Assessing Uncorrected Insertions and Deletions - Flye 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/flye-largestCon
unpolished <- "/Users/Taruna/Dropbox/Wilbanks-lab/MinION/annie-row/genome-assemblies/blast/flye-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 IOPLDOMB_00001 276 sp|Q88EM0|DADA1_PSEPK 432
## 2 IOPLDOMB_00002 449 sp|A4WPF7|SYDND_RHOS5 591
## 3 IOPLDOMB_00003  33 sp|A1B431|SYDND_PARDP 591
## 4 IOPLDOMB_00004 158  sp|Q8YS92|NICK_NOSS1 351
## 5 IOPLDOMB_00005  97   sp|C1F3C4|DXS_ACIC5 627
## 6 IOPLDOMB_00006 152  sp|P0DM78|PHOP_SALTY 224
```

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 Flye Assembly",
                             xlim = c(0.0,4))
```

Histogram for Unpolished Flye Assembly
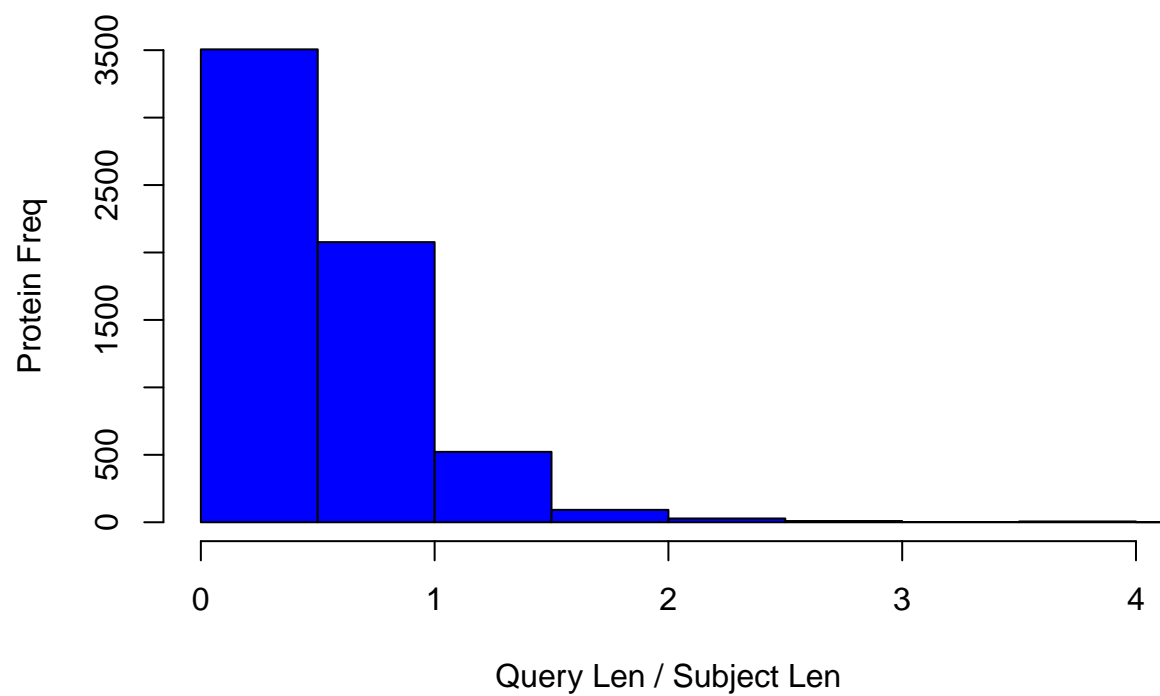
Create histogram for the Unpolished Assembly

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

Histogram for Polished Flye Assembly

Protein Freq

Query Len / Subject Len