Demultiplexing Illumina Sequencing Data of 10X Single Cell Genomics on UGA Clusters (Sapelo2)

Mengrui Zhang, Madgy Alabady

March 23, 2021

1 Introduction

This tutorial describes how to download sequencing runs from Illumina BaseSpace and demultiplex the raw BCL data into 10X genomics single cell data on the UGA GACRC Clusters.

2 Data Download from Illminia BaseSpace

The BaseSpace Command Line Interface (CLI) environment is installed on both the GACRC transfer (xfer) and computational (Sapelo2) nodes. We will use the transfer nodes to download the run since the file transfer node will provide a faster download speed.
First, log in to the xfer node using your GACRC credentials with the following command:

```
$ ssh myID@xfer.gacrc.uga.edu
```

Next, open up an internet browser on your local computer, log into your BaseSpace account, and navigate to the target run page. Figure 1 shows an example of the target run dashboard. Copy the URL of this website as it will be used later in the command line. For example, the URL link for the run in the Figure 1 is "https://basespace.illumina.com/run/196794705/details".

Go back to the terminal where we logged in the xfer server, we need to login to the Basespace from our terminal first.

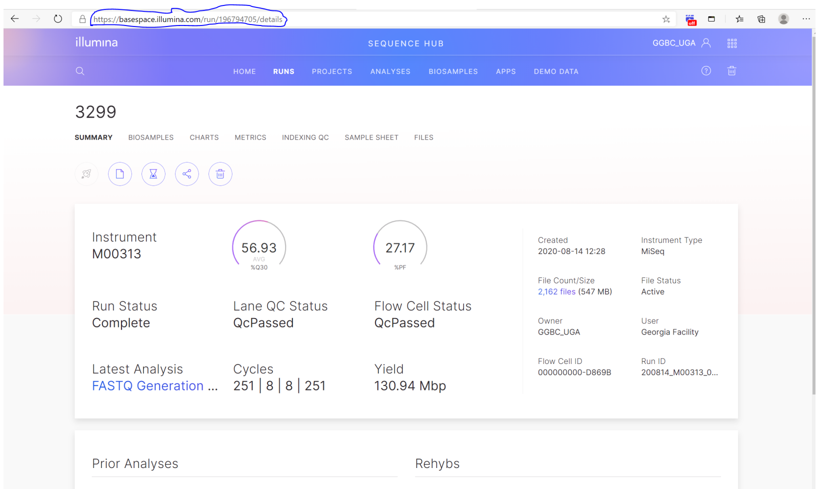
```
$ bs auth
```

A link will show up. Copy the link to a browser and login.
Modify the following code with your own working directory to download the BCL dataset.

```
$ cd #dirctory where the data should be downloaded
$ bs download run -i 196794705 -o ./3219

# this is the url that you copied from BaseSpace without the "/details" part.
```

After running the above code, the run data will begin downloading immediately from basespace.



3299

Instrument
M00313

56.93
AVG
%Q30

27.17
%PF

Run Status
Complete

Lane QC Status
QcPassed

Flow Cell Status
QcPassed

Latest Analysis
FASTQ Generation ...

Cycles
251 | 8 | 8 | 251

Yield
130.94 Mbp

Created
2020-08-14 12:28

Instrument Type
MiSeq

File Count/Size
2,162 files (547 MB)

File Status
Active

Owner
GGBC_UGA

User
Georgia Facility

Flow Cell ID
000000000-D869B

Run ID
200814_M00313_0...

Prior Analyses

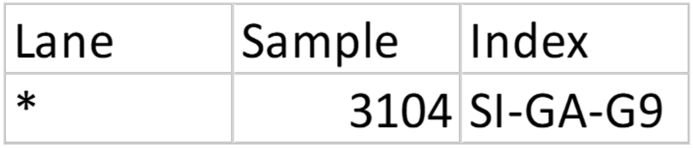
Rehybs

Figure 1: Example run sample sheet from BaseSpace

3 Create Sample Sheet

Now, let's create the sample sheet that's required for the demultiplexing process. Figure 2 is an example sample sheet. When creating the sample sheet in Excel, you will need to save it as a ".csv" file . The lane column indicates which lane(s) of the flowcell to process. It can be either a single lane, a range of lanes (e.g., 2-4), or use * to process all lanes in the flowcell. The sample column indicates the name of the sample. The index column indicates the 10x sample index that was used in library construction.

```
Lane,Sample,Index
*,3104,SI-GA-G9
```



Lane	Sample	Index
*	3104	SI-GA-G9

Figure 2: Example sample sheet format

After making the sample sheet, upload your .csv file to the directory containing the downloaded BCL data. The file can be transferred using either scp command or FileZilla software. Please check the following URL for transferring the file to the server. An example scp command on your local computer can be:

```
  $ scp <SampleSheet.csv> myID@xfer.gacrc.uga.edu:<#directory of downloaded data>
```

Please check and make sure the sample sheet file is in the folder of downloaded BCL data before going to the next step.

4 Submit Job to UGA GACRC Cluster

We need to create a bash script file with the conversion to fastq and demultiplexing commands and submit it to the queue of the computer cluster (sapelo2). The following code is an example of this bash script. Please make sure to modify the following. 1. Change the email address to yours and change the BCL files directory. We can name this script as mkfastq.sh

```
#!/bin/bash
#SBATCH --job-name=mkfastq           # Job name
#SBATCH --partition=batch           # Partition (queue) name
#SBATCH --nodes=1                     # Number of nodes
#SBATCH --ntasks=12                   # Number of MPI ranks
#SBATCH --mem=100gb                    # Job memory request
#SBATCH --time=12:00:00               # Time limit hrs:min:sec
#SBATCH --output=mkfastq.out       # Standard output log
#SBATCH --error=mkfastq.err        # Standard error log
#SBATCH --mail-type=ALL          # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=<User_Email>  # Where to send mail

cd $SLURM_SUBMIT_DIR

module load CellRanger/4.0.0
module load bcl2fastq2/2.20.0-foss-2019b

cellranger mkfastq --id=<Name of output folder> \
--qc --run=<Directory of downloaded run. #For example ./3299> \
--csv=<Directory of the sample sheet file>
```

The following command line can be used on your local computer to upload this script to the server:

```
$ scp <mkfastq.sh> myID@xfer.gacrc.uga.edu:<directory of downloaded data>
```

Please make sure the data folder contains both the sample sheet and the script (example name: mkfastq.sh) file before submitting jobs. To submit the job, run the following command:

```
$ sbatch  demultiplex10x.sh
```

To check the status of the job. Run:

```
$ qstat -u <myID>
```

There is a queue for job submissions and depending on the usage load of the batch queue, your job may take awhile before it runs. Once the demultiplexing process has finished, the newly created folder has the same name as the id name in the job scripts. For example,

```
$ ./<id_name>/outs/fastq_path/
```

The above directory contains the demultiplexed fastq files for all the samples.