FastUniq

2022-01-20

20193852 문유빈

1) Goal : FastUniq로 시퀀서(시퀀싱 기기) 때문에 생기는 에러(duplicate) 제거

2) FastUniq

▷Introduction:

==========

FastUniq as an ultrafast de novo tool for removal of duplicates in paired short DNA sequence reads in FASTQ format. FastUniq identifies duplicates by comparing sequences between read pairs and does not require complete genome sequences as prerequisites. FastUniq is capable of simultaneously handling reads with different lengths and results in highly efficient running time.

▷Parameter

==========

-i : The input file list of paired FSATQ sequence files [FILE IN]

Maximum 1000 pairs

This parameter is used to specify a list of paired sequence files in FASTQ format as input, in which two adjacent files with reads in the same order belong to a pair.

ex> inputlist.txt

-t : Output sequence format [q/f/p]

q : FASTQ format into TWO output files

f : FASTA format into TWO output files

p : FASTA format into ONE output file

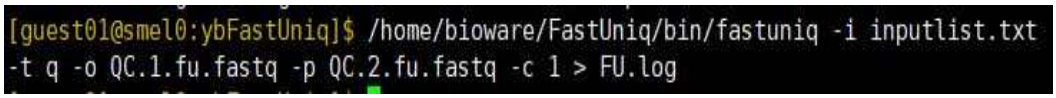
default = q

This parameter is used to specify sequence format in output file(s). FastUniq could output read pairs into two files in either FASTQ [q] or FASTA [f] format, in which reads in the same order belonging to a pair. FastUniq could also output read pairs into a single file in FASTA format [p], in which adjacent reads belonging to a pair.

-o : The first output file [FILE OUT]
>QC.1.trimmed.fastq 안에서의 duplicate 제거
-p : The second output file [FILE OUT]
Optional. ONLY required when output sequence format(-t) is specify as [q] or [f].
>QC.2.trimmed.fastq 안에서의 duplicate 제거

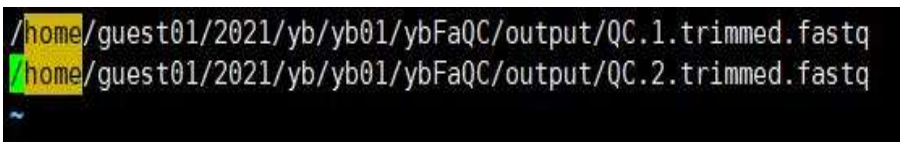
-c : Types of sequence descriptions for output [0/1]
0 : The raw descriptions
1 : New serial numbers assigned by FastUniq
default = 0
> 원래는 0 많이 쓰는데 진주 언니 할 때 1 씀

3) 입력 코드
① ② ③ ④
/home/bioware/FastUniq/bin/fastuniq -i inputlist.txt -t q -o QC.1.fu.fastq -p QC.2.fu.fastq -c 1 > FU.log
⑤ ⑥



① /home/bioware/FastUniq/bin/fastuniq => fastuniq 프로그램 절대 위치

② -i inputlist.txt => input file 리스트



** fastuniq 돌릴 pair된 fastq 파일 두 개를 한 txt 파일안에 적기
** 해당 파일의 위치는 xshell에 pwd 입력하면 나옴(절대 경로 적기)
** txt 파일 만들 때는 cat > 또는 vi 사용

③ -t q => FASTQ format 2개의 output files

④ -o QC.1.fu.fastq => QC.1.trimmed.fastq 파일의 fastuniq 돌린 결과 파일

⑤ -p QC.2.fu.fastq => QC.2.trimmed.fastq 파일의 fastuniq 돌린 결과 파일

⑥ -c 1 => 새로운 번호 붙이기

4) Output

>ybFastUniq 디렉토리에 생성된 파일들

```
[guest01@smel0:ybFastUniq]$ ll
total 15428164
-rw-rw-r-- 1 guest01 guest01          0 Jan 20 11:41 FU.log
-rw-rw-r-- 1 guest01 guest01        120 Jan 20 11:40 inputlist.txt
-rw-rw-r-- 1 guest01 guest01 7901693604 Jan 20 11:45 QC.1.fu.fastq
-rw-rw-r-- 1 guest01 guest01 7896739512 Jan 20 11:45 QC.2.fu.fastq
```

FU.log, QC.1.fu.fastq, QC.2.fu.fastq 파일 생성됨

>QC.1.fu.fastq 파일

```
@1
AAAAAAAAAAAAAAAAAAAAAGAGGGGAGGGGAAGAAAAGTGGAAATACCTTTGGGCACCAGACACCGCAGCAATTCGACG
AGGGTGGGTGCAAACGATGTTGCAATCTGCTCCTTGTAGCCTCTGTATCATAGATTCCATGATATATTGT
+1
AAFFFFAJ<JFAJJJJFJJ<<J-FFJF7J<F<--7FA-FF-FJFF<--F<FFJJFJ<<-7F-F<AJJJF<-7-F7FFAAJJ
<JF-77FJAJJ<FJJJJ-A<-<<<<<AA<J7<F77<FJFJJAJJFFF-FJJFA-7-FJFAJFAAJJJJA-
@2
AAAAAAAAAAAAAAAAAAAAGAACCTGACGAAACACGCGTAAATAAAGCAACCTGACACGCGTATATATACGTAAGGGAAAA
TGGACAGAATGAAACGGAGAGGGGGTGGGGGAGGGGGGGTCGCGTGATTGTGGGATAGTTAGAAGTCACG
+2
AAAAFJFF-AFFFJJJFFA7<7F<FAJJA-7F-FA7-7F<-<FF-<-<-F7-<<FJJF7---7<F7AF--<AJA-<JFA<-
7----<-A<<FJ<J---7---7FFA-7<---7-7F-7-AF-7---7-7AA7-F-<----7-7FJF<<<FF
@3
AAAAAAAAAAAAAAAAAAAAGAACCTGACGAAACACGCGTATATATAGCAACCTGACACGCGTATATATACGTAAGGGAAAC
TGGACAGAATGAAACGAAGAGGGGGTGGGGGAGGGGGGGTCGCGAGTTTGTGGGATAATTGGAAGTCACG
+3
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ<JJJJJFJJFJAJFJJJ7JAJFJJJJJFJJFJJFJJJ
JJJJJJJJJJFJJJJJFJJJJJJJJ7AJJJJ<JJJJJJJ-AJ-J<J<JJF-AJJJFJJFJJJJFFFFJAJ
```

> QC.2.fu.fastq 파일

```
@1
TTACAGATTTCCAGAGCAGGCAGAAATTGGATGGTTATGTCAAACTCAAGGCCGTGCGCGAGAAGTTGCCAATTTTCCCGA
CTCGCGACAAGTTCTTGGACGTTTTGGCGAGAAATCAAGTAGTGGTCTTGACTGGAGAGACAGGTTCTGG
+1
AAFF-A-AFJAAFFJAJF<J<AFF-<FFJAA<J<FJAAJFAJ7-F-FFJJJJAJ<FJAJFJ77-AJJF<JJF<-FFJJF-<
J<AAJJAJAJFA-AF-F<JJJF7J7AFJJF<<-7AF-A<<FF-F7--A--7FF-)-7<JJFAF)-FF-7<
@2
CGGGGCATTCCGGGCCGTTGGCGTGTATTTGTCTATTTCCGCCACCTCCCGCCCCATTCTCTCCGGCGAGGATAACGCCAT
ACATTCCGGAGATAACTCCCACCATTCGCGCCTCCCAGCTGGCTGTCGTCAGCCCGGGTCGCATCATCTC
+2
AA-------<FF<--77-7-7-----<<FJJJFFJJJJJJFJFFAA<JJ<77-AA<JJA<JJJ7--7-<7-7FJJ---77J
FJFJJJJA7FAJ-77A-7FJJ7FJFJA<F-A<7<<J<A-<JFJ7)))7-AJJA)7-)-))---7<--7<7
@3
CTGGCCCTCCAATTGCCCAATCCGTCCCCAATCCACGCCCAATAATTCCCTCCATTGGGCATCGGTTAGTCGCCATTAGTA
GCTTTTGGTCACCCACTAGTAACTTTGGGCACCCATAACCGTGATTTTCCCGTGACTTCCAATTATCCCA
+3
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
FJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJJJJJJJJJJJJJJJJJJJ7
```

** duplicate 제거 전 / 후 비교

> before

```
-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
```

>after

```
-rw-rw-r-- 1 guest01 guest01 7901693604 Jan 20 11:45 QC.1.fu.fastq
-rw-rw-r-- 1 guest01 guest01 7896739512 Jan 20 11:45 QC.2.fu.fastq
```

=> 두 파일 약 1GB 정도 차이남(9.2GB, 7.9GB)
이는 많은 양으로 시퀀싱 기계(HiSeq)의 품질이 안 좋은 것을 의미