Amplicon Sequencing QA

Jiahao Huang 03/31/2019

Q1: What is the constitution of each read, for example (adapter + primer + amplified region)?
A: According to the sequencing protocol, workflow done with illumina MiSeq platform. The constitution of each read should be: sequencing adapter + index + primer + amplified region + reverse primer.

Q2: Which gene region did those reads mapped to?
A: In order to find out the mapped region, samtools is needed. Here are the commands and the comments of the workflow:

```
# Get the mapped reads from the original bam file and check the mapped reads number
samtools view -f 4 -b a1n.bam > a1n_mapped.bam
samtools view a1n_mapped.bam | wc -l
    1912681

# Do a brief overview about mapped reads
samtools view a1n_mapped.bam | cut -f1 | sort | uniq -c | sort -nr
    2 M01_0
 19663 2
 8894 19
 3471 22
 2919 18
 1966 20
 1632 10
 1541 5
 1126 8
 1006 9
 1004 3
 1031 4
  969 11
  852 6
  645 16
  632 21
  503 21
  495 17
  394 13
  298 12
  223 15
  189 14
  174 7
   16 X
   12 GL000221.1
    2 GL000228.1
    1 GL000191.1
    1 GL000192.1

# Using bleamv.count and bplf.ptf to find the mapped region
bleamv.count -f bam -m intermediate-nonmapq a1n_mapped.bam Homo_sapiens.GRCh37.87.chr.gtf \
| awk '{if(0 | exp) exp=1; exp = exp.count}'

# Check the exp.count
head exp.count
ENSG00000146648.9924235
ENSG00000141510.9924235
__no_feature 1972
ENSG00000117448 301
ENSG00000141956 301
ENSG00000184271 73
ENSG00000224458 43
ENSG00000247737 56
ENSG00000182142 56
ENSG00000110651 29
```

As for the results, there are 992420 reads, about 94.25% mapped reads, mapped to the gene ENSG00000146648. After checking this gene in IGV, I would like to conclude that most of the reads mapped to the EGFR gene on chr 7.