COMS 4761, Assignment 4

Jing He:jh3283

March 1, 2013

1.Sol:

To find the index of original position in genome is in fact a tracing back question. To start with storing a vector of **??** $= [1..3 * 10^9]$, each number need 32 bits, and the total storage would be $S_0 = 32 * 3 * 10^9$. To save space, we adopt the similar thrifty storage strategy: index every I rows, so the total bits stored in total would be $log(3 * 10^9/I) * S_0/I$, so $log(3 * 10^9/I) * 3 * 10^9 \leq I * 3 * 10^9$

when the interval I ¿ 27, the storage space would be less than $3 * 10^9$, the each searching, we can tracing the index. the expectation # of index one reads will hit is $32/28 = 8/7$

pseudo code:

```
step 1 store:
for ( 1:3*10^9)
if( i modular 28 == 0)
  store index into a vector S_{1} = {1...107142857}
step 2 trace:
index_flag =0;
n = length of reads;
while (n!=0)    //trace from the end to the start
{
   if( (read[n] have index) && (index_flag ==0) )
   {  index_flag =1;
      original_index = index_val - n;
   }
```

2.a.Sol:

Heterozygote:

$$Pr(C) = Pr(A) = (1 - 1\%)/2 + 0.5 * \frac{1}{300} \tag{1}$$
$$= 0.4966667 \tag{2}$$
$$Pr(T) = Pr(G) = \frac{1}{2}\frac{1}{300} * 2 \tag{3}$$
$$= 0.003333333 \tag{4}$$

Homozygote:

$$Pr(A) = (1 - 1\%) \tag{5}$$
$$= 99\% \tag{6}$$
$$Pr(T) = P(C) = P(G) \tag{7}$$
$$= \frac{1}{300} \tag{8}$$
$$= 0.003333333 \tag{9}$$

likelihood ratio:

$$= \frac{P(Data \| Hete)}{P(Data \| Homo)} \tag{10}$$
$$= \frac{(0.4966667^{10} * 0.4966667^4 * 0.003333333}{0.99^{10} * (0.003333333^4 * 0.003333333} \tag{11}$$
$$= \frac{1.852633e - 07}{3.721735e - 13} \tag{12}$$
$$= 497787.5 \tag{13}$$

2.b.Sol:

For: $P_{homo}(P \| data)$

$$P(AC) = 1.8526 * 10{-}7 \tag{14}$$
$$P(AA) = 3.721735e - 13 \tag{15}$$
$$P(CC) = 0.99^4 * 0.003333333^{10} * 0.003333333 \tag{16}$$
$$= 5.422587e - 28 \tag{17}$$

$P_{hete}(P \| data)$

$$= \frac{P_{hete}(data \| P) * P_{hete}(P)}{\sum_{i=1}^{3} P(data \| P_i)} \tag{18}$$
$$= \frac{1.8526 * 10e - 7 * 0.81}{(1.8526 * 10e - 7 * 0.81 + 3.721735 * 10e - 13 * 0.18 + 5.422587 * 10e - 28 * 0.01)} \tag{19}$$
$$= 0.9999993 \tag{20}$$

$P_{homo}(P\|data)$

$$= \frac{P_{homo}(data\|P) * P_{homo}(P)}{\sum_{i=1}^{3} P(data\|P_i)} \tag{21}$$

$$= \frac{3.721735 * 10e-13 * 0.18}{(1.8526 * 10e-7 * 0.81 + 3.721735e-13 * 0.18 + 1.755967e-19 * 0.01)} \tag{22}$$

$$= \frac{1.500606e-06}{1.500607e-06} \tag{23}$$

$$= 7e-07 \tag{24}$$

2.c.Sol: Noted that,

$$P(\text{novel heterozygote}) = \frac{300,000}{3 * 10^9 - 3 * 10^7} \tag{25}$$

$$= 0.0001010101 \tag{26}$$

$$P(\text{homo}) = 1 - 0.0001010101 \tag{27}$$

$$= 0.999899 \tag{28}$$

$P(data|Hete)$

$$= \frac{P(Hete|data) * P(Hete)}{P(data)} \tag{29}$$

$$= \frac{1.8526 * 10e-7 * 0.0001010101}{(1.8526 * 10e-7 * 0.0001010101 + 3.721735e-13 * 0.999899)} \tag{30}$$

$$= \frac{1.871313e-10}{1.875034e-10} \tag{31}$$

$$= 0.9980155 \tag{32}$$

$P(data|Homo)$

$$= \frac{P(Homo|data) * P(Homo)}{P(data)} \tag{33}$$

$$= \frac{3.721735e-13 * 0.999899}{(1.8526 * 10e-7 * 0.0001010101 + 3.721735e-13 * 0.999899)} \tag{34}$$

$$= \frac{3.721359e-13}{1.875034e-10} \tag{35}$$

$$= 0.001984688 \tag{36}$$

2.d.Sol:

for heterozygote, per 2.a

$$P(non-referenceallel) = \frac{1}{300} \qquad P(A) = 0.99 \tag{37}$$

Using Possion model for the read number, $\lambda = x/300$,
$X \sim \text{Pois}(\lambda)$

for interval [11,a], the prob of calling a false positive when observing x reads is :
when i = [11,a],

$$P(x \geq 4) = \sum_{i=11}^{i=a} Pois(\lambda) \tag{38}$$

$$\tag{39}$$

when i = [a+1,b],

$$P(b \geq x \geq (a+1)) = \sum_{i=(a+1)}^{i=b} Pois(\lambda) \tag{40}$$

$$\tag{41}$$

when i = [b+1,29],

$$P(29 \geq x \geq (b+1)) = \sum_{i=(b+1)}^{i=29} Pois(\lambda) \tag{42}$$

$$\tag{43}$$

```
mybinom <- function(x, i) {
    p <- 0.01
    return(sum(choose(i, x) * (p)^x * (1 - p)^(i - x)))
}
for (a in 11:28) {
    res <- 0
    x <- c(0:3)
    for (i in 11:a) {
        res <- res + 1 - mybinom(x, i)
    }
    for (b in (a + 1):28) {
        x <- c(0:4)
        for (i in (a + 1):b) {
            res <- res + 1 - mybinom(x, i)
        }
        for (c in (b + 1):29) {
            x <- c(0:5)
            res <- res + 1 - mybinom(x, i)
        }
```

```
        if (res > 0.01) {
            print(c(as.character(a), as.character(b), res))
        }
    }
}
```

So, basically, we can choose random a, b in [11,29]

2.e.Sol:

For a possion with $\lambda = 20$, $P(30 > X > 10) \sim \text{Pois}(\lambda)$

```
ppois(29, 20, lower.tail = T) - ppois(9, 20, lower.tail = T)
```

```
## [1] 0.9732
```

0.9731864 0.0268136 ignore **2.f.Sol:**

For possion distribution, $\lambda = 20, whena = 15, b = 19$

```
p_hete <- function(i, cutoff) {
    1 - ppois(cutoff - 1, i * 0.4966667)
}
p_1 <- sum(unlist(lapply(11:15, FUN = function(x) {
    dpois(x, 20) * p_hete(x, 4)
})))
p_2 <- sum(unlist(lapply(16:19, FUN = function(x) {
    dpois(x, 20) * p_hete(x, 4)
})))
p_3 <- sum(unlist(lapply(20:29, FUN = function(x) {
    dpois(x, 20) * p_hete(x, 4)
})))
p_1 + p_2 + p_3
```

```
## [1] 0.9419
```

2.g.Sol:

Heterozygote AC:

$$
\begin{aligned}
& P(data|hete) && (44)\\
& = (1 - 10^{-3.7}) * (1 - 10^{-3.3}) * (1 - 10^{-3.0}) && (45)\\
& * (1 - 10^{-2.7}) * (1 - 10^{-2.3}) * (1 - 10^{-2.0}) && (46)\\
& * (1 - 10^{-1.7}) * (1 - 10^{-1.3}) * (1 - 10^{-1.0}) && (47)\\
& * (1 - 10^{-0.7}) * (1 - 10^{-1.7}) * (1 - 10^{-1.3}) && (48)\\
& * (1 - 10^{-1.0}) * (1 - 10^{-0.7}) * 10^{-0.3} && (49)\\
& = 0.2212375 && (50)
\end{aligned}
$$

Homozygote AA:

likelihood ratio:

$$= \frac{P(Data\|Hete)}{P(Data\|Homo)} \tag{51}$$

$$= \frac{0.2212375}{6.581923e-06} \tag{52}$$

$$= 33612.9 \tag{53}$$

3.a.Sol: The average distance between two reads $1000 - 2 * 100 = 800$ bp, with the $std = 200$ bp.

$$Z = \frac{(x-u)}{std/\sqrt{n}} = \frac{(1250-800)}{200/\sqrt{n}} \tag{54}$$

$$P(Z < z_\alpha) = 1 - \alpha \tag{55}$$

$$p^n < 10e-6 \tag{56}$$

$$nlog(p) < log(10e-6) \tag{57}$$

```
n <- 0
repeat {
    n <- n + 1
    z <- (1250 - 800)/(200/sqrt(n))
    p <- 1 - pnorm(z)
    if (p < 10^(-6)) {
        print(p)
        print(n)
        break
    }
}
```

```
## [1] 2.438e-07
## [1] 5
```

n= 5

3.b.Sol:

according to Possion distrition, with $\lambda = 5$

$$p(x) = \frac{\lambda^x exp(-\lambda)}{x} \tag{58}$$

$$p(x < 5) = \sum_{i=0}^{4} \frac{5^x exp{-5}}{5} \tag{59}$$

$$= 0.4404933 \tag{60}$$

Disclosure: Discuss with kuixi zhu, Boris, Ola, nanfang,xu