Assignment 3

Author: "LuchaoQi" Email: lqi9@jhu.edu

Q1a.

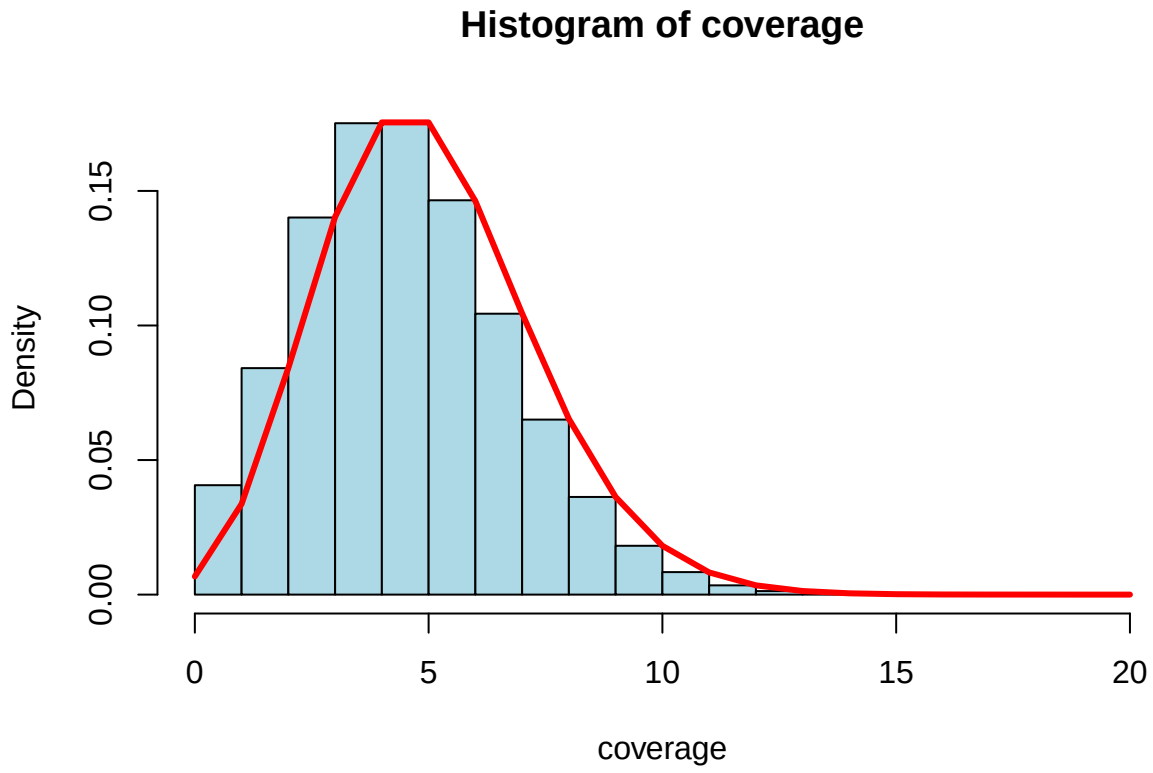
We need $5 * 10^4$ 100bp reads.

$$
\begin{aligned}
n * 100bp &= 1Mbp * 5 \\
n &= \frac{5 * 10^6}{10^2} \\
n &= 5 * 10^4
\end{aligned}
$$

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
```

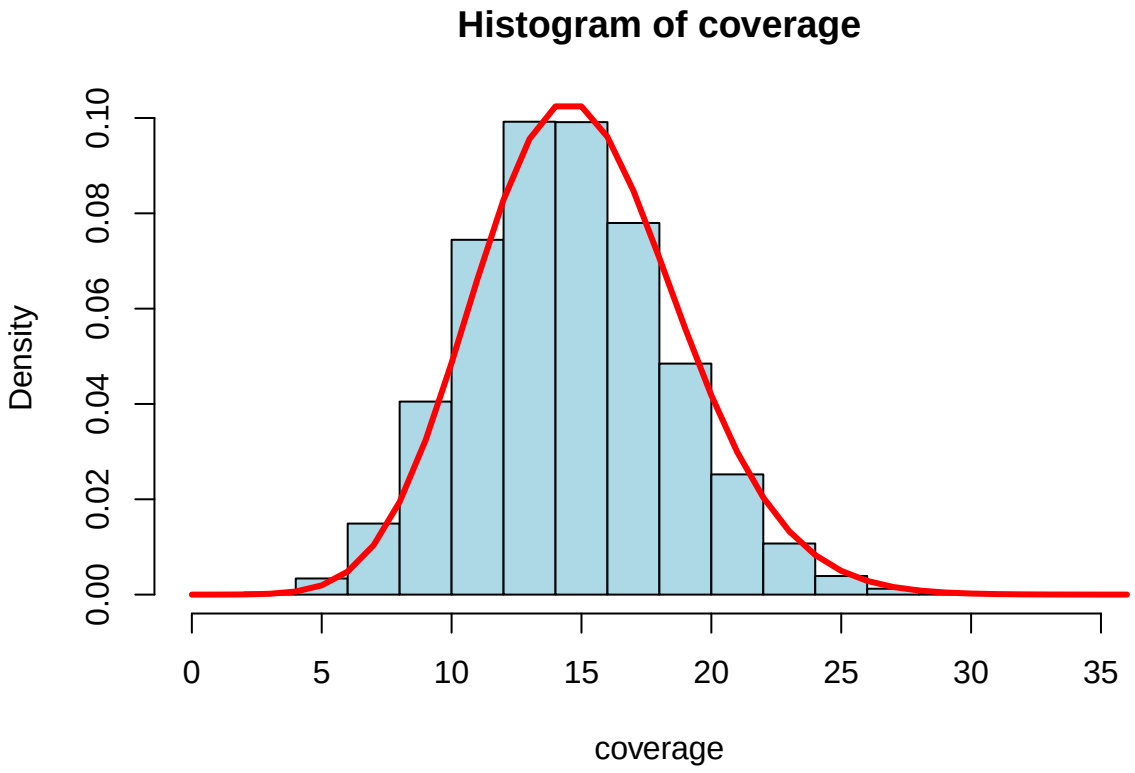
Q1c.

Theoretically, the Poisson expecation 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)
```

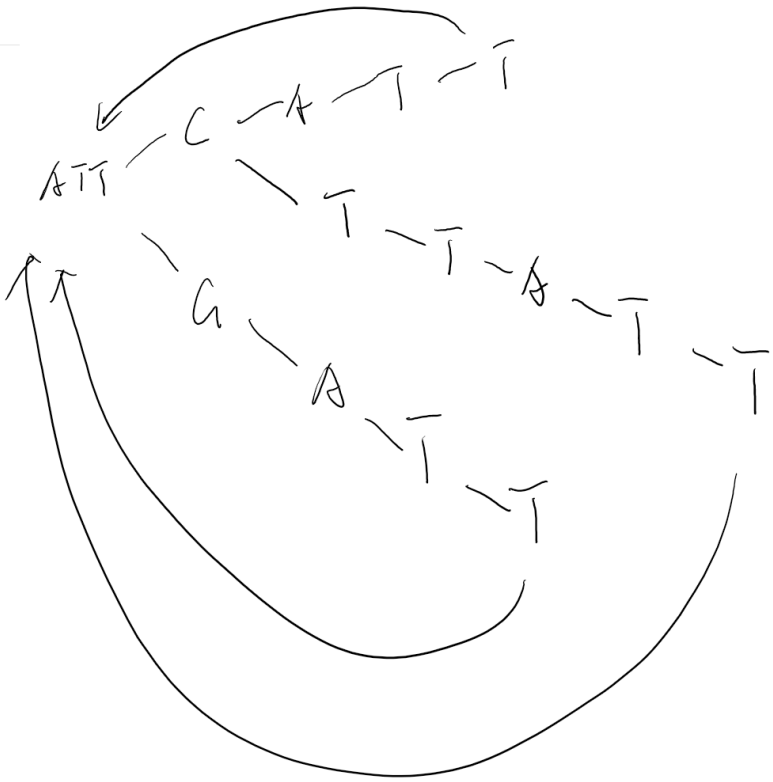
```
## [1] 15
```

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

Q2a.

de Bruijn Graph construction

~~ATTC~~
~~ATTG~~
~~CATT~~
~~CTTA~~
~~GATT~~
~~TATT~~
~~TCAT~~
~~TCTT~~
~~TGAT~~
~~TTAT~~
~~TTCA~~
~~TTCT~~
~~TTGA~~

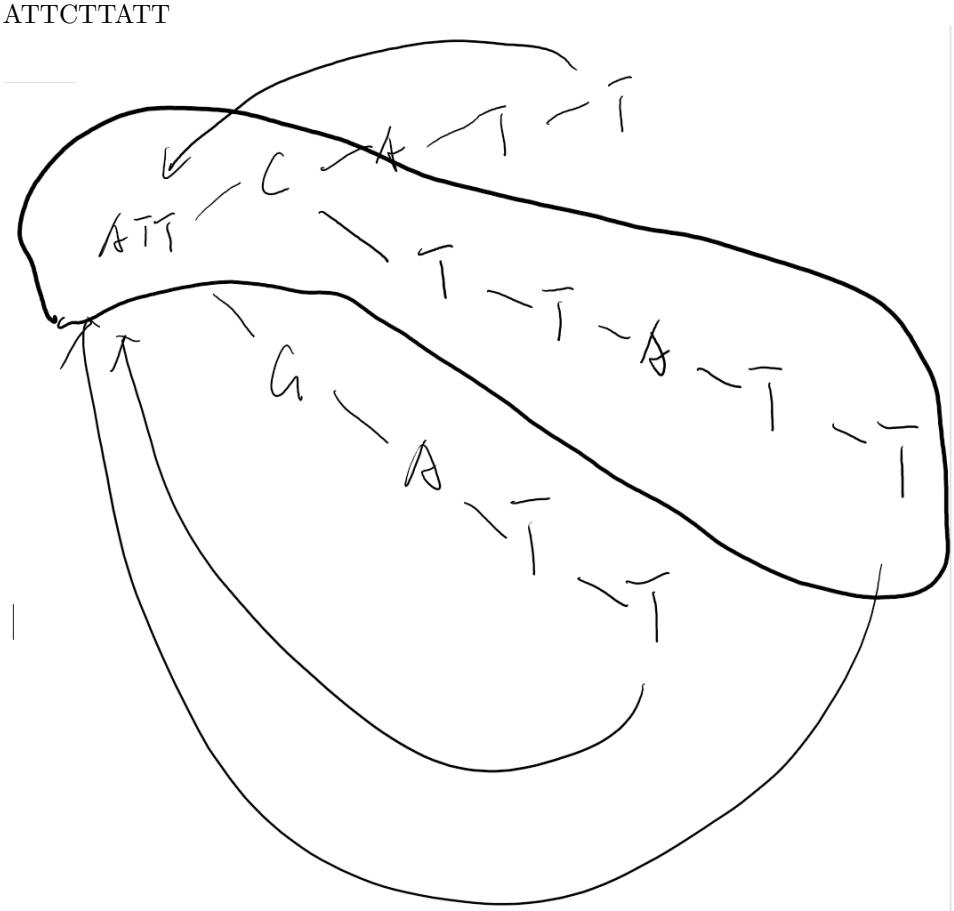


Q2b.

One possible genome sequence could be:
ATTCATTCTTATTG

Q2c.

The longest repeat should be:
ATTCTTATT



Q3a.

Through following code, We could see the results ouputted at the end of code.

```
def computeBWT(s):
    s = s +'$'
    rows = sorted(s[i:] + s[:i] for i in range(len(s)))
    bwt = [row[-1:] for row in rows]
    print("".join(bwt))
    return "".join(bwt)
computeBWT('I_am_fully_convinced_that_species_are_not_immutable;_but_that_those_belonging_to_wha
t_are_called_the_same_genera_are_lineal_descendants_of_some_other_and_generally_extinct_species,
_in_the_same_manner_as_the_acknowledged_varieties_of_any_one_species_are_the_descendants_of_that
_species._Furthermore,_I_am_convinced_that_natural_selection_has_been_the_most_important,_but_no
t_the_exclusive,_means_of_modification.')
```

```
## .etsense__$.,eIIrftassrse,;emyleeymedntt,ee,fetetssssyeeelftttedfdtsetndntgdort_ercr__ss_metd
d_____vh_hhhhhcn__a___innsseeeeax__neneeeenn__eorvlhsrhhmrmhrmhhnrnmpppplcclglbsbeccgghnnnhiiiii
ddi__ooooi_nd__n_wttttttttttttfcccctrd___vvtlgttscaabelw_auaecllaa_aao_i_r_imooei_iiiaeeoineeio
a__kaaaaooatm____sii_lccmphmnn_nssssmeeeueoaaaaaeoueeeeaentta__eeo_u_____onaosaucaauoaur_________
__or_exac_nnaftFlbbm_inn_oeelln
```

Q4.

Through following code, We could see the results ouputted at the end of code.

```
def decodeBWT(r):
    rows = [""] * len(r)
    for i in range(len(r)):
        rows = sorted(r[i] + rows[i] for i in range(len(r)))
    s = [row for row in rows if row.endswith("$")][0]
    print(s.rstrip("$").strip())
    return s.rstrip("$").strip()
decodeBWT('.uspe_gexr_______$..,e.orrs,sdddeedkdsuoden-tf,tyewtktttt,sewteb_ce__ww__h_PPsm_u_nas
eueeennnrrlmwwhWcrskkmHwhttv_no_nnwttzKt_l_ocoo_be___aaaooaAakiiooett_oooi_sslllfyyD__uouuueceet
enagan___rru_aasanIiatt__c__saacooor_ootjeae______ir__a')
```

```
## We_went_up,_saw_the_structure,_we_came_back_to_Kings_and_looked_at_our_Pattersons,_and_every_
section_of_our_Pattersons_we_looked_at_screamed_at_you,_Double_Helix._And_it_was_just_there._-_o
nce_you_knew_what_to_look_for._It_was_amazing.
```