Problem Set 3

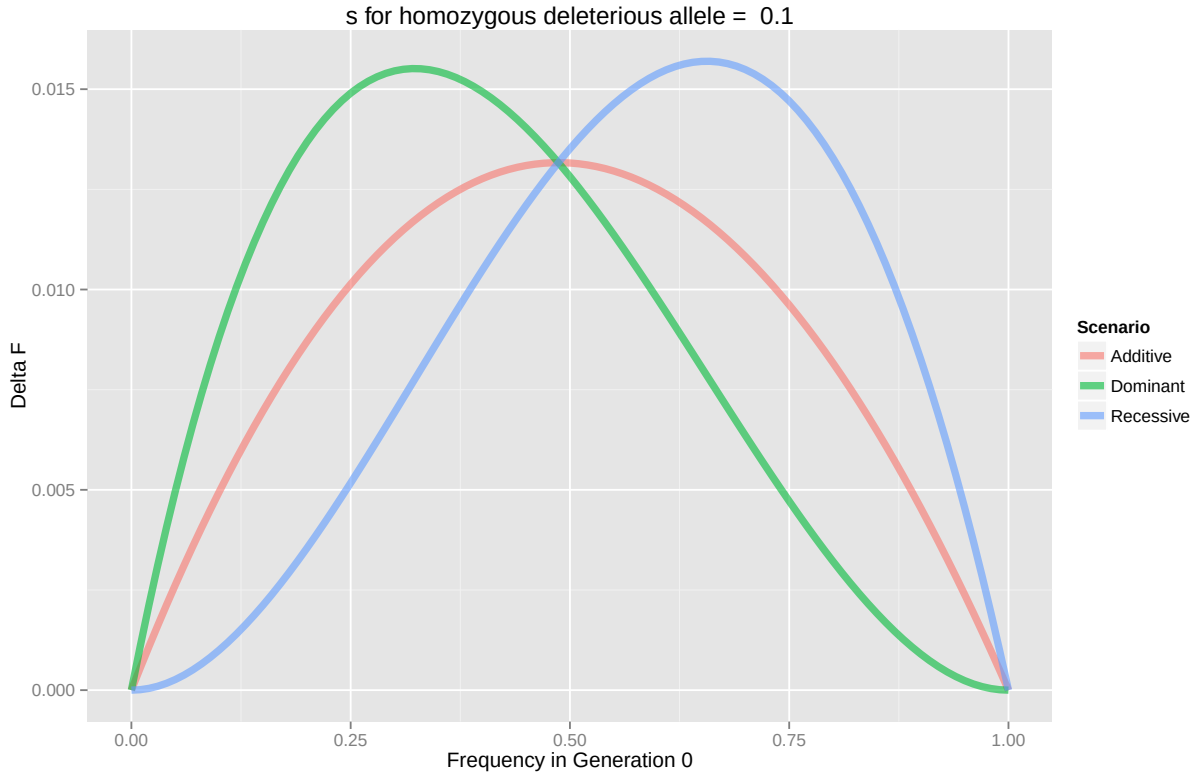
Gitanshu Munjal

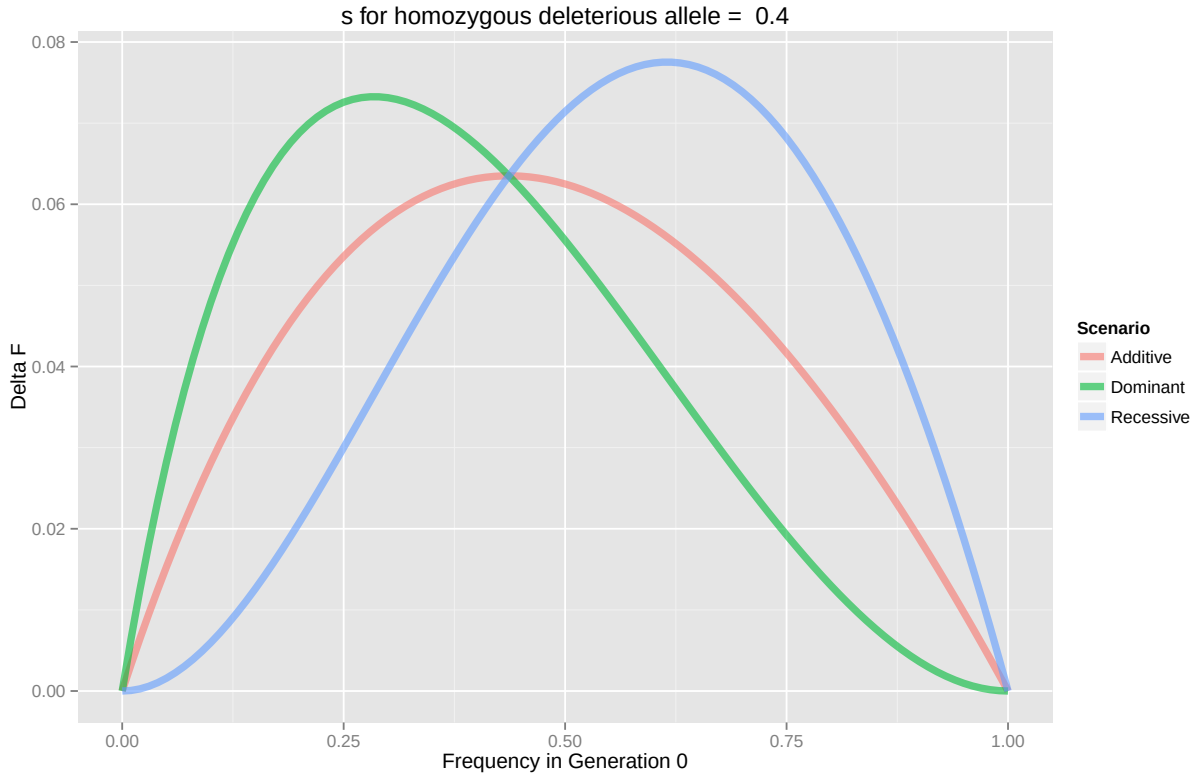
Saturday, May 9, 2015

Problem 1

Generate a specific graph to display the properties of selection in large populations.

```
############################
# Selection in Large Pops
############################

#Given Information (s, step size, scenarios)
for(s in c(0.1,0.4)){

plotdata <- data.frame(G0Freq=rep(seq(0,1,0.01),3),
                       Scenario = rep(c("Dominant","Recessive","Additive"),
                                      c(101,101,101)),
                       s_homo=0,                                    # s for adv. homozygote
                       s_hetero=rep(c(0,s,s/2),c(101,101,101)))   # s for heterozygote

#Frequency of advantageous allele in G1
plotdata$G1Freq =
#Numerator  = f(adv_hom)*(1-s_adv_hom) + (1/2)*f(het)*(1-s_het)
  (
    (plotdata$G0Freq)^2 * (1-plotdata$s_homo) +
      (0.5 * (2*(plotdata$G0Freq)*(1 - plotdata$G0Freq))  * (1-plotdata$s_hetero))
    ) /

#Denominator = f(adv_hom)*(1-s_adv_hom) + f(het)*(1-s_het) + f(del_hom)*(1-s_del_hom)
  (
    (plotdata$G0Freq)^2 * (1-plotdata$s_homo) +
      ((2*(plotdata$G0Freq)*(1 - plotdata$G0Freq))  * (1-plotdata$s_hetero)) +
      (1 - plotdata$G0Freq)^2 * (1 - s)

    )

#Delta F = f(adv_gen1) - f(adv_gen0)
plotdata$deltaF = plotdata$G1Freq - plotdata$G0Freq

library(ggplot2)
a <- ggplot(plotdata,aes(x=G0Freq, y=deltaF, colour=Scenario)) +
      geom_line(size=2, alpha=0.6) +
      labs(title= paste("s for homozygous deleterious allele = ",s),
           x= "Frequency in Generation 0", y= "Delta F")

print(a)
}
```


s for homozygous deleterious allele = 0.1
0.015
0.010
0.005
0.000
Delta F
0.00
0.25
0.50
0.75
1.00
Frequency in Generation 0
Scenario
Additive
Dominant
Recessive


s for homozygous deleterious allele = 0.4
0.08
0.06
0.04
0.02
0.00
Delta F
0.00
0.25
0.50
0.75
1.00
Frequency in Generation 0
Scenario
Additive
Dominant
Recessive

Problem 2

Four of the six plots from above are asymmetric. Explain the biological reason behind these asymmetric patterns.

The assymetric plots from above fall under two scenarios: (1) when the advantageous allele is dominant and (2) when the advantageous allele is recessive. Based on our knowledge of Hardy-Weinberg Equilibrium (HWE), we know that rare alleles in a population are found mostly in the heterozygotic state. Keeping this in mind, lets think about our scenarios from above one at a time.

Dominant Scenario

When the advantageous allele is dominant, selection only acts on (and stops from contributing to the next generation) the deleterious homozygotic genotype since the heterozygote is indistinguishable (in terms of fitness) from the advantageous homozygote. Thus, only the advantageous homozygotic and the heterozygotic genotypes contribute alleles to the next generation if s(against deleterious homozygote) = 1.

Note: if $0 < s < 1$, a (1-s) proportion of deleterious homozygotes are able to contribute also.

When the advantageous allele is at low frequency, its homozygotic state is much rare compared to its heterozygotic state (according to HWE). At this time (low frequency of advantageous allele), the population is composed mostly of deleterious recessive homozygotes and heterozygotes. Selection acts quickly (even more so if s is increased) to remove the abundant deleterious homozygotes and this drives the rapid change in allele frequency of the advantageous allele (rise of the green plots above) up until recessive homozygotes become rare and the population is mostly heterozygotes and advantageous homozygotes. The change/rise in allele frequency of the advantageous allele is much slower after that as the materials for selection (recessive homozygotes) are increasingly rarer and selection has to wait for them to appear as the heterozygotes segregate (fall of the green plots above).

Recessive Scenario

When the advantageous allele is recessive, selection acts on (and stops from contributing to the next generation) the deleterious homozygotic and heterozygotic genotype. Thus, the advantageous homozygotic genotype is the sole contributor to the next generation if $s = 1$.

Note: if $0 < s < 1$, a (1-s) proportion of deleterious genotypes are able to contribute also.

When the advantageous allele is at low frequency, its homozygotic state is much rare compared to its heterozygotic state (according to HWE). At this time (low frequency of advantageous allele), the population is composed mostly of deleterious homozygotes and heterozygotes. Even if the intensity of selection is the same as in the case for a dominant advantageous allele, the frequency of the advantageous allele does not change much (dip in the rise of the blue plot above) till the advantageous recessive homozygotes become easier to find. Subsequently, the frequency of the advantageous allele rises faster than previous but still slower compared to the dominance scenario as only one genotype contributes to the increase (rise of the blue plot). The fall of the blue plot is steep because the frequency of the advantageous allele is high and selection quickly drives this to fixation by removing a propotion (s) of the deleterious genotypes.

Problem 3

Explain the difference between coalescent effective population size (Ne) and instantaneous Ne? What are two ways to estimate coalescent Ne? What is one way to estimate instantaneous Ne?

Instantaneous N_e

Instantaneous effective population size of a population is the size of a Wright-Fisher population that experiences the same magnitude of genetic drift as the actual population under consideration. Instantaneous N_e "enables us to draw inferences concerning the evolutionary effects of finite population size by providing a mechanism for incorporating factors that result in deviations from the ideal [1]." Factors here referring to scenarios of unbalanced sex ratios, variation in offspring number, etc.

One way to estimate instantaneous N_e is to collect random individuals from successive generations of a population and genotype these at a large number of loci. Subsequently, one could estimate the change in frequencies at these different loci between successive generations. These data from all the loci genotyped could then be used to estimate the maximum likelihood size of a Wright-Fisher population that fit these data.

Coalescent N_e

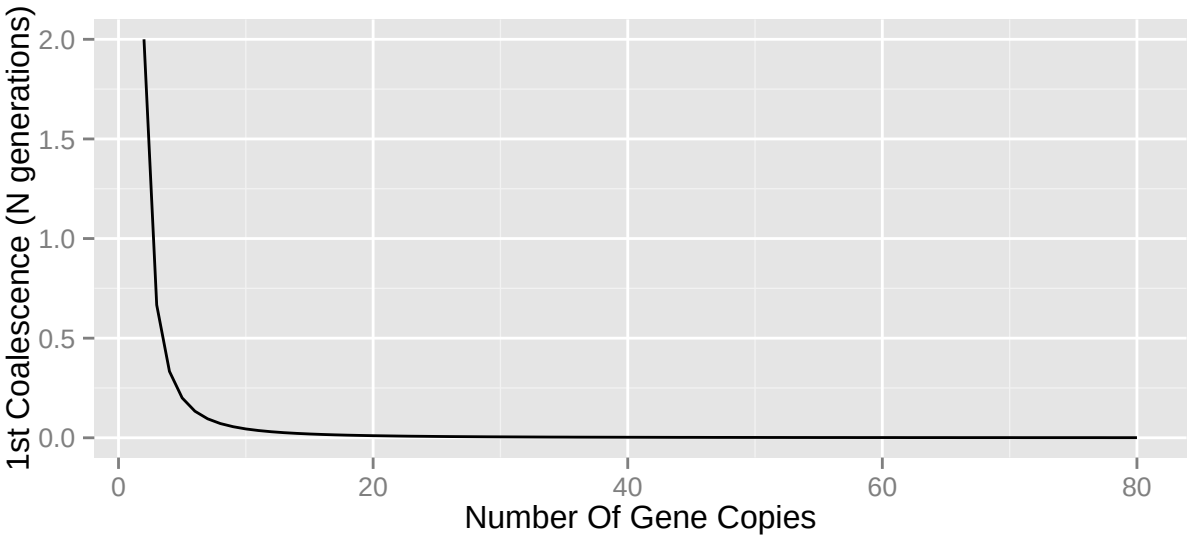
Coalescent effective population size of a population is "an estimate of the equilibrium or average N_e over a long period... it is primarily determined by the cumulative impacts of mutation and genetic drift [1]." For the sake of explanation, consider a population going through bottlenecks and other demographic events over evolutionary time (so that instantaneous N_e fluctuates through time) and yielding a coalescent N_e estimate of 2N. Then, the amount of diversity retained by this population will be equal in amount to the amount of diversity retained by a population with a constant $N_e = 2\text{N}$.

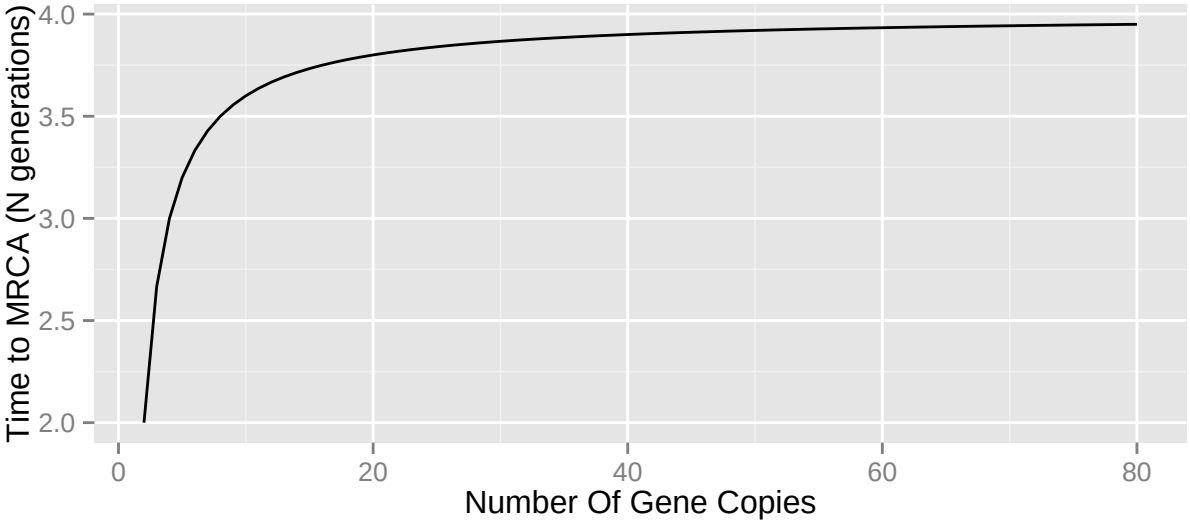
Coalescent N_e can be estimated using the relation $\theta = 4N_e\mu$. The value for θ can be estimated using Tajima's logic ($\pi = \theta_\pi = 4N_e\mu$, where π is the observed mean mutations between multiple pairs of 2 random gene copies) or using Watterson's logic ($s/(\sum_{k=1}^{n-1} \frac{1}{k} = \theta_s = 4N_e\mu$, where s is the number of segregating sites)

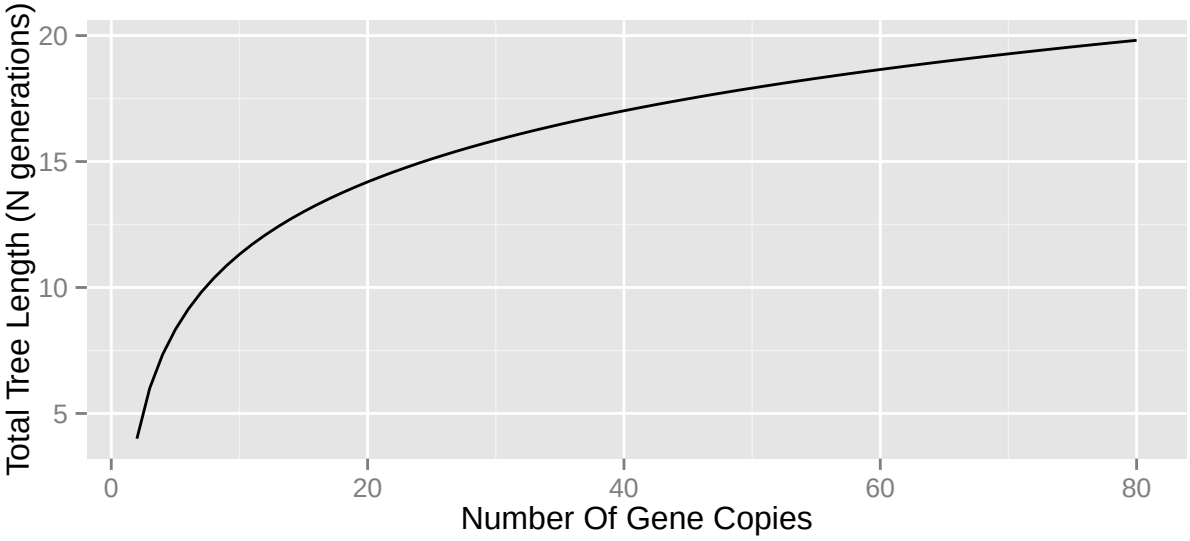
Problem 4

Generate a specific graph to display the properties of the coalescent process in a Wright-Fisher population. The x-axis should be number of gene copies and range from 2-80. The y-axis should be number of generations in N units. Perform calculations in steps of one gene copy and plot the following three expectations: (1) time to the first coalescent event; (2) time to the most recent common ancestor of all gene copies; (3) total tree length.

```
rm(list=ls())

#Given Information (step size)
plotdata <- data.frame(NumberOfGeneCopies = seq(2,80,1))

#Time to 1st Coalescence = 4 * (1/( (k)*(k-1) ))
plotdata$TimeTo1stCoalescence <-

  4 / ((plotdata$NumberOfGeneCopies)*(plotdata$NumberOfGeneCopies - 1))

#Time to MRCA = 4 * summationover2:k(1/( (k)*(k-1) ))
plotdata$TimeToMRCA <- NA
for (i in 1:nrow(plotdata)){

  k <- 2:plotdata$NumberOfGeneCopies[i]
  plotdata$TimeToMRCA[i] <- 4 * sum( 1 / ((k)*(k-1)) )

}

#Total Tree Length = 4 * summationover2:k(k/( (k)*(k-1) ))
plotdata$TotalTreeLength <- NA
for (i in 1:nrow(plotdata)){

  k <- 2:plotdata$NumberOfGeneCopies[i]
  plotdata$TotalTreeLength[i] <- 4 * sum( k / ((k)*(k-1)) )

}


library(gridExtra)
library(ggplot2)
c <- ggplot(data=plotdata, aes(x=NumberOfGeneCopies,y=TimeTo1stCoalescence)) +
      geom_line() +
      labs(x="Number Of Gene Copies", y="1st Coalescence (N generations)")
m <- ggplot(data=plotdata, aes(x=NumberOfGeneCopies,y=TimeToMRCA)) + geom_line() +
      labs(x="Number Of Gene Copies", y="Time to MRCA (N generations)")
t <- ggplot(data=plotdata, aes(x=NumberOfGeneCopies,y=TotalTreeLength)) + geom_line() +
      labs(x="Number Of Gene Copies", y="Total Tree Length (N generations)")
grid.arrange(c,m,t,nrow=3)
```


1st Coalescence (N generations)
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
Number Of Gene Copies


Time to MRCA (N generations)
4.0
3.5
3.0
2.5
2.0
0
20
40
60
80
Number Of Gene Copies


Total Tree Length (N generations)
20
15
10
5
0
20
40
60
80
Number Of Gene Copies

Problem 5

You sequence a 5.6 kb locus in 5 diploid individuals and observe 11 segregating sites. What is your estimate of theta in this population? What property of the expected coalescent tree is this estimate based on? What is your estimate of coalescent Ne assuming a mutation rate of 10^-8 per bp per generation?

```
rm(list=ls())

#Given Information
locussize <- 5.6 * 1000                    #5.6kb to bp
k <- 5 * 2                                 #chromosomes in 5 diploid individuals
s <- 11                                    #segregating sites
mu <- (10)^(-8)                            #per bp mutation rate

#Theta_s = s / summationover1:k-1(1/k)
stheta <- (s) / ( sum( 1/(1:(k-1)) ) )
stheta
```

```
## [1] 3.888343
```

Thus, θ_s = **3.89**.

This estimate derives from the total tree length (formula mentioned in problem 4). Any mutations in the tree space give rise to segregating sites so that $E[s] = TTL * \mu$ and thus, $\mu = E[s]/TTL$ (1). We also know that $\mu = \theta/4N_e$ (2). From here one can solve (1) and (2), to derive:

$$\frac{s}{\sum_{k=1}^{n-1} \frac{1}{k}} = \theta_s$$

```
#Coalescent_Ne = Theta_s/(4*mu)
mulocus <- mu * locussize                  #mutation rate for locus
coalNe <- stheta / (4 * mulocus)
round(coalNe,0)
```

```
## [1] 17359
```

Thus, coalescent N_e = **17359**

Problem 6

If you sample 100 sets of four gene copies, each set has an actual time to the first coalescent event. Do you expect the number of sets that have an actual first coalescent before 2N/6 to be approximately equal to the number of sets to have an actual first coalescent after 2N/6? Explain your answer.

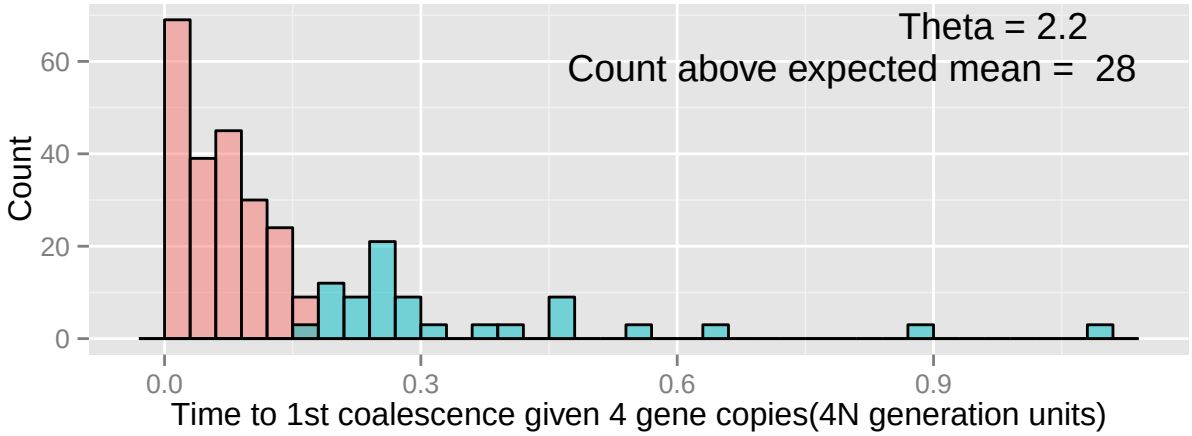
No, I do not expect the number of sets that have an actual first coalescent before 2N/6 to be approximately equal to the number of sets to have an actual first coalescent after 2N/6.

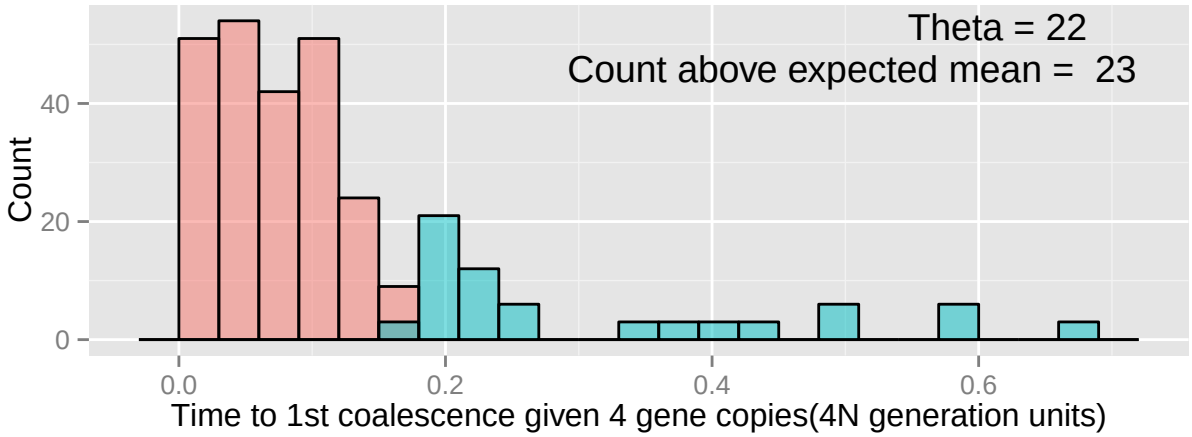
The response variable underlying this experiment (time to 1st coalescence event for 4 gene copies; replicated 100 times) essentially consists of **counts** (number of times it takes x generations) from **randomly chosen coalescence events**. The expected distribution for such a variable is a Poisson distribution as by definiton a Poisson distribution is a **discrete probability distribution for counts of occurence of rare events over a time/space interval**. Both the most frequent (mode) and middle(median) values of time for 1st coalescence would lie to the left of the mean (2N/6) and more of the area under the distribution lies to the left of the mean. Given these facts, **I expect the number of sets that have an actual first coalescent before 2N/6 to be higher** than the number of sets that have an actual first coalescent after 2N/6. This conclusion was also supported by ms simulations [2]

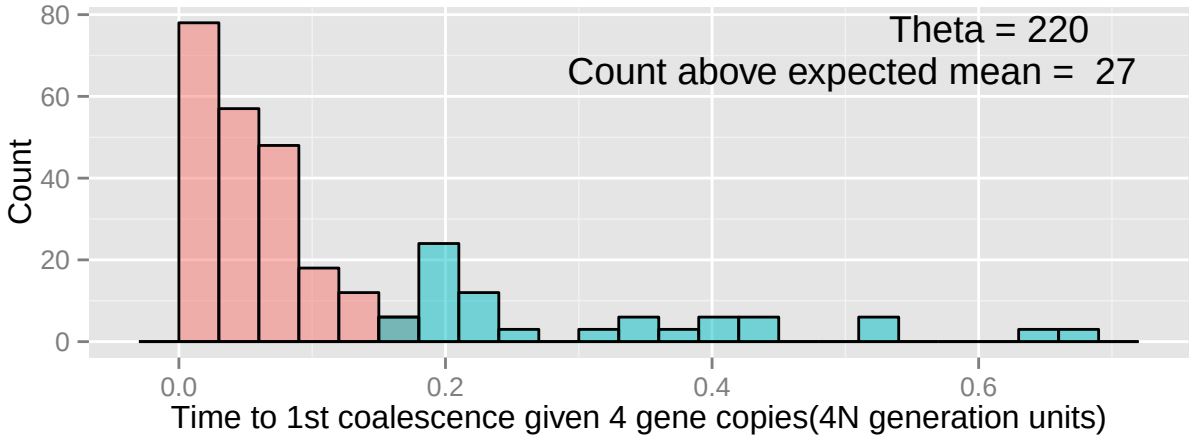
Command used for simulations: ./ms 4 100 -t 2.2 -T
The tree output was parsed to retrieve time to 1st coalesence event.

... CONTINUED ON NEXT PAGE ...

```
#Read in filtered results (theta, time) from simulations
ms <- read.table("C:\\Users\\gitanshu\\Desktop\\ms.txt",
                 header=T, sep="\t")
ms <- data.frame(ms)

#Classify simulations with time greater than average time for a given sim(theta/popsize)

ms$above <- NA
ms[ms$theta=="theta = 2.2",3] <- ms[ms$theta=="theta = 2.2",2] >
                                                    0.17

ms[ms$theta=="theta = 22",3] <- ms[ms$theta=="theta = 22",2] >
                                                    0.17

ms[ms$theta=="theta = 220",3] <- ms[ms$theta=="theta = 220",2] >
                                                    0.17
library(ggplot2)
library(gridExtra)

a <- ggplot(ms, aes(x=ms[ms$theta=="theta = 2.2",2],
                    fill=ms[ms$theta=="theta = 2.2",3])) +
      labs(title =paste("Theta = 2.2", "\n","Count above expected mean = ",
                        table(ms[ms$theta=="theta = 2.2",3])[2]))

b <- ggplot(ms, aes(x=ms[ms$theta=="theta = 22",2],
                    fill=ms[ms$theta=="theta = 22",3])) +
      labs(title =paste("Theta = 22", "\n","Count above expected mean = ",
                        table(ms[ms$theta=="theta = 22",3])[2]))

c <- ggplot(ms, aes(x=ms[ms$theta=="theta = 220",2],
                    fill=ms[ms$theta=="theta = 220",3])) +
      labs(title =paste("Theta = 220", "\n","Count above expected mean = ",
                        table(ms[ms$theta=="theta = 220",3])[2]))

#Formatting
z1 <- geom_histogram(binwidth=.03, alpha=.5, colour="black",position="identity")

z2 <- labs(x= "Time to 1st coalescence given 4 gene copies(4N generation units)",
           y= "Count")
z3 <- theme(legend.position = "none", plot.title = element_text(vjust=-4, hjust=0.9))

a <- a + z1 + z2 + z3
b <- b + z1 + z2 + z3
c <- c + z1 + z2 + z3
grid.arrange(a, b, c,nrow=3)
```


Theta = 2.2
Count above expected mean = 28
Count
60
40
20
0
0.0
0.3
0.6
0.9
Time to 1st coalescence given 4 gene copies(4N generation units)


Theta = 22
Count above expected mean = 23
Count
40
20
0
0.0
0.2
0.4
0.6
Time to 1st coalescence given 4 gene copies(4N generation units)


Theta = 220
Count above expected mean = 27
Count
80
60
40
20
0
0.0
0.2
0.4
0.6
Time to 1st coalescence given 4 gene copies(4N generation units)

References

[1] Hedrick, P. *Genetics of populations*. Jones & Bartlett Learning, 2011.

[2] Hudson, R. R. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics 18*, 2 (2002), 337–338.