CDC/21_1st QC

2022-02-11

20193852 문유빈

1> 진주 언니한테 받은 셀 스크립트

```
inPath="/data/Original_data/CDC/21_1st/"
outPath="/home/bbang9/Project/2020/CDC/21_1st/Novaseq/Analysis/"
mkdir $outPath

for file in ${inPath}*_1.fastq.gz
do
    stub=${file%_1.fastq.gz}
    stub2=${stub#$inPath}
    ID="$(cut -d'_' -f4<<<"$stub2")"
    sample=$stub2
    echo $sample
    mkdir $outPath$sample
    mkdir $outPath$sample/QcReads
    /home/bioware/FaQCs/FaQCs.pl -p $inPath$stub2\_1.fastq.gz $inPath$stub2\
_2.fastq.gz  -q 20 -min_L 50 -avg_q 0 -n 1 -lc 0.85 -5end 0 -3end 0 -split_s
ize 100000 -d $outPath\/$sample/QcReads -t 8 -adapter > $outPath$sample\/QC.
log
done
```

2> 스크립트 작성

1) 우선 샘플 이름만 출력해 보기(echo)

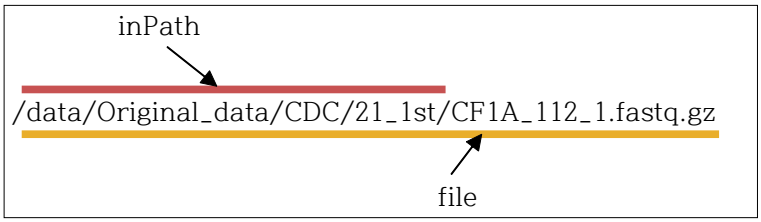
```
inPath="/data/Original_data/CDC/21_1st/"
outPath="/home/guest01/2021/yb/yb01/CDC_JJ/Analysis"

for file in ${inPath}*_1.fastq.gz
do
        stub=${file%_1.fastq.gz}
        stub2=${stub#$inPath}
        ID="$(cut -d'_' -f4<<<"$stub2")"
        sample=$stub2
        echo $sample

done
```

▷ *의 의미는 모든 것

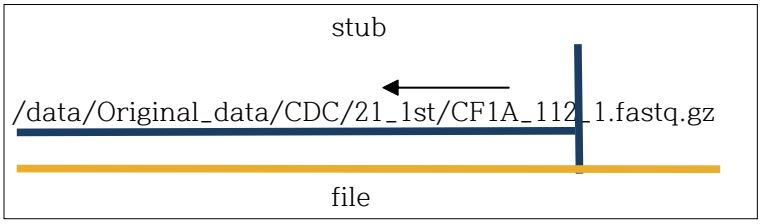
${inPath}*는 경로 디렉토리 안 모든 파일의 _1.fastq.gz를 file로 지정



▷ %의 의미는 오른쪽 앞에 것을 가져 온다

stub=${file%_1.fastq.gz}는

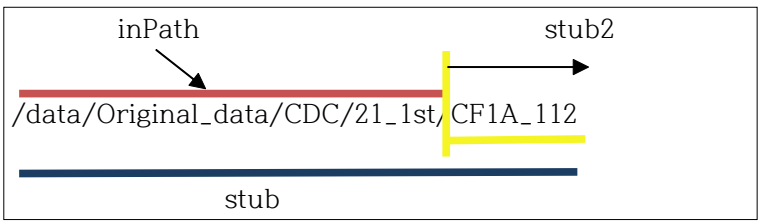
파일에서 _1.fastq.gz의 앞 부분이 stub



▷ #의 의미는 오른쪽 뒤에 것을 가져 온다

stub2=${stub#$inPath}이라면 stub2는 파일명만

stub에서 /data/Original_data/CDC/21_1st/의 뒷 부분이 stub2



>> 실행시 해당 디렉토리내 파일 명이 출력된다.

```
[guest01@smel0:script]$ sh 01.QC_C_yb.sh
CF1A_112
CF1A_114
CF1A_122
CF1A_124
CF1A_1312
CF1A_1314
CF1A_132
CF1A_1324
CF1A_1334
CF1A_134
CF1A_1352
CF1A_1362
CF1A_1372
CF1A_1382
CF1A_142
CF1A_152
CF1A_312
CF1A_314
CF1A_324
CF1A_3312
```

★ 셀 스크립트 생성 법

1) vi 스크립트명.sh

```
[guest01@smel-cluster:script]$ vi test.sh
```

2) 스크립트 작성

```
echo "abcde"
```

저장하고 나오면

```
[guest01@smel-cluster:script]$ ll
total 12
-rwxr-xr-x 1 guest01 guest01 583 Feb 11 11:21 01.QC_C.sh
-rwxrwxr-x 1 guest01 guest01 521 Feb 11 12:09 01.QC_C_yb.sh
-rw-rw-r-- 1 guest01 guest01  13 Feb 11 15:08 test.sh
```

ㄴ아직 초록색 x

3) chmod +x 스크립트명.sh (실행 파일로 바꿔줌)
4) ./스크립트명.sh (실행하기)

```
[guest01@smel-cluster:script]$ chmod +x test.sh
[guest01@smel-cluster:script]$ ./test.sh
abcde
```

abcde 출력됨

```
[guest01@smel-cluster:script]$ ll
total 12
-rwxr-xr-x 1 guest01 guest01 583 Feb 11 11:21 01.QC_C.sh
-rwxrwxr-x 1 guest01 guest01 521 Feb 11 12:09 01.QC_C_yb.sh
-rwxrwxr-x 1 guest01 guest01  13 Feb 11 15:08 test.sh
```

초록색으로 바뀜

★ 주의

```
inPath="/data/Original_data/CDC/21_1st/"
outPath="/home/guest01/2021/yb/yb01/CDC_JJ/Analysis"

for file in ${inPath}*_1.fastq.gz
do
        stub=${file%_1.fastq.gz}
        stub2=${stub#$inPath}
        sample=$stub2
        echo $sample
        mkdir $outPath$sample
        mkdir $outPath$sample/QcReads

        /home/bioware/FaQCs/FaQCs.pl -p $inPath$stub2\_1.fas
QcReads -t 8 -adapter > $outPath$sample\/QC.log

done
~
```

outPath="~/" <-
폴더니 맨 마지막에
/(슬래시) 넣어야함

tab으로 띄어쓰기 말고
스페이스 4번으로 띄워쓰기

stub = ${file$ ~} -> 이렇게 쓰면 변수 인식 못함

```
01.QC_C_yb.sh: line 6: stub: command not found
01.QC_C_yb.sh: line 7: stub2: command not found
01.QC_C_yb.sh: line 9: sample: command not found
```

stub=${file$ ~} <- 붙여줘야 인식 함

3> 전체 스크립트

```
inPath="/data/Original_data/CDC/21_1st/"
outPath="/home/guest01/2021/yb/yb01/CDC_JJ/Analysis/"

for file in ${inPath}*_1.fastq.gz
do
    stub=${file%_1.fastq.gz}
    stub2=${stub#$inPath}
    sample=$stub2
    echo $sample
    mkdir $outPath$sample
    mkdir $outPath$sample/QcReads

    /home/bioware/FaQCs/FaQCs.pl -p $inPath$stub2\_1.fastq.gz $inPath$stub2\_2.fastq.gz -q 20 -min_L 50 -av
g_q 0 -n 1 -lc 0.85 -5end 0 -3end 0 -split_size 100000 -d $outPath\/$sample/QcReads -t 8 -adapter > $outPat
h$sample\/QC.log

done
```

xshell에서 실행시

```
[guest01@smel0:script]$ ./01.QC_C_yb.sh
CF1A_112
There were 50 or more warnings (use warnings() to see the first 50)
CF1A_114
```

warnings 발생

but> CF1A_112 폴더 들어가서 보면 별다른 문제가 없음

- CF1A_112의 log

```
Bwa extension trimming algorithm is used.
Processing /data/Original_data/CDC/21_1st/CF1A_112_1.fastq.gz /data/Origi
nal_data/CDC/21_1st/CF1A_112_2.fastq.gz file
Processed 200000/85519166
 Post Trimming Length(Mean, Std, Median, Max, Min) of 198882 reads with O
verall quality 35.61
 (150.36, 5.25, 151.0, 151, 50)
Processed 400000/85519166
 Post Trimming Length(Mean, Std, Median, Max, Min) of 198841 reads with O
verall quality 35.57
 (150.34, 5.30, 151.0, 151, 50)
Processed 600000/85519166
 Post Trimming Length(Mean, Std, Median, Max, Min) of 198839 reads with O
verall quality 35.59
 (150.35, 5.33, 151.0, 151, 50)
Processed 800000/85519166
 Post Trimming Length(Mean, Std, Median, Max, Min) of 198890 reads with O
verall quality 35.69
 (150.39, 5.11, 151.0, 151, 50)
```

- CF1A_112의 QcReads 폴더

```
[guest01@smel-cluster:QcReads]$ ll
total 30688444
-rw-rw-r-- 1 guest01 guest01 15640312587 Feb 11 13:54 QC.1.trimmed.fastq
-rw-rw-r-- 1 guest01 guest01 15627711017 Feb 11 13:55 QC.2.trimmed.fastq
-rw-rw-r-- 1 guest01 guest01          94 Feb 11 12:18 QC.fastqCount.txt
-rw-rw-r-- 1 guest01 guest01      261376 Feb 11 13:56 QC_qc_report.pdf
-rw-rw-r-- 1 guest01 guest01        1126 Feb 11 13:55 QC.stats.txt
-rw-rw-r-- 1 guest01 guest01   156667309 Feb 11 13:55 QC.unpaired.trimmed.fastq
```

- stat

```
Before Trimming
Reads #: 85519166
Total bases: 12913394066
Reads Length: 151.00

After Trimming
Reads #: 85091188 (99.50 %)
Total bases: 12799150438 (99.12 %)
Mean Reads Length: 150.42
  Paired Reads #: 84665072 (99.50 %)
  Paired total bases: 12735402787 (99.50 %)
  Unpaired Reads #: 426116 (0.50 %)
  Unpaired total bases: 63747651 (0.50 %)

Discarded reads #: 427978 (0.50 %)
Trimmed bases: 114243628 (0.88 %)
  Reads Filtered by length cutoff (50 bp): 218993 (0.26 %)
  Bases Filtered by length cutoff: 4978106 (0.04 %)
  Reads Filtered by continuous base "N" (1): 30957 (0.04 %)
  Bases Filtered by continuous base "N": 4668724 (0.04 %)
  Reads Filtered by low complexity ratio (0.8): 178028 (0.21 %)
  Bases Filtered by low complexity ratio: 21459667 (0.17 %)
  Reads Trimmed by quality (20.0): 8856611 (10.36 %)
  Bases Trimmed by quality: 69265011 (0.54 %)
  Reads Trimmed with Adapters/Primers: 254271 (0.30 %)
  Bases Trimmed with Adapters/Primers: 13872120 (0.11 %)
    Nextera-primer-adapter-1 118385 reads (0.14 %) 6456074 bases (0.05 %)
    Nextera-primer-adapter-2 135886 reads (0.16 %) 7416046 bases (0.06 %)
QC.stats.txt (END)
```

** 진주 언니께 물어보니 이상 없다고 하셨음

4> 백그라운드에서 실행하기

```
[guest01@smel0:CDC_JJ]$ rm -r AnalysisCF1A*
[guest01@smel0:CDC_JJ]$ ll
```

중단한 후 다시 실행할 때는 생긴 파일 삭제 필요
-> *는 전체 CF1A로 시작하는 모든 파일 삭제

▷ nohup 와 &

```
[guest01@smel0:script]$ nohup ./01.QC_C_yb.sh > /home/guest01/2021/yb/yb01/C
DC_JJ/Analysis/01.QC_C.nhlog &
[1] 8452
[guest01@smel0:script]$ nohup: ignoring input and redirecting stderr to stdo
ut
```

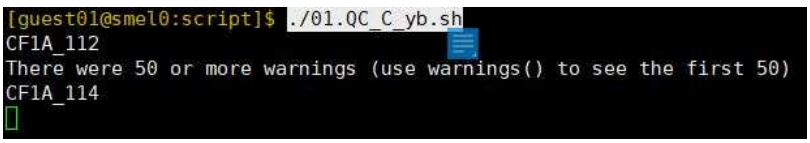
▷nohup ./XXX.sh > XX.nhlog 2>&1 &

① 실행할 셸 스크립트
② 로그 적을 파일
③ 표준 출력과 표준 에러를 로그 파일에 같이 써라(내 실행에는 안 적음)
④ 백그라운드로 돌려라

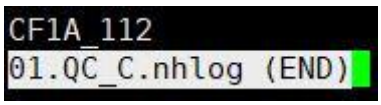
** 여기 좋은 정보
https://joonyon.tistory.com/98

** 백그라운드로 돌리면 내가 xshell 프로그램을 끄거나 컴퓨터를 꺼도 서버에서 실행 다 할 때까지 돌아감

** 여기서 로그에 쌓이는 출력은 xshell 터미널상에 기록되는 이런 출력



nhlog



*** 2022-02-11 5:11 pm/last check