Demultiplexing: Assignment the First

7-26-2023

We will be de-multiplexing sequences with dual-matched indexes (same index on both ends of the read).
N=unknown and put the read into “unknown” category for the simplest version of quality filtering.
Index hopping has occurred when index reverse complement does not match but is actually another index from the list of provided indexes known to be in the multiplexed samples.

Notes:
With open files as fhread1, fhread2, fhindex1, fhindex2
If startswith (@) in fh:
Read1 = fhread1.readline
Read 2 = fhread2.readline
Index 1 = fhindex1.readline
Index 2 = fhindex2.readline
Compare index to list of indexes
Append to header: “index - RC of index”
Eg, TCTTCGAC-TTCTCGAC (index seq and reverse complement of index seq: easy to look at and see that they match).

Bash commands for initial file exploration:

```
>>zcat <filename> | head -4 | grep -A1 "^@" --no-group-separator | grep -v "^@" | wc -m
102
```

102 -1(newline character) = 101nt per read.
9-1(newline character)=8nt per index.

Determine the encoding (and Q-score binning):
Illumina 1.8+ uses Phred+33 encoding for quality scores (J scores are present).

Made test files

```
>> zcat <filename> | head -40 > <testfilename>
```

Installed matplotlib to my bgmp_py311 conda environment so I can run matplotlib on it.

First sbatch didn’t work because I didn’t unzip the files. Added gzip open to .py
Second sbatch didn’t work because I was trying to write to the bgmp shared folder. Oops.
Ran a third time and it worked.
Ran a fourth time to include the read name specifically in the title of my plots.

Resource summary:

For R1
Maximum resident set size (kbytes): 68000
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 1:58:31
Exit status: 0

For R2
Maximum resident set size (kbytes): 70352
Percent of CPU this job got: 98%
Elapsed (wall clock) time (h:mm:ss or m:ss): 16:09.91
Exit status: 0

For R3
Maximum resident set size (kbytes): 65888
Percent of CPU this job got: 98%
Elapsed (wall clock) time (h:mm:ss or m:ss): 15:51.88
Exit status: 0

For R4
Maximum resident set size (kbytes): 70264
Percent of CPU this job got: 99%
Elapsed (wall clock) time (h:mm:ss or m:ss): 1:58:10
Exit status: 0