

# Demultiplexing Illumina Sequencing Data 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 Illumina 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 run ID from the link address 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", then the run ID is 196794705.

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 #directory 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.

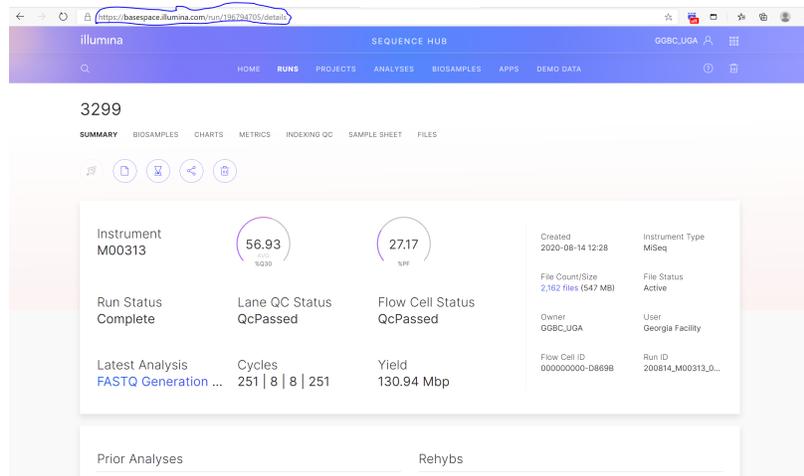


Figure 1: Example run dashboard from BaseSpace

### 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. Below is the same template when creating from a plain text file. Please put the sample name under *sample\_ID* column, i7 barcode under *index1* column and i5 barcode at *index2* column, etc.

```
[Header] ,,,,,,
IEMFileVersion,4,,,,,
Experiment Name,2309,,,,,
Date,3/26/2019,,,,,
Workflow,GenerateFASTQ,,,,,
Application,FASTQ Only,,,,,
Assay,TruSeq HT,,,,,
Description,,,,,
Chemistry,Amplicon,,,,,
,,,,,
[Reads] ,,,,,,
251,,,,,
251,,,,,
[Settings] ,,,,,,
ReverseComplement,0,,,,,
Adapter ,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA,,,,,
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT,,,,,
,,,,,
[Data] ,,,,,,
Sample_ID,Sample_Name,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,
index2,Sample_Project,Description
109-1_ChIP_dCAC-1_H3K27me2me3_Rep2,,,D701,ATTACTCG,D501,TATAGCCT,,
```

109-2\_ChIP\_dCAC-1\_H3K27me3\_Rep1 , , , , D702, TCCGGAGA, D501, TATAGCCT, ,

[Header]									
MEMFileVersion									
Experiment Name									
Date									
Workflow	GenerateFASTQ								
Application	FASTQ Only								
Assay	TruSeq HT								
Description									
Chemistry	Amplicon								
[Reads]									
	251								
	251								
[Settings]									
ReverseComplement									
Adapter1	AGATCGGAAGAGCACACGTCTGAATCCAGTCA								
AdapterRead2	AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT								
[Data]									
Sample_ID	Sample_Name	Sample_Plate	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2	Sample_Project	Description
109-1_ChIP_dCAC-1_H3K27me3_Rep2				D701	ATTACTCG	D501	TATAGCCT		
109-2_ChIP_dCAC-1_H3K27me3_Rep1				D702	TCCGGAGA	D501	TATAGCCT		
109-3_ChIP_dCAC-2_H3K27me3_Rep1				D703	CGCTGATT	D501	TATAGCCT		
109-4_ChIP_dCAC-2_H3K27me3_Rep1				D704	GAGATTCC	D501	TATAGCCT		
109-5_ChIP_dHDA-1_H3K27me3_Rep2				D705	ATTCAGAA	D501	TATAGCCT		
109-6_ChIP_dHDA-1_H3K36me3_Rep2				D706	GAATTCGT	D501	TATAGCCT		
109-7_CutRun_WT_H3K36me3_Rep1				D707	CTGAAGCT	D501	TATAGCCT		
109-25_ChIP_WT_input_rep1				D701	ATTACTCG	D503	CCTATCCT		
109-26_ChIP_WT_GFP_rep1				D702	TCCGGAGA	D503	CCTATCCT		
109-27_ChIP_WT_K23me1_rep1				D703	CGCTGATT	D503	CCTATCCT		
109-28_ChIP_ACX135-1_GFP_rep2				D704	GAGATTCC	D503	CCTATCCT		

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. Please name the sample sheet as "SampleSheet.csv" so that the software can import it correctly.

## 4 Submit Job to UGA GACRC Cluster

Now, we need to create a bash script file that has 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 change the email address to yours and uploaded to the same folder where the BCL files were downloaded.

```
#!/bin/bash
#SBATCH --job-name=bcl2fastq           # 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=bcl2fastq.out       # Standard output log
#SBATCH --error=bcl2fastq.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 bcl2fastq2/2.20.0-foss-2019b
```

```
bcl2fastq --no-lane-splitting --processing-threads <threads> \  
--run-folder-dir <directory of downloaded data> \  
--output-dir <output directory>
```

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

```
$ scp bcl2fastq.sh myID@xfer.gacrc.uga.edu:<directory of downloaded data>
```

Make sure the data folder contains both the sample sheet .csv file and this script (example name: bcl2fastq.sh) before submit jobs. To submit the job, run the following code:

```
$ sbatch bcl2fastq.sh
```

To check the status of the job. Run:

```
$ qstat -u <myID> # change uga MyID
```

There is usually a waiting line for jobs. Your job might not be run immediately after you submitted. Once the demultiplexing process has finished, the newly created "fastq" folder contains the demultiplexed fastq files for all the samples.