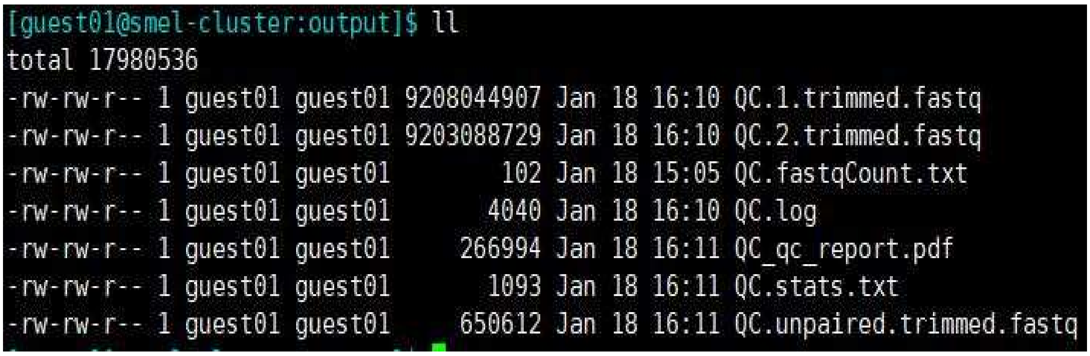
FaQC 분석

2022-01-19(수)

20193852 문유빈

1) 목표 : FaQC output file 분석

2) QC output



```
[guest01@smel-cluster:output]$ ll
total 17980536
-rw-rw-r-- 1 guest01 guest01 9208044907 Jan 18 16:10 QC.1.trimmed.fastq
-rw-rw-r-- 1 guest01 guest01 9203088729 Jan 18 16:10 QC.2.trimmed.fastq
-rw-rw-r-- 1 guest01 guest01        102 Jan 18 15:05 QC.fastqCount.txt
-rw-rw-r-- 1 guest01 guest01       4040 Jan 18 16:10 QC.log
-rw-rw-r-- 1 guest01 guest01     266994 Jan 18 16:11 QC_qc_report.pdf
-rw-rw-r-- 1 guest01 guest01       1093 Jan 18 16:11 QC.stats.txt
-rw-rw-r-- 1 guest01 guest01     650612 Jan 18 16:11 QC.unpaired.trimmed.fastq
```

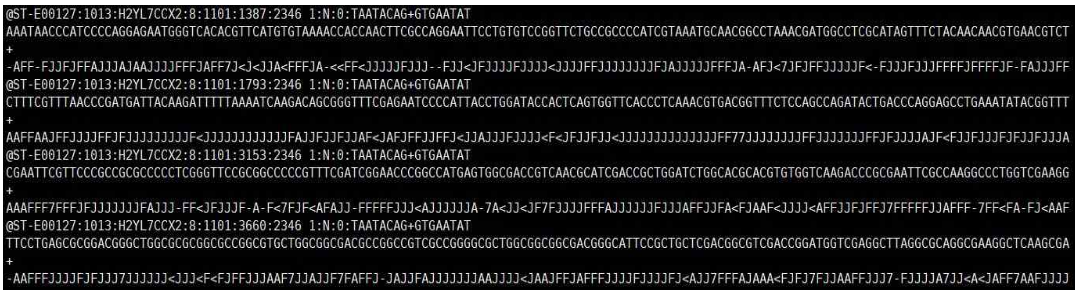
① QC.1.trimmed.fastq & QC.2.trimmed.fastq

Paired-ends 파일이 input 됐을 때 나오는 두 개의 paired-ends files

*QC.1~ 파일은 페어의 첫 번째 멤버

*QC.2~ 파일은 페어의 두 번째 멤버

*파일1, 파일2가 paired된 상태



```
@ST-E00127:1013:H2YL7CCX2:8:1101:1387:2346 1:N:0:TAATACAG+GTGAATAT
AAATAACCCATCCCCAGGAGAATGGGTCACACGTTCATGTGTAAAACCACCAACTTCGCCAGGAATTCCTGTGTCCGGTTCTGCCGCCCCATCGTAAATGCAACGGCCTAAACGATGGCCTCGCATAGTTTCTACAACAACGTGAACGTCT
+
-AFF-FJJFJFFAJJJAJAAJJJJFFFJAFF7J<J<JJA<FFFJA-<<FF<JJJJJFJJJ--FJJ<JFJJJJFJJJJ<JJJJFFJJJJJJJJFJAJJJJJFFFJA-AFJ<7JFJFFJJJJJF<-FJJJFJJJFFFFJFFFFJF-FAJJJFF
@ST-E00127:1013:H2YL7CCX2:8:1101:1793:2346 1:N:0:TAATACAG+GTGAATAT
CTTTCGTTTAACCCGATGATTACAAGATTTTTAAAATCAAGACAGCGGGTTTCGAGAATCCCCATTACCTGGATACCACTCAGTGGTTCACCCTCAAACGTGACGGTTTCTCCAGCCAGATACTGACCCAGGAGCCTGAAATATACGGTTT
+
AAFFAAJFFJJJJFFJFJJJJJJJJJF<JJJJJJJJJJJJFAJJFJJFJJAF<JAFJFFJJFFJ<JJAJJJFJJJJ<F<JFJJFJJ<JJJJJJJJJJJJJJFF77JJJJJJJJFFJJJJJJJFFJFJJJJAJF<FJJFJJJFJFJJFJJJA
@ST-E00127:1013:H2YL7CCX2:8:1101:3153:2346 1:N:0:TAATACAG+GTGAATAT
CGAATTCGTTCCCGCCGCGCCCCCTCGGGTTCCGCGGCCCCCGTTTCGATCGGAACCCGGCCATGAGTGGCGACCGTCAACGCATCGACCGCTGGATCTGGCACGCACGTGTGGTCAAGACCCGCGAATTCGCCAAGGCCCTGGTCGAAGG
+
AAAFFF7FFFJFJJJJJJJFAJJJ-FF<JFJJJF-A-F<7FJF<AFAJJ-FFFFFJJJ<AJJJJJJA-7A<JJ<JF7FJJJJFFFAJJJJJJFJJJAFFJJFA<FJAAF<JJJJ<AFFJJFJFFJ7FFFFFJJAFFF-7FF<FA-FJ<AAF
@ST-E00127:1013:H2YL7CCX2:8:1101:3660:2346 1:N:0:TAATACAG+GTGAATAT
TTCCTGAGCGCGGACGGGCTGGCGCGCGGCGCCGGCGTGCTGGCGGCGACGCCGGCCGTCGCCGGGGCGCTGGCGGCGGCGACGGGCATTCCGCTGCTCGACGGCGTCGACCGGATGGTCGAGGCTTAGGCGCAGGCGAAGGCTCAAGCGA
+
-AAFFFJJJJFJFJJJ7JJJJJJ<JJJ<F<FJFFJJJAAF7JJAJJF7FAFFJ-JAJJFAJJJJJJJAAJJJJ<JAAJFFJAFFFJJJJFJJJJFJ<AJJ7FFFAJAAA<FJFJ7FJJAAFFJJJ7-FJJJJA7JJ<A<JAFF7AAFJJJJ
```

그림 1) QC.1.trimmed.fastq

```
@ST-E00127:1013:H2YL7CCX2:8:1101:1387:2346 2:N:0:TAATACAG+GTGAATAT
AATGAGGAATTATGAAACCTTTAGTATTAGTGCGCGGTGGCGGAGATATTGCGTCTGGCGCGATTTATCGATTGAAACGAGCAGGCTATAAGGTTATTGTCAATGAAATCGCTATTCCTACTATGATTCGCCGTGAAGTTTGCTAGGGAA
+
AAAFFJJJFJFJJJJJJJ-FJJJJJJJFJJJFFJJAFFJJJJFJAF<JFJAJJAJ-FJJJJJJFJFJAJJFF-FJJJJJJA7<JJA<<-77AFJJJJAAFAFF<<7AF7-AA-7J--7-7-7---7AJ77AAA--)77AA7<<---7)-7
@ST-E00127:1013:H2YL7CCX2:8:1101:1793:2346 2:N:0:TAATACAG+GTGAATAT
TTCACAAAACGGTTTTTTTGGTGCCATTAAAAGTGCAGGATTACAGCCGCTATTTTTTGCAGGTATTGTCGGCAATCCATTCTGTTTTCAGCAACAGCGAAAACGAAAGTATTGTAATTGAAATAATCAACTCGGTTTACAATGCCGTGGA
+
AAFFFFJJJJJJAJJJJJJAJJJJJJJJJJJJJJJJJJJJJJJJJFJFJJJJJJJJJJJJJJJF<FJJJ7F<<J-FFAJFJFJJAFJJJAJFF-FJAFFFJFAAAAFJJJFAAJJ-<JF77-AAA-FFF-<J--7A<7F-<-7<7A7-7FJ)<
@ST-E00127:1013:H2YL7CCX2:8:1101:3153:2346 2:N:0:TAATACAG+GTGAATAT
CGGGGCGACGTTCTCGGGCGTTTCGGCCTCGCGCGGAGGCGGTGGCGGGCTCAGGTCCTCGAACAGGGTGGCGGCGAGCTCGGCGGAGCCGCGCCGTTCGAGGAAGCCGAGGATGCGCAGGACGCGGACACGGGCGGGCAGGGTGATGGTC
+
AAF-F<77-7-AF--7-AF7---<-<FJFJJJFJJJJJJJJJFJJJJFJ<-AJFFJJJJFFJJJJJJAFJJJFF-AJJAJFJJ<AJAJAFJJJJJFJJAF-7FFAA-FJF<JJA<FFAJJJFJFFFF)<)A)7)<FF<--<F)7)<FFJ<
@ST-E00127:1013:H2YL7CCX2:8:1101:3660:2346 2:N:0:TAATACAG+GTGAATAT
GTCGAGATTGACGGCGCAGTTCACCACCGTGGTGATGCCGTGCCGCTTCAGGAGCGCGATGTCCGACGCACCCTCGCGGCCGCCGACATAGAGATCGACGCCGTAGCCGGGCAGGGCGGCGTAGACGCGGGAGATGTGCGGCCGGTCATGA
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFAJJ<FFJJJJJJJJJJJJJJJJJJJJJJJJ-A<AJJ<JJF<J<J<FJAJFA<JJJ7AF<FJJ-7FJF----FFAJFAA)AJAJ-)AAF77)-7--7<<)<7F)7A7<-
```

그림 2) QC.2.trimmed.fastq

-> QC.1.trimmed.fastq

@ST-E00127:1013:H2YL7CCX2:8:1101:1387:2346 1:N:0:TAATACAG+GTGAATAT

fw

-> QC.2.trimmed.fastq

@ST-E00127:1013:H2YL7CCX2:8:1101:1387:2346 2:N:0:TAATACAG+GTGAATAT

bw

** 1: vs 2: => the member of a pair, 1 or 2 (paired-end or mate-pair reads only)

②QC.fastqCount.txt

```
KP-WWTP-1512_1.fastq.gz 24590939        3713231789      151.00
KP-WWTP-1512_2.fastq.gz 24590939        3713231789      151.00
QC.fastqCount.txt (END)
```

=> input file 1 reads bases reads length

=> input file 2 reads bases reads length

** 위 아래를 합하면 전체 reads, bases

③ QC.log

```
Bwa extension trimming algorithm is used.
Processing /data/Original_data/KOPRI/WWTP/KP-WWTP-1512_1.fastq.gz /data/Original_data/KOPRI/WWTP/KP-WWTP-1512_2.fastq.gz file
Processed 2000000/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999948 reads with Overall quality 37.92
 (150.88, 2.78, 151.0, 151, 50)
Processed 4000000/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999836 reads with Overall quality 37.77
 (150.86, 2.96, 151.0, 151, 50)
Processed 6000000/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999939 reads with Overall quality 37.97
 (150.86, 3.01, 151.0, 151, 50)
Processed 8000000/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999954 reads with Overall quality 37.73
 (150.87, 2.86, 151.0, 151, 50)
Processed 10000000/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999965 reads with Overall quality 37.82
 (150.85, 3.14, 151.0, 151, 50)
```

```
Processed 40000000/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999912 reads with Overall quality 37.21
 (150.90, 2.51, 151.0, 151, 50)
Processed 41181878/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1181839 reads with Overall quality 37.46
 (150.91, 2.32, 151.0, 151, 51)
Processed 43181878/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999936 reads with Overall quality 37.79
 (150.91, 2.39, 151.0, 151, 50)
Processed 45181878/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999916 reads with Overall quality 37.37
 (150.90, 2.47, 151.0, 151, 50)
Processed 47181878/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999929 reads with Overall quality 37.46
 (150.90, 2.55, 151.0, 151, 50)
Processed 49181878/49181878
 Post Trimming Length(Mean, Std, Median, Max, Min) of 1999910 reads with Overall quality 37.58
 (150.90, 2.49, 151.0, 151, 50)
(END)
```

* 진행상황 기록
* 진행된 reads/전체 reads
* overall quality = quality score?

④ QC_qc_report.pdf => Quality report pdf file

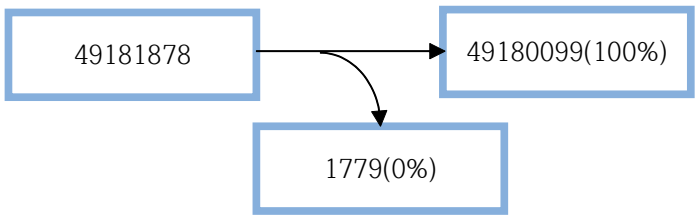
⑤ QC.stats.txt

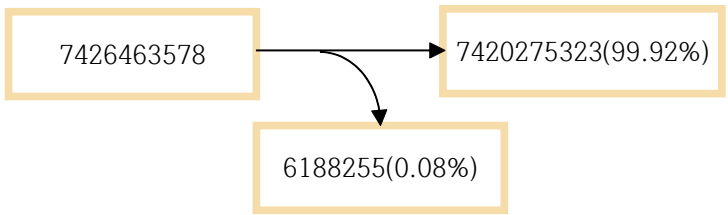
```
Before Trimming
Reads #: 49181878
Total bases: 7426463578
Reads Length: 151.00

After Trimming
Reads #: 49180099 (100.00 %)
Total bases: 7420275323 (99.92 %)
Mean Reads Length: 150.88
  Paired Reads #: 49178332 (100.00 %)
  Paired total bases: 7420014071 (100.00 %)
  Unpaired Reads #: 1767 (0.00 %)
  Unpaired total bases: 261252 (0.00 %)

Discarded reads #: 1779 (0.00 %)
Trimmed bases: 6188255 (0.08 %)
  Reads Filtered by length cutoff (50 bp): 97 (0.00 %)
  Bases Filtered by length cutoff: 3921 (0.00 %)
  Reads Filtered by continuous base "N" (2): 181 (0.00 %)
  Bases Filtered by continuous base "N": 27106 (0.00 %)
  Reads Filtered by low complexity ratio (0.8): 1501 (0.00 %)
  Bases Filtered by low complexity ratio: 226584 (0.00 %)
  Reads Trimmed by quality (5.0): 53669 (0.11 %)
  Bases Trimmed by quality: 53674 (0.00 %)
  Reads Trimmed with Adapters/Primers: 103313 (0.21 %)
  Bases Trimmed with Adapters/Primers: 5876970 (0.08 %)
    Nextera-primer-adapter-2 70588 reads (0.14 %) 4159953 bases (0.06 %)
    Nextera-primer-adapter-1 32725 reads (0.07 %) 1717017 bases (0.02 %)
QC.stats.txt (END)
```

Trimming 전 reads, bases, reads length

Trimming 후 reads, bases / pairedor or unpaired reads, bases

다양한 filter, trimming 방법으로 버려지는 read, bases





* reads x reads length = bases
* length cutoff(50bp) -> 최소 50bp 넘어야 하는데 못 넘어서 자른거
* continuous base “N” -> N은 A,T,G,C, 전부 될 수 있음. NNNNNN... 계속 이어진 시퀀스는 제대로 됐는지 모름, 에러일 확률 높음

⑥QC.unpaired.trimmed.fastq => unpaired 된 reads file

페어 안 된 것들만