Demultiplexing Illumina Sequencing Data on UGA Clusters (Sapelo2)

Mengrui Zhang, Madgy Alabady

October 2020

1 Introduction

This tutorial is for downloading sequencing runs from Illumina BaseSpace. The tutorial also shows the conversion and demultiplexing of the raw BCL data on the UGA GACRC Clusters.

2 Data Download from Illminia BaseSpace

The basespace-cli environment is installed on the GACRC transfer (xfer) and computational (sapelo2) nodes. We will use the transfer nodes to download the run because it is faster.
Log in to the xfer node using your GACRC credentials, as following:

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

Open up an Internet browser on your local computer and log into your BaseSpace account and navigate to the run that needs download. Figure 1 is an example run dashboard. Copy the Url of this website, you will use the link address later in the command line. As an example, here, the Url link for this example run is "https://basespace.illumina.com/run/196794705/details".
Go back to your terminal where we logged in the xfer server, modify the following code with your own working directory to download the BCL dataset.

```
$ cd #dirctory where the data should be downloaded
$ bc cp https://basespace.illumina.com/run/196794705 ./3299

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

After running the above code, you have to authenticate the download process by copying the link that the software will provide and paste to a browser and confirm it. After confirming the link, The run data will begin downloading immediately.

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, please save it as the ".csv" file

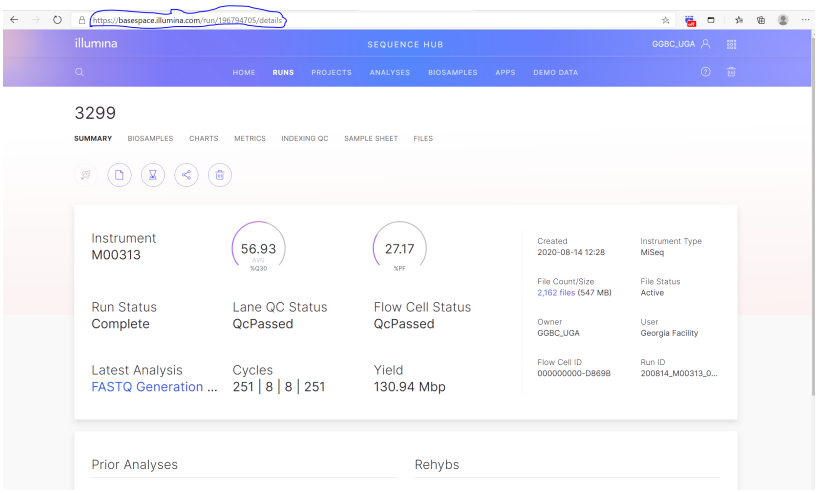

Figure 1: Example run dashboard from BaseSpace

format. 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,,
```

After making the sample sheet, please upload the sample sheet file to the directory that has the downloaded BCL data. You can use either scp command or FileZilla software. Please check the

[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	Descriptio
109-1_ChIP_dCAC-1_H3K27me2me3_Rep2				D701	ATTACTCG	D501	TATAGCCT		
109-2_ChIP_dCAC-1_H3K27me3_Rep1				D702	TCCGGAGA	D501	TATAGCCT		
109-3_ChIP_dCAC-2_H3K27me2me3_Rep1				D703	CGCTCATT	D501	TATAGCCT		
109-4_ChIP_dCAC-2_H3K27me3_Rep1				D704	GAGATTCC	D501	TATAGCCT		
109-5_ChIP_dHDA-1_H3K27me2me3_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	CGCTCATT	D503	CCTATCCT		
109-28_ChIP_ACx135-1_GFP_rep2				D704	GAGATTCC	D503	CCTATCCT		

Figure 2: Example run dashboard from Basespace

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.

```
#PBS -S /bin/bash
#PBS -q batch
#PBS -N bcl2fastq
#PBS -l nodes=1:ppn=6
#PBS -l walltime=12:00:00
#PBS -l mem=40gb
#PBS -M myID@uga.edu
#PBS -m abe

cd $PBS_O_WORKDIR
module load bcl2fastq2/2.20.0-foss-2016b-Python-2.7.14

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:

```
$ qsub bcl2fastq.sh
```

To check the status of the job. Run:

```
$ qstat_me
```

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.