ABCexercise

Clive Lau

November 28, 2015

Estimation of pop size

Exercise 1: ABC rejection sampling algorithm

a) Assume that N can be any value between 100 and 100,000. Draw 1 million values from the prior distribution of N. Let's use a uniform distribution for N. Assume that N~U[100,100000].

```
priorNN <- runif(1000000, min = 100, max = 100000)
```

b) For each value of N, simulate a TMRCA. Note, you should use the same approach as on last week's assignment (i.e. Draw times from an exponential distribution). Second hint: The total number of Y chromosomes in the population is N, rather than 2N (because the Y chromosome is haploid). Adjust your rates of coalescence accordingly.

```
CoalesTime <- rep(NA, 1000000)
for (ii in 1:1000000){
  CoalesTime[ii] <- rexp(1, 1/(priorNN[ii]))
}
mean(CoalesTime)
```

```
## [1] 50033.49
```

c) For each TMRCA, add a Poisson number of mutations with the appropriate mutation rate. Again, this will follow from what you did last week.

```
bp <- 100000
mu <- 1e-8*bp
MutaRate <- 2*mu*CoalesTime
NumbSNP <- rep(NA, 1000000)
for (ii in 1:1000000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
mean(NumbSNP)
```

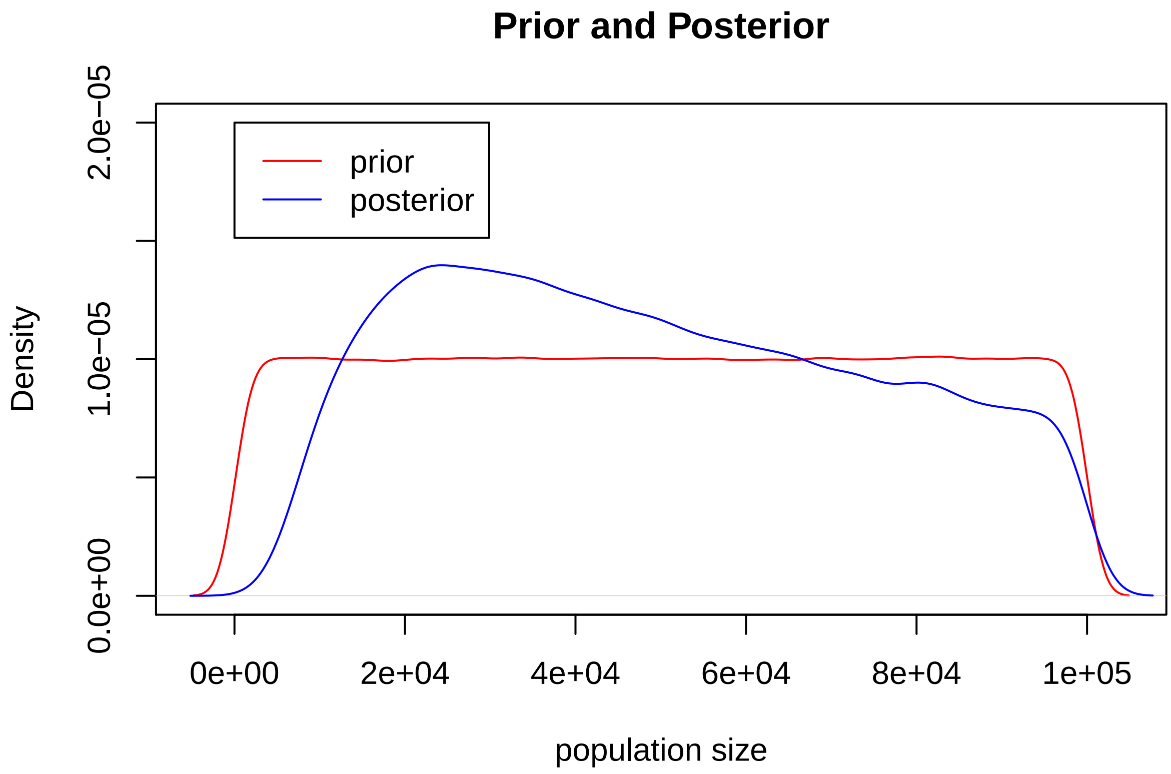
```
## [1] 100.0765
```

d) We now need to decide which of the million draws from the prior give data that are "close" to the observed number of SNPs in the actual data and should be accepted. To do this, let's accept all values of N that give somewhere between 45-55 SNPs.

```
sims <- cbind(priorNN, NumbSNP)
acceptNN <- subset(sims, sims[, 2] >= 45 & sims[, 2] <= 55)
posteriorNN <- acceptNN[, 1]
```

Exercise 2: Make a density plot of your prior distribution and your posterior distribution of N. Please plot them on the same axes and be sure to label which line corresponds to the prior and which corresponds to the posterior.

```
plot(density(priorNN),col = 2, ylim = c(0, 2e-5), lwd = 1,xlab = "population size", main = "Prior and Po
lines(density(posteriorNN),col = 4,lwd = 1)
legend(0, 2e-5 ,c("prior","posterior"),lwd=1,col=c(2,4),cex=1)
```



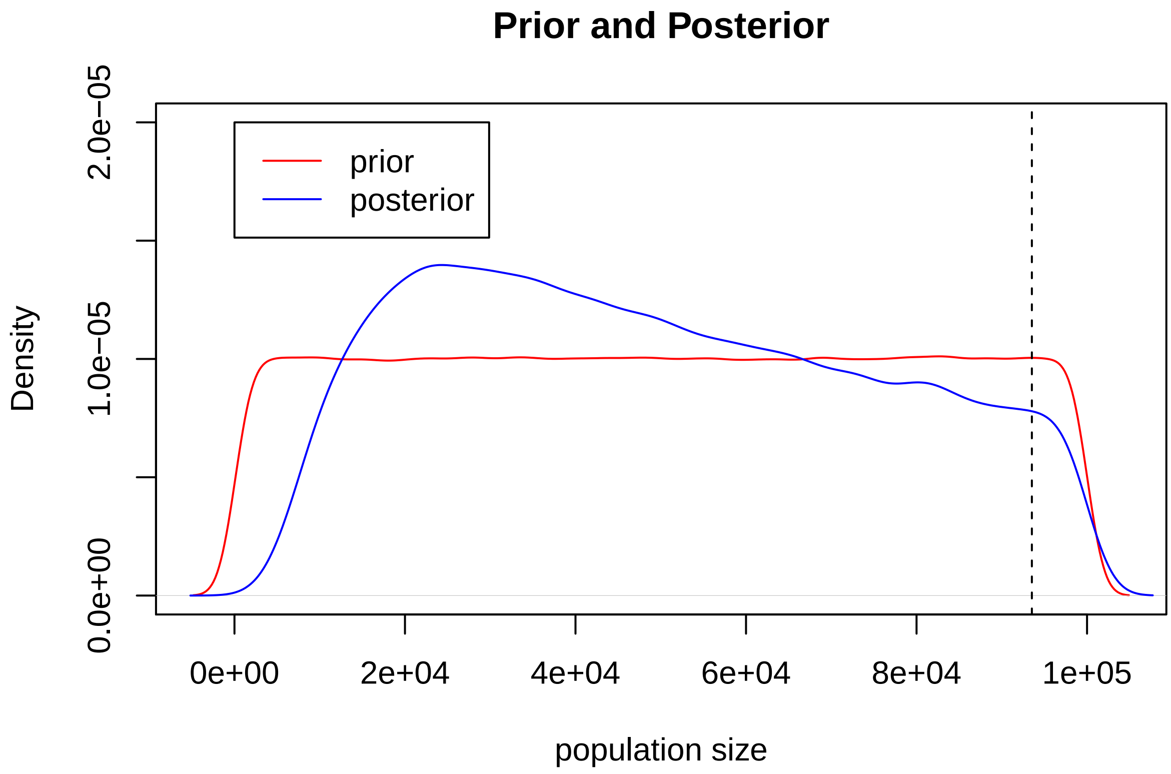
Exercise 3: What is the median value of the posterior distribution of N?

```
median(posteriorNN)
```

```
## [1] 47138.87
```

Exercise 4: Generate a 95% credible interval for the posterior distribution of N (like a confidence interval, but for Bayesians. Note, in this framework, the there actually is a 95% chance of the true parameter value falling in this region. This is not the case for normal frequentist confidence intervals). Hint, use the "quantile" function in R.

```
plot(density(priorNN),col = 2, ylim = c(0, 2e-5), lwd = 1,xlab = "population size", main = "Prior and Po
lines(density(posteriorNN),col = 4,lwd = 1)
legend(0, 2e-5 ,c("prior","posterior"),lwd=1,col=c(2,4),cex=1)
abline(v = quantile(posteriorNN, 0.95), lty = 2, lwd = 1, col = 1)
```



Exercise 5: How does the posterior distribution differ from the prior distribution? A descriptive answer here will suffice. The degree to which the posterior differs from the prior relates to the amount of information in the data.

The posterior distribution is more skewed to the left than the prior distribution.

Exercise 6: Repeat your ABC analysis, but change the prior distribution of N to be U~[1000,1000000]. What is the mean, median, and 95% credible interval for the posterior distribution. How does this differ from what you computed in questions 3-4 for the original prior distribution? What does this tell you about the effect of the prior distribution for in Bayesian statistics?

```
#prior distribution of N~U[1000,1000000]
priorNN <- runif(1000000, min = 1000, max = 1000000)

CoalesTime <- rep(NA, 1000000)
for (ii in 1:1000000){
  CoalesTime[ii] <- rexp(1, 1/(priorNN[ii]))
}

bp <- 100000
mu <- 1e-8*bp
MutaRate <- 2*mu*CoalesTime
NumbSNP <- rep(NA, 1000000)
for (ii in 1:1000000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
mean(NumbSNP)
```

```
## [1] 999.1094
```

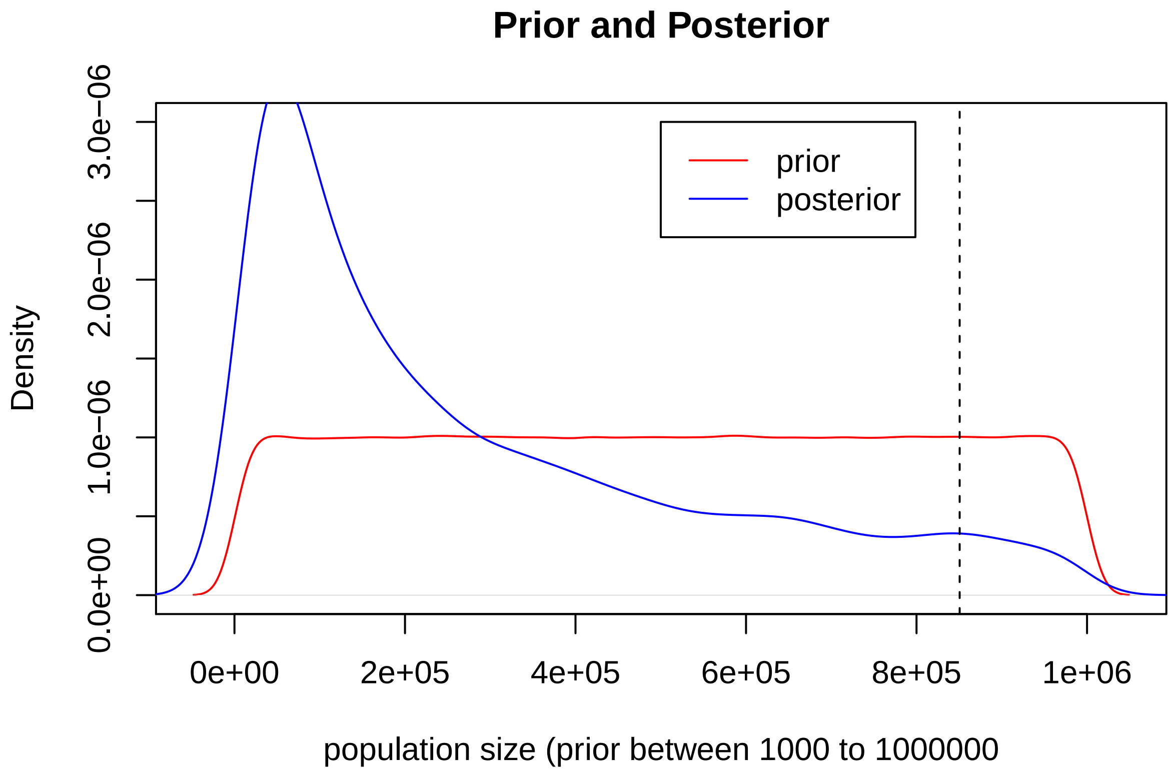
```
sims <- cbind(priorNN, NumbSNP)
acceptNN <- subset(sims, sims[, 2] >= 45 & sims[, 2] <= 55)
posteriorNN <- acceptNN[, 1]

plot(density(priorNN),col = 2, ylim = c(0, 3e-6), lwd = 1,xlab = "population size (prior between 1000 t
lines(density(posteriorNN),col = 4,lwd = 1)
legend(5e5, 3e-6 ,c("prior","posterior"),lwd=1,col=c(2,4),cex=1)

#median posterior
median(posteriorNN)
```

```
## [1] 183677.9
```

```
#95% credible interval
abline(v=quantile(posteriorNN,0.95),lty=2,lwd=1,col=1)
```



The posterior is now even more skewed to the left compared to the prior. The median is much bigger than the previous example. This shows that the prior heavily affects the statistical inferrence in Bayesian statistics.

Exercise 7: If you wanted a more precise estimate of the population size (assume that there is no sex biased demographic history so that you can easily extrapolate the total population size from the Y chromosome population size and vice versa), would you be better off: A) sequencing a bigger region of the Y chromosome, or B) sequencing the same amount on the autosomes? Why?

Sequencing the same amount on the autosomes, because it is subjected to recombination, yielding new genealogies. Whereas sequencing more of the non-recombining Y chromosome would not represent new genealogies.

Estimation of population split times

Exercise 1: Once the two chromosomes make it back into the ancestral population (of size N=100,000), what is the expected amount of additional time we have to wait until they coalesce?

The expected time of coalescence before split is N=100,000

Exercise 2: What is the expected time until the two chromosomes coalesce with each other? You should write this formula (use tsplit as the split time, rather than a specific number, because you haven't estimated this number yet!). Hint: Break the total time into 2 parts: the time from the present day till tsplit and the time for what happens in the ancestral population (i.e. your answer for part 1) of this question).

The total expected time of coalescence is t_split + 100,000

Exercise 3: ABC approach to estimate t_split

a) Assume that tsplit can be any value between 50,000 and 1,000,000 generations. Draw 1 million values from the prior distribution of tsplit. Assume that tsplit ~U[50000,1000000].

```r
PriorTT_split <- runif(1000000, min = 50000, max = 1000000)
```

b) For each value of tsplit, simulate a TMRCA. Note, you should use the same approach as on the first part of this assignment. But, keep in mind that the coalescent time is a function of both the split time (tsplit, which is drawn from your prior) and the coalescent time in the ancestral population.

```r
NN <- 100000
CoalesTime <- rep(NA, 1000000)
for (ii in 1:1000000){
  CoalesTime[ii] <- rexp(1, 1/NN)
}
TMRCA <- CoalesTime + PriorTT_split
```

c) For each TMRCA, add a Poisson number of mutations with the appropriate mutation rate. Again, this will follow from what you did last week.

```r
bp <- 100000
mu <- 1e-8*bp
MutaRate <- 2*mu*TMRCA
NumbSNP <- rep(NA, 1000000)
for (ii in 1:1000000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
mean(NumbSNP)
```

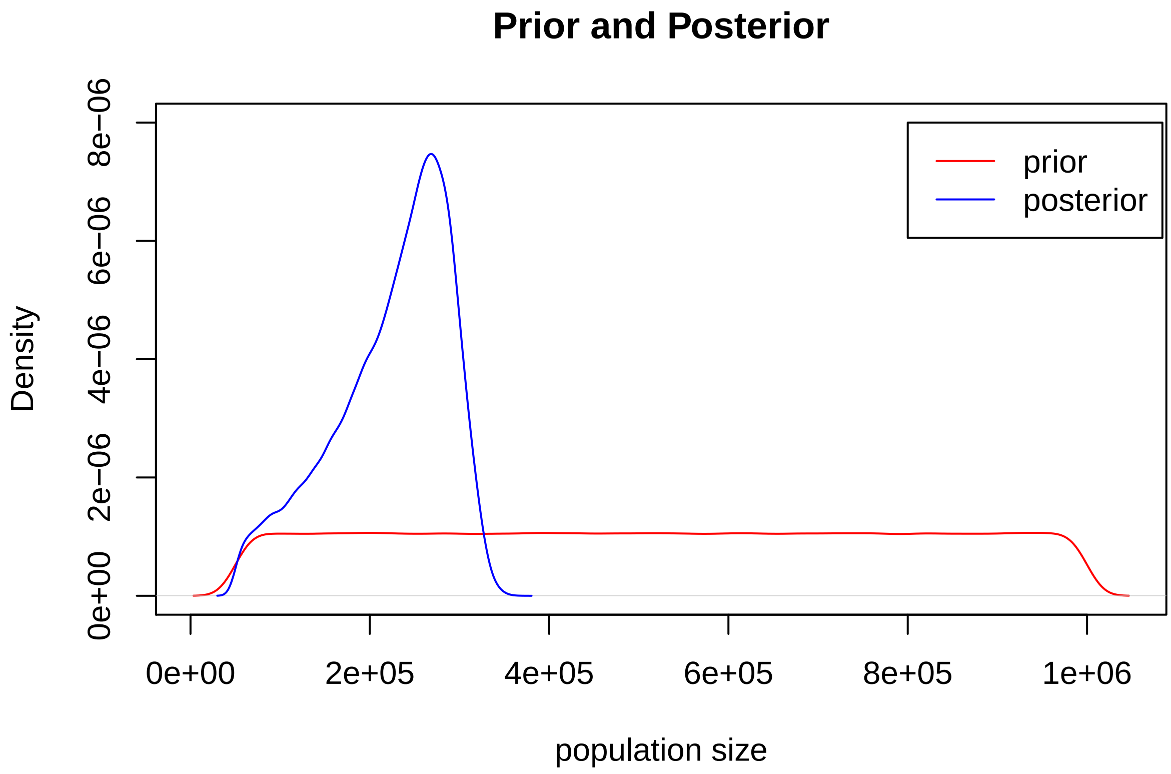
```
## [1] 1250.081
```

d) We now need to decide which of the million draws from the prior generate data that are "close" to the observed number of SNPs in the actual data and should be accepted. To do this, let's accept all values of tsplit that give somewhere between 550 and 650 SNPs.

```r
sims <- cbind(PriorTT_split, NumbSNP)
acceptTT_split <- subset(sims, sims[, 2] >= 550 & sims[, 2] <= 650)
posteriorTT_split <- acceptTT_split[, 1]
```

Exercise 4: Make a density plot of your prior distribution and your posterior distribution of tsplit. Again, plot both the prior and the posterior on the same axes, and label which is which.

```
plot(density(PriorTT_split),col = 2, ylim = c(0, 8e-6), lwd = 1,xlab = "population size", main = "Prior
lines(density(posteriorTT_split),col = 4,lwd = 1)
legend(8e5, 8e-6 ,c("prior","posterior"),lwd = 1, col=c(2,4),cex=1)
```



Exercise 5: What is the median value of the posterior distribution of tsplit?

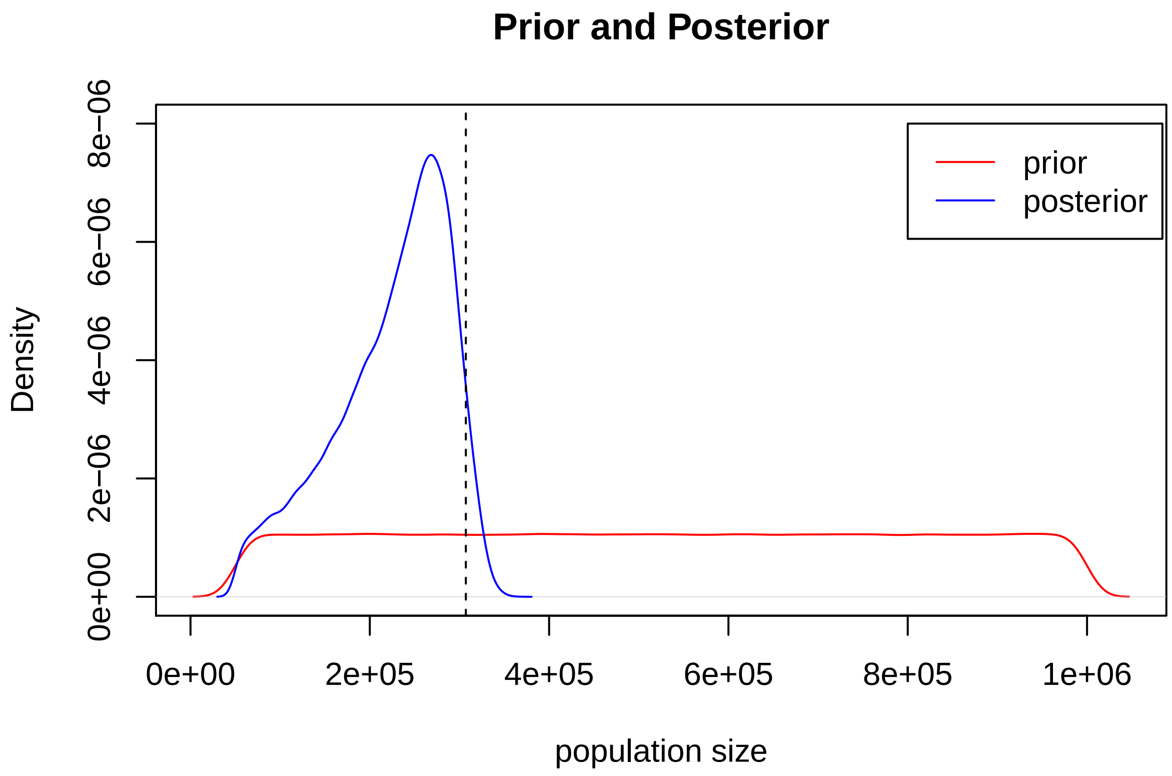
```
median(posteriorTT_split)
```

```
## [1] 238000.7
```

Exercise 6: Generate a 95% credible interval for the posterior distribution of tsplit. Hint, use the “quantile” function in R.

```
plot(density(PriorTT_split),col = 2, ylim = c(0, 8e-6), lwd = 1,xlab = "population size", main = "Prior
lines(density(posteriorTT_split),col = 4,lwd = 1)
```

```r
legend(8e5, 8e-6 ,c("prior","posterior"),lwd = 1, col=c(2,4),cex=1)
abline(v=quantile(posteriorTT_split,0.95),lty=2,lwd=1,col=1)
```



Exercise 7: How does the posterior distribution differ from the prior distribution? A descriptive answer here will suffice. The degree to which the posterior differs from the prior relates to the amount of information in the data.

The posterior distribution is more narrow and more skewed to the left.

Exercise 8: What if we did not know the true value of N? Repeat the ABC approach to estimate tsplit, but this time, rather than assuming that N=100,000, draw N from a N~U[1000,1000000] distribution.

```r
priorNN <- runif(1000000, min = 1000, max = 1000000)
PriorTT_split <- runif(1000000, min = 50000, max = 1000000)

CoalesTime <- rep(NA, 1000000)
for (ii in 1:1000000){
  CoalesTime[ii] <- rexp(1, 1/priorNN[ii])
}
TMRCA <- CoalesTime + PriorTT_split
```

```r
bp <- 100000
mu <- 1e-8*bp
MutaRate <- 2*mu*TMRCA
NumbSNP <- rep(NA, 1000000)
for (ii in 1:1000000){
  NumbSNP[ii] <- rpois(1, MutaRate[ii])
}
mean(NumbSNP)
```

```
## [1] 2052.106
```

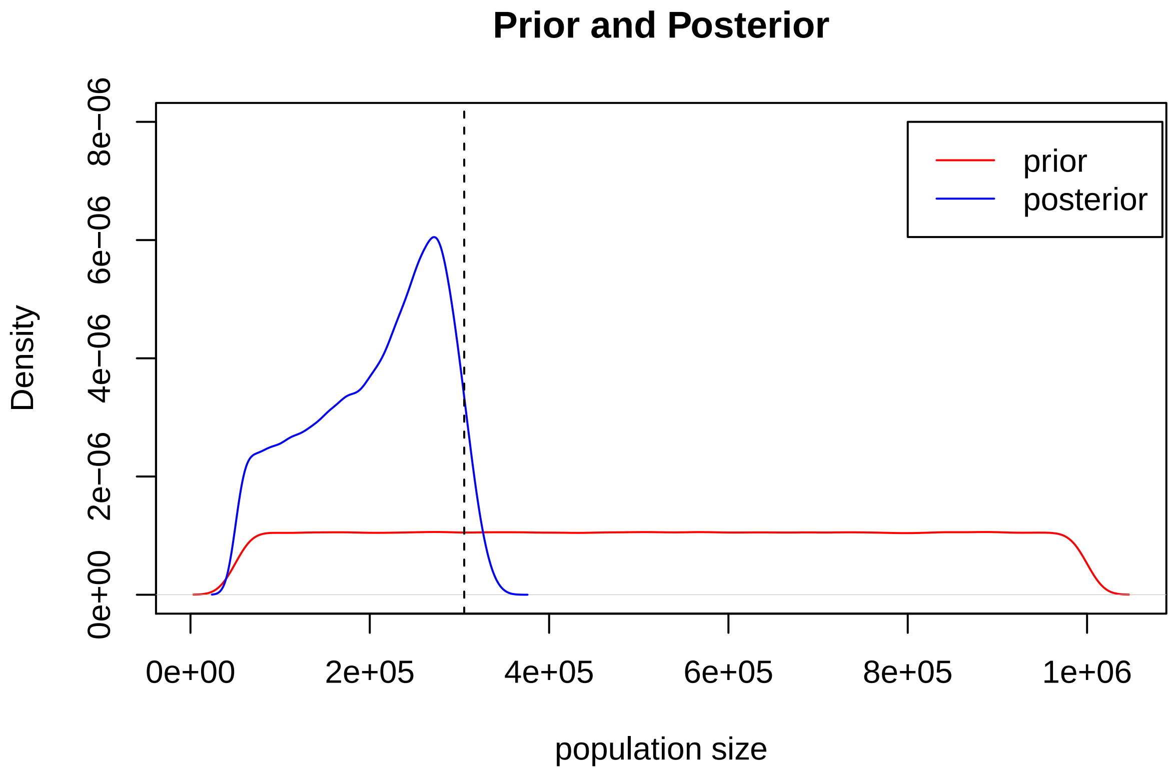
```r
sims <- cbind(PriorTT_split, NumbSNP)
acceptTT_split <- subset(sims, sims[, 2] >= 550 & sims[, 2] <= 650)
posteriorTT_split <- acceptTT_split[, 1]

plot(density(PriorTT_split),col = 2, ylim = c(0, 8e-6), lwd = 1,xlab = "population size", main = "Prior
lines(density(posteriorTT_split),col = 4,lwd = 1)
legend(8e5, 8e-6 ,c("prior","posterior"),lwd = 1, col=c(2,4),cex=1)
abline(v=quantile(posteriorTT_split,0.95),lty=2,lwd=1,col=1)

#median posterior
median(posteriorTT_split)
```

```
## [1] 219306.8
```

```r
#95% credible interval
abline(v=quantile(posteriorTT_split,0.95),lty=2,lwd=1,col=1)
```



Exercise 9: Do your estimates of the median and credible interval of tsplit differ from above?

The median is slightly lower than before. The peak of the posterior distribution is slightly shorter and less narrow, and the credible interval remains more or less the same.