Gulishana Adilijiang (adiliji2)

STAT 530 Bioinformatics: Homework 5

Problem1:

Since the support of random variable Y is within (0,1), thus the value of E(Y|X) will be restricted within (0,1) as well. However, in this linear regression model, the right side of the equation $X^T ß$ could be out of this range.

Therefore, a better model for conditional mean of Y is a logistic transformed model:

$$\log(E(Y|X)/(1-E(Y|X))) = X^T ß$$

$$\text{or} \quad E(Y|X)/(1-E(Y|X)) = \exp\{ X^T ß \}$$

$$\text{or} \quad E(Y|X) = \exp\{ X^T ß \} / (1+ \exp\{ X^T ß \})$$

Problem2:

The low-quality bases will cause more false-alignment, thus decrease the quality and liability of the sequence analysis, as well as the downstream interpretations of data.
Another way to deal with low-quality bases for illumine reads is not to trim them but to correct them after superimposing them to each other. Use the most frequent sequences to modify and correct the low frequency ones. However, this read correction method is not appropriate for dif ferential expression measurement through RNA-seq, where the sequence abundance is intrinsic -ally different. Thus for RNA-seq, we should trim the low-quality bases for more accurate interpr -etation of data.

Problem3:

Part (a)

```
$ cat HW5_rna-seq.txt | fastx_barcode_splitter.pl --bcfile HW5_barcodes.txt \--bol --prefix dem
ultiplex_ --suffix ".fastq" --mismatches 1 > demultiplex.log
$ cat demultiplex.log
```

Barcode	Count	Location
1	0	demultiplex_1.fastq
10	0	demultiplex_10.fastq
11	0	demultiplex_11.fastq
12	0	demultiplex_12.fastq
13	0	demultiplex_13.fastq

14	0	demultiplex_14.fastq
15	0	demultiplex_15.fastq
16	0	demultiplex_16.fastq
17	1	demultiplex_17.fastq
18	0	demultiplex_18.fastq
19	0	demultiplex_19.fastq
2	0	demultiplex_2.fastq
20	0	demultiplex_20.fastq
21	0	demultiplex_21.fastq
22	0	demultiplex_22.fastq
23	0	demultiplex_23.fastq
24	0	demultiplex_24.fastq
25	0	demultiplex_25.fastq
26	0	demultiplex_26.fastq
27	0	demultiplex_27.fastq
28	0	demultiplex_28.fastq
29	0	demultiplex_29.fastq
3	0	demultiplex_3.fastq
30	0	demultiplex_30.fastq
31	0	demultiplex_31.fastq
32	0	demultiplex_32.fastq
33	1	demultiplex_33.fastq
34	0	demultiplex_34.fastq
35	0	demultiplex_35.fastq
36	0	demultiplex_36.fastq
37	0	demultiplex_37.fastq
38	0	demultiplex_38.fastq
39	0	demultiplex_39.fastq
4	1	demultiplex_4.fastq
40	0	demultiplex_40.fastq
41	0	demultiplex_41.fastq
42	0	demultiplex_42.fastq
5	0	demultiplex_5.fastq
6	0	demultiplex_6.fastq
7	0	demultiplex_7.fastq
8	0	demultiplex_8.fastq
9	2	demultiplex_9.fastq
unmatched	995	demultiplex_unmatched.fastq
total	1000	

The smallest number of reads is **0**, the largest number of reads is **2**.
And there are **955** of 1000 reads are unmatched.

Part (b)

Every sequence read in the rna-seq.txt file has an “N” at the location of the first nucleotide. This will induce extra mismatches, which ends up with much more unmatched reads while the mism -atch tolerance is only 1.

To fix the problem, we need to trim the first “N”s in the rna-seq.txt file, and then demultiplex th -e new file again.

```
$ fastx_trimmer -f 2 -i HW5_rna-seq.txt -o trim.fastq
$ cat trim.fastq | fastx_barcode_splitter.pl --bcfile HW5_barcodes.txt \--bol --prefix demultiplex
_ --suffix “.fastq” --mismatches 1 > demultiplex.log
$ cat demultiplex.log
```

Barcode	Count	Location
1	30	demultiplex_1.fastq
10	26	demultiplex_10.fastq
11	29	demultiplex_11.fastq
12	27	demultiplex_12.fastq
13	23	demultiplex_13.fastq
14	37	demultiplex_14.fastq
15	24	demultiplex_15.fastq
16	44	demultiplex_16.fastq
17	36	demultiplex_17.fastq
18	58	demultiplex_18.fastq
19	23	demultiplex_19.fastq
2	38	demultiplex_2.fastq
20	42	demultiplex_20.fastq
21	16	demultiplex_21.fastq
22	38	demultiplex_22.fastq
23	25	demultiplex_23.fastq
24	6	demultiplex_24.fastq
25	15	demultiplex_25.fastq
26	8	demultiplex_26.fastq
27	13	demultiplex_27.fastq
28	12	demultiplex_28.fastq
29	12	demultiplex_29.fastq
3	16	demultiplex_3.fastq
30	36	demultiplex_30.fastq
31	27	demultiplex_31.fastq
32	30	demultiplex_32.fastq
33	26	demultiplex_33.fastq
34	8	demultiplex_34.fastq
35	24	demultiplex_35.fastq
36	14	demultiplex_36.fastq
37	13	demultiplex_37.fastq

38	23	demultiplex_38.fastq
39	25	demultiplex_39.fastq
4	8	demultiplex_4.fastq
40	14	demultiplex_40.fastq
41	16	demultiplex_41.fastq
42	3	demultiplex_42.fastq
5	32	demultiplex_5.fastq
6	36	demultiplex_6.fastq
7	15	demultiplex_7.fastq
8	3	demultiplex_8.fastq
9	33	demultiplex_9.fastq
unmatched	16	demultiplex_unmatched.fastq
total	1000	

Now, the smallest number of reads is **3**, the largest number of reads is **58**.
And there are **16** of 1000 reads are unmatched.

Problem4:

$ fastq-dump HW5_SRR4017758.sra --split-files --gzip --outdir ./fastq > HW5_SRR4017758_fastq-dump.log

$ cd fastq

$ zcat HW5_SRR4017758_1.fastq.gz | head -2

@HW5_SRR4017758.1 HWI-ST1155:189:C2BA6ACXX:1:1101:1475:2039 length=100

NCAGTTACTAACACTCCATCATTCTGAGCACGTATATGCTCATTATGTGACGCTATGAATTTAATAATGCTCTTACACTCAGGAATACGATTAGGCATAG

$ zcat HW5_SRR4017758_2.fastq.gz | head -2

@HW5_SRR4017758.1 HWI-ST1155:189:C2BA6ACXX:1:1101:1475:2039 length=100

TTGATTTGTTGAATTTACCTAGATTGTATTATGGTGGTCTCGCAGGAGAGGAGTCGTTTGATAGTAATATCGTGCTTGTGACTATGCCTAATCGTATTCC

The first 10 bases of the first read of the SRR4017758_1.fast.gz file is “**NCAGTTACTA**”.

The first 10 bases of the first read of the SRR4017758_2.fast.gz file is “**TTGATTTGTT**”.