URLs to read 2	
<b>reference</b>	<b>Reference to align reads to</b> <ul style="list-style-type: none"> <li>• <b>Pre-created genome references:</b> <ul style="list-style-type: none"> <li>– “GRCh38_ens93filt” for human, genome version is GRCh38, gene annotation is generated using human Ensembl 93 GTF according to <a href="#">cellranger mkgtf</a>;</li> <li>– “GRCm38_ens93filt” for mouse, genome version is GRCm38, gene annotation is generated using mouse Ensembl 93 GTF according to <a href="#">cellranger mkgtf</a>;</li> </ul> </li> <li>• Create a custom genome reference using <a href="#">smart-seq2_create_reference workflow</a>, and specify its Google bucket URL here.</li> </ul>	
aligner	Which aligner to use for read alignment. Options are “hisat2-hca”, “star” and “bowtie”	“star”
output_genome_bam	Whether to output bam file with alignments mapped to genomic coordinates and annotated with their posterior probabilities.	false

## Outputs:

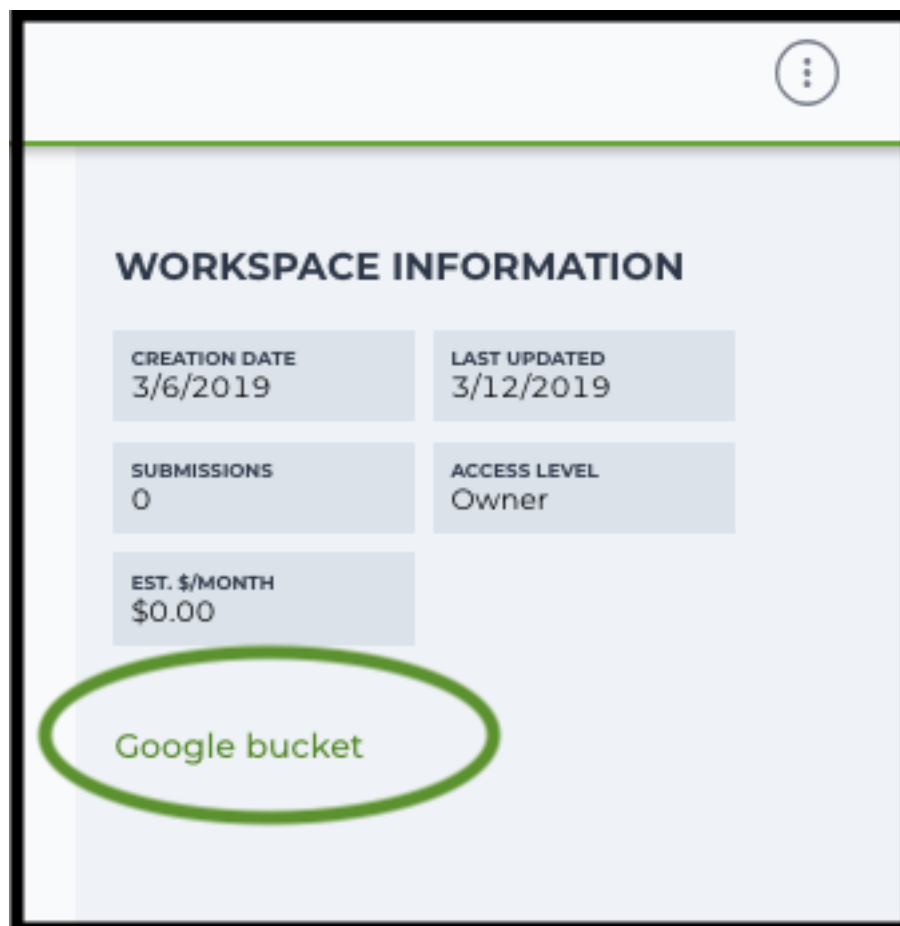
Name	Description
rsem_gene	RSEM gene expression estimation.
rsem_isoform	RSEM isoform expression estimation.
rsem_trans_bam	RSEM transcriptomic BAM.
rsem_genome_bam	RSEM genomic BAM files if <code>output_genome_bam</code> is <code>true</code> .
rsem_time	RSEM execution time log.
aligner_log	Aligner log.
rsem_cnt	RSEM count.
rsem_model	RSEM model.
rsem_theta	RSEM theta.

### 1.1.16 Drop-seq pipeline

This workflow follows the steps outlined in the [Drop-seq alignment cookbook](#) from the [McCarroll lab](#), except the default STAR aligner flags are `-limitOutSJcollapsed 1000000 -twopassMode Basic`. Additionally the pipeline provides the option to generate count matrices using [dropEst](#).

1. Copy your sequencing output to your workspace bucket using [gsutil](#) in your unix terminal.

You can obtain your bucket URL in the dashboard tab of your Terra workspace under the information panel.



Note: Broad users need to be on an UGER node (not a login node) in order to use the `-m` flag

Request an UGER node:

```
reuse UGER
qcrsh -q interactive -l h_vmem=4g -pe smp 8 -binding linear:8 -P regevlab
```

The above command requests an interactive node with 4G memory per thread and 8 threads. Feel free to change the memory, thread, and project parameters.

Once you're connected to an UGER node, you can make `gsutil` available by running:

```
reuse Google-Cloud-SDK
```

Use `gsutil cp [OPTION]... src_url dst_url` to copy data to your workspace bucket. For example, the following command copies the directory at `/foo/bar/nextseq/Data/VK18WBC6Z4` to a Google bucket:

```
gsutil -m cp -r /foo/bar/nextseq/Data/VK18WBC6Z4 gs://fc-e0000000-0000-
↪0000-0000-000000000000/VK18WBC6Z4
```

`-m` means copy in parallel, `-r` means copy the directory recursively.

2. Non Broad Institute users that wish to run `bcl2fastq` must create a custom docker image.

See [bcl2fastq](#) instructions.

3. Create a sample sheet.

Please note that the columns in the CSV must be in the order shown below and does not contain a header line. The sample sheet provides either the FASTQ files for each sample if you’ve already run bcl2fastq or a list of BCL directories if you’re starting from BCL directories. Please note that BCL directories must contain a valid bcl2fastq sample sheet (SampleSheet.csv):

Column	Description
Name	Sample name.
Read1	Location of the FASTQ file for read1 in the cloud (gsurl).
Read2	Location of the FASTQ file for read2 in the cloud (gsurl).

Example using FASTQ input files:

```
sample-1,gs://fc-e0000000-0000-0000-0000-000000000000/dropseq-1/sample1-1_
↪L001_R1_001.fastq.gz,gs://fc-e0000000-0000-0000-0000-000000000000/
↪dropseq-1/sample-1_L001_R2_001.fastq.gz
sample-2,gs://fc-e0000000-0000-0000-0000-000000000000/dropseq-1/sample-2_
↪L001_R1_001.fastq.gz,gs://fc-e0000000-0000-0000-0000-000000000000/
↪dropseq-1/sample-2_L001_R2_001.fastq.gz
sample-1,gs://fc-e0000000-0000-0000-0000-000000000000/dropseq-2/sample1-1_
↪L001_R1_001.fastq.gz,gs://fc-e0000000-0000-0000-0000-000000000000/
↪dropseq-2/sample-1_L001_R2_001.fastq.gz
```

Note that in this example, sample-1 was sequenced across two flowcells.

Example using BCL input directories:

```
gs://fc-e0000000-0000-0000-0000-000000000000/flowcell-1
gs://fc-e0000000-0000-0000-0000-000000000000/flowcell-2
```

Note that the flow cell directory must contain a bcl2fastq sample sheet named SampleSheet.csv.

#### 4. Upload your sample sheet to the workspace bucket.

Example:

```
gsutil cp /foo/bar/projects/sample_sheet.csv gs://fc-e0000000-0000-0000-
↪0000-000000000000/
```

#### 5. Import *dropseq\_workflow* workflow to your workspace.

See the Terra documentation for [adding a workflow](#). The *dropseq\_workflow* is under Broad Methods Repository with name “**cumulus/dropseq\_workflow**”.

Moreover, in the workflow page, click the Export to Workspace... button, and select the workspace you want to export *dropseq\_workflow* workflow in the drop-down menu.

#### 6. In your workspace, open *dropseq\_workflow* in WORKFLOWS tab. Select Run workflow with inputs defined by file paths as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click the SAVE button.

## Inputs

Please see the description of important inputs below.

Name	Description
input_csv_file	CSV file containing sample name, read1, and read2 or a list of BCL directories.
output_directory	Pipeline output directory (gs URL e.g. "gs://fc-e0000000-0000-0000-0000-000000000000/dropseq_output")
reference	hg19, GRCh38, mm10, hg19_mm10, mmul_8.0.1 or a path to a custom reference JSON file
run_bcl2fastq	Whether your sample sheet contains one BCL directory per line or one sample per line (default false)
run_dropseq_tool	Whether to generate count matrixes using Drop-Seq tools from the <a href="#">McCarroll lab</a> (default true)
run_dropest	Whether to generate count matrixes using <a href="#">dropeSt</a> (default false)
cellular_barcode_whitelist	Optional whitelist of known cellular barcodes
drop_seq_tools_force_cell_id	If supplied, bypass the cell detection algorithm (the elbow method) and use this number of cells.
dropest_cells_max	Maximal number of output cells
dropest_genes_min	Minimal number of genes for cells after the merge procedure (default 100)
dropest_min_merge_threshold	Threshold for the merge procedure (default 0.2)
dropest_max_cell_distance	Max cell distance between barcodes (default 2)
dropest_max_umi_distance	Max cell distance between UMIs (default 1)
dropest_min_genes_before_merge	Minimal number of genes for cells before the merge procedure. Used mostly for optimization. (default 10)
dropest_merge_strategy	Barcode merge strategy (can be slow), recommended to use when the list of real barcodes is not available (default true)
dropest_velocity	Save separate count matrices for exons, introns and exon/intron spanning reads (default true)
trim_sequence	The sequence to look for at the start of reads for trimming (default "AAGCAGTGGTATCAACGCAGAGTGAATGGG")
trim_num_bases	How many bases at the beginning of the sequence must match before trimming occur (default 5)
umi_base_range	The base location of the molecular barcode (default 13-20)
cellular_barcode_base_range	The base location of the cell barcode (default 1-12)
star_flags	Additional options to pass to STAR aligner

Please note that run\_bcl2fastq must be set to true if you're starting from BCL files instead of FASTQs.

## Custom Genome JSON

If you're reference is not one of the predefined choices, you can create a custom JSON file. Example:

```
{
  "refflat": "gs://fc-e0000000-0000-0000-0000-000000000000/human_mouse/
↪hg19_mm10_transgenes.refFlat",
  "genome_fasta": "gs://fc-e0000000-0000-0000-0000-000000000000/human_mouse/
↪hg19_mm10_transgenes.fasta",
  "star_genome": "gs://fc-e0000000-0000-0000-0000-000000000000/human_mouse/
↪STAR2_5_index_hg19_mm10.tar.gz",
  "gene_intervals": "gs://fc-e0000000-0000-0000-0000-000000000000/human_
↪mouse/hg19_mm10_transgenes.genes.intervals",
  "genome_dict": "gs://fc-e0000000-0000-0000-0000-000000000000/human_mouse/
↪hg19_mm10_transgenes.dict",
  "star_cpus": 32,
  "star_memory": "120G"
}
```

The fields `star_cpus` and `star_memory` are optional and are used as the default cpus and memory for running STAR with your genome.

## Outputs

The pipeline outputs a list of google bucket urls containing one gene-count matrix per sample. Each gene-count matrix file produced by Drop-seq tools has the suffix `'dge.txt.gz'`, matrices produced by dropEst have the extension `.rds`.

## Building a Custom Genome

The tool **dropseq\_bundle** can be used to build a custom genome. Please see the description of important inputs below.

Name	Description
<code>fasta_file</code>	Array of fasta files. If more than one species, fasta and gtf files must be in the same order.
<code>gtf_file</code>	Array of gtf files. If more than one species, fasta and gtf files must be in the same order.
<code>genomeSAindexNbases</code>	Number (bases) of the SA pre-indexing string. Typically between 10 and 15. Longer strings will use much more memory, but allow faster searches. For small genomes, must be scaled down to $\min(14, \log_2(\text{GenomeLength})/2 - 1)$

## dropseq\_workflow Terra Release Notes

### Version 11

- Added `fastq_to_sam_memory` and `trim_bam_memory` workflow inputs

### Version 10

- Updated workflow to WDL version 1.0

### Version 9

- Changed input `bcl2fastq_docker_registry` from optional to required

### Version 8

- Added additional parameters for `bcl2fastq`

### Version 7

- Added support for multi-species genomes (Barnyard experiments)

### Version 6

- Added `star_extra_disk_space` and `star_disk_space_multiplier` workflow inputs to adjust disk space allocated for STAR alignment task.

### Version 5

- Split preprocessing steps into separate tasks (FastqToSam, TagBam, FilterBam, and TrimBam).

### Version 4

- Handle uncompressed fastq files as workflow input.
- Added optional `prepare_fastq_disk_space_multiplier` input.

### Version 3

- Set default value for `docker_registry` input.

**Version 2**

- Added docker\_registry input.

**Version 1**

- Renamed sccloud to cumulus
- Added use\_bases\_mask option when running bcl2fastq

**Version 18**

- Created a separate docker image for running bcl2fastq

**Version 17**

- Fixed bug that ignored WDL input star\_flags (thanks to Carly Ziegler for reporting)
- Changed default value of star\_flags to the empty string (Prior versions of the WDL incorrectly indicated that basic 2-pass mapping was done)

**Version 16**

- Use cumulus dockerhub organization
- Changed default dropEst version to 0.8.6

**Version 15**

- Added drop\_deq\_tools\_prep\_bam\_memory and drop\_deq\_tools\_dge\_memory options

**Version 14**

- Fix for downloading files from user pays buckets

**Version 13**

- Set GCLOUD\_PROJECT\_ID for user pays buckets

**Version 12**

- Changed default dropEst memory from 52G to 104G

**Version 11**

- Updated formula for computing disk size for dropseq\_count

**Version 10**

- Added option to specify merge\_bam\_alignment\_memory and sort\_bam\_max\_records\_in\_ram

**Version 9**

- Updated default drop\_seq\_tools\_version from 2.2.0 to 2.3.0

**Version 8**

- Made additional options available for running dropEst

**Version 7**

- Changed default dropEst memory from 104G to 52G

**Version 6**

- Added option to run dropEst

**Version 5**

- Specify full version for bcl2fastq (2.20.0.422-2 instead of 2.20.0.422)



**Version 4**

- Fixed issue that prevented bcl2fastq from running

**Version 3**

- Set default run\_bcl2fastq to false
- Create shortcuts for commonly used genomes

**Version 2**

- Updated QC report

**Version 1**

- Initial release

**dropseq\_bundle Terra Release Notes****Version 4**

- Added create\_intervals\_memory and extra\_star\_flags inputs

**Version 3**

- Added extra disk space inputs
- Fixed bug that prevented creating multi-genome bundles

**Version 2**

- Added docker\_registry input

**Version 1**

- Renamed sccloud to cumulus

**Version 1**

- Changed docker organization

**Version 1**

- Initial release

**1.1.17 bcl2fastq****License**

[bcl2fastq license](#)

**Workflows**

Workflows such as **cellranger\_workflow** and **dropseq\_workflow** provide the option of running `bcl2fastq`. We provide dockers containing `bcl2fastq` that are accessible only by members of the Broad Institute. Non-Broad Institute members will have to provide their own docker images. Please note that if you're a Broad Institute member and are not able to pull the docker image, please check <https://app.terra.bio/#groups> to see that you're a member of the `all_broad_users` group. If not, please contact Terra support and ask to be added to the `all_broad_users@firecloud.org` group.

## Docker

Read [this tutorial](#) if you are new to Docker.

Then for a Debian based docker (e.g. [continuumio/miniconda3](#)), create the Dockerfile as follows:

```
RUN apt-get update && apt-get install --no-install-recommends -y alien unzip
ADD bcl2fastq2-v2-20-0-linux-x86-64.zip /software/
RUN unzip -d /software/ /software/bcl2fastq2-v2-20-0-linux-x86-64.zip && alien -i /
↳software/bcl2fastq2-v2.20.0.422-Linux-x86_64.rpm && rm /software/bcl2fastq2-v2*
```

Next, download `bcl2fastq` from [the Illumina website](#), which requires registration. Choose the Linux rpm file format and download `bcl2fastq2-v2-20-0-linux-x86-64.zip` to the same directory as your Dockerfile.

You can host your private docker images in the [Google Container Registry](#).

## Example

In this example we create a docker image for running `cellranger mkfastq` version 3.0.2.

1. Create a GCP project or reuse an existing project.
2. Enable the Google Container Registry
3. Clone the cumulus repository:

```
git clone https://github.com/lilab-bcb/cumulus.git
```

4. Add the lines to `cumulus/docker/cellranger/3.0.2/Dockerfile` to include `bcl2fastq` (see [Docker](#)).
5. Ensure you have [Docker installed](#)
6. Download `cellranger` from <https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/3.0>
7. Build, tag, and push the docker. Remember to replace `PROJECT_ID` with your GCP project id:

```
cd cumulus/docker/cellranger/3.0.2/
docker build -t cellranger-3.0.2 .
docker tag cellranger-3.0.2 gcr.io/PROJECT_ID/cellranger:3.0.2
gcr.io/PROJECT_ID/cellranger:3.0.2
```

8. Import `cellranger_workflow` workflow to your workspace (see [cellranger\\_workflow steps](#)), and enter your docker registry URL (in this example, `"gcr.io/PROJECT_ID/"`) in `cellranger_mkfastq_docker_registry` field of [cellranger\\_workflow inputs](#).

### 1.1.18 Cell Ranger alternatives to generate gene-count matrices for 10X data

This `count` workflow generates gene-count matrices from 10X FASTQ data using alternative methods other than Cell Ranger.

#### Prepare input data and import workflow

##### 1. Run `cellranger_workflow` to generate FASTQ data

You can skip this step if your data are already in FASTQ format.

Otherwise, you need to first run *cellranger\_workflow* to generate FASTQ files from BCL raw data for each sample. Please follow [cellranger\\_workflow manual](#).

Notice that you should set **run\_mkfastq** to `true` to get FASTQ output. You can also set **run\_count** to `false` if you want to skip Cell Ranger count, and only use the result from *count* workflow.

For Non-Broad users, you'll need to build your own docker for *bcl2fastq* step. Instructions are [here](#).

## 2. Import count

Import *count* workflow to your workspace.

See the Terra documentation for [adding a workflow](#). The *count* workflow is under Broad Methods Repository with name "**cumulus/count**".

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export *count* workflow in the drop-down menu.

## 3. Prepare a sample sheet

### 3.1 Sample sheet format:

The sample sheet for *count* workflow should be in TSV format, i.e. columns are separated by tabs not commas. Please note that the columns in the TSV can be in any order, but that the column names must match the recognized headings.

The sample sheet describes how to identify flowcells and generate channel-specific count matrices.

A brief description of the sample sheet format is listed below (**required column headers are shown in bold**).

Column	Description
<b>Sample</b>	Contains sample names. Each 10x channel should have a unique sample name.
<b>Flowcells</b>	Indicates the Google bucket URLs of folder(s) holding FASTQ files of this sample.

The sample sheet supports sequencing the same 10x channel across multiple flowcells. If a sample is sequenced across multiple flowcells, simply list all of its flowcells in a comma-separated way. In the following example, we have 2 samples sequenced in two flowcells.

Example:

```
Sample Flowcells
sample_1      gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4/
↪sample_1_fastqs,gs://fc-e0000000-0000-0000-0000-000000000000/VK10WBC9Z2/
↪sample_1_fastqs
sample_2      gs://fc-e0000000-0000-0000-0000-000000000000/VK18WBC6Z4/
↪sample_2_fastqs
```

Moreover, if one flowcell of a sample contains multiple FASTQ files for each read, i.e. sequences from multiple lanes, you should keep your sample sheet as the same, and *count* workflow will automatically merge lanes altogether for the sample before performing counting.

### 3.2 Upload your sample sheet to the workspace bucket:

Use `gsutil` (you already have it if you've installed Google cloud SDK) in your unix terminal to upload your sample sheet to workspace bucket.

Example:

```
gsutil cp /foo/bar/projects/sample_sheet.tsv gs://fc-e0000000-0000-0000-0000-000000000000/
```

#### 4. Launch analysis

In your workspace, open **Count** in **WORKFLOWS** tab. Select the desired snapshot version (e.g. latest). Select **Process single workflow from files** as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click SAVE button. Select Use call caching and click INPUTS. Then fill in appropriate values in the Attribute column. Alternative, you can upload a JSON file to configure input by clicking Drag or click to upload json.

Once INPUTS are appropriated filled, click RUN ANALYSIS and then click LAUNCH.

## Workflow inputs

Below are inputs for *count* workflow. Notice that required inputs are in bold.

Name	Description	Example	Default
<b>input_tsv_file</b>	Input TSV sample sheet describing metadata of each sample.	“gs://fc-e0000000-0000-0000-0000-000000000000/sample_sheet.tsv”	
<b>genome</b>	Genome reference name. Current support: GRCh38, mm10.	“GRCh38”	
<b>chemistry</b>	10X genomics’ chemistry name. Current support: “tenX_v3” (for V3 chemistry), “tenX_v2” (for V2 chemistry), “dropseq” (for Drop-Seq).	“tenX_v3”	
<b>output_directory</b>	URL of output directory.	“gs://fc-e0000000-0000-0000-0000-000000000000/count_result”	
<b>run_count</b>	If you want to run count tools to generate gene-count matrices.	true	true
<b>count_tool</b>	Count tool to generate result. Options: <ul style="list-style-type: none"> <li>• “StarSolo”: Use <a href="#">STARsolo</a>.</li> <li>• “Optimus”: Use <a href="#">Optimus</a> pipeline, developed by the Data Coordination Platform team of the Human Cell Atlas.</li> <li>• “Bustools”: Use <a href="#">Kallisto BUSTools</a>.</li> <li>• “Alevin”: Use <a href="#">Salmon Alevin</a>.</li> </ul>	“StarSolo”	“StarSolo”
<b>docker_registry</b>	Docker registry to use. Notice that docker image for Bustools is separate. <ul style="list-style-type: none"> <li>• “quay.io/cumulus” for images on Red Hat registry;</li> <li>• “cumulusprod” for backup images on Docker Hub.</li> </ul>	“quay.io/cumulus”	“quay.io/cumulus”
<b>config_version</b>	Version of config docker image to use. This docker is used for parsing the input sample sheet for downstream execution. Available options: 0.2, 0.1.	“0.2”	“0.2”
<b>zones</b>	Google cloud zones to consider for execution.	“us-east1-d us-west1-a us-west1-b”	“us-central1-a us-central1-b us-central1-c us-central1-f us-east1-b us-east1-c us-east1-d us-west1-a us-west1-b us-west1-c”
<b>num_cpu</b>	Number of CPUs to request for count per channel. Notice that when use Optimus for count, this input only affects steps of copying files. Optimus uses CPUs due to its own strategy.	32	32
<b>disk_space</b>		500	500

## Workflow outputs

See the table below for *count* workflow outputs.

Name	Type	Description
output_folder	String	Google Bucket URL of output directory. Within it, each folder is for one sample in the input sample sheet.

### 1.1.19 Topic modeling

#### Prepare input data

Follow the steps below to run **topic\_modeling** on [Terra](#).

1. Prepare your count matrix. **Cumulus** currently supports the following formats: ‘zarr’, ‘h5ad’, ‘loom’, ‘10x’, ‘mtx’, ‘csv’, ‘tsv’ and ‘fcs’ (for flow/mass cytometry data) formats
2. Upload your count matrix to the workspace.

Example:

```
gsutil cp /foo/bar/projects/dataset.h5ad gs://fc-e0000000-0000-0000-0000-000000000000/
```

where `/foo/bar/projects/dataset.h5ad` is the path to your dataset on your local machine, and `gs://fc-e0000000-0000-0000-0000-000000000000/` is the Google bucket destination.

3. Import *topic\_modeling* workflow to your workspace.

See the Terra documentation for [adding a workflow](#). The *cumulus* workflow is under Broad Methods Repository with name “**cumulus/topic\_modeling**”.

Moreover, in the workflow page, click the `Export to Workspace...` button, and select the workspace to which you want to export *topic\_modeling* workflow in the drop-down menu.

4. In your workspace, open *topic\_modeling* in WORKFLOWS tab. Select `Run workflow with inputs defined by file paths` as below

- ☒ Run workflow with inputs defined by file paths
- ☐ Run workflow(s) with inputs defined by data table

and click the `SAVE` button.

#### Workflow input

Inputs for the *topic\_modeling* workflow are described below. Required inputs are in bold.

Name	Description	Example	Default
<b>input_file</b>	Google bucket URL of the input count matrix.	“gs://fc-e0000000-0000-0000-0000-000000000000/my_dataset.h5ad”	
<b>number_of_topics</b>	Number of number of topics.	[10,15,20]	
<b>prefix_exclude</b>	Comma separated list of features to exclude that start with prefix.	“mt-,Rpl,Rps”	“mt-,Rpl,Rps”
<b>min_percent_expressed</b>	Percent of features expressed below min_percent.	2	
<b>max_percent_expressed</b>	Percent of features expressed below min_percent.	98	
<b>random_number_seed</b>	Random number seed for reproducibility.	0	0

## Workflow output

Name	Type	Description
coherence_plot	File	Plot of coherence scores vs. number of topics
perplexity_plot	File	Plot of perplexity values vs. number of topics
cell_scores	Array[File]	Topic by cells (one file for each topic number)
feature_topics	Array[File]	Topic by features (one file for each topic number)
report	Array[File]	HTML visualization report (one file for each topic number)
stats	Array[File]	Computed coherence and perplexity (one file for each topic number)
model	Array[File]	Serialized LDA model (one file for each topic number)
corpus	File	Serialized corpus
dictionary	File	Serialized dictionary

### 1.1.20 Contributions

We welcome contributions to our repositories that make up the Cumulus ecosystem:

- [pegasus](#)
- [pegasusio](#)
- [demuxEM](#)
- [cumulus](#)
- [cumulus\\_feature\\_barcoding](#)
- [cirrocumulus](#)
- [altocumulus](#)
- [stratocumulus](#)

In addition to the Cumulus team, we would like to sincerely thank the following contributors:

Name	Note
Kirk Gosik	Assistance with topic modeling workflow

### 1.1.21 Contact us

If you have any questions related to Cumulus, please feel free to contact us via one of the following ways:

- Report new [issues](#) on our GitHub repository.
- Send emails to [Cumulus Support Google Group](#).
- Join [Cumulus Support](#) Google Group for discussion and release update.