Assessing Uncorrected Insertions and Deletions - Canu Assemblies

Taruna Schuelke

6/14/2019

Load the packages

Read files into R

```
polished <- "/Users/Taruna/Dropbox/Wilbanks-lab/MinION/annie-row/genome-assemblies/blast/canu-largestCon
unpolished <- "/Users/Taruna/Dropbox/Wilbanks-lab/MinION/annie-row/genome-assemblies/blast/canu-largestC
polishedTable <- read.table(file.path(polished), sep = "\t", header = FALSE, check.names = FALSE)
unpolishedTable <- read.table(file.path(unpolished), sep = "\t", header = FALSE, check.names = FALSE)
head(unpolishedTable)
```

```
##               V1  V2                  V3  V4
## 1 PNMPKLKM_00001 164 sp|P37735|DCTP_RHOCA 333
## 2 PNMPKLKM_00002 134 sp|A4Z0A1|NAPA_BRASO 834
## 3 PNMPKLKM_00003  59 sp|B1GVX6|BOA5_BOTFB 375
## 4 PNMPKLKM_00004 370 sp|D7REY3|CDHA_PSEU3 791
## 5 PNMPKLKM_00005 214 sp|P19913|DCML_HYDPS 803
## 6 PNMPKLKM_00006  61 sp|Q4J6M3|CUTA_SULAC 748
```

Assign colnames to each table

```
colnames(polishedTable) <- c("queryID", "qLen", "subjectID", "sLen" )
colnames(unpolishedTable) <- c("queryID", "qLen", "subjectID", "sLen" )
```

Divide the qLen with sLen and create a new column

```
newPolish <- transform(polishedTable, ratio = qLen/sLen)
newUnpolish <- transform(unpolishedTable, ratio = qLen/sLen)
```

Create histogram for the Polished Assembly

```
unp_hist <- hist(newUnpolish$ratio, col = "blue",
                             xlab = "Query Len / Subject Len",
                             ylab = "Protein Freq",
                             main = "Histogram for Unpolished Canu Assembly",
                             xlim = c(0.0,4))
```

Histogram for Unpolished Canu Assembly
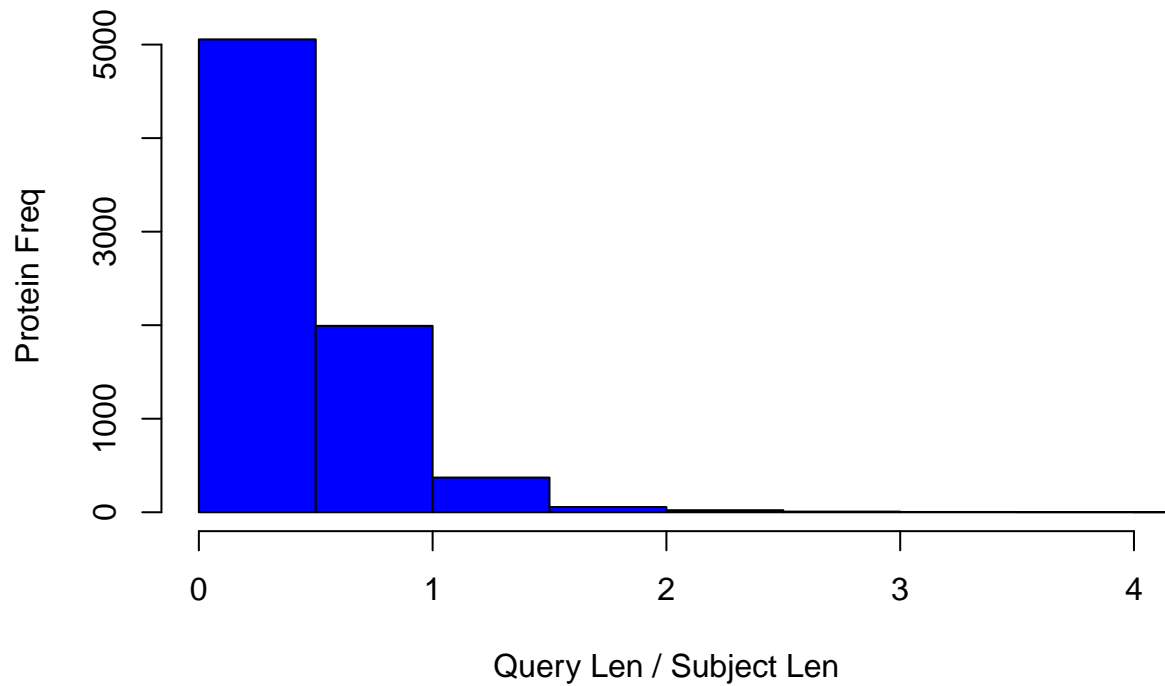
Create histogram for the Unpolished Assembly

```r
p_hist <- hist(newPolish$ratio, col = "gold",
                                 xlab = "Query Len / Subject Len",
                                 ylab = "Protein Freq",
                                 main = "Histogram for Polished Canu Assembly",
                                 xlim = c(0.0,4))
```

Histogram for Polished Canu Assembly

Protein Freq

Query Len / Subject Len