

Assignment 3

Author: “LuchaoQi” Email: lqi9@jhu.edu

Q1a.

We need $5 * 10^4$ 100bp reads.

$$n * 100bp = 1Mbp * 5$$

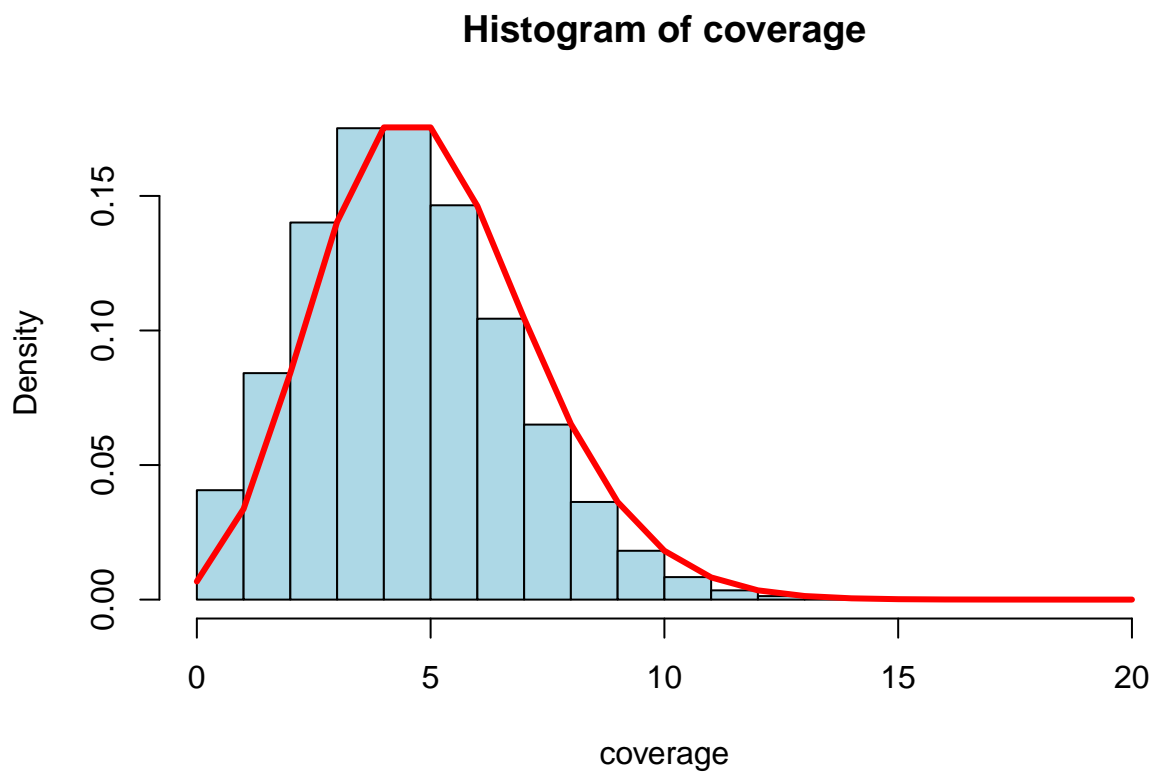
$$n = \frac{5 * 10^6}{10^2}$$

$$n = 5 * 10^4$$

Q1b.

Use following R code to simulate 5x coverage of a 1Mbp genome:

```
set.seed(100)
s = 1000000
n = 5
a = sample(1:s, n*s, replace=TRUE)
coverage = rep(0, s)
for (i in a){coverage[i] = coverage[i]+1}
hist(coverage, prob=T, col="light blue")
xfit<-seq(min(coverage), max(coverage))
yfit<-dpois(xfit, n)
lines(xfit, yfit, col="red", lwd=3)
```



```
length(which(coverage==0))
```

```
## [1] 6871
```

```
mean(coverage)
```

```
## [1] 5
```

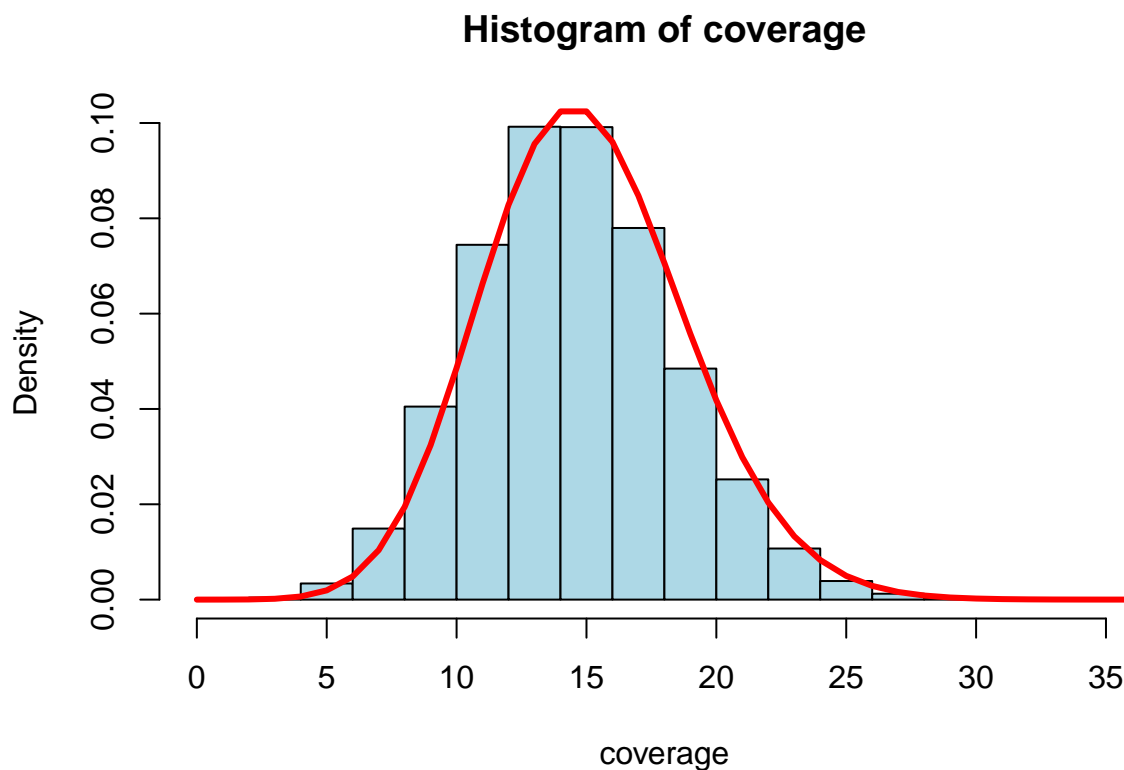
Q1c.

Theoretically, the Poisson expectation should be the value of coverage:5, which equals to exactly the mean of our simulations.

Q1d.

Use following R code to simulate 15x coverage:

```
set.seed(100)
s = 1000000
n = 15
a = sample(1:s, n*s, replace=TRUE)
coverage = rep(0, s)
for (i in a){coverage[i] = coverage[i]+1}
hist(coverage, prob=T, col="light blue")
xfit<-seq(min(coverage), max(coverage))
yfit<-dpois(xfit, n)
lines(xfit, yfit, col="red", lwd=3)
```



```
length(which(coverage==0))
```

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

```
mean(coverage)
```

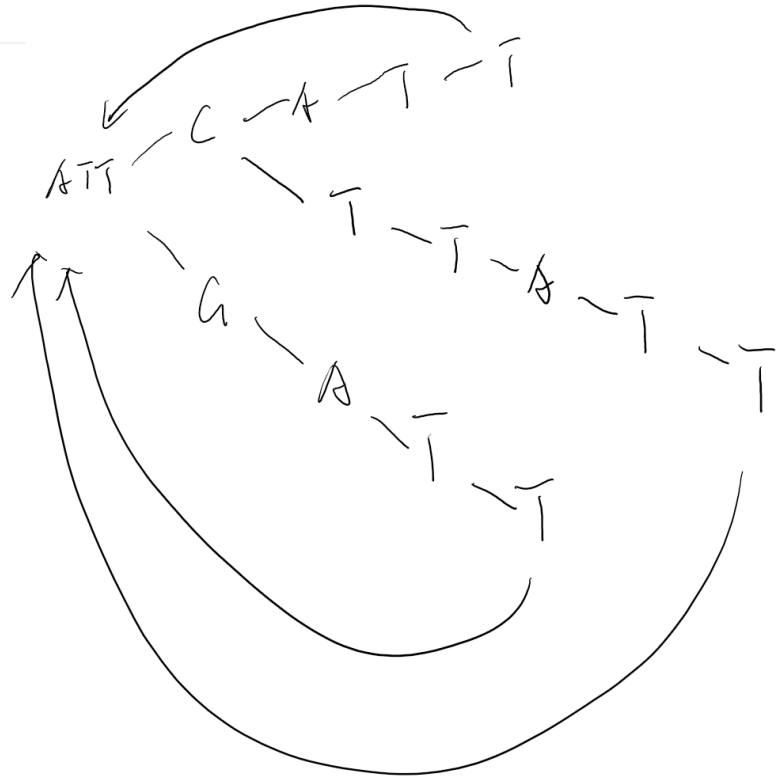
```
## [1] 15
```

Theoretically, the Poisson expectation should be the value of coverage: 15, which equals to exactly the mean of our simulations.

Q2a.

de Bruijn Graph construction

~~ATT~~C
~~CATT~~G
~~TCAT~~T
C~~TTA~~
~~GATT~~
~~TATT~~
~~TCAT~~
~~TCTT~~
~~TGAT~~
~~TTAT~~
~~TTCA~~
~~TTCT~~
~~TTGA~~



Q2b.

One possible genome sequence could be:
ATTCATTCTTATTG

Q2c.

The longest repeat should be:
ATTCTTATT

