R Notebook, Week of 8/11

First we simulate data according to MAMBA and calculate the PPR values (from the MAMBA package) and the PRP values (from the PRP package). The job R files used for this are located in the code/data directories. The packages are calculated for all 50000 SNPS. Here, we'll look at the case where the nonoutlier study rate is 0.975.

```
load(file = "data/mamba_data/sim_mamba_mod_p975.rda")
load(file = "data/mamba_data/mamba_data_p975.rda")

pprs <- sim_mod$ppr
```

We take the indices for the SNPs that have at least one outlier study.

```
# indices for snps w/ and w/o outliers
out_studies <- mamba_data$Ojk
out_rows_ind <- which(rowSums(out_studies == 0) > 0) # indices of snps with outlier studies
no_out_rows_ind <-which(rowSums(out_studies) == 10) # indices of rows w/o outliers
```

Then we load in the data containing the PRPs (from PRP package) for each SNP.

```
# loading PRP data

load(file = "data/prp_data/post_prp_data_pval_p975.rda") # post_prp_data_pval
```

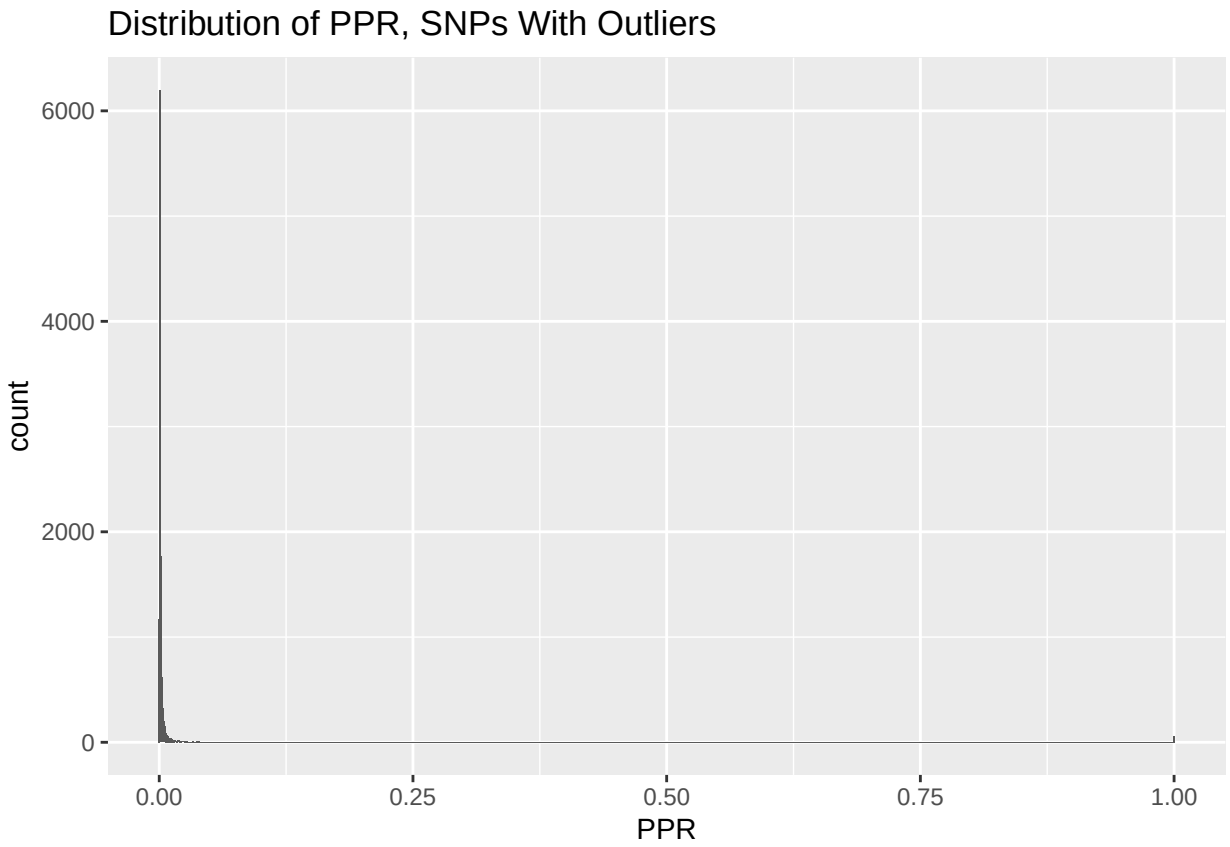
Below are the PPRs and PRPs from the the SNPs with and without outlier studies.

```
# MAMBA pprs for snps w/ and w/o outliers
out_ppr <- pprs[out_rows_ind]
nonout_ppr <- pprs[no_out_rows_ind]

# prps for snps w/ and w/o outliers
out_prp <- post_prp_data_pval[out_rows_ind]
nonout_prp <- post_prp_data_pval[no_out_rows_ind]
```

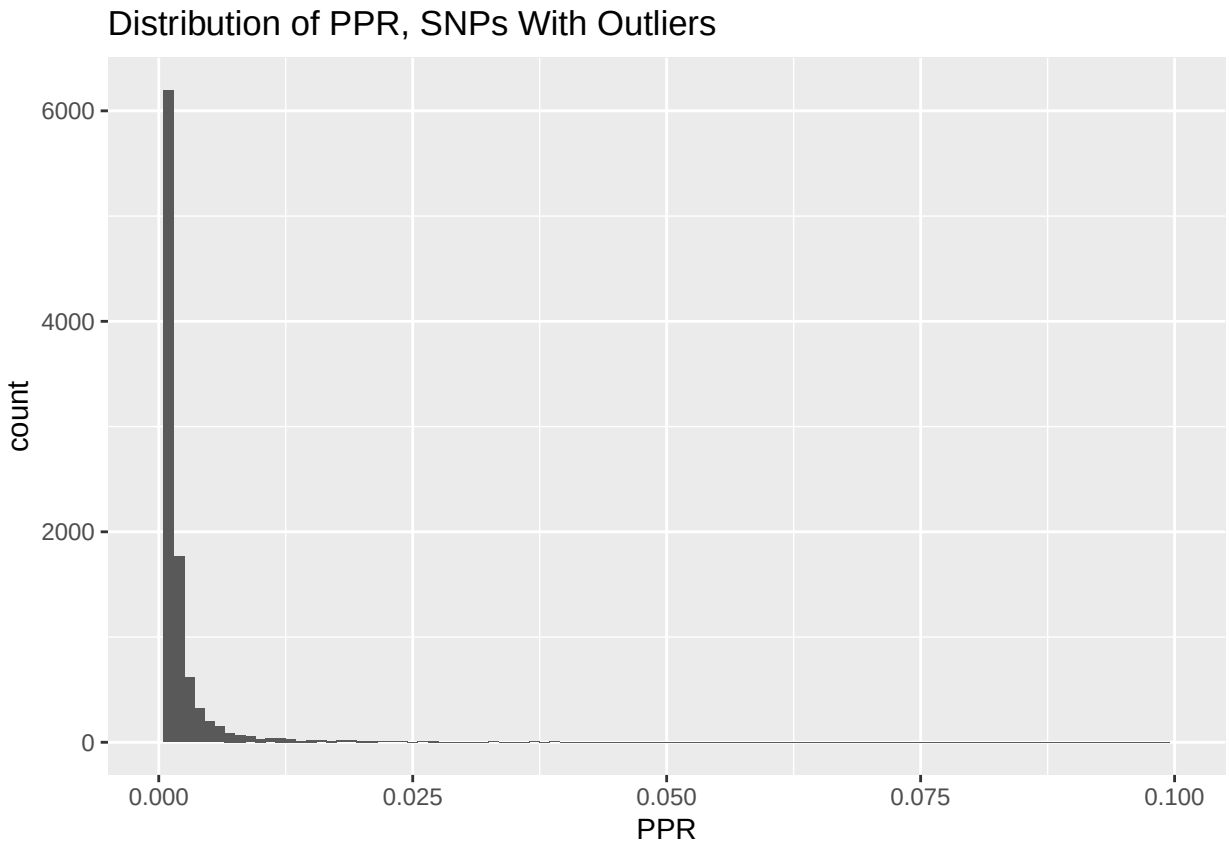
For our SNPs **with outliers**, the distribution of our PPRs looks as follows:

```
ggplot(data = as.data.frame(out_ppr), aes(out_ppr)) +
  geom_histogram(binwidth = 0.001) +
  ggtitle("Distribution of PPR, SNPs With Outliers") +
  xlab("PPR")
```



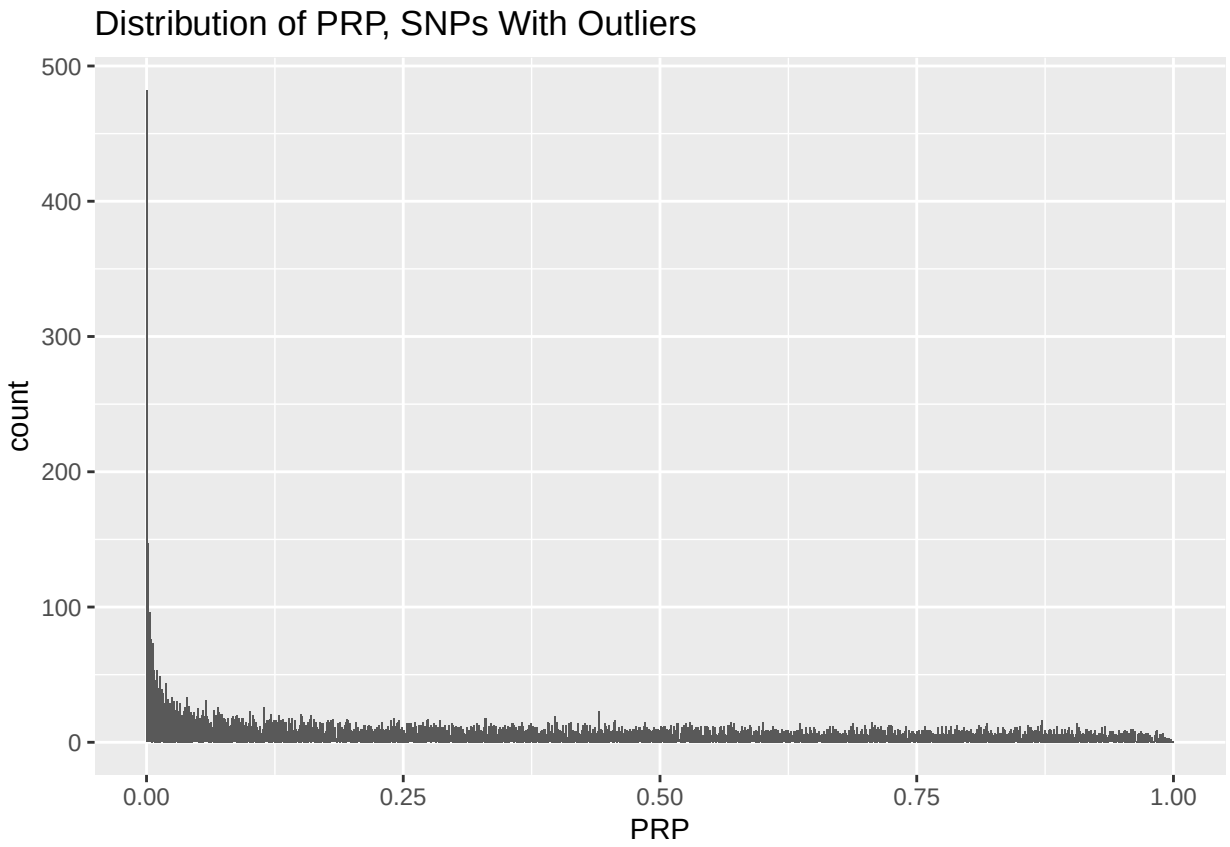
Rescaling the graph by ignoring some outliers, we get the histogram below.

```
ggplot(data = as.data.frame(out_ppr), aes(out_ppr)) +
  geom_histogram(binwidth = 0.001) +
  ggtitle("Distribution of PPR, SNPs With Outliers") +
  xlab("PPR") +
  xlim(0, 0.1)
```



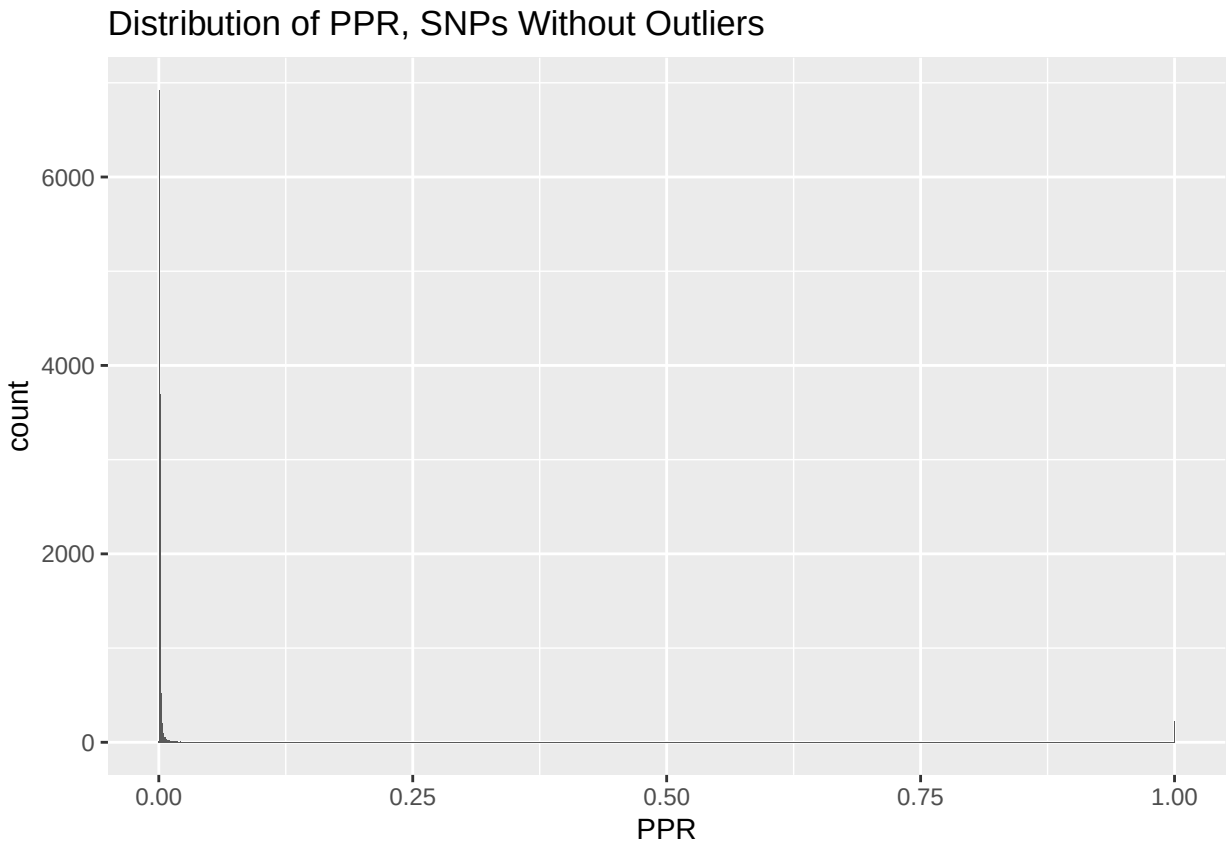
In contrast, this is what the distribution of our PRPs for outlier SNPs looks like. Keep in mind the differently scaled x-axis.

```
ggplot(data = as.data.frame(out_prp), aes(out_prp)) +
  geom_histogram(binwidth = 0.001) +
  ggtitle("Distribution of PRP, SNPs With Outliers") +
  xlab("PRP")
```



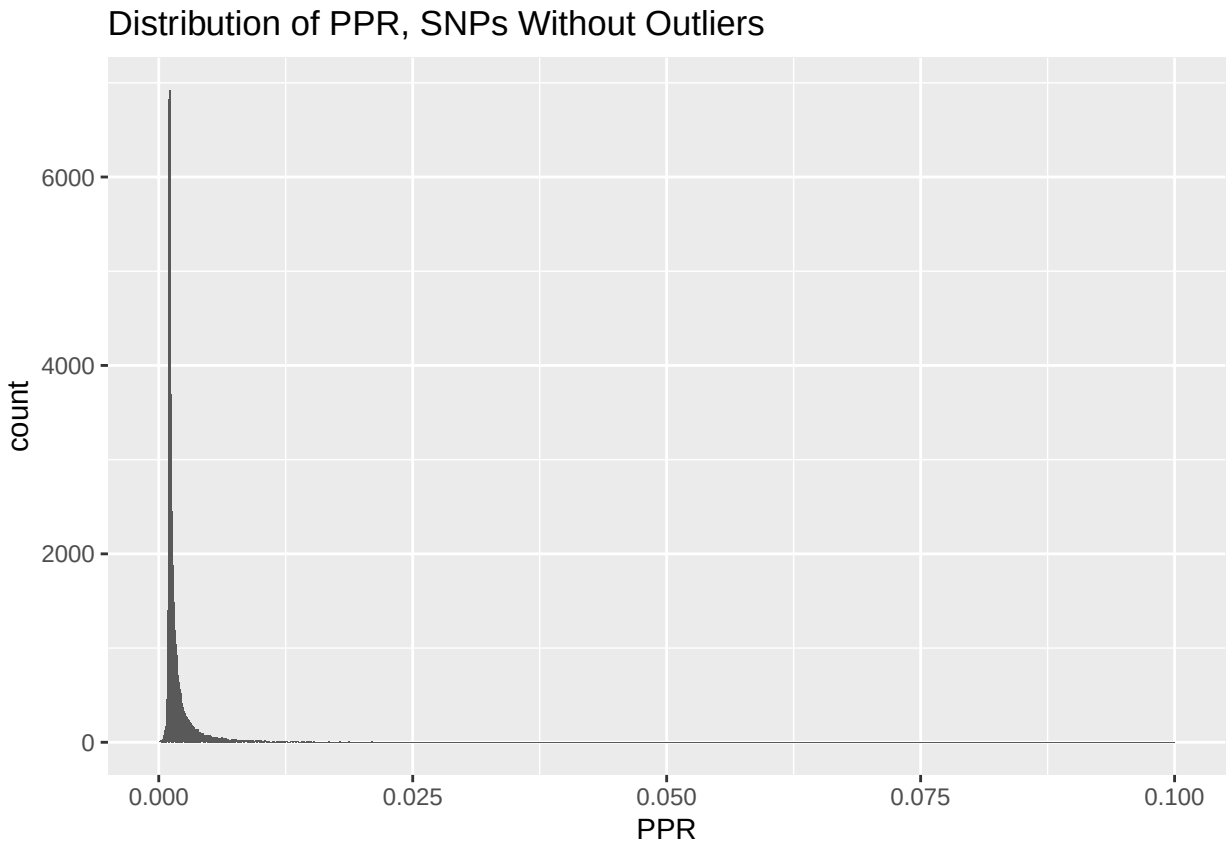
For our SNPs **without outliers**, the PRP distribution is given in the following histogram.

```
ggplot(data = as.data.frame(nonout_ppr), aes(nonout_ppr)) +
  geom_histogram(binwidth = 0.0001) +
  ggtitle("Distribution of PPR, SNPs Without Outliers") +
  xlab("PPR")
```



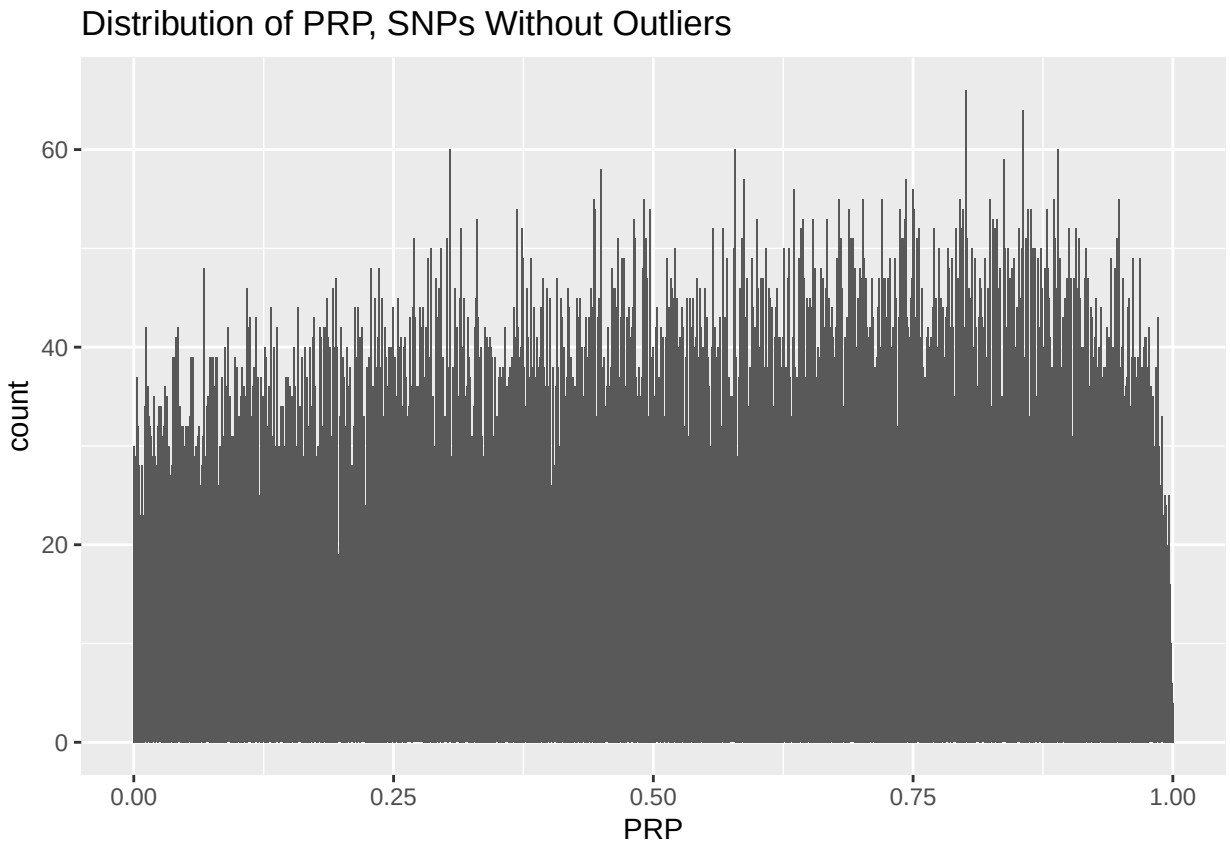
Like before, if we ignore extreme PPR values, our histogram looks different.

```
ggplot(data = as.data.frame(nonout_ppr), aes(nonout_ppr)) +
  geom_histogram(binwidth = 0.0001) +
  ggtitle("Distribution of PPR, SNPs Without Outliers") +
  xlab("PPR") +
  xlim(0, 0.1)
```



In contrast, this is what the distribution of our PRP values looks like.

```
ggplot(data = as.data.frame(nonout_prp), aes(nonout_prp)) +
  geom_histogram(binwidth = 0.001) +
  ggtitle("Distribution of PRP, SNPs Without Outliers") +
  xlab("PRP")
```


Distribution of PRP, SNPs Without Outliers
count
60
40
20
0
0.00
0.25
0.50
0.75
1.00
PRP

Different Population Sampling (this hasn't been updated yet from last week)

We sample from 5000 individuals. For each "study," we sample at different allele frequencies (0.01, 0.1, and 0.25).

```
# allele sampling
n = 5000
allele0.01 <- rbinom(n, 2, 0.01)
allele0.1 <- rbinom(n, 2, 0.1)
allele0.25 <- rbinom(n, 2, 0.25)

# noise
epsilon <- rnorm(5000, mean = 0, sd = 1)
```

Using a linear equation, we get our phenotype values.

```
beta = rnorm(1, mean = 0, sd = 1)
y0.01 = (beta * allele0.01) + epsilon
y0.1 = (beta * allele0.1) + epsilon
y0.25 = (beta * allele0.25) + epsilon
```

Finally we fit a linear model to get an effect size estimate.

```
lm_model0.01 <- lm(y0.01 ~ allele0.01)
bhat0.01 <- lm_model0.01$coefficients[2]
se0.01 <- summary(lm_model0.01)$sigma

lm_model0.1 <- lm(y0.1 ~ allele0.1)
bhat0.1 <- lm_model0.1$coefficients[2]
se0.1 <- summary(lm_model0.1)$sigma

lm_model0.25 <- lm(y0.25 ~ allele0.25)
bhat0.25 <- lm_model0.25$coefficients[2]
se0.25 <- summary(lm_model0.25)$sigma
```

The effect size estimates are given in the table below. Our true effect size is -0.7760756.

allele0.01	allele0.1	allele0.25
-0.6745	-0.7719	-0.775

Their respective standard errors are given in the table below.

0.991, *0.9911* and *0.9911*

Using these predictions, we can see how MAMBA and the PRP library behave. We treat each allele frequency as its own study, and so we compare the three studies with different allele frequencies to each other.

```
# list of our (predicted) effect size and variance
snpeffect <- c(bhat0.01, bhat0.1, bhat0.25)
snpeffect_prp <- rbind(snpeffect, snpeffect)
snpvar <- c(se0.01**2, se0.1**2, se0.25**2)
snpvar_prp <- rbind(snpvar, snpvar)
```

First we fit these results to MAMBA.

```
diff_pop <- mamba(beta = snpeffect_prp,
                  sjk2 = snpvar_prp)
```

```
diff_pop$outliermat
```

```
##    snp       Oj_1      Oj_2       Oj_3         ppr
## 1:   1 0.03128935 0.0296398 0.02958526 0.003348493
## 2:   2 0.03128935 0.0296398 0.02958526 0.003348493
```

Next we fit to the PRP model.

```
diff_pop_prp <- posterior_prp(beta = snpeffect,
                              se = sqrt(snpvar))
```

Using MAMBA, the PPR is 0.0033485. The p-value given by the PRP package is 0.772.

Giant Consortium Studies

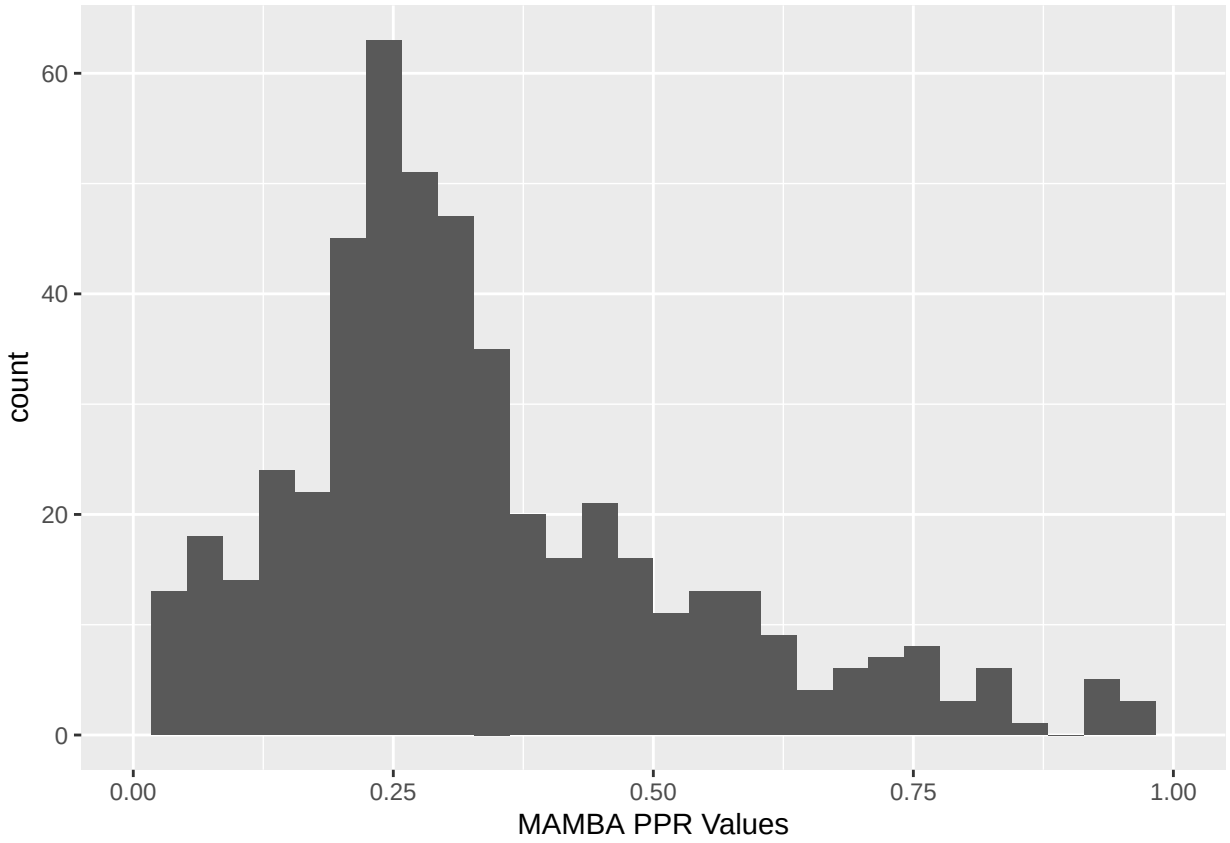
BMI

First we look at the BMI.

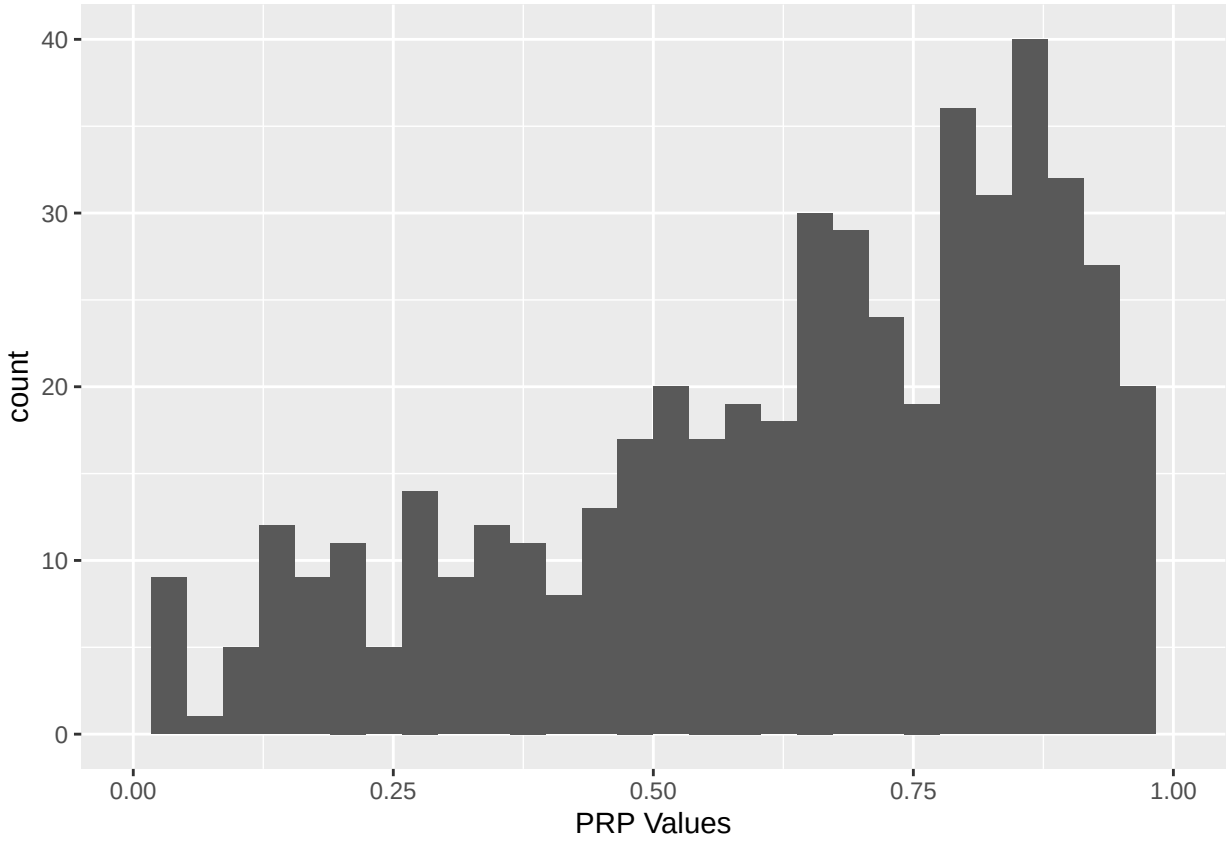
BMI Graphs

Now we get a better look at the BMI through graphical representations.

We take a look at our PPR and PRP values. The histogram of our PPR values for BMI is given below.



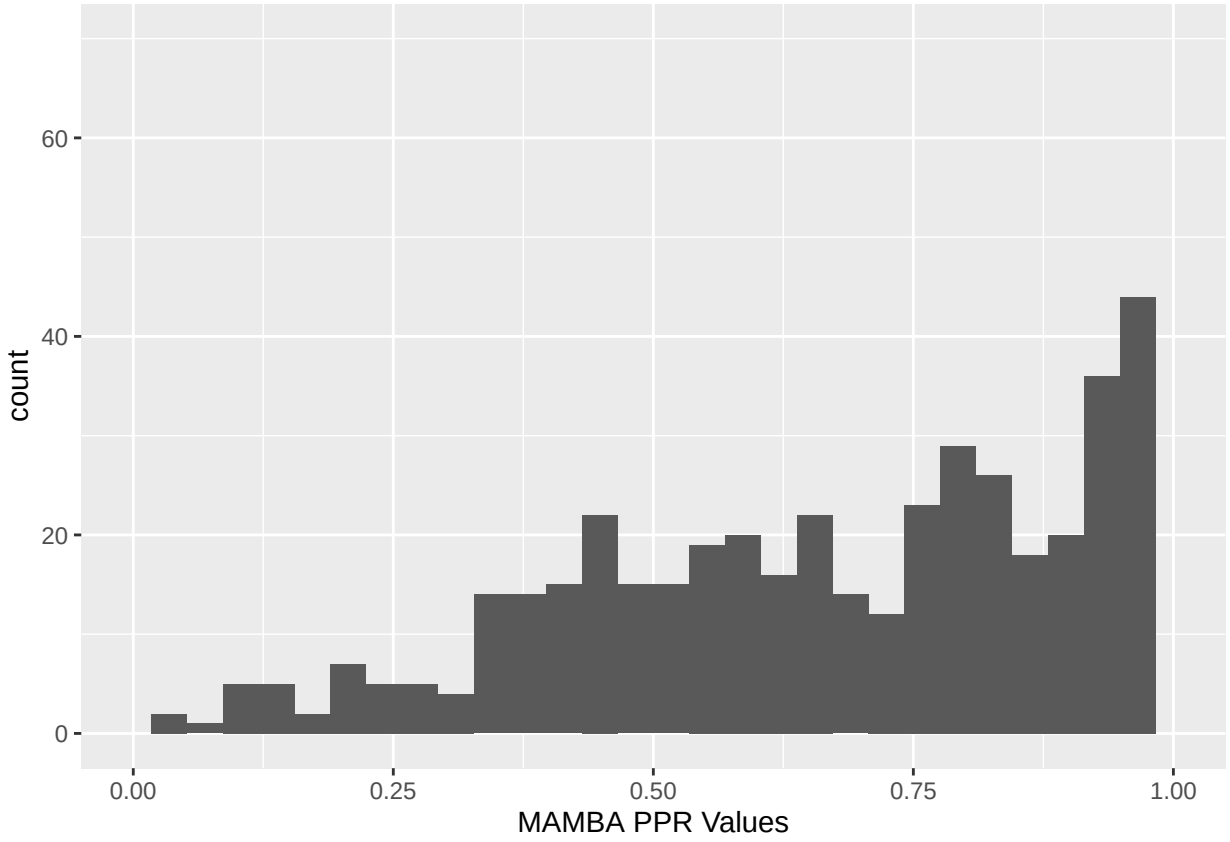
In contrast, here is our histogram of PRP values.


40
30
20
10
0
count
0.00
0.25
0.50
0.75
1.00
PRP Values

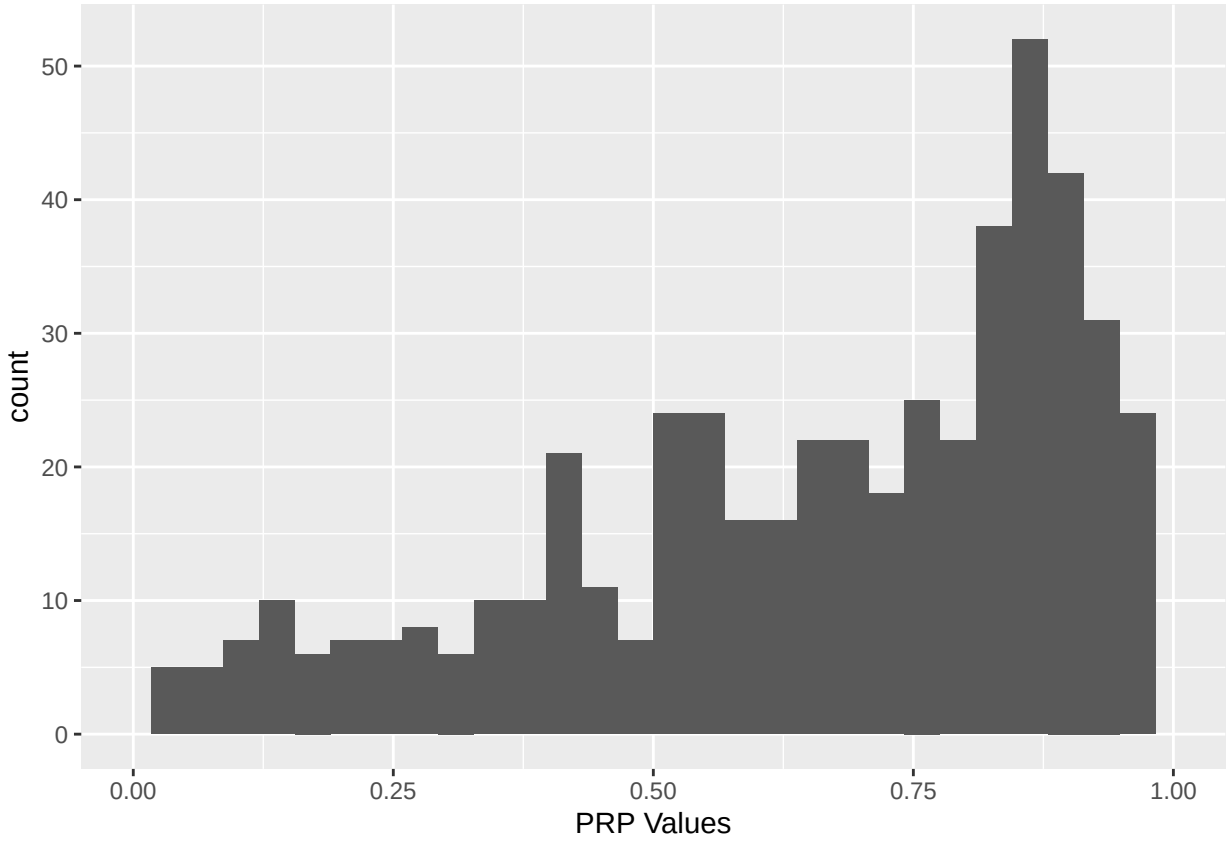
Height

Next we look at height.

We take a look at our PPR and PRP values. The histogram of our PPR values is given below.



In contrast, we have our histogram for PRP values.


50
40
30
20
10
0
count
0.00
0.25
0.50
0.75
1.00
PRP Values