Whole genome assembly workshop

Exercise 1. Estimate genome size

1. Using Putty (Windows) or Terminal (Mac) to connect to your assigned computer.
   Create a directory /workdir/myUserID (replace myUserID with you BioHPC ID), copy the fastq.gz file to the working directory, then de-compress the file.

```
mkdir /workdir/myUserID

cd /workdir/myUserID

cp /shared_data/assembly_workshop_2015/SRR1554178.fastq.gz ./

gunzip SRR1554178.fastq.gz
```

2. Get kmer count from the fastq file.

```
/programs/jellyfish-2.1.4/bin/jellyfish count -s 1G -m 21 -t 4 -o kmer -C SRR1554178.fastq

/programs/jellyfish-2.1.4/bin/jellyfish dump -c -t  kmer >  kmer21.txt
```

3. Get a histogram of the kmer count

```
awk '{print $2}' kmer21.txt |sort -n |uniq -c | awk '{print $2 "\t" $1}'>  histogram
```

4. Plot the kmer distribution, and estimate the genome size.

   Start R program by enter “R”.

   Run the following R code.

   The genome size is printed on the screen. A PDF file “`kmerplot.pdf`” is created with the kmer distribution. The black link is the actual distribution, the green line is the poisson fitted model.

```
data <- read.table(file="histogram",header=F)
poisson_expeact_K=function(file,start=3,end=200,step=0.1){
diff<-1e20
min<-1e20
pos<-0
total<-sum(as.numeric(file[start:dim(file)[1],1]*file[start:dim(file)[1],2]))
singleCopy_total <- sum(as.numeric(file[10:500,1]*file[10:500,2]))
for (i in seq(start,end,step))
 {
    singleC <- singleCopy_total/i
    a<-sum((dpois(start:end, i)*singleC-file[start:end,2])^2)
    if (a < diff){
      pos<-i
      diff<-a
    }
  }
  print(paste("Total kmer bases: ", total))
  print(paste("Kmer depth: ", pos))
  print(paste("Genome size: ", total/pos))
  pdf("kmerplot.pdf")
  plot(1:200,dpois(1:200, pos)*singleC, type = "l", col=3, lwd=2, lty=2,
ylim=c(0,150000), xlab="K-mer coverage",ylab="K-mer counts frequency")
  lines(file[1:200,],type="l",lwd=2)
  dev.off()
}
poisson_expeact_K(data)
```

Modified from http://arxiv.org/ftp/arxiv/papers/1308/1308.2012.pdf.