

Solutions

Data upload and inspection

```
library(ShortRead)
# d<-readFastq(dirPath='C:/Users/daniel/Documents/science/teaching/cesky
# krumlov 2016/course.materials/R.files',
# pattern='illumina.SE100.fastq', withIds=T)
id(d)[1000:1002]

## A BStringSet instance of length 3
## width seq
## [1] 53 BS-DSFCNTROL03:312:C3...1101:6287:2090 1:Y:0:
## [2] 53 BS-DSFCNTROL03:312:C3...1101:6319:2094 1:Y:0:
## [3] 53 BS-DSFCNTROL03:312:C3...1101:6383:2100 1:Y:0:

sread(d)[1000:1002]

## A DNASTringSet instance of length 3
## width seq
## [1] 100 AAGGGTGCAGGAGCAGCCATGA...TGAAGTANNGAATACTAATTA
## [2] 100 AATTTTGCAGGTGGTTGTTACC...ACAGCTTNNCTCGAGAGGTT
## [3] 100 CAGCTTTAGCAATAGCACCAGA...TTTAATANNAGCATCACCCAT

quality(d)[1000:1002]

## class: FastqQuality
## quality:
## A BStringSet instance of length 3
## width seq
## [1] 100 <<<<?222<222=?<?>????...#####
## [2] 100 <<<@@@2<@.24@@@#@@@@...#####
## [3] 100 <<<@?22:.@2@??>??@??...#####
```

New object X containing subset of d. Count and remove N-reads

```
X <- d[10001:20000]
length(grep("N", sread(X)))/length(X)
```

```
## [1] 0.0043
```

```
X <- clean(X)
```

Proportion of X derived from individual CGATA, and new object Y containing only these data

```
length(grep("CGATA", narrow(sread(X), start = 1, width = 5)))/length(X)
```

```
## [1] 0.0141
```

```
Y <- X[grep("CGATA", narrow(sread(X), start = 1, width = 5))]
sread(Y)[1:5]
```

```
## A DNASTringSet instance of length 5
```

```
## width seq
```

```
## [1] 100 CGATATGCAGGACTTCTCAACG...CAAGAGAAGCCGTCTATCTGG
```

```
## [2] 100 CGATATGCAGGAATCCTAAAAA...AAGAGAAAATGTTTCGTTTAC
```

```
## [3] 100 CGATATGCAGGCTGTTTCCTCC...TCTTTATACCGTGTTGGTCAG
```

```
## [4] 100 CGATATGCAGGAGTTTTGATGG...TGGCCTGTCATGACTTTTTAT
```

```
## [5] 100 CGATATGCAAGAGTACTATAAA...ATGATGGAAACTGCTTCTGGA
```

Proportion of Y with ok restriction overhang, and new object Z containing only these data

```
length(grep("TGCAGG", narrow(sread(Y), start = 6, width = 6)))/length(Y)

## [1] 0.871

Z <- Y[grep("TGCAGG", narrow(sread(Y), start = 6, width = 6))]
```

Clip barcodes and write out to fastq

```
Z <- narrow(Z, start = 6, end = 100)
# writeFastq(Z,
# file='C:/Users/daniel/Documents/science/teaching/cesky
# krumlov 2016/course.materials/R.files/CGATA.fastq')
```

Ok, data ready for alignment!