Day 4 Worksheet – Part 1.1 FastQC to evaluate sequence library quality

Author: Jessica Westfall: jessica.westfall@colorado.edu

FASTQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
TRIMMOMATIC: http://www.usadellab.org/cms/?page=trimmomatic

Introduction: Today's practice will be to evaluate the sequence quality of our fastq files by using fastqc. Once we have evaluated our library, we will apply the appropriate trimming and reevaluate our sequence library in the next worksheet.

! Note: The directory and username used in the screenshot will be for my working directory and username and will be different than yours.

Make working directories

The first step is to making some working directories that we will be using for our sbatch scripts and outdirectories for our error and output logs, fastQC output files.

1. Use command **pwd** to determine what directory you are in.
2. You can either from here enter in cd to go to your home directory, use relative path to go to the directory you want relative to your starting point, or use an absolute path. I want to create these files in my scratch directory so I will use **cd** an absolute path. Afterwards I will use command **pwd** to confirm that I am in the correct directory.

```
-bash-4.2$ cd /scratch/Users/jewe1055/sread/
-bash-4.2$ pwd
/scratch/Users/jewe1055/sread
```

3. Now that I am in the directory I want to work in, I shall make 3 more folders using the **mkdir** command; fastqc, scripts, eofiles. These folders will be used to store the output results from fastQC (fastqc), the error and output logs (eofiles), and the sbatch scripts (scripts)..Use command **ls -lsh** to confirm the folders are present.

```
-bash-4.2$ mkdir fastqc; mkdir scripts; mkdir eofiles
-bash-4.2$ ls -lsh
total 0
0 drwxrwxr-x 2 jewe1055 dowelldegrp 0 Jul 15 10:20 eofiles
0 drwxrwxr-x 2 jewe1055 dowelldegrp 0 Jul 15 10:20 fastqc
0 drwxrwxr-x 2 jewe1055 dowelldegrp 0 Jul 15 10:20 scripts
```

FastQC

Now that we have made the directories, we will next locate our fasta files and do a sequence quality analysis.

! Remember, do not alter your fasta files. That is do not use a text editor on the fasta file

4. Change directory (**cd**) into /scratch/Workshop/sread/data_files/day4/fastq.

directory. This directory will have the fastq files that we will be use to practice. Use command **ls -lsh** to see the files in the directory. These files are unzipped, but note that sometimes fastq files are zipped (.fastq.gz). You do not need to unzip the files to run fastQC.

```
-bash-4.2$ cd /scratch/Shares/dowell/sread/data_files/day4/fastq
-bash-4.2$ ls -lsh
total 854M
 71M -rw-rwxr--+ 1 qiya9811 dowelldegrp  71M Jul  6 12:33 chr21Eric_repA.RNA.end1.fastq
 71M -rw-rwxr--+ 1 qiya9811 dowelldegrp  71M Jul  6 12:33 chr21Eric_repA.RNA.end2.fastq
 37M -rw-rwxr--+ 1 qiya9811 dowelldegrp  37M Jul  6 12:33 chr21Eric_repB.RNA.end1.fastq
 37M -rw-rwxr--+ 1 qiya9811 dowelldegrp  37M Jul  6 12:33 chr21Eric_repB.RNA.end2.fastq
 33M -rw-rwxr--+ 1 qiya9811 dowelldegrp  33M Jul  6 12:33 chr21Eric_repC.RNA.end1.fastq
 33M -rw-rwxr--+ 1 qiya9811 dowelldegrp  33M Jul  6 12:33 chr21Eric_repC.RNA.end2.fastq
114M -rw-rwxr--+ 1 qiya9811 dowelldegrp 114M Jul  6 12:33 chr21Ethan_repA.RNA.end1.fastq
114M -rw-rwxr--+ 1 qiya9811 dowelldegrp 114M Jul  6 12:33 chr21Ethan_repA.RNA.end2.fastq
 85M -rw-rwxr--+ 1 qiya9811 dowelldegrp  85M Jul  6 12:33 chr21Ethan_repB.RNA.end1.fastq
 85M -rw-rwxr--+ 1 qiya9811 dowelldegrp  85M Jul  6 12:33 chr21Ethan_repB.RNA.end2.fastq
 90M -rw-rwxr--+ 1 qiya9811 dowelldegrp  90M Jul  6 12:33 chr21Ethan_repC.RNA.end1.fastq
 90M -rw-rwxr--+ 1 qiya9811 dowelldegrp  90M Jul  6 12:33 chr21Ethan_repC.RNA.end2.fastq
```

5. Select one of the fastq file. I will be using **chr21Ethan_repA.RNA.end1.fastq**.
6. Copy (**rsync** or **cp**) the **d4_fastqc.sbatch** script into your script directory that you made in the previous steps. Below I am copying from the workshop directory to my directory. I then use **ls** to confirm the file is present in the directory.
To copy, the command syntax is **rsync <input><output>**

```
-bash-4.2$ rsync /scratch/Shares/dowell/sread/scripts/day4/d4_fastqc.sbatch /scra
tch/Users/jewe1055/sread/scripts/.
-bash-4.2$ ls /scratch/Users/jewe1055/sread/scripts/
d4_fastqc.sbatch
```

7. Edit the sbatch script by using **vim <sbatch>** to open a text editor on your sbatch script. Then type **i** to toggle into edit mode. You will need to change the job name, user email, and the standard output and error log directories. I will be changing the **–job-name=<JOB NAME>** to something related to the job that I am running, for example 'fastqc'. Additionally you will want to change the **–mail-user=<YOUR_EMAIL>** to your email, as well as the path to your eofiles directory for the standard output (**--output)** and error log (**--error)**. The **%x** will be replace by your **-job-name** and the **%j** will be replace by the job id that will be assigned by the job manager when you run your sbatch script.

```
bin/bash
ATCH --job-name=<JOB_NAME>                                  # Job name
ATCH --mail-type=FAIL                                       # Mail events (NONE, BEGIN, END, FAIL,

ATCH --mail-user=<YOUR_EMAIL>                               # Where to send mail
ATCH --nodes=1                                              # Numbers of nodes
ATCH --ntasks=1                                             # Number of CPU (tasks)
ATCH --time=00:15:00                                        # Time limit hrs:min:sec
ATCH --partition=short                                      # Job queue
ATCH --mem=2gb                                              # Memory limit
ATCH --output=/scratch/Users/<USERNAME>/eofiles/%x_%j.out
ATCH --error=/scratch/Users/<USERNAME>/eofiles/%x_%j.err
```

Lastly, check the following to make sure the amount of CPU and nodes that you request

is appropriate for the job. For fastQC, I am using 1 node, 1 task or processor, 4gb for memory and 15 minutes for wall time.

8. Setting variables will make your scripts easier to read. In addition, this makes it easier to reference to a given path and utilize it in your scripts. For the **INDIR=** change the path to where the fastq files are located (step 4). For the **OUTDIR=**, point to your fastqc directory from step 3.

```
################ SET REQUIRED VARIABLES ################################################
## the fastq files will be used as input to fastqc.
## output will be a fastqc file used to assess quality

INDIR=/scratch/Workshop/sread/data_files/day4/fastq
OUTDIR=/scratch/Users/<USERNAME>/fastqc
```

9. The remainder of this sbatch script involves (1) loading the appropriate module necessary to run fastQC and (2) writing the command to run fastQC. The variables that we edited in step 8 will be called here in the script so there should not be many edits that are further needed in the sbatch script below.

```
########################################################################################
############### LOAD REQUIRED MODULES ##################################################

module load fastqc/0.11.5

########################################################################################
################## PRINT JOB INFO ######################################################

printf "\nfastq Directory: $INDIR"
printf "\nOutput Directory: $OUTDIR"
printf "\nRun on: $(hostname)"
printf "\nRun from: $(pwd)"
printf "\nScript: $0\n"
date

printf "\nYou've requested $SLURM_CPUS_ON_NODE core(s).\n"

########################################################################################
################## RUN JOB #############################################################
mkdir -p ${OUTDIR}

fastqc \
 ${INDIR}/*.fastq.gz \
 -o ${OUTDIR}
```

Some notes about the sbatch script.

- **module load**: We will need to load the modules necessary to run a job. Here we will use fastqc. If you need to find the available modules on the computer cluster you can use the command **module spider <string>**. On the terminal you will see the versions available for loading.
- **Printing job info**: We use this section to log information about our run. This file will be found in your **eofiles** directory.
- **mkdir -p**: This syntax will generate the directory if it does not yet exist
- **fastqc <input_file> --outdir <output_file>**: This is the syntax that is used to run fastQC. Here we will use the variable we defined by calling it with the syntax

${variable_name}. Notice how the script above also has a '\' at the end. We can write this syntax as a single line but it is harder to read. This breaks up the code to make it easier to read however either way will run. If your '\' does not change color as you see above, you may have an extra space after the '\' . Remove that space or your code will not run properly.

10. Once everything looks good and ready to go, we want to save and quit. Press **esc** to exit out of edit mode, then type **:wq**. This will write/save (w) and quit (q) the script.
11. Last step is to run the script. We will submit the script to the job manager SLURM using the command **sbatch <sbatch_file>**. The job manager will assign a job id to your run. You can check the status of your job using the command **squeue -u <username>**. You can also **tail** your eofiles to see the running log while the job is running. The eofiles will have the associated job name you assign it in step 7 and the job number that the job manager assigned.

```
bash-4.2$ sbatch d4_fastqc.sbatch
Submitted batch job 7750390
bash-4.2$ tail ../eofiles/fastqc_7750390.err
Approx 50% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 55% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 60% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 65% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 70% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 75% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 80% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 85% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 90% complete for chr21Ethan_repA.RNA.end1.fastq
Approx 95% complete for chr21Ethan_repA.RNA.end1.fastq
```

12. The error and output log in your eofiles will contain different information about the job. These files can vary depending on the task you are running. You should see two different files if you **-ls -lahtr** the directory. You can view these files using the command **more**, **less**, **cat**, or **head**.

Output file: <job_name>.<job_id>.out. This file has the information that we requested under the 'Print Job Info' section of our sbatch script.

```
bash-4.2$ more ../eofiles/fastqc_7750390.out

fastq Directory: /scratch/Shares/dowell/sread/data_files/day4/fastq
Output Directory: /scratch/Users/jewe1055/sread/fastqc
Run on: fijinode-01
Run from: /scratch/Users/jewe1055/sread/scripts
Script: /tmp/slurmd/job7750390/slurm_script
Thu Jul 15 23:21:06 MDT 2021

You've requested 1 core(s).
Analysis complete for chr21Ethan_repA.RNA.end1.fastq
```

Error file: <job_name>.<job_id>.err. This file has additional information on the run. Since we already tailed it in step 11, we won't repeat it here.

13. The last part of this section is to open the fastqc file that contains information on the quality of our sequence library. First, locate the output files by **cd <fastqc>** directory.

```
bash-4.2$ cd ../fastqc/
bash-4.2$ ls -lahtr
total 728K
drwxrwxr-x 5 jewe1055 dowelldegrp    3 Jul 15 10:45 ..
-rw-rw-r-- 1 jewe1055 dowelldegrp 417K Jul 15 23:21 chr21Ethan_repA.RNA.end1_fastqc.zip
-rw-rw-r-- 1 jewe1055 dowelldegrp 310K Jul 15 23:21 chr21Ethan_repA.RNA.end1_fastqc.html
drwxrwxr-x 2 jewe1055 dowelldegrp    2 Jul 15 23:21 .
```

There are two files; zip and html. Since we cannot view the html file from the compute cluster terminal, we will copy the file over to your local directory so you can open it on a web browser. Use the command **pwd** to get the absolute path of your directory that contains the html file on the cluster. Next open a new terminal which will be on your local computer and not the compute cluster. Use the **rsync <username@cluster>:<path/to/input_file> <path/to/output>** to copy the file to your local computer.

```
(base) cu-genvpn-tcom-10:~ user$ rsync jewe1055@fiji.colorado.edu:/scratch/Users
 /jewe1055/sread/fastqc/chr21Ethan_repA.RNA.end1_fastqc.html ~/Desktop/.
```

14. Open the file on your browser of choice and congratulations on writing and running your first sbatch script.

This is what you should see when you open the html file.

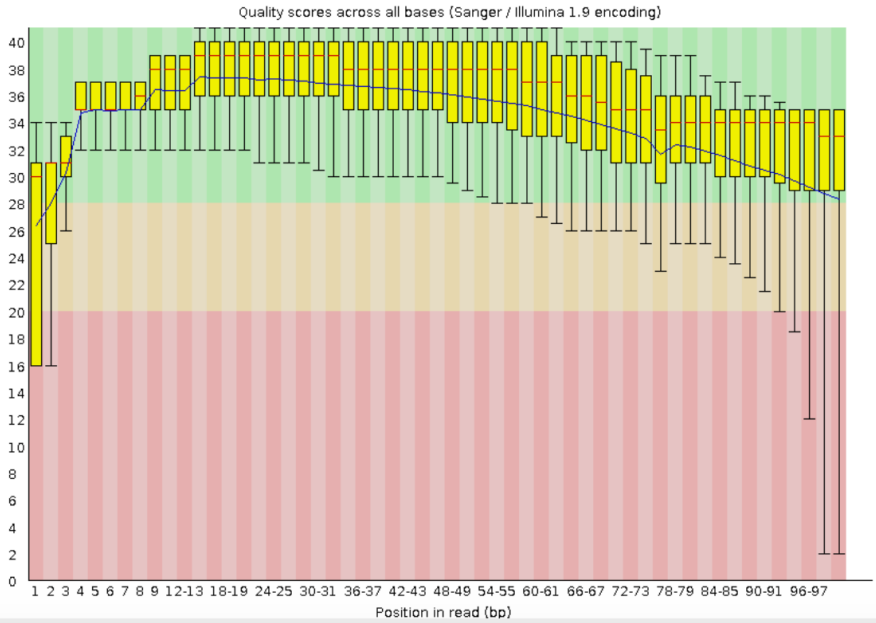
FastQC Report

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Filename	chr21Ethan_repA.RNA.end1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	477054
Sequences flagged as poor quality	0
Sequence length	100
%GC	42

Per base sequence quality



For examples of good and bad Illumina fastQC, go to https://www.bioinformatics.babraham.ac.uk/projects/fastqc/