R Notebook, Week of 8/18

Darren Lin

Contents

MAMBA Simulations **2**

SNPs With Outliers 3

SNPs Without Outliers 6

SNP Counts 9

Giant Consortium Studies **10**

Height 10

MAMBA Simulations

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 for the generated MAMBA simulations is 0.975.

```
# loading mamba and prp data
load(file = "data/mamba_data/sim_mamba_mod_p975.rda")
load(file = "data/mamba_data/mamba_data_p975.rda")
load(file = "data/prp_data/post_prp_data_pval_p975.rda") # post_prp_data_pval

pprs <- sim_mod$ppr
```

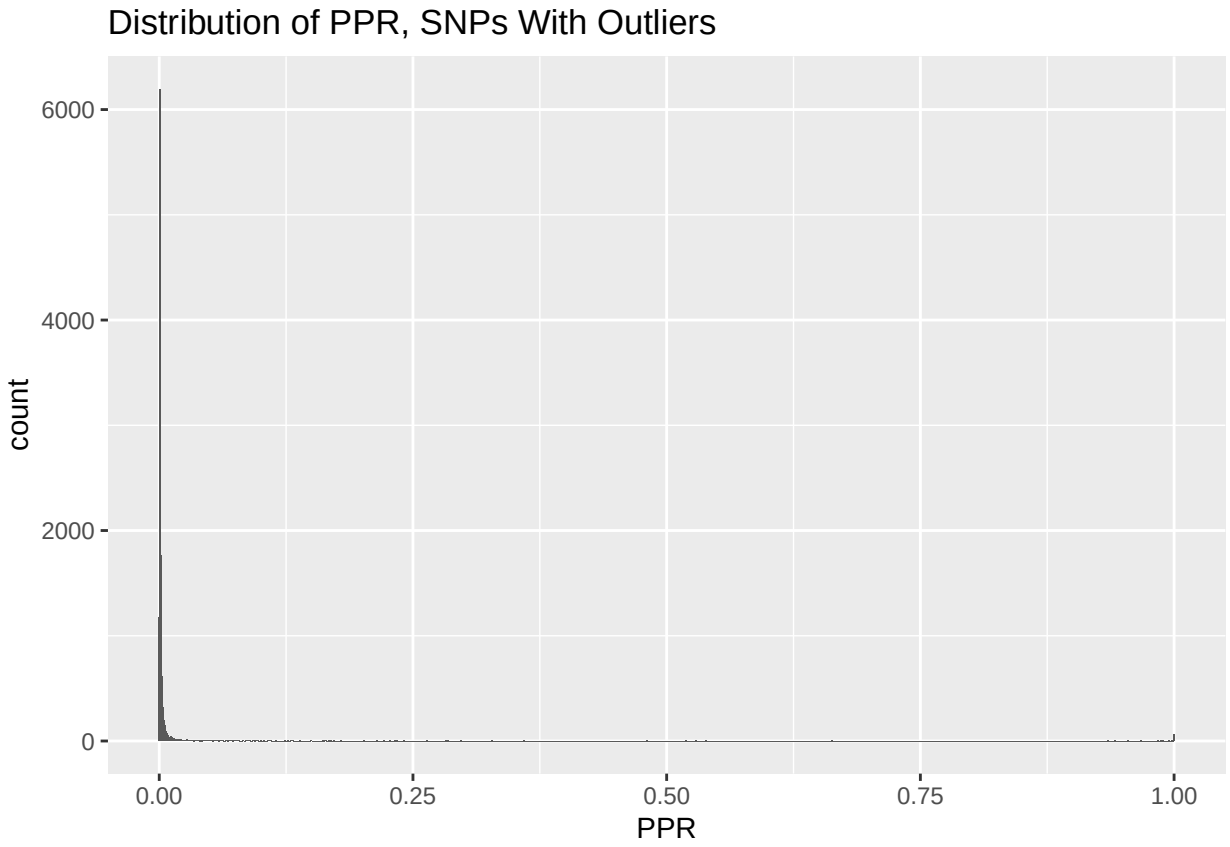
We take the indices for the SNPs that have at least one outlier study. We use these indices to compare against the results of the MAMBA and PRP libraries.

```
# 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

# 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]
```

SNPs With Outliers

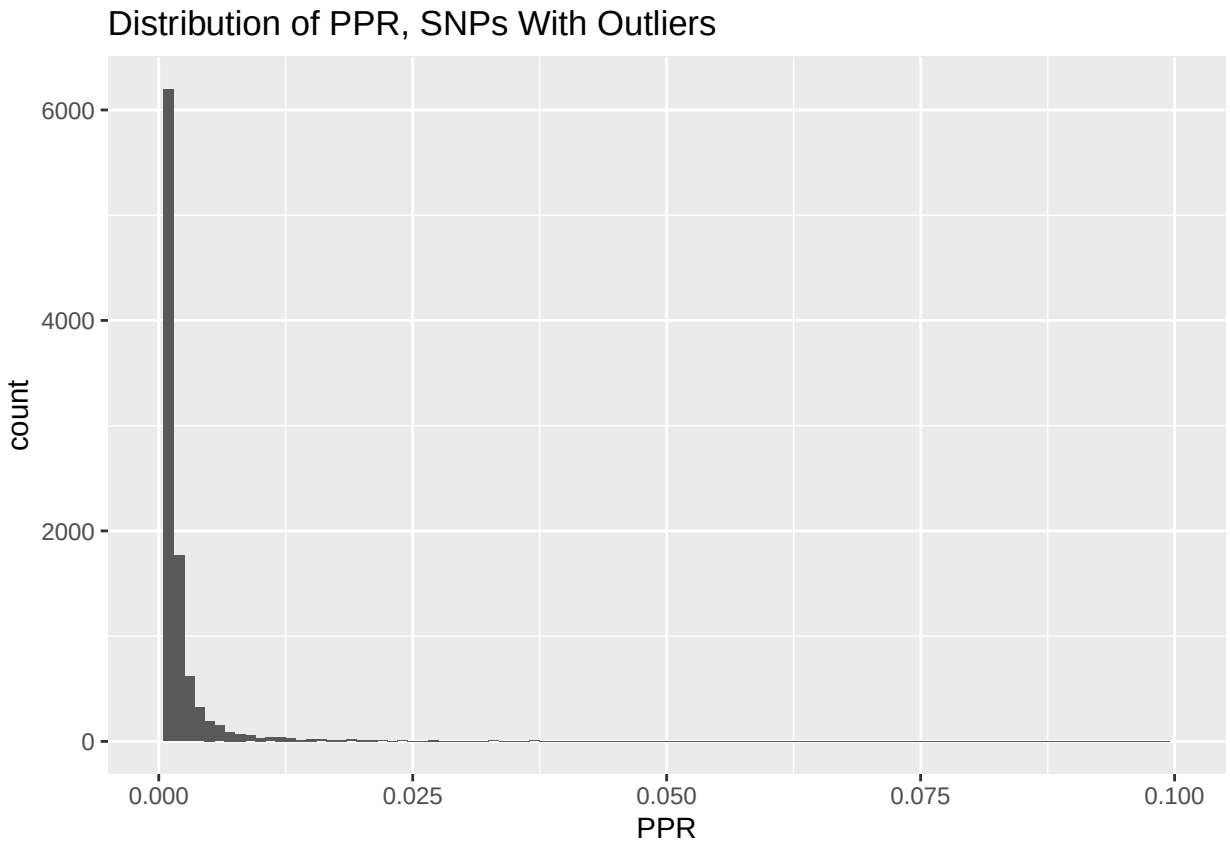
For our SNPs **with outliers**, the distribution of our PPR values from MAMBA is given below. We can see that our distribution is heavily right skewed, with most PPR values below 0.1.

```
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. We can see it is still heavily right skewed.

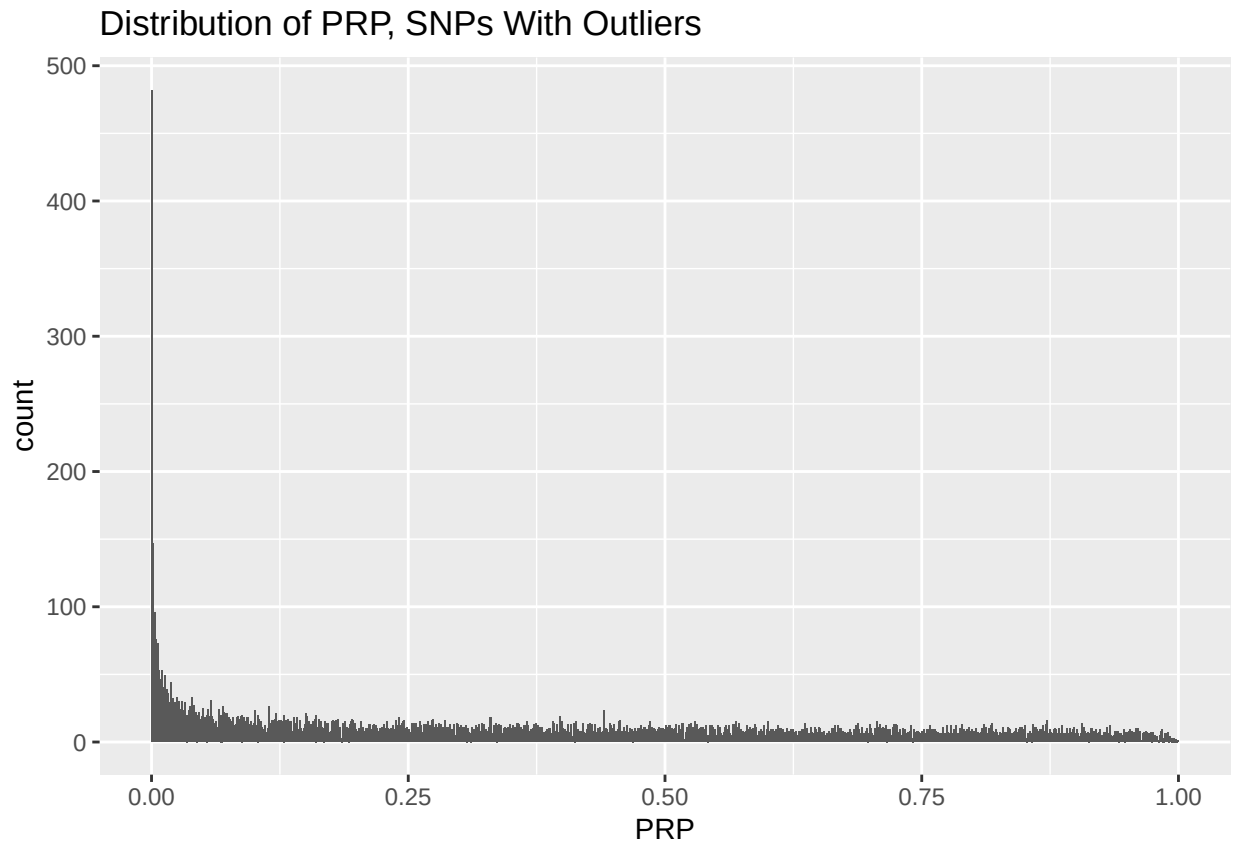
```
ggplot(data = as.data.frame(out_ppr), aes(out_ppr)) +
  geom_histogram(binwidth = 0.001) +
  ggtitle("Distribution of PPR, SNPs With Outliers") +
  xlab("PPR") +
  xlim(0, 0.1)
```



With outliers, there are around 11004 SNPS with PPR values less than or equal to 0.05 (our significant SNPs).

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")
```

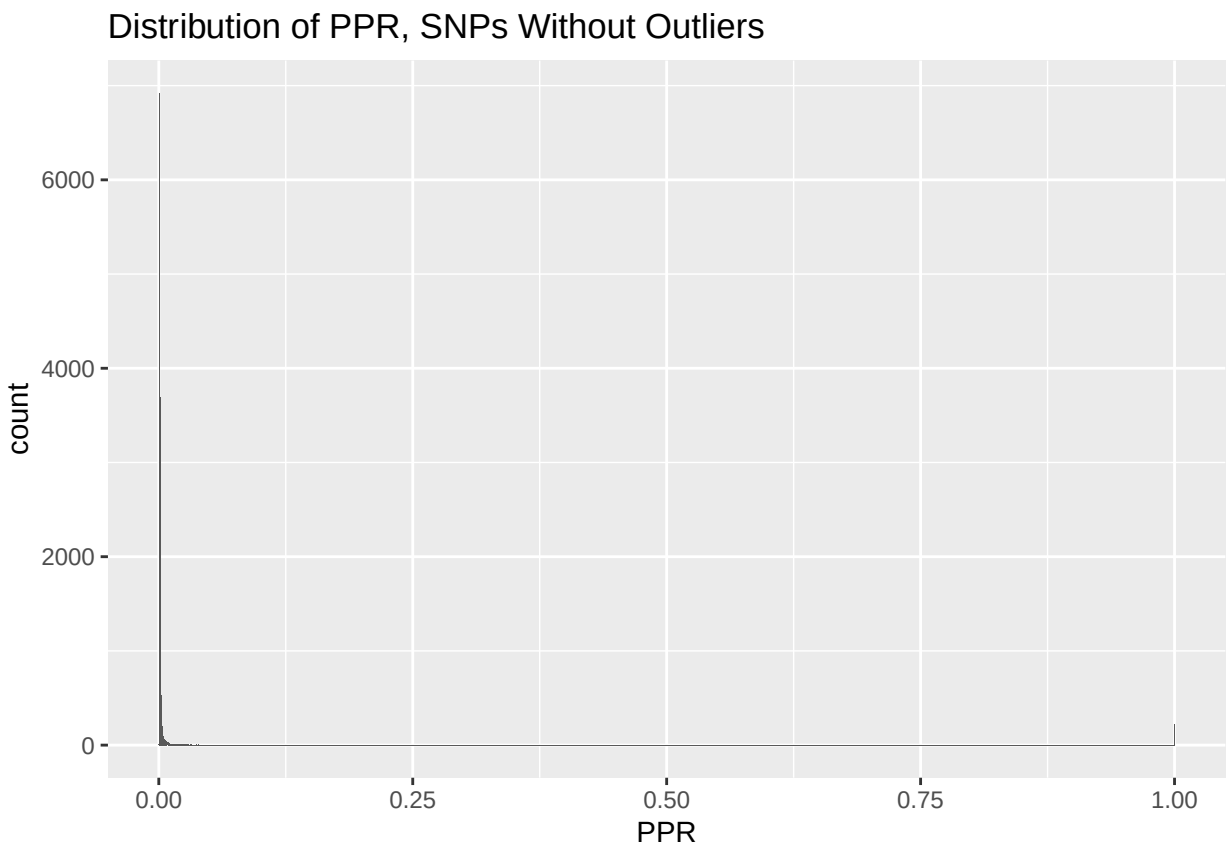


With outliers, there are around 2302 SNPS with PPR values less than or equal to 0.05.

SNPs Without Outliers

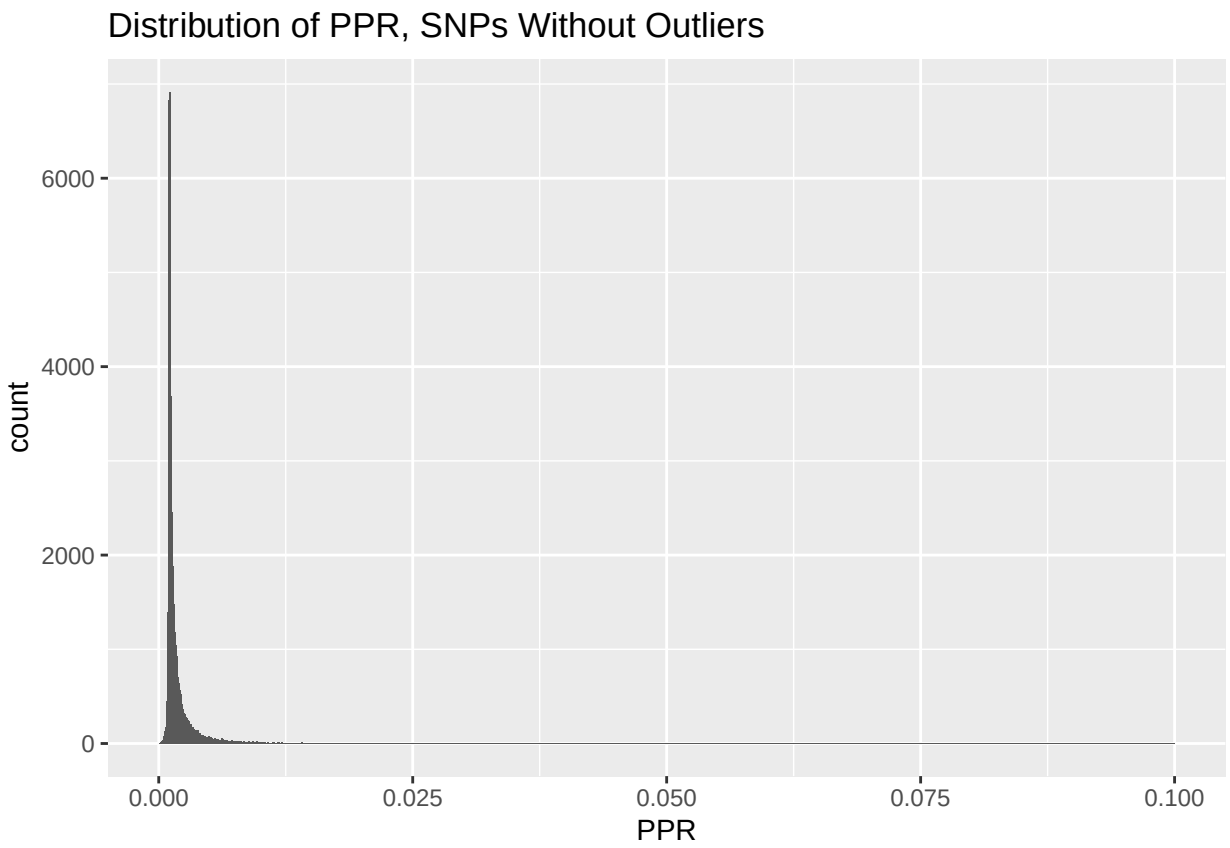
For our SNPs **without outliers**, the PRP distribution is given in the following histogram.

```
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, we can see that much of our PPR values are below 0.025.

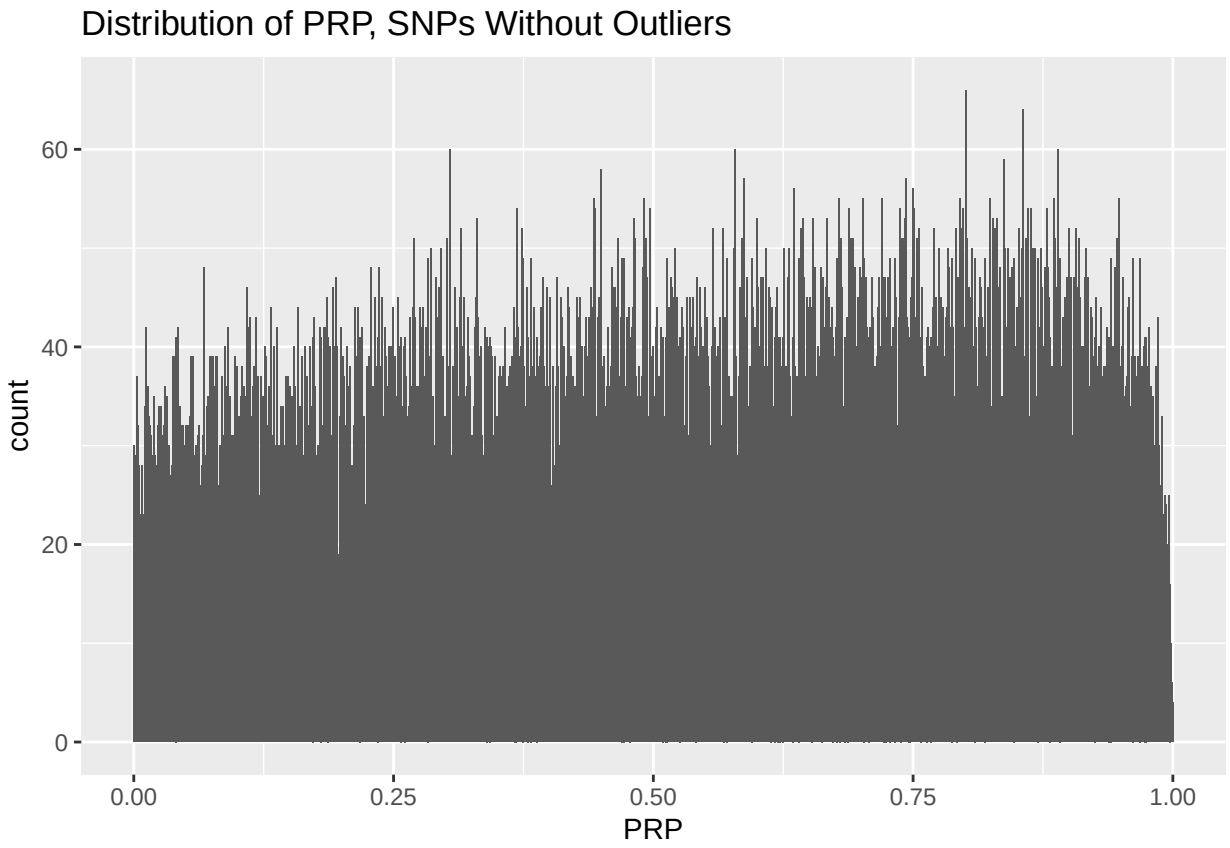
```
ggplot(data = as.data.frame(nonout_ppr), aes(nonout_ppr)) +
  geom_histogram(binwidth = 0.0001) +
  ggtitle("Distribution of PPR, SNPs Without Outliers") +
  xlab("PPR") +
  xlim(0, 0.1)
```



The number of known nonoutlier SNPs with PPR values less than or equal to 0.05 is 38376. The total number of SNPs with PPRs less than or equal to 0.05 using the MAMBA method is 49380.

In contrast, this is what the distribution of our PRP values looks like. We can see that it is much more uniform compared to the other graphs above.

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



The number of known nonoutlier SNPs with PPR values less than or equal to 0.05 is 38376. The total number of SNPs with PPRs less than or equal to 0.05 using the MAMBA method is 3852. Recall that here, our nonoutlier study rate is 0.975.

SNP Counts

A table of the number of **outlier** SNPs that are considered nonreplicable (a PRP or PPR value of 0.05 or less) is given below for each replicability method.

	MAMBA Outlier	PRP Outlier
Significant	11004	2302
Non-Significant	167	8869

For comparison, a table of the number of **nonoutlier** SNPs that are considered nonreplicable is below.

	MAMBA Nonoutlier	PRP Nonoutlier
Significant	38376	1550
Non-Significant	453	37279

A table of the total significant and nonsignicant SNPs is given below. For reference, there were a total of 50000 SNPs in the simulated data.

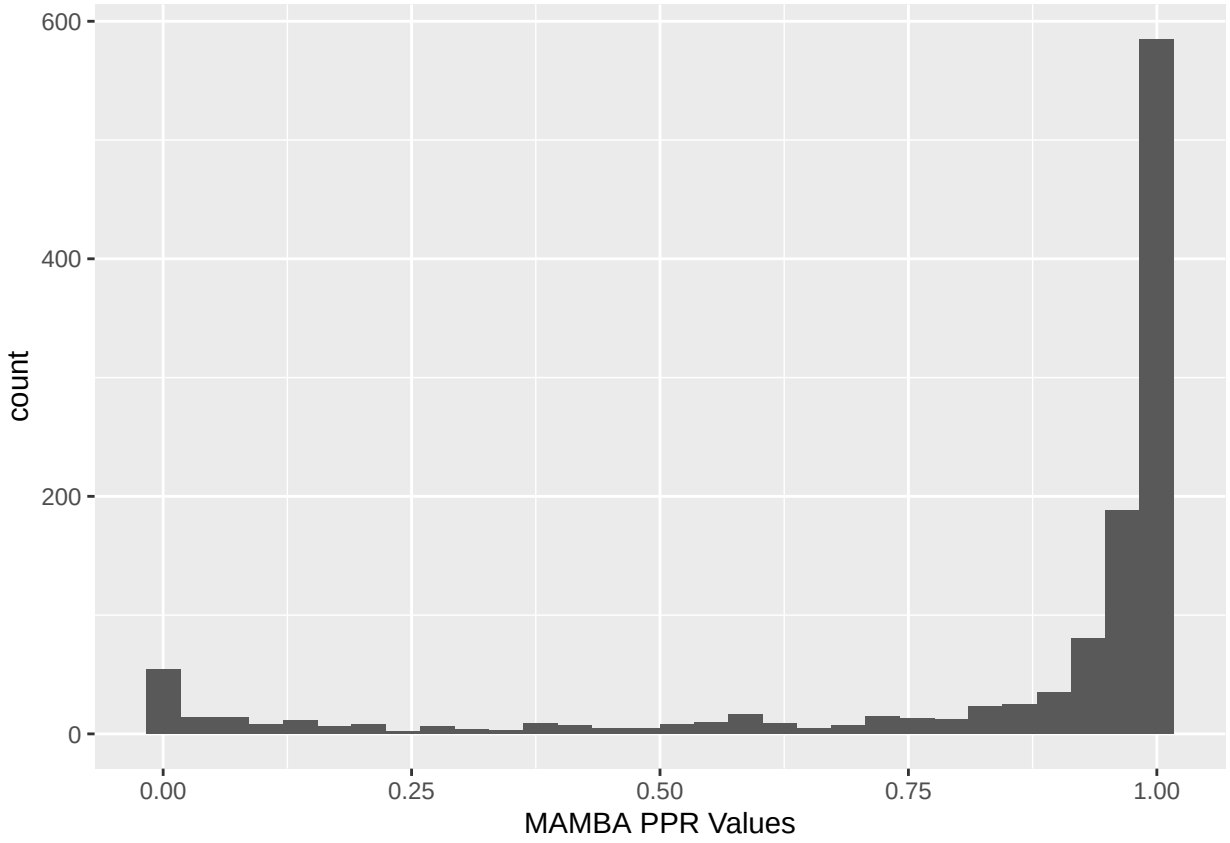
	MAMBA	PRP
Total Significant SNPs	49380	3852
Total Nonsignificant SNPs	620	46148

Giant Consortium Studies

Height

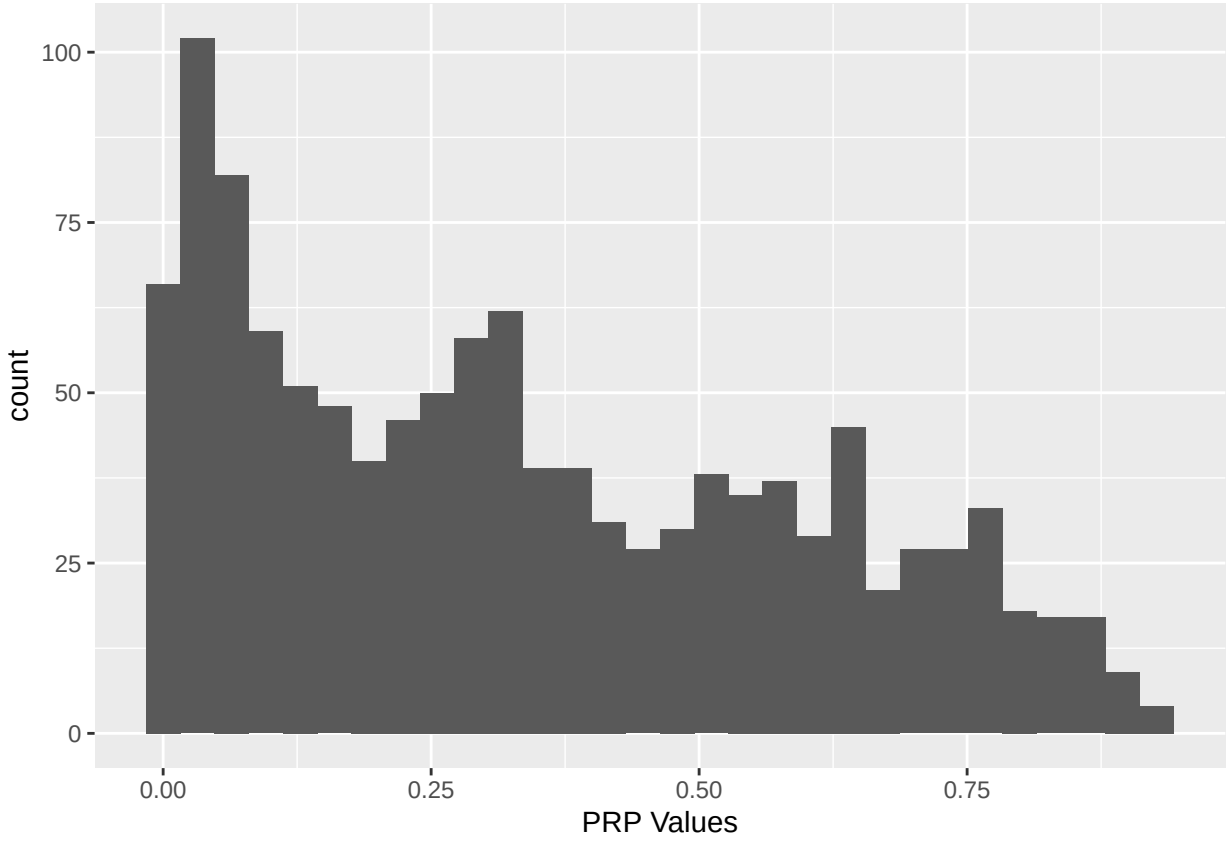
Height Graphs

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

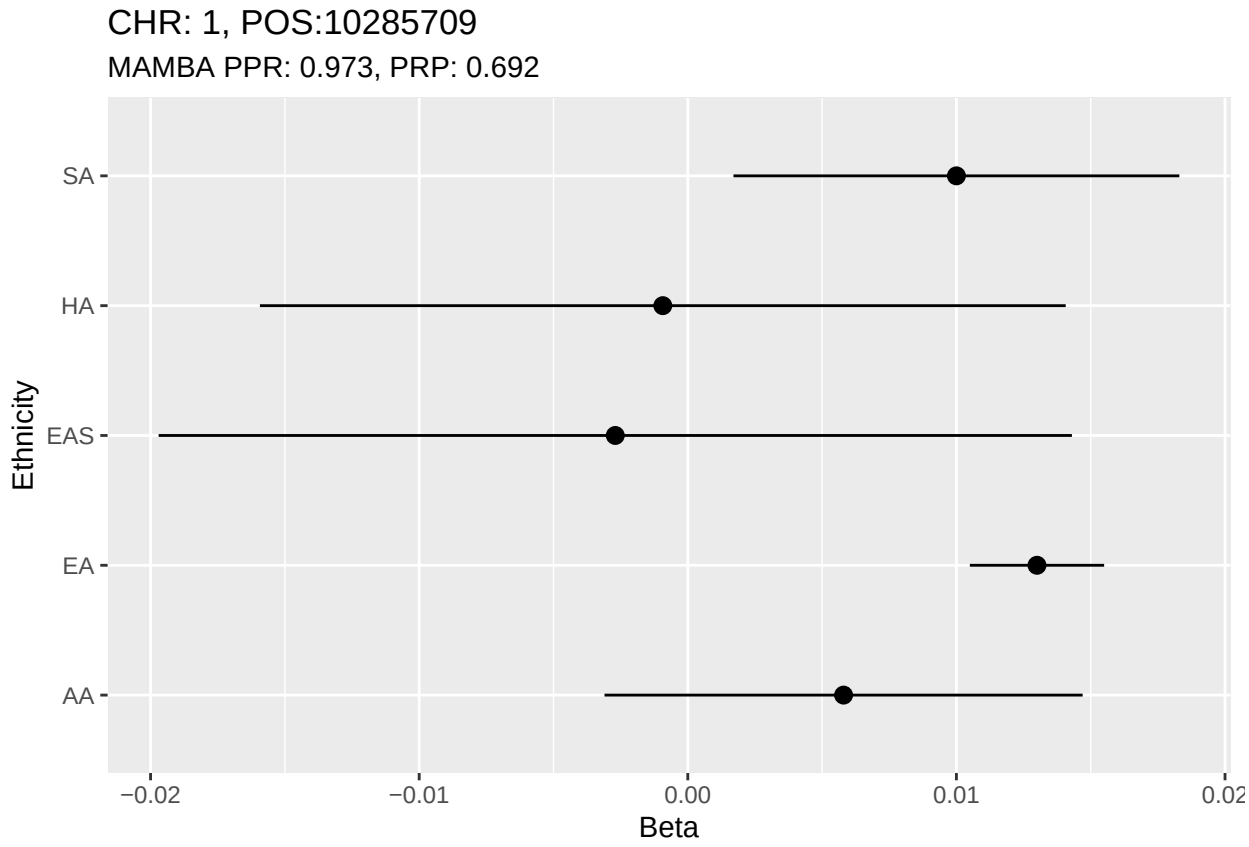


We can see that the distribution of the PPR values is skewed left. Many values relatively high, with only a few that seem significant.

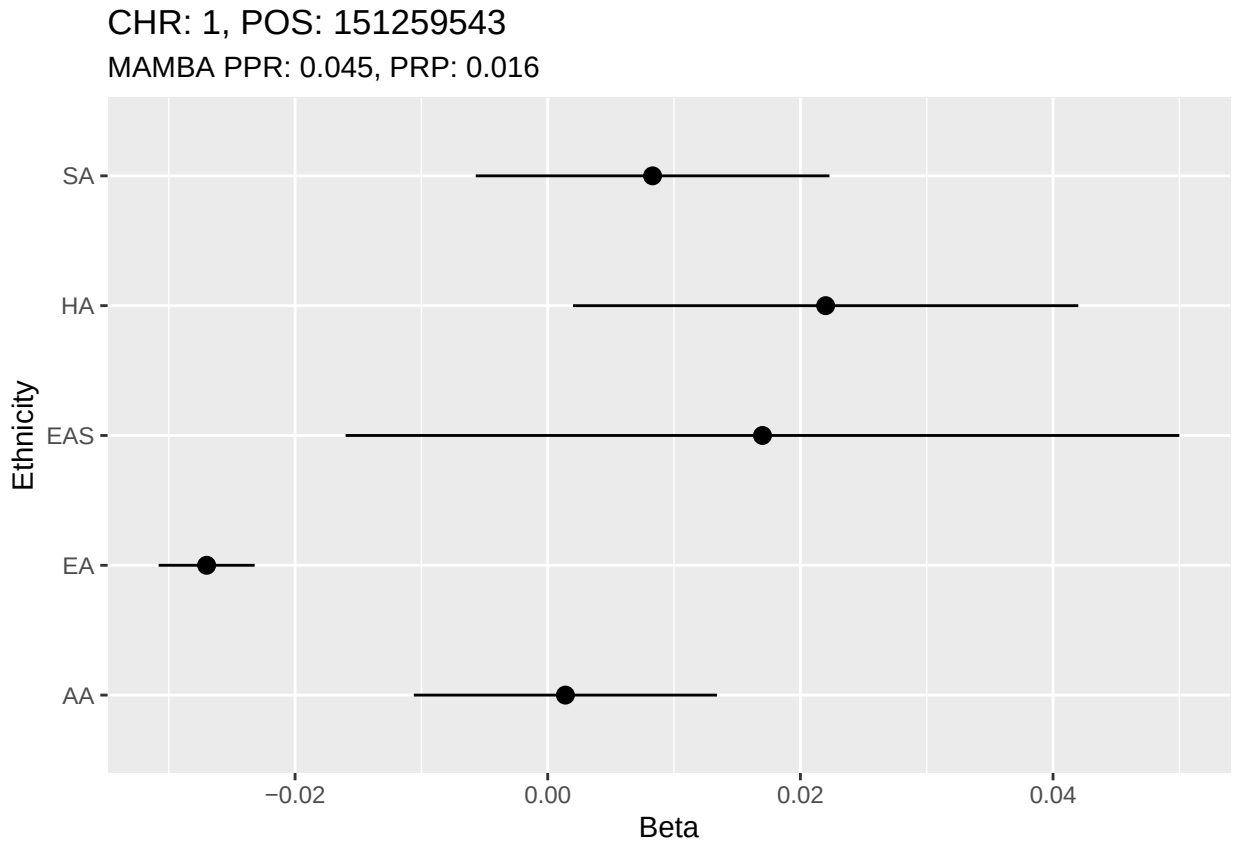
In contrast, our histogram for PRP values seems more evenly distributed, with a slight right skew More values are considered significant.



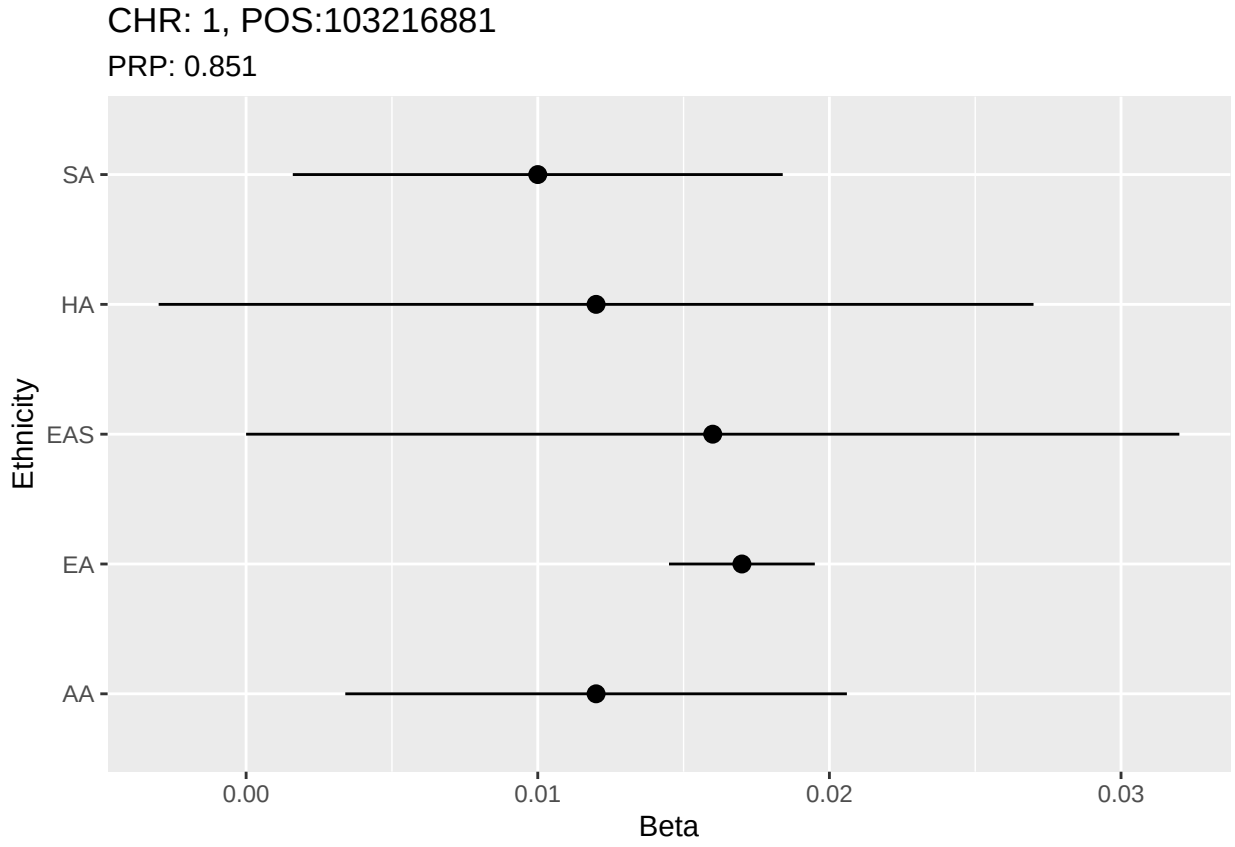
Now we'll look at how the effect sizes look between SNPs that MAMBA and PRP have deemed significant and not signficant. Below, we have a forest plot that shows the effect size estimates between various ethnic groups. We can see that generally, there is a lot of overlap in the effect sizes and their standard error ranges.



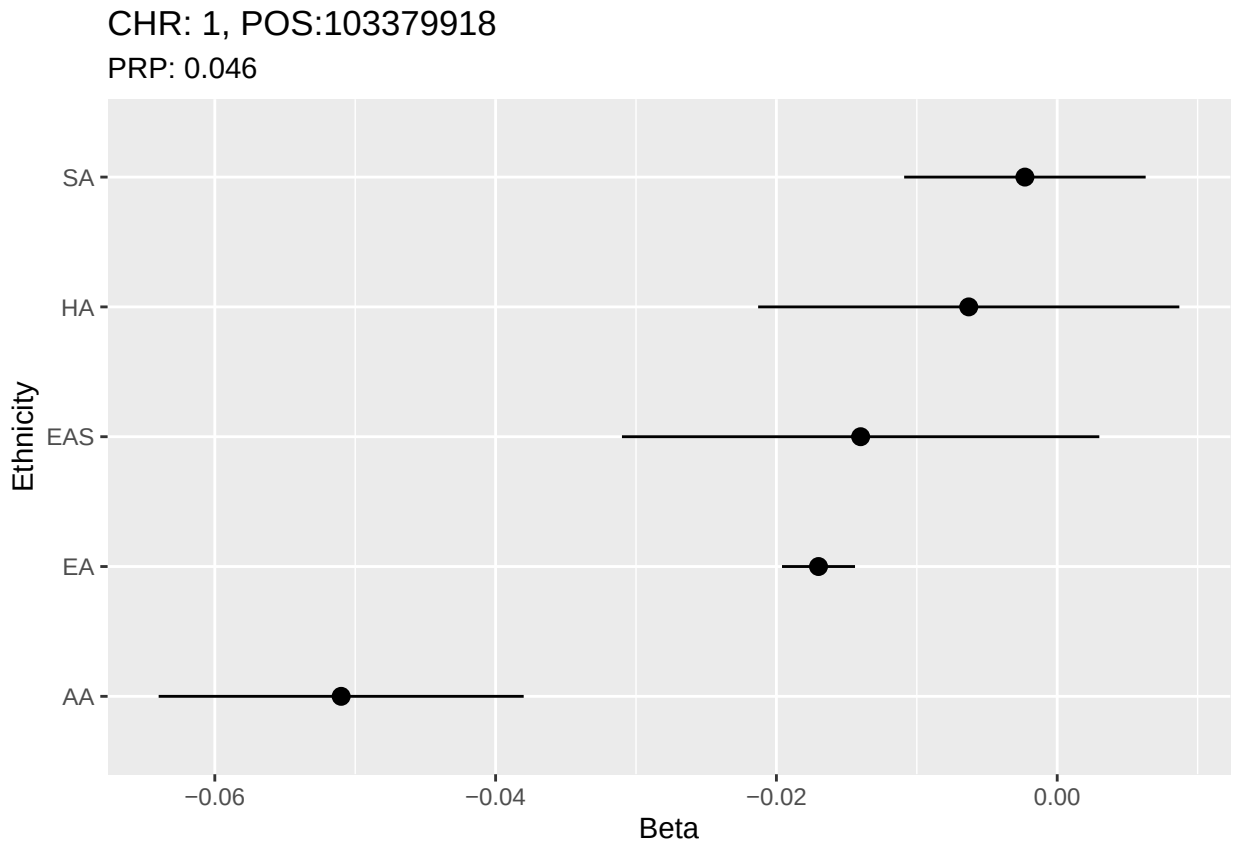
In contrast, below we have a SNP that the replicability packages have determined as not replicable. There is significantly less overlap between the estimated effect sizes and their standard error ranges for each ethnicity group. This generally matches what is expected from a nonreplicable SNP.



Since MAMBA seems to struggle a bit in determining replicability in simulations, we'll look at PRP replicability determined SNPs to get an idea of how it determines replicability. Below is a graph of a SNP that is replicable according to the PRP library. Like before, we can see that there is a lot of overlap in the effect size estimate.



Similarly, the nonreplicable SNP has less overlap.



A shiny app version of these plots is also provided that let's you compare different SNPs. It gives a better idea of how and why MAMBA and PRP classify each SNP's replicability.