# Genetic Drift and Coalescence

**Exercise 1: simulate binomial distribution with probability of successs p=0.1 in a sample of size 10. Find expected number of successes**

```
expected <- mean(rbinom(10000, 10, 0.1))
expected
```

```
## [1] 0.9989
```

**Exercise 2: write a function to simulate T generations of genetic drift for L independent SNPs. Keep track of allele freq of each SNP in each gen. All SNPs should start in the initial gen at freq q.**

```
#genetic drift
pp <- 0.1
NN <- 10 # number of diploid indiv; 2*NN = number of chromosome
count <- rbinom(1, 2*NN, pp)
count
```

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

```
count/(2*NN)
```

```
## [1] 0.05
```

```
#function
GenDriftFunc <- function(LL, NN, p0, TT){
  #matrix of allele freq of each SNP (col) in each gen (row)
  freqs <- matrix(nrow = TT, ncol = LL)
  #fill first row with p0
  freqs[1, ] <- p0
  #for loop to iterate rbinom for each subsequent row
  for (tt in 2:TT){
    for(ll in 1:LL){
        freqs[tt, ll] <- rbinom(1, 2*NN, freqs[tt-1, ll])/(2*NN)
    }
  }
  return(freqs)
}
```

**Exercise 3: use the function to simulate drift**

```
NN=100
LL=1000
TT=10000
p0=0.1
sim1 <- GenDriftFunc(LL, NN, p0, TT)
```

a) How many of the 1000 SNPs are at freq 0 after 10K generations?

```
#extinct?
sum(sim1[10000, ] == 0)
```

```
## [1] 903
```

b) How many are at freq 1?

```
#fixation?
sum(sim1[10000, ] == 1)
```

```
## [1] 97
```

c) Does value agree with prediction for prob of fixation of neutral allele?

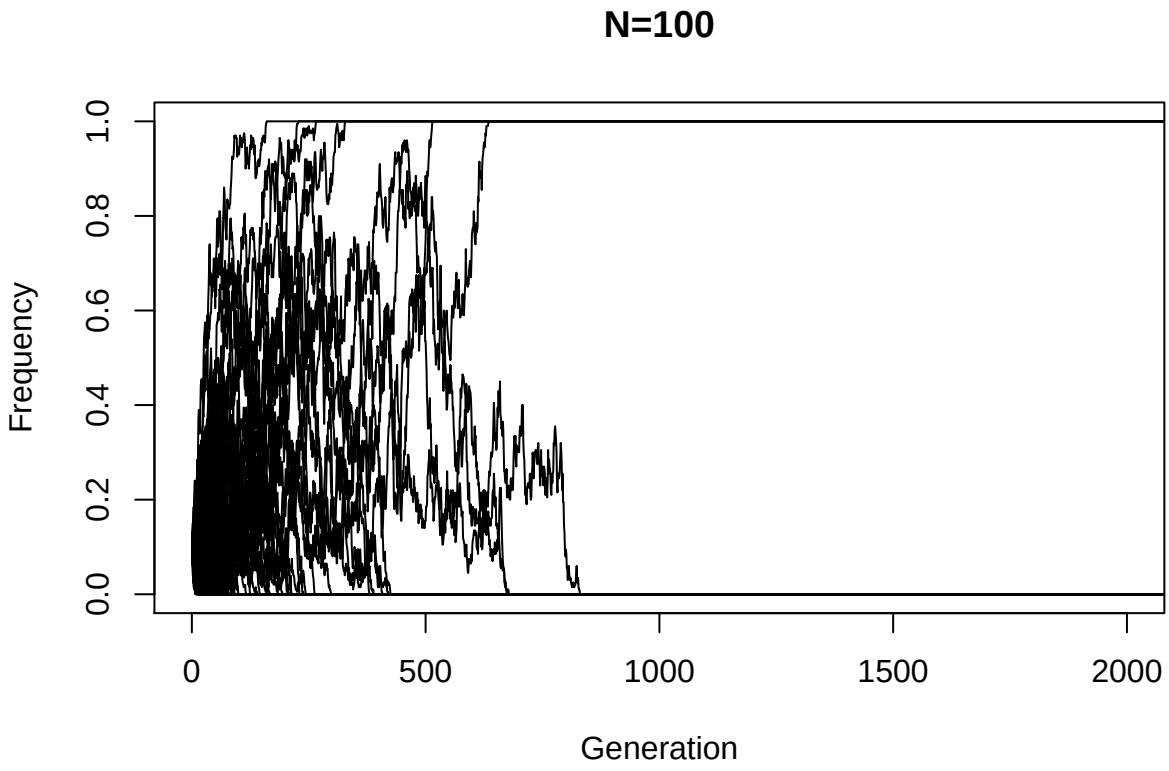
```
#agree?
sum(sim1[10000, ] == 1)/1000
```

```
## [1] 0.097
```

Yes, the probability of fixation is close to 10%.

d) Plot trajectories for 100 of the SNPs

```
#plot
plot(sim1[ , 1], type = "l", ylim = c(0, 1), xlim = c(0, 2000), main = "N=100", xlab = "Generation", yla
for(ii in 2:100){
    lines(sim1[ , ii])
}
```




e) Repeat simulation with p=0.6. How many of the 1000 SNP are at 0?

```
#repeat
NN=100
LL=1000
TT=10000
p0=0.6
sim2 <- GenDriftFunc(LL, NN, p0, TT)
#extinct?
sum(sim2[10000, ] == 0)
```

```
## [1] 393
```

f) How many are at freq 1?

```
#fixation?
sum(sim2[10000, ] == 1)
```

```
## [1] 607
```

g) Does value agree with prediction for prob of fixation of neutral allele?

```r
#agree?
sum(sim2[10000, ] == 1)/1000
```

```
## [1] 0.607
```

Yes, the probability of fixation is close to 60%.

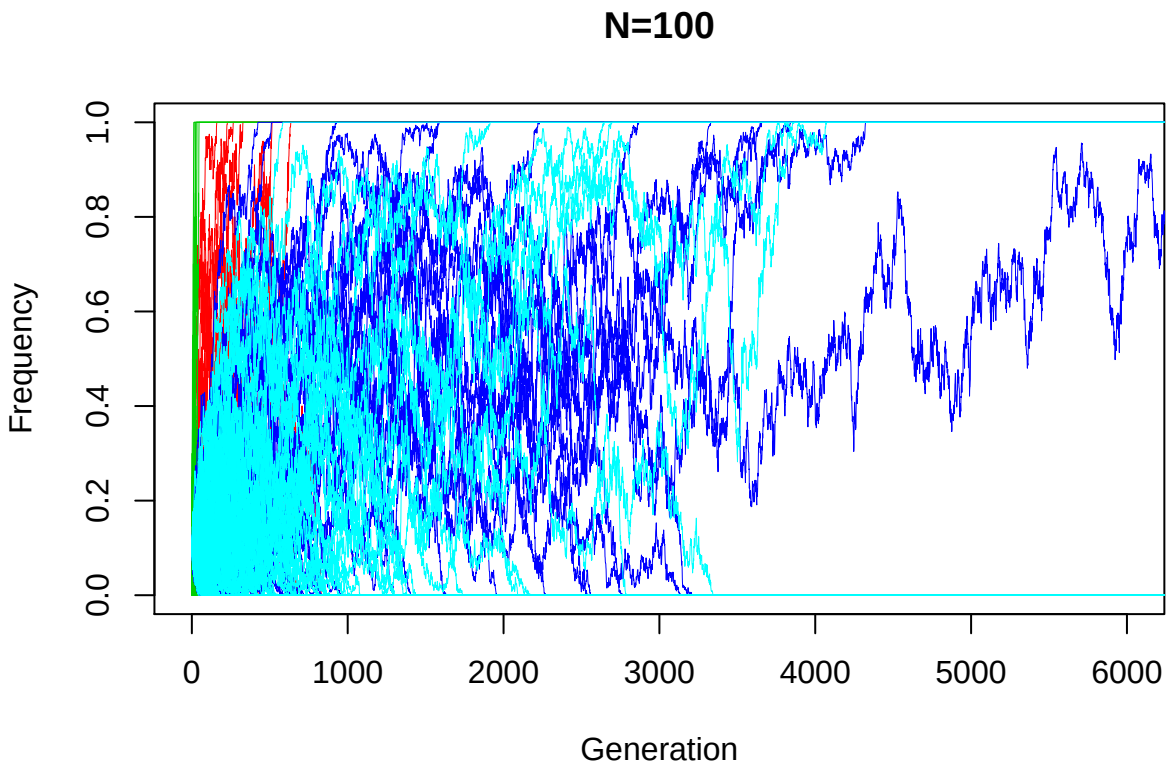
### Exercise 4: Effect of pop size on gen drift. Repeat simulation, set N=10, 500, and 1000. Keep other parameters the same. (N is number of diploids)

```r
# N=10
NN=10
LL=1000
TT=10000
p0=0.1
sim3 <- GenDriftFunc(LL, NN, p0, TT)

# N=500
NN=500
LL=1000
TT=10000
p0=0.1
sim4 <- GenDriftFunc(LL, NN, p0, TT)

# N=1000
NN=500
LL=1000
TT=10000
p0=0.1
sim5 <- GenDriftFunc(LL, NN, p0, TT)
```

a) Plot all 4 pop sizes on the same page.

```r
#plot sim1, sim3, sim4, sim5
#100=red; 10=green; 500=blue; 1000=cyan
plot(sim1[ , 1], type = "l", ylim = c(0, 1), xlim = c(0, 6000), main = "N=100", xlab = "Generation", yla
for(ii in 2:100){
  lines(sim1[ , ii], col = 2, lwd = 0.1)
  adjustcolor(col = 2, alpha.f = 0.8)
}
for(ii in 1:100){
  lines(sim3[ , ii], col = 3, lwd = 0.1)
  adjustcolor(col = 3, alpha.f = 1)
}
for(ii in 1:100){
  lines(sim4[ , ii], col = 4, lwd = 0.1)
  adjustcolor(col = 4, alpha.f = 0.6)
}
for(ii in 1:100){
  lines(sim5[ , ii], col = 5, lwd = 0.1)
```

```
  adjustcolor(col = 5, alpha.f = 0.6)
}
```




b) How does pop size affect allele freq change?

The larger the population size, the longer it takes for alleles to either reach fixation or go extinct.

c) For each pop size, in what proportion of simulation replicates did the derived allele become fixed by the end?

```
#fixation?
fix1 <- sum(sim1[10000, ] == 1)/1000
fix3 <- sum(sim3[10000, ] == 1)/1000
fix4 <- sum(sim4[10000, ] == 1)/1000
fix5 <- sum(sim5[10000, ] == 1)/1000
popsize <- c(10, 100, 500, 1000)
fixedproportion <- c(fix3, fix1, fix4, fix5)
table <- rbind(popsize, fixedproportion)
table
```

```
##                    [,1]    [,2]  [,3]  [,4]
## popsize          10.000 100.000 5e+02 1e+03
## fixedproportion   0.091   0.097 9e-02 9e-02
```

d) How is this probability affected by the pop size?

It isn't affect by the population size.

e) How does this probability of fixation estimated from the simulations match with theoretical prediction?

They are all close to 10%, matching the theoretical prediction.

## Coalescent simulations

### Exercise 5: what is the rate of coalescence for a sample size of 2 chromosomes in a population of size 2N chromosomes?

The probability of coalescence in 1 generation will be 1/(2N) for pop size of 2N chromosomes.

### Exercise 6: Perform 10,000 simulations of coalescent times for a sample size of 2 chromosomes from this pop size of 2N=20,000.

a) What is the average TMRCA?

```
#10,000 sims; 2N=20,000
NN <- 10000
CoalesTime <- rexp(10000, 1/(2*NN))
mean(CoalesTime)
```

```
## [1] 20112.84
```

b) What is the theoretical expectation?

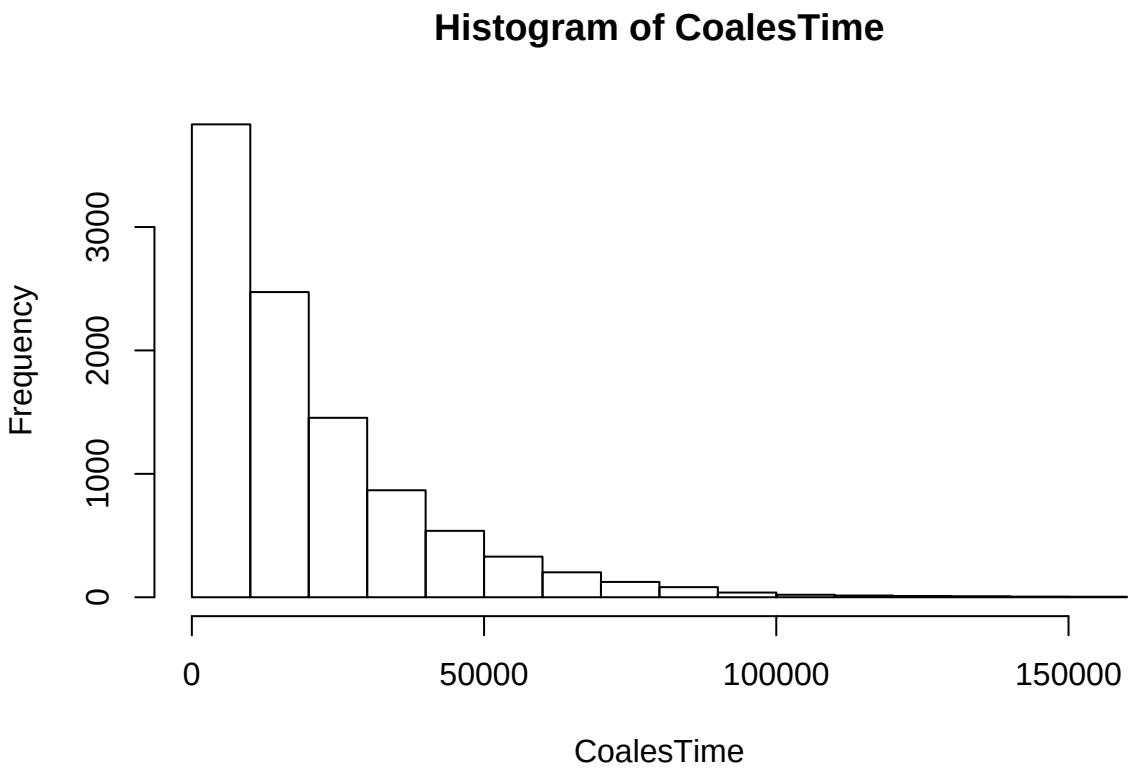
Expected Coalescent time=2N=20,000.

c) How do the two values compare?

They are very close.

d) Make density plot of simulated coalescent times.

```
#density plot
hist(CoalesTime)
```




e) What is the standard devation of the coalescent times?

```
sd(CoalesTime)
```

```
## [1] 19927.17
```

### Exercise 7: Repeat simulation; set 2N=2,000.

a) What is the average TMRCA?

```
#10,000 sims, 2N=2,000
NN <- 1000
CoalesTime <- rexp(10000, 1/(2*NN))
mean(CoalesTime)
```

```
## [1] 1978.916
```

b) What is the theoretical expectation?

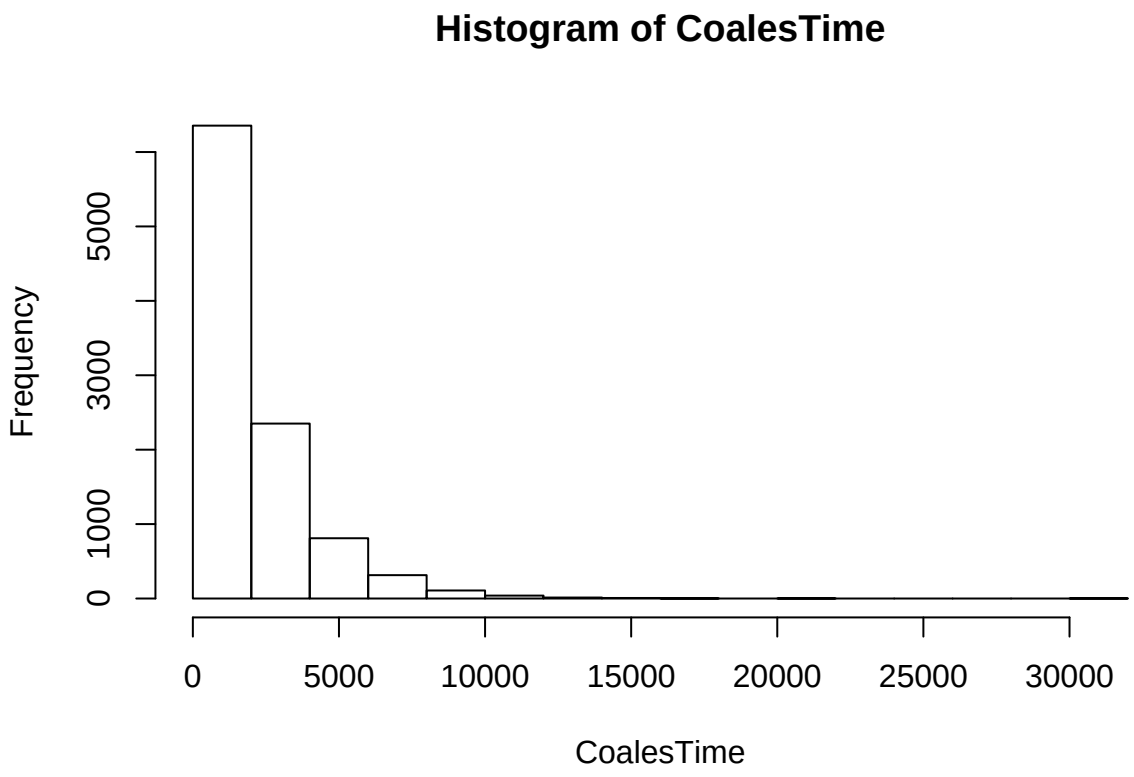
Expected Coalescent time=2N=2,000.

c) How do the two values compare?

They are very close.

d) Make density plot of simulated coalescent times.

```
#density plot
hist(CoalesTime)
```




e) What is the standard devation of the coalescent times?

```
sd(CoalesTime)
```

```
## [1] 1983.282
```

f) How does the average TMRCA from the simulations in question 6 compare to the average TMRCA from the simulations in question 7?

The average TMRCA for 2N=20,000 is 10 times that of the average TMRCA for 2M=2,000.

g) What can you conclude about how the population size affects the expected coalescent time?

The expected coalescent time increases linearly with respect to the population size.

### EXercise 8: Add mutations to the genealogies simulated. Set mu=1e-4

a) Add mutations to the genealogies in question 6 (2N=20,000)

```
#add mutations, 10,000 sims, 2N=20,000
mu <- 1e-4
NN <- 10000
CoalesTime <- rexp(10000, 1/(2*NN))
MutaRate <- 2*mu*CoalesTime
mean(MutaRate)
```

```
## [1] 3.913343
```

```
NumbSNP <- rep(NA, NN)
for (ii in 1:10000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
```

b) What is the average number of SNPs per genealogy?

```
mean(NumbSNP)
```

```
## [1] 3.9202
```

c) What is theta predicted to be in this example?

```
#theta=4*N*mu
theta <- 4*NN*mu
theta
```

```
## [1] 4
```

d) How does theta compare to the average number of SNPs in simulations?

They are similar.

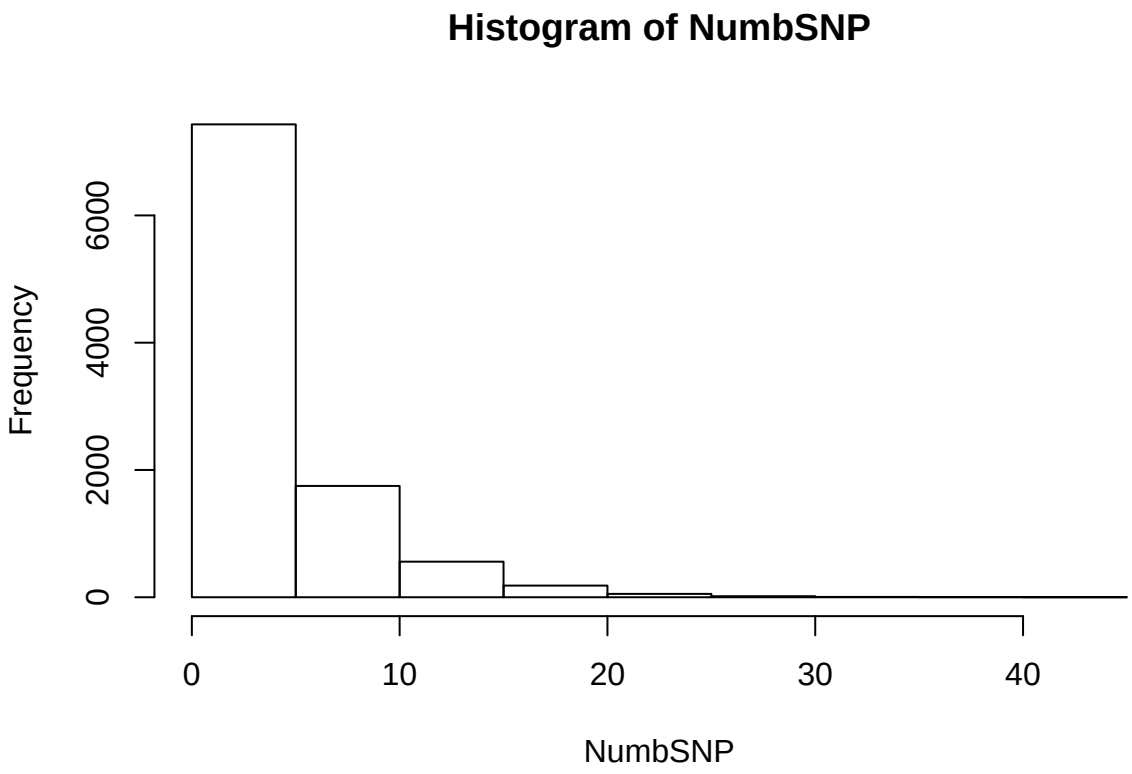
e) Pretty neat, huh?

```
print("Yes, pretty neat!")
```

```
## [1] "Yes, pretty neat!"
```

f) Make density plot on the number of SNPs per simulation replicate

```
#density plot
hist(NumbSNP)
```




### Exercise 10: repeat all parts in question 8 with 2N=2,000

a) Add mutations to the genealogies in question 6 (2N=20,000)

```
#add mutations, 10,000 sims, 2N=2,000
mu <- 1e-4
NN <- 1000
CoalesTime <- rexp(10000, 1/(2*NN))
MutaRate <- 2*mu*CoalesTime
mean(MutaRate)
```

```
## [1] 0.3998185
```

```
NumbSNP <- rep(NA, NN)
for (ii in 1:10000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
```

b) What is the average number of SNPs per genealogy?

```
mean(NumbSNP)
```

```
## [1] 0.4043
```

c) What is theta predicted to be in this example?

```
#theta=4*N*mu
theta <- 4*NN*mu
theta
```

```
## [1] 0.4
```

d) How does theta compare to the average number of SNPs in simulations?

They are similar.

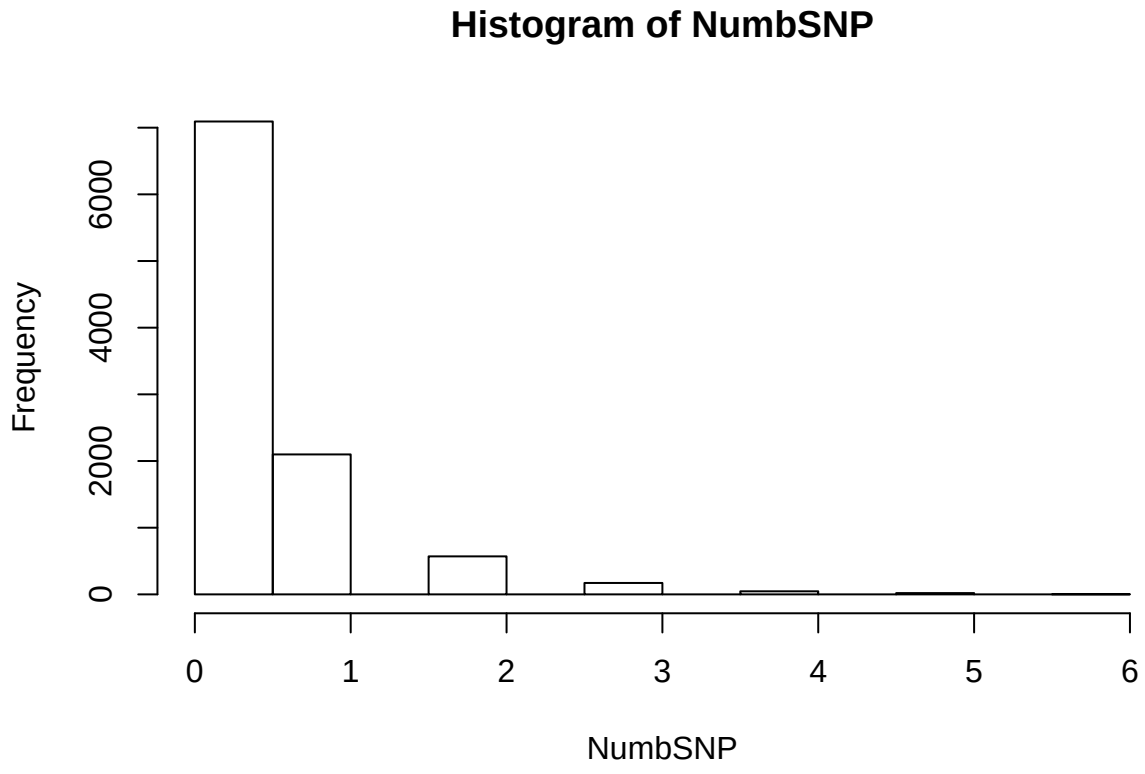
e) Pretty neat, huh?

```
print("Yes, pretty neat!")
```

```
## [1] "Yes, pretty neat!"
```

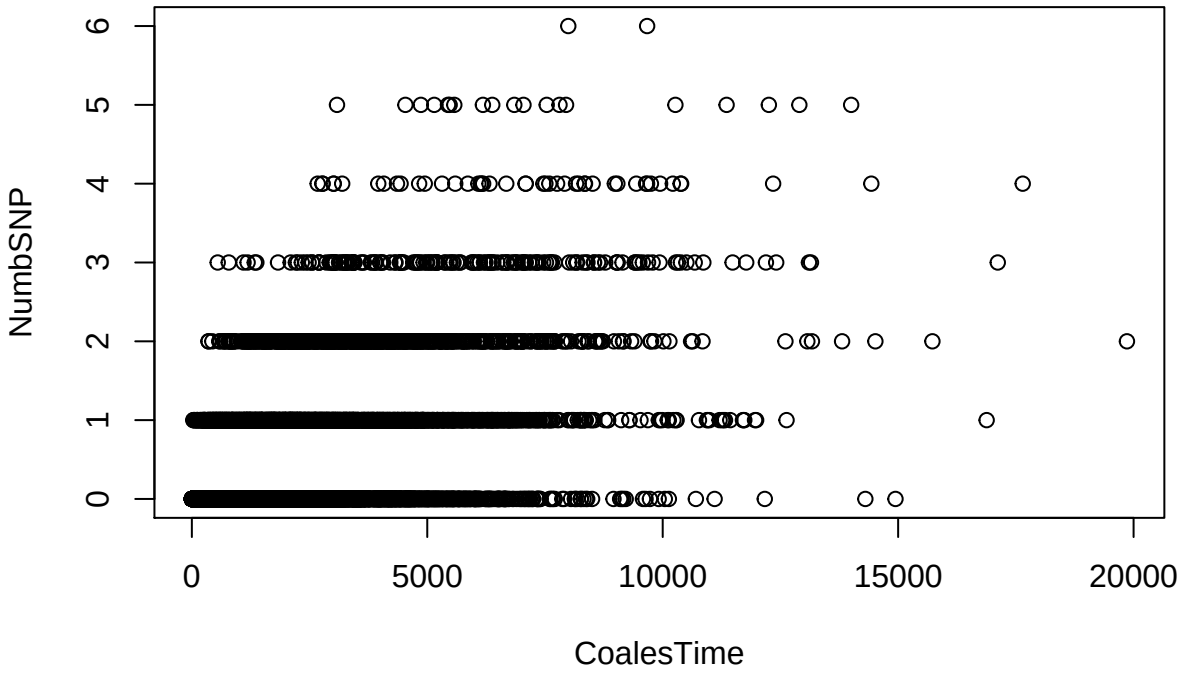
f) Make density plot on the number of SNPs per simulation replicate

```
#density plot
hist(NumbSNP)
```



**Exercise 11: Make scatter plot of number of SNPs in each sim replicate vs TMRCA for that sim replicate. How are these variables related to each other? Does the number of SNPs tell you anything about the TMRCA?**

```
plot(CoalesTime, NumbSNP)
```




Higher number of SNPs occur in longer TMRCA.

**Exercise 12: imagine you sequenced a 10kb region of DNA from a diploid individual. You observed 10 SNPs in that 10kb interval. Independent evidence suggests mutation rate=1e-8 per bp per gen.**

a) Use coalescent sims to evaluate whether a model with N=10,000 is compatible with your observed data. Or, what proportion of simulation replicates from this model have >= 10 SNPs?

```
mu <- 1e-8*10000
NN <- 10000
CoalesTime <- rexp(10000, 1/(2*NN))
MutaRate <- 2*mu*CoalesTime
mean(MutaRate)
```

```
## [1] 3.996684
```

```
NumbSNP <- rep(NA, NN)
for (ii in 1:10000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
sum(NumbSNP >= 10)
```

```
## [1] 1092
```

Yes, it is compatible.

b) Is the model with N=1,000 compatible with your observed data? What proportion of simulation replicates from this model have >= 10 SNPs?

```
mu <- 1e-8*10000
NN <- 1000
CoalesTime <- rexp(10000, 1/(2*NN))
MutaRate <- 2*mu*CoalesTime
mean(MutaRate)
```

```
## [1] 0.4023142
```

```
NumbSNP <- rep(NA, NN)
for (ii in 1:10000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
sum(NumbSNP >= 10)
```

```
## [1] 0
```

No, it is not compatible.