Approximate Bayesian Computation exercise

Gaurav Kandlikar

November 23, 2015

Estimation of population size

Imagine that you resequenced a 100kb region of two Y chromosomes, which are non-recombining. You observe 50 SNPs within this region. You would like to use this dataset to estimate the effective population size of males. You know that the mutation rate per base pair on the Y chromosome is $\mu = 1e^{-8}$ per generation.

Let's develop an ABC approach to this:

Question 1

Assume that N can be any value between 100 and 100000. Draw 1 million values from the prior distribution of N, assuming N $[100, 100000]$

```r
# Draw Ns from distribution --------
reps <- 1e6
min = 100; max = 100000
N <- runif(reps, min, max)
# Mean should be min+max/2
(min+max)/2; mean(N)
```

```
## [1] 50050
```

```
## [1] 50039.4
```

For each value of N, simulate a TMRCA.

```r
# Generate coalescent times ---------
rates <- 1/(N) # Only 1 Y chromosome in males...
coal_times <- rexp(reps, rates)
# Mean should be the same as mean(N)
mean(coal_times)
```

```
## [1] 49937.89
```

For each TMRCA, add a Poisson distributed number of mutations

```r
# Generate SNPs -----------
mu <- 1e-8
bps <- 100000
net_mutation_rates <- 2*mu*bps*coal_times
num_snps <- rpois(reps, net_mutation_rates)
# Mean should be approx 2Nu
mean(2*(N)*mu*bps); mean(num_snps)
```

```
## [1] 100.0788
## [1] 99.87501
```

We now need to decide which of the million draws from the priors to accept i.e. which are "close enough". Here, close enough is defined as "between 45 and 55"

```
lineages <- cbind(N, coal_times, num_snps)

# Subset to the draws of N that gave rise to 45-55 SNPs
accept <- subset(lineages, lineages[,3] < 56 & lineages[,3] > 44)

# Make sure it worked
min(accept[,3]); max(accept[,3])
```

```
## [1] 45
```

```
## [1] 55
```

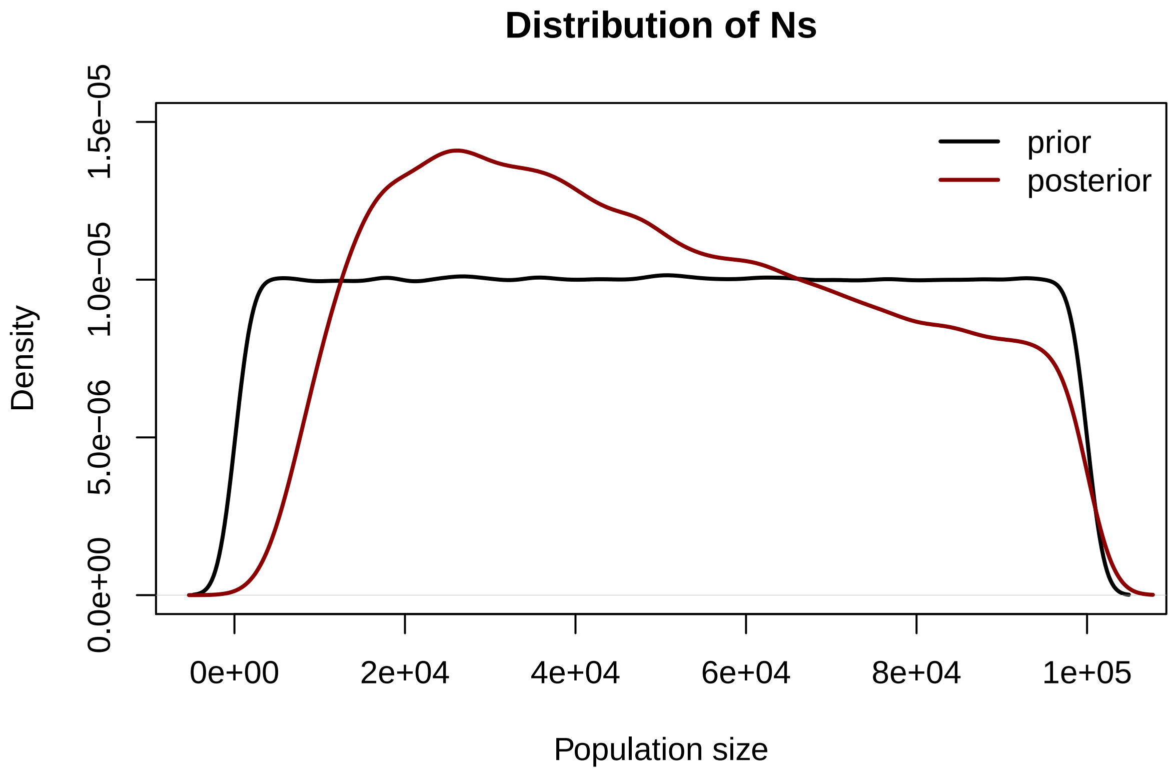
```
# Proportion of samples we keep
nrow(accept)/reps
```

```
## [1] 0.057362
```

Question 2

Make a density plot of the prior and the posterior.

```
# Plot
plot(density(N), lwd = 2, ylim = c(0, 0.000015), main = "Distribution of Ns", xlab = "Population size")
lines(density(accept[,1]), col = "darkred", lwd = 2)
legend("topright", lty = 1, col = c("black", "darkred"), legend = c("prior", "posterior"),
       bty = "n", lwd = 2)
```



Question 3 and 4

What is the median value of the posterior distribution of N? Generate a 95 credible interval for the posterior distribution of N

```
# Metrics on accepted Ns
med <- median(accept[,1]); med
```

```
## [1] 46958.67
```

```
quantile(accept[,1], c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 10013.96 96862.28
```

Question 5

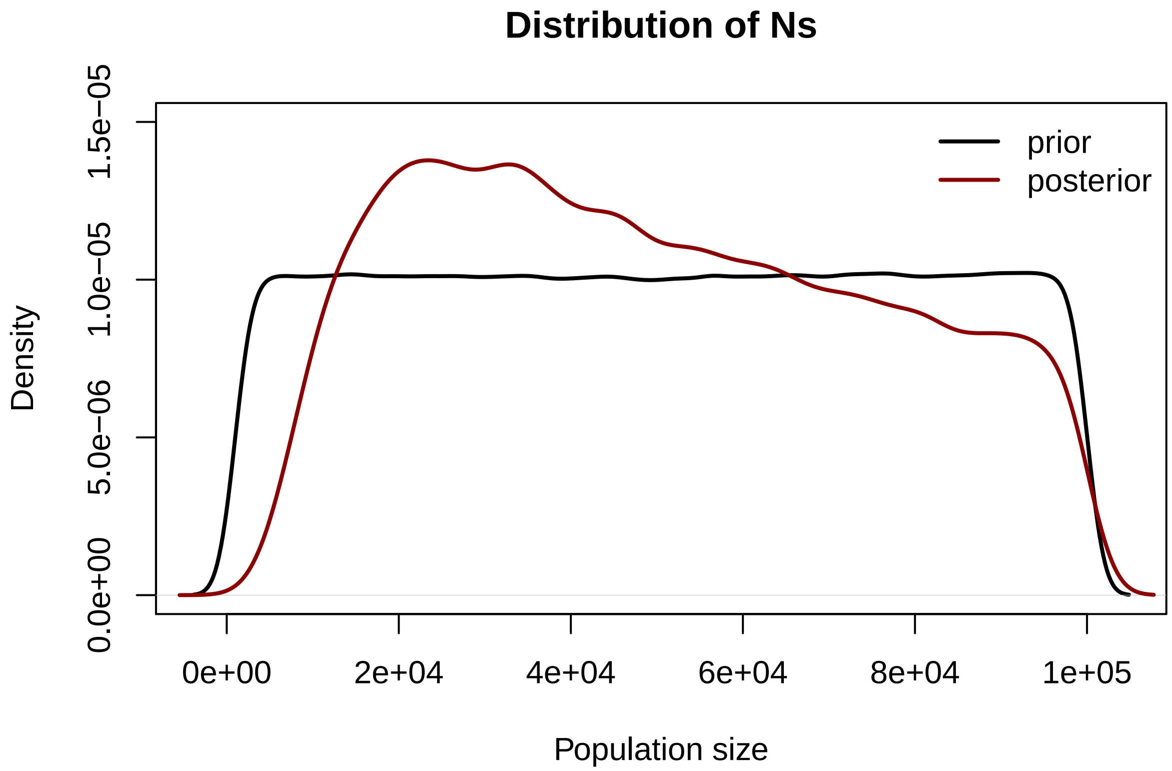
How does the posterior distribution differ from the prior distribution?

The posterior distribution of N is narrower than the uniform prior from which Ns were originally drawn. The posterior is centered around the median above with a 95% credible interval from ~9800 to ~97000.

Question 6

Repeat the ABC analysis with N $[1000, 100000]$

```r
# Draw Ns from distribution --------
reps <- 1e6
min = 1000; max = 100000
N <- runif(reps, min, max)
# Mean should be min+max/2
print("Expected and real Mean N");(min+max)/2; mean(N)
# Generate coalescent times ---------
rates <- 1/(N) # Only 1 Y chromosome in males...
coal_times <- rexp(reps, rates)
# Mean should be the same as mean(N)
print("Mean coal time"); mean(coal_times)
# Generate SNPs -----------
mu <- 1e-8
bps <- 100000
net_mutation_rates <- 2*mu*bps*coal_times
num_snps <- rpois(reps, net_mutation_rates)
# Mean should be approx 2Nu
print("Expected and real mean Num SNPs");mean(2*(N)*mu*bps); mean(num_snps)
lineages <- cbind(N, coal_times, num_snps)

# Subset to the draws of N that gave rise to 45-55 SNPs
accept <- subset(lineages, lineages[,3] < 56 & lineages[,3] > 44)

# Make sure it worked
print("Min and Max SNPs in accepts");min(accept[,3]); max(accept[,3])

# Proportion of samples we keep
print("Proportion of samples accepted");nrow(accept)/reps

# Plot
plot(density(N), lwd = 2, ylim = c(0, 0.000015), main = "Distribution of Ns", xlab = "Population size")
lines(density(accept[,1]), col = "darkred", lwd = 2)
legend("topright", lty = 1, col = c("black", "darkred"), legend = c("prior", "posterior"),
       bty = "n", lwd = 2)
```



```
# Metrics on accepted Ns
print("Median value of accept"); med <- median(accept[,1]); med
print("95 CI of median"); quantile(accept[,1], c(0.025, 0.975))
```

```
## [1] "Expected and real Mean N"
## [1] 50500
## [1] 50559.69
## [1] "Mean coal time"
## [1] 50546.56
## [1] "Expected and real mean Num SNPs"
## [1] 101.1194
## [1] 101.1062
## [1] "Min and Max SNPs in accepts"
## [1] 45
## [1] 55
## [1] "Proportion of samples accepted"
## [1] 0.057567
## [1] "Median value of accept"
## [1] 47234.62
## [1] "95 CI of median"
##      2.5%     97.5%
##  9876.411 96851.223
```

What is the mean, median, and 95 CI for the posterior? How dies this differ from the values computed in q3-4? What does this tell you about the effect of the prior distribution in Bayesian statistics?

```r
# Draw Ns from distribution --------
reps <- 1e6
min = 1000; max = 100000
N <- runif(reps, min, max)
N <- 2*N
# Mean should be min+max/2
print("Expected and real Mean N");(min+max)/2; mean(N)
```

```
## [1] "Expected and real Mean N"

## [1] 50500

## [1] 100954.8
```

```r
# Generate coalescent times ---------
rates <- 1/(N) # Only 1 Y chromosome in males...
coal_times <- rexp(reps, rates)
# Mean should be the same as mean(N)
print("Mean coal time"); mean(coal_times)
```

```
## [1] "Mean coal time"

## [1] 100912.8
```

```r
# Generate SNPs -----------
mu <- 1e-8
bps <- 100000
net_mutation_rates <- 2*mu*bps*coal_times
num_snps <- rpois(reps, net_mutation_rates)
# Mean should be approx 2Nu
print("Expected and real mean Num SNPs");mean(2*(N)*mu*bps); mean(num_snps)
```

```
## [1] "Expected and real mean Num SNPs"

## [1] 201.9095

## [1] 201.8251
```

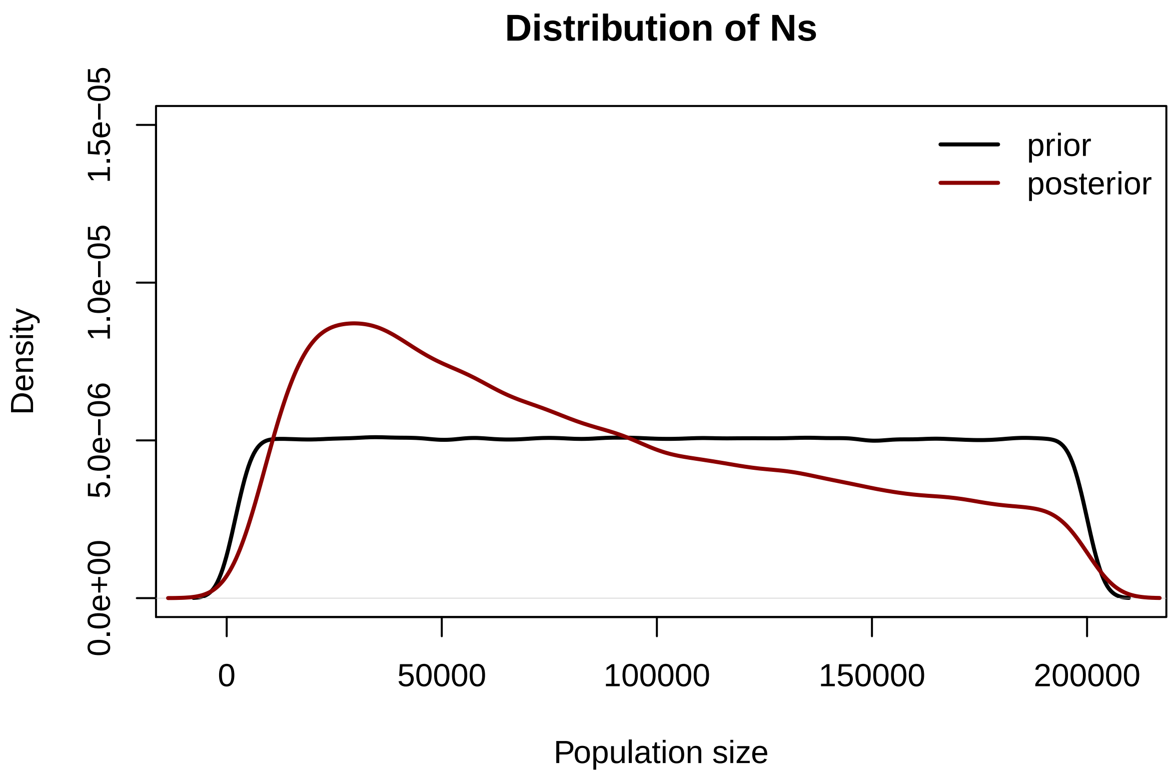
```r
lineages <- cbind(N, coal_times, num_snps)

# Subset to the draws of N that gave rise to 45-55 SNPs
accept <- subset(lineages, lineages[,3] < 56 & lineages[,3] > 44)

# Make sure it worked
print("Min and Max SNPs in accepts");min(accept[,3]); max(accept[,3])
```

```
## [1] "Min and Max SNPs in accepts"

## [1] 45

## [1] 55
```

```
# Proportion of samples we keep
print("Proportion of samples accepted");nrow(accept)/reps
```

```
## [1] "Proportion of samples accepted"

## [1] 0.044548
```

```
# Plot
plot(density(N), lwd = 2, ylim = c(0, 0.000015), main = "Distribution of Ns", xlab = "Population size")
lines(density(accept[,1]), col = "darkred", lwd = 2)
legend("topright", lty = 1, col = c("black", "darkred"), legend = c("prior", "posterior"),
       bty = "n", lwd = 2)
```



```
# Metrics on accepted Ns
print("Median value of accept"); med <- median(accept[,1]); med
```

```
## [1] "Median value of accept"

## [1] 73437.66
```

```
print("95 CI of median"); quantile(accept[,1], c(0.025, 0.975))
```

```
## [1] "95 CI of median"

##      2.5%     97.5%
##  11346.04 191315.09
```

Question 7

If you wanted a more precise estimate of the population size, would you be better off A) sequencing a bigger region of the Y chromosome, or B) sequencing the same amount on the autosomes? Why?

Seems like the second (higher)

Estimates of population split times

Question 1

Once the two chromosomes make it back into the ancestral population of size 100000, what is the expected amount of additional time we have to wait until they coalesce?

Since we are dealing with the Y chromosome, of which there is only 1 copy in diploid men, the expected coalescent time is equal to N- here, $E[T] = 100000$.

Question 2

What is the expected time until the two chromosomes coalesce with each other?

The net expected time to coalescence is $E[T_{coal}] = T_{split} + N$.

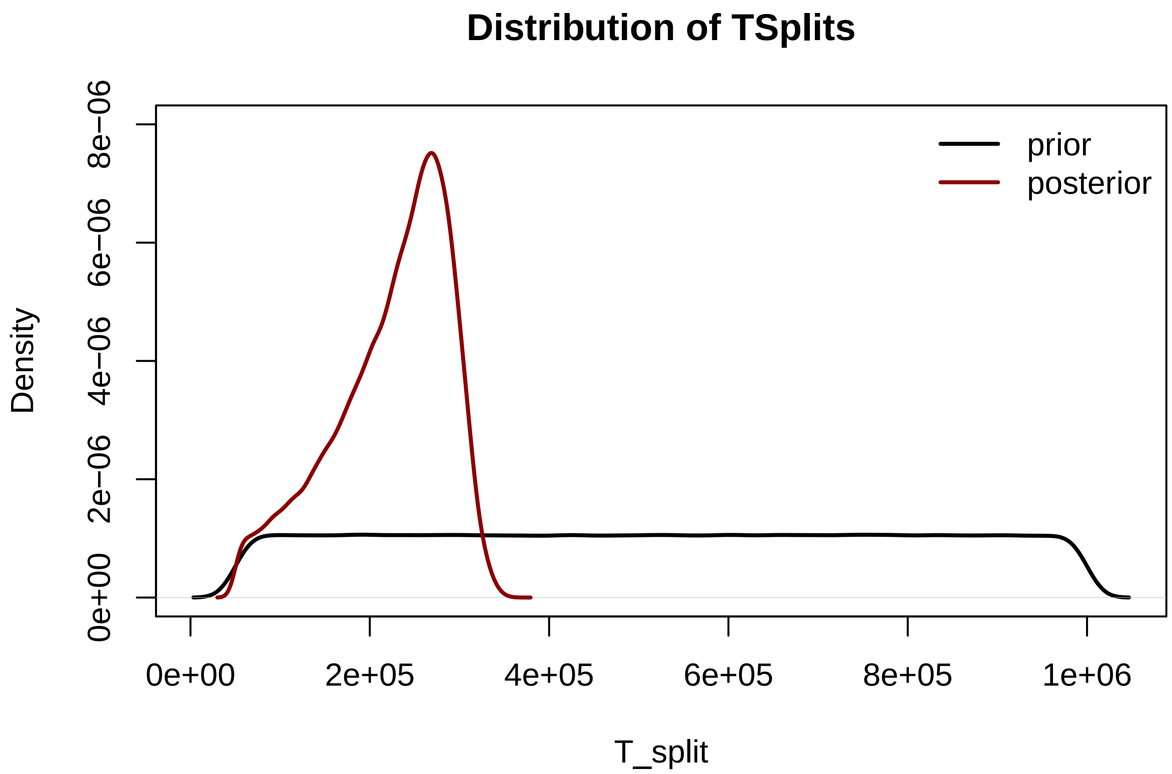
Question 3

Implement the following ABC approach to estimate T_{split}.

```
reps <- 1e6
min <- 50000
max <- 1000000
t_splits <- runif(reps, min, max)

N_anc <- 100000
rate <- 1/(N_anc) # Only 1 Y chromosome in males...
coal_times_pre_split <- rexp(reps, rate)

length(coal_times_pre_split) == length(t_splits)
```

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

```
coal_times <- coal_times_pre_split + t_splits



mu <- 1e-8
bps <- 100000
net_mutation_rates <- 2*mu*coal_times*bps
num_snps <- rpois(reps, net_mutation_rates)

lineages <- cbind(t_splits, coal_times, num_snps)
accepts <- subset(lineages, lineages[,3] > 549 & lineages[,3] < 651)
plot(density(lineages[,1]), ylim = c(0, 0.000008), lwd = 2, xlab = "T_split",
     main = "Distribution of TSplits")
lines(density(accepts[,1]), lwd = 2, col = "darkred")
legend("topright", lty = 1, col = c("black", "darkred"), legend = c("prior", "posterior"),
       bty = "n", lwd = 2)
```



```
mean(accepts[,1]); median(accepts[,1])
```

```
## [1] 223475.5
```

```
## [1] 237440.3
```

```
quantile(accepts[,1], probs = c(0.025, 0.975))
```

```
##     2.5%    97.5%
##  74346.7 314545.1
```

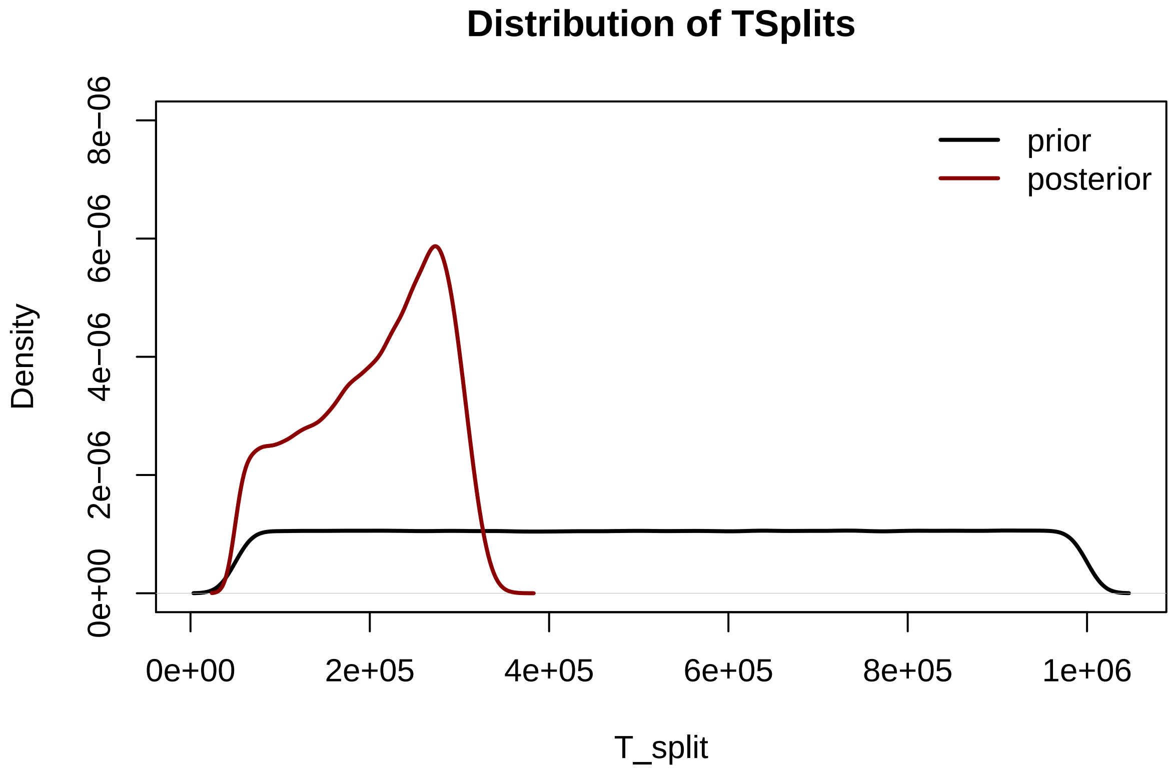
What if ancenstral N is unknown

```
reps <- 1e6
min <- 50000
max <- 1000000
```

```r
t_splits <- runif(reps, min, max)

N_anc <- runif(reps, min = 1000, max = 1000000) # This is different
rates <- 1/(N_anc) # Only 1 Y chromosome in males...
coal_times_pre_split <- rexp(reps, rates)

length(coal_times_pre_split) == length(t_splits)
```

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

```r
coal_times <- coal_times_pre_split + t_splits


mu <- 1e-8
bps <- 100000
net_mutation_rates <- 2*mu*coal_times*bps
num_snps <- rpois(reps, net_mutation_rates)

lineages <- cbind(t_splits, coal_times, num_snps)
accepts <- subset(lineages, lineages[,3] > 549 & lineages[,3] < 651)
plot(density(lineages[,1]), ylim = c(0, 0.000008), lwd = 2, xlab = "T_split",
     main = "Distribution of TSplits")
lines(density(accepts[,1]), lwd = 2, col = "darkred")
legend("topright", lty = 1, col = c("black", "darkred"), legend = c("prior", "posterior"),
       bty = "n", lwd = 2)
```



```r
mean(accepts[,1]); median(accepts[,1])
```

```
## [1] 204956.3
```

```
## [1] 217452.1
```

```r
quantile(accepts[,1], probs = c(0.025, 0.975))
```

```
##      2.5%     97.5% 
##  60579.68 314460.26
```