

CARMA Tutorial

Zikun Yang

10/20/2022

Introduction

This document describes a complete walk through the usage of the package ‘CARMA’ with an application to computing the posterior inclusion probability (PIP) of variants at loci of interest. In this document, we will illustrate:

- Summary statistics and in-sample linkage disequilibrium (LD) matrix.
- Summary statistics and LD matrix extracted from external panels.

Environment preparation

Install the package ‘CARMA’.

```
devtools::install_github("ZikunY/CARMA")
library("CARMA")
```

Create a folder ‘CARMA’, then download the example datasets on OSF through command wget.

```
mkdir CARMA
cd CARMA
##### Download and save the demo data in folder `CARMA'
wget -O Sample_data.tar.gz https://osf.io/5gqz8/download
##### or download the file from https://osf.io/4t2bz/
tar -zxvf Sample_data.tar.gz
```

Install several packages needed to run CARMA.

```
pkgs = c("data.table", "magrittr", "dplyr", "devtools", "R.utlis")
pkgs.na = pkgs[!pkgs %in% installed.packages()[, "Package"]]
if (length(pkgs.na) > 0) {
  install.packages(pkgs.na)
}
if (!"CARMA" %in% installed.packages()[, "Package"]) {
  devtools::install_github("ZikunY/CARMA")
}
```

Individual level data

Simulating data

We simulate individual level data for the purpose of this demonstration. We use the R package ‘sim1000G’ (Dimitromanolakis et al. 2019) to simulate genotypes based on the 1000 Genomes Project data (phase 3, European population). The phenotype is simulated through

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{X} are the genotypes, β is a sparse coefficient vector such as $\beta_i \neq 0$ if the i th SNP is a causal SNP, and ϵ is the standard Gaussian error. Causal variants are selected according to a probability proportional to the linear predictor $\mathbf{w}_i'\boldsymbol{\theta}$, where \mathbf{w}_i is the vector of annotations associated with the i th SNP and $\boldsymbol{\theta}$ is the coefficients vector of the annotations.

Example at locus chr1: 200,937,832-201,937,832

In this section, we use the simulated data based on the locus chr1:200,937,832-201,937,832. We computed the summary statistics (Z-scores) and the in-sample LD matrix. The three causal SNPs are highlighted in the plot below (the 287th, 1275th, and 2572th SNPs at the locus; the left, middle, right red points respectively).

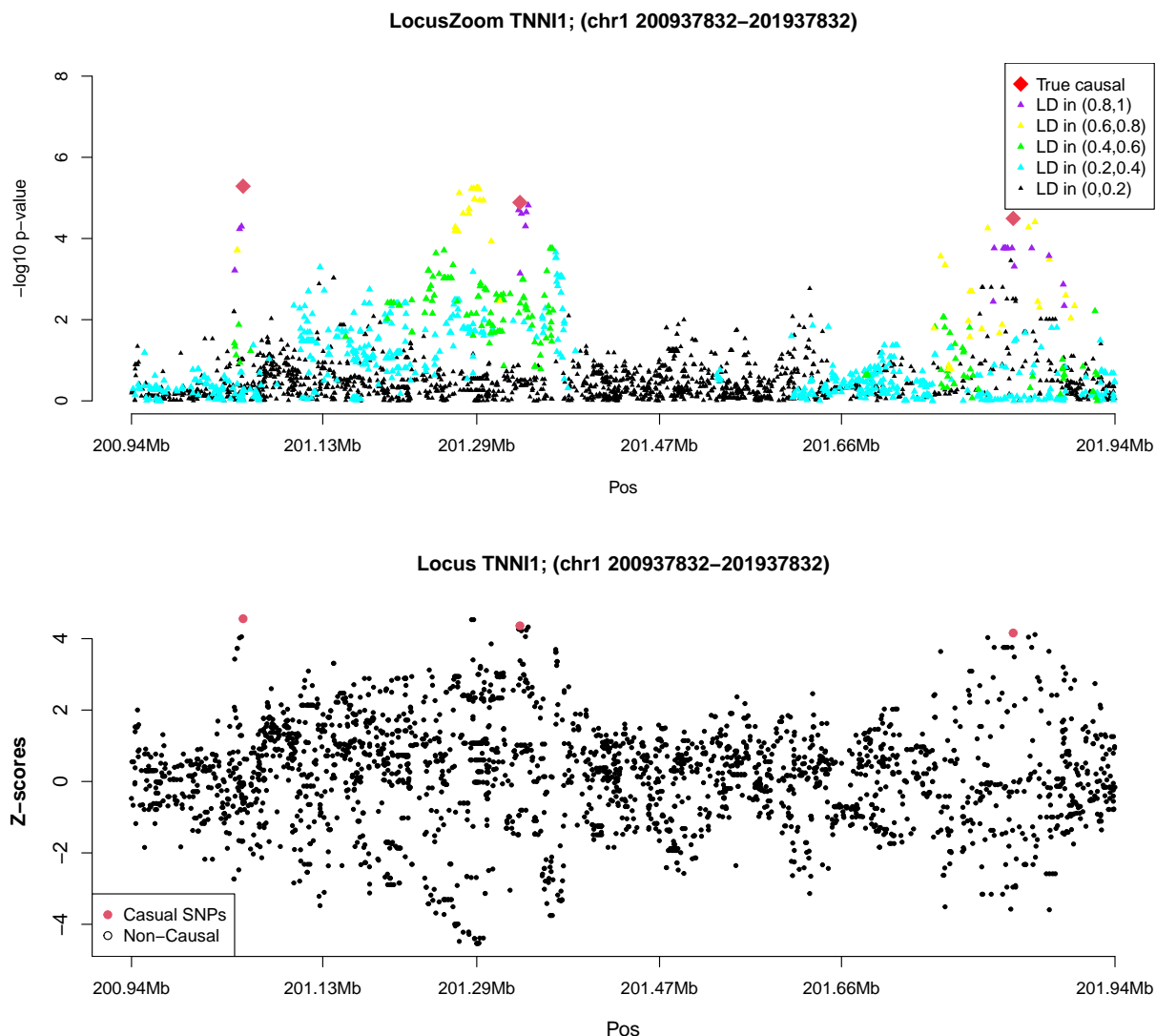


Figure 1: LocusZoom and Z-score plots for locus chr1:200,937,832-201,937,832

As shown in Figure 1, one of the causal SNPs has few highly correlated SNPs and larger Z-scores, whereas the other two SNPs are highly correlated to the surrounding SNPs with similar Z-scores.

Running CARMA without annotations. We run CARMA without annotations first. The input format of CARMA is the list class. We use the "CARMA_fixed_sigma" function in the package. As recommended

in the paper, we choose the dimensional hyperparameter $\eta = 1$, which can be interpreted as the prior expectation of the number of causal SNPs resided in the testing locus. Notice that without annotations, all SNPs have identical prior probabilities of being causal generated by the Poisson prior distribution, which assigns prior probability on the model size and provides false discovery control.

```
library(data.table)
library(magrittr)
library(dplyr)
library(devtools)
library(R.utils)
##### setting up the working directory or the wd where the data are stored
setwd('CARMA')
##### load the GWAS summary statistics
sumstat<- fread(file = "Sample_data/sumstats_chr1_200937832_201937832.txt.gz",
                 sep = "\t", header = T, check.names = F, data.table = F,
                 stringsAsFactors = F)
##### load the pair-wise LD matrix (assuming the variants are sorted in the same order
##### as the variants in sumstat file)
ld = fread(file = "Sample_data/sumstats_chr1_200937832_201937832_ld.txt.gz",
           sep = "\t", header = F, check.names = F, data.table = F,
           stringsAsFactors = F)

print(head(sumstat))
```

The input data ‘sumstat’ are typical results containing summary statistics, such as

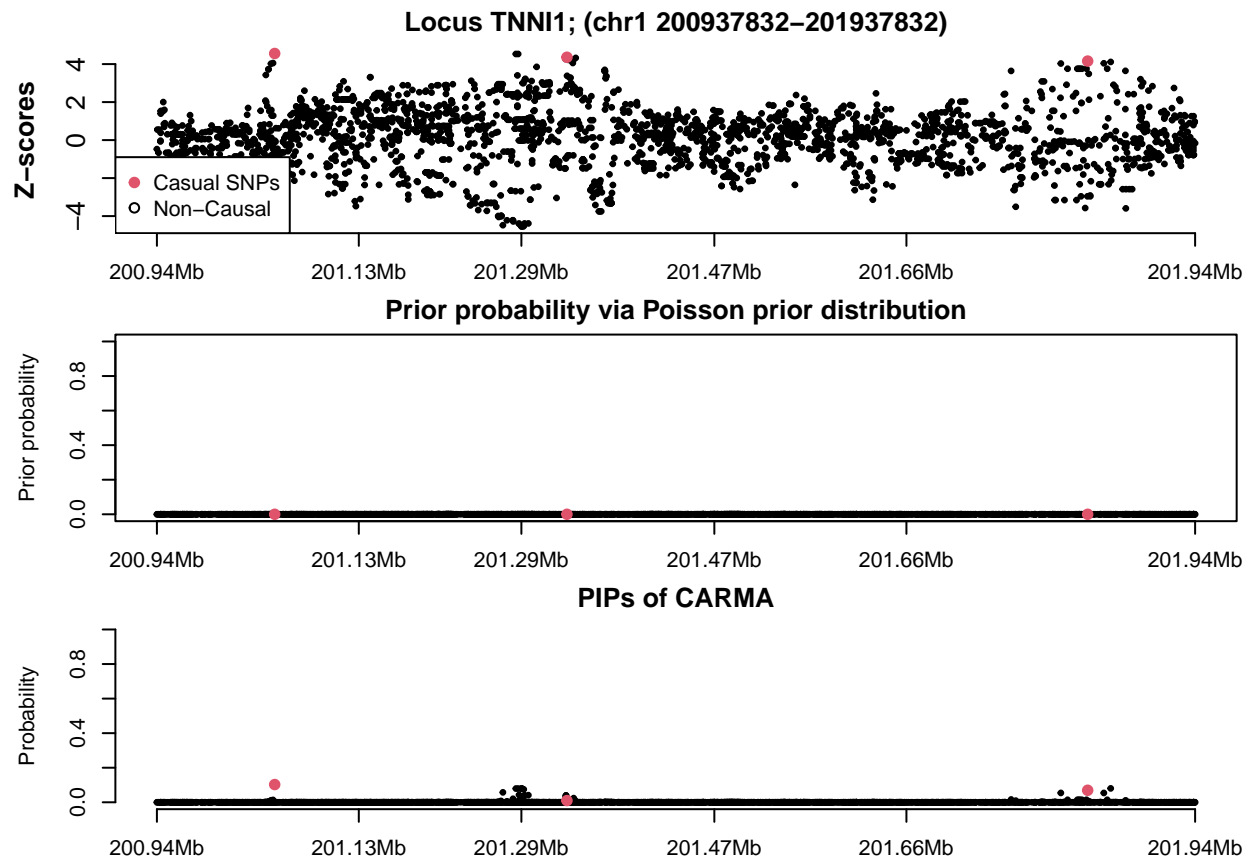
##	ID	CHR	POS	Ref	Alt	SNP	N	MAF	Z	Pval
## 1	1:200938029	G:A	1 200938029	G	A	rs10494829	10000	0.26	-0.475	1.37
## 2	1:200938474	G:T	1 200938474	G	T	rs4915210	10000	0.29	0.550	0.58
## 3	1:200939642	C:T	1 200939642	C	T	rs3208703	10000	0.14	-0.666	1.49
## 4	1:200940180	G:A	1 200940180	G	A	rs3198583	10000	0.29	0.550	0.58
## 5	1:200941423	C:A	1 200941423	C	A	rs56368827	10000	0.26	-0.071	1.06
## 6	1:200941549	T:C	1 200941549	T	C	rs296570	10000	0.92	-0.031	1.02

Next, we run CARMA with the input summary statistics Z and the LD matrix with the hyperparameter $\eta = 1$ as the default setting. Notice that given that the LD matrix is in-sample, we do not turn on the outlier detection ‘outlier.switch=F’.

```
z.list<-list()
ld.list<-list()
lambda.list<-list()
z.list[[1]]<-sumstat$Z
ld.list[[1]]<-as.matrix(ld)
lambda.list[[1]]<-1
CARMA.results<-CARMA_fixed_sigma(z.list,ld.list,lambda.list=lambda.list,
                                outlier.switch=F)
##### Posterior inclusion probability (PIP) and credible set (CS)
sumstat.result = sumstat %>% mutate(PIP = CARMA.results[[1]]$PIPs, CS = 0)
if(length(CARMA.results[[1]]$`Credible set`[[2]])!=0){
  for(1 in 1:length(CARMA.results[[1]]$`Credible set`[[2]])){
    sumstat.result$CS[CARMA.results[[1]]$`Credible set`[[2]][[1]]]=1
  }
}
##### write the GWAS summary statistics with PIP and CS
fwrite(x = sumstat.result,
       file = "Sample_data/sumstats_chr1_200937832_201937832_carma.txt.gz",
```

```
sep = "\t", quote = F, na = "NA", row.names = F, col.names = T,
compress = "gzip")
```

We can check the results.



##	SNPs.index	Causal.status	Z.scores	PIPs
##	287	True causal	4.6	0.103
##	2603	Non-causal	4.1	0.080
##	1095	Non-causal	4.5	0.079
##	1105	Non-causal	4.5	0.079
##	1120	Non-causal	-4.5	0.079
##	1121	Non-causal	-4.5	0.079
##	1125	Non-causal	-4.5	0.079
##	1134	Non-causal	-4.5	0.075
##	2572	True causal	4.2	0.069
##	1057	Non-causal	-4.5	0.057
##	2530	Non-causal	4.0	0.054
##	2591	Non-causal	4.0	0.054
##	1110	Non-causal	-4.4	0.043
##	1141	Non-causal	-4.4	0.041
##	1158	Non-causal	-4.4	0.041
##	1272	Non-causal	4.4	0.040
##	1273	Non-causal	4.4	0.040
##	1268	Non-causal	4.3	0.027
##	1298	Non-causal	4.2	0.024
##	1131	Non-causal	-4.5	0.021

We can observe that the 287th SNP (the causal SNP at left) received a larger PIP value to the surrounding SNPs. On the other hand, the other two causal SNPs, which are highly correlated to surrounding SNPs with similar Z-scores, shared the PIPs with the highly correlated SNPs. We can also check the credible sets and credible models.

```
CARMA.results[[1]]$`Credible set`[[2]]
```

```
## list()
```

```
CARMA.results[[1]]$`Credible model`[[3]]
```

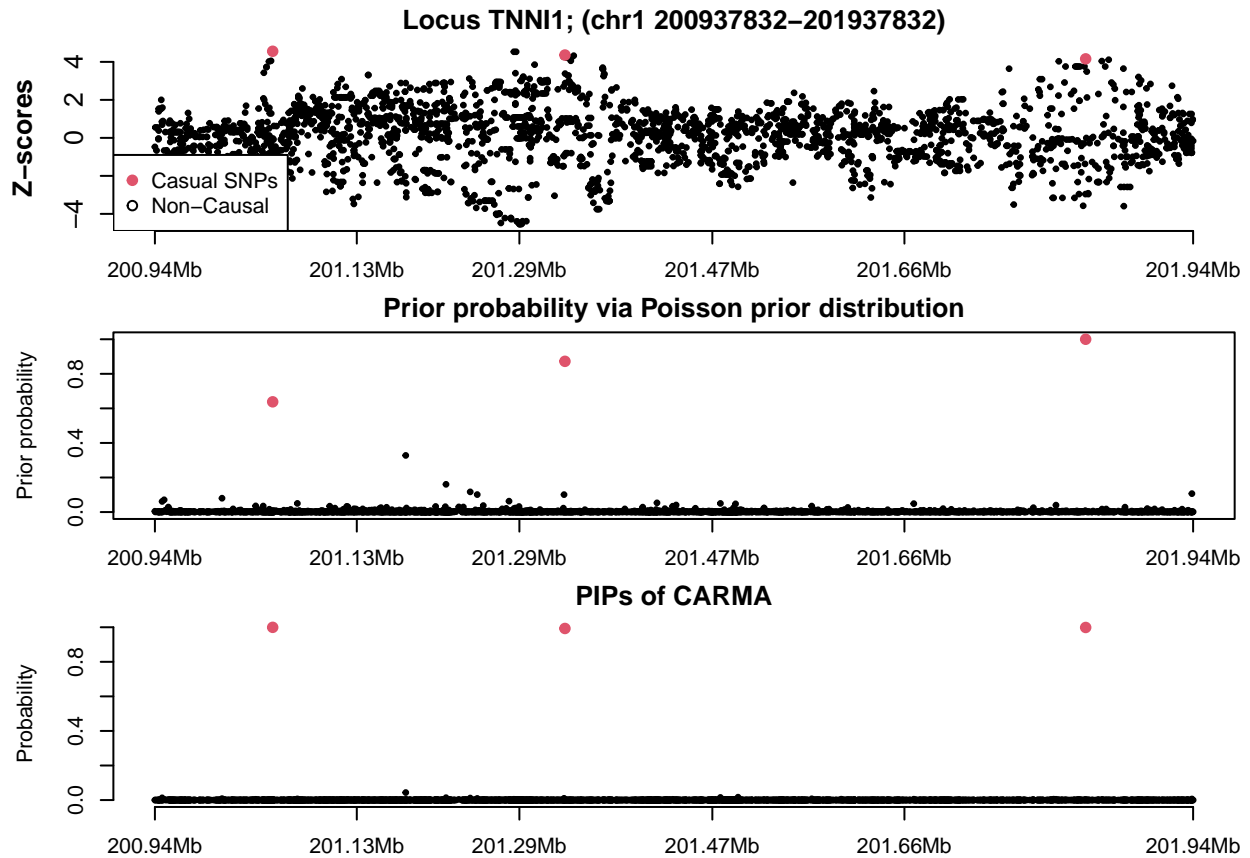
```
## [1] 287 1120 1121 1125 1131 1095 1105 1134 1057 1109 1110 1141 1158 1272 1273
## [16] 1275 1304 2603 1086 1268 2572 1298 1063 1082 1084 1281 2530 2591 1294 269
## [31] 1043 257 1041 1048 1050 1051 1053
```

Due to relatively weak signal strength, none of the signals have enough PIPs to formulate credible sets. On the other hand, credible model still identified 22 candidate SNPs, which include all three true causal SNPs.

Running CARMA with annotations We can include functional annotations into CARMA:

```
##### load the functional annotations for the variants included in GWAS summary
##### statistics (assuming the variants are sorted in the same order as the
##### variants in sumstat file)
annot=fread(file = "Sample_data/sumstats_chr1_200937832_201937832_annotations.txt.gz",
            sep="\t", header = T, check.names = F, data.table = F,
            stringsAsFactors = F)
##### z.list and ld.list stay the same with the previous setting,
##### and we add annotations this time.
annot.list<-list()
annot.list[[1]]<-annot
CARMA.results<-CARMA_fixed_sigma(z.list,ld.list,lambda.list=lambda.list,w.list=annot.list,
                                outlier.switch=F)
##### Posterior inclusion probability (PIP) and credible set (CS)
sumstat.result = sumstat %>% mutate(PIP = CARMA.results[[1]]$PIPs, CS = 0)
if(length(CARMA.results[[1]]$`Credible set`[[2]])!=0){
  for(l in 1:length(CARMA.results[[1]]$`Credible set`[[2]])){
    sumstat.result$CS[CARMA.results[[1]]$`Credible set`[[2]][[l]]]=1
  }
}
##### write the GWAS summary statistics with PIP and CS
fwrite(x = sumstat.result,
      file = "Sample_data/sumstats_chr1_200937832_201937832_carma_annot.txt.gz",
      sep = "\t", quote = F, na = "NA", row.names = F, col.names = T,
      compress = "gzip")
```

We can first check the resulting PIPs. This time, we also include the prior probability of a variant being causal estimated by CARMA.



##	SNPs.index	Causal.status	Z.scores	PIPs
## 287	287	True causal	4.56	0.9996
## 2572	2572	True causal	4.16	0.9988
## 1275	1275	True causal	4.36	0.9930
## 724	724	Non-causal	0.38	0.0436
## 1827	1827	Non-causal	-2.58	0.0166
## 1709	1709	Non-causal	0.70	0.0152
## 898	898	Non-causal	0.47	0.0138
## 16	16	Non-causal	0.69	0.0129
## 967	967	Non-causal	1.14	0.0118
## 1273	1273	Non-causal	4.36	0.0101
## 985	985	Non-causal	1.22	0.0094
## 142	142	Non-causal	-0.28	0.0083
## 2018	2018	Non-causal	-0.50	0.0072
## 363	363	Non-causal	-0.25	0.0066
## 2796	2796	Non-causal	-0.77	0.0062
## 2083	2083	Non-causal	1.68	0.0051
## 2521	2521	Non-causal	-3.16	0.0051
## 2249	2249	Non-causal	-1.23	0.0049
## 1095	1095	Non-causal	4.53	0.0046
## 2523	2523	Non-causal	-1.63	0.0044

As observed from the figure above, the inclusion of annotations helps CARMA distinguish the true causal variants from the highly correlated SNPs, such as the 1275th and 2572th SNP, which in the absence of functional annotations cannot be distinguished from other highly correlated SNPs. Also, the 287th SNP receives a larger PIP this time. We can also examine the credible sets and credible models of CARMA.

```
CARMA.results[[1]]$`Credible set`[[2]]
```

```
## [[1]]
## [1] 287
##
## [[2]]
## [1] 2572
##
## [[3]]
## [1] 1275
```

```
CARMA.results[[1]]$`Credible model`[[3]]
```

```
## [1] 287 1275 2572
```

The number of SNPs in credible models has been reduced significantly. Also, the credible sets identified the three true causals.

Summary statistics with LD matrix extracted from reference panels

Usually, individual level data are not available in meta-analysis GWAS studies. Instead, summary statistics are made available and an external LD matrix is used. These complex meta-analysis settings create inconsistencies between summary statistics and LD values which can lead to biased PIP values.

We use summary statistics from a meta-analysis for Alzheimer's disease (AD) (Jansen et al. 2019). The meta-analysis of AD is based on clinically diagnosed AD and AD-by-proxy with 71,880 cases and 383,378 controls of European ancestry. The clinically diagnosed AD case-control data are from 3 consortia (PGC-ALZ, IGAP, and ADSP), and the AD-by-proxy data are based on 376,113 individuals of European ancestry from UK BioBank (UKBB). We use the LD matrix extracted from the UKBB. For the CARMA model, we include 187 annotations provided by PolyFun plus PolyFun prior causal probability (Weissbrod et al. 2020).

Demonstration with loci ADAMTS4 and CR1

We illustrate CARMA on two loci, ADAMTS4 and CR1 on chromosome 1. We extract data at locus ADAMTS4/CR1, and extract the corresponding LD matrices from the UKBB (provided by PolyFun).

Sample of data at the locus ADAMTS4

##	uniqID.a1a2	CHR	BP	A1	A2	SNP	Z	P	Nsum	Neff
## 1	1:160656603_A_T	1	160656603	A	T	rs6702441	0.15	0.881	429975	423497
## 2	1:160657127_T_C	1	160657127	T	C	rs143426473	0.81	0.417	435185	428660
## 3	1:160657137_G_A	1	160657137	G	A	rs11589131	0.05	0.960	429757	423281
## 4	1:160657197_C_G	1	160657197	C	G	rs10908797	1.75	0.079	377075	375815
## 5	1:160657356_C_T	1	160657356	C	T	rs145169682	-0.28	0.776	17477	17477
## 6	1:160658364_G_A	1	160658364	G	A	rs7539434	0.24	0.814	380902	379635

##	dir	EAF	BETA	SE
## 1	?+--+	0.370	0.00034	0.0023
## 2	?+++	0.034	0.00486	0.0060
## 3	?+--+	0.369	0.00011	0.0023
## 4	??++	0.097	0.00685	0.0039
## 5	???-	0.012	-0.01407	0.0495
## 6	??++	0.028	0.00163	0.0069

Sample of data at the locus CR1

##	uniqID.a1a2	CHR	BP	A1	A2	SNP	Z	P	Nsum	Neff
## 1	1:207287187_T_C	1	207287187	T	C	rs2808470	0.755	0.45	433909	427395
## 2	1:207288258_C_T	1	207288258	C	T	rs147553990	-0.627	0.53	364527	364527

```
## 3 1:207288297_T_C 1 207288297 T C rs17020983 1.136 0.26 434723 428202
## 4 1:207288309_T_G 1 207288309 T G rs79498904 0.879 0.38 364051 364051
## 5 1:207288392_G_A 1 207288392 G A rs17020993 1.108 0.27 436498 429961
## 6 1:207288897_T_C 1 207288897 T C rs12031629 0.094 0.93 71639 71639
## dir EAF BETA SE
## 1 ?-++ 0.1940 0.0021 0.0027
## 2 ??-? 0.0033 -0.0127 0.0203
## 3 ?-++ 0.0960 0.0042 0.0037
## 4 ??+? 0.0128 0.0092 0.0104
## 5 ?+++ 0.1571 0.0033 0.0030
## 6 ?-?+ 0.4299 0.0005 0.0053
```

Notice that the sample size values in the column “Nsum” can vary from 9,703 to 444,006.

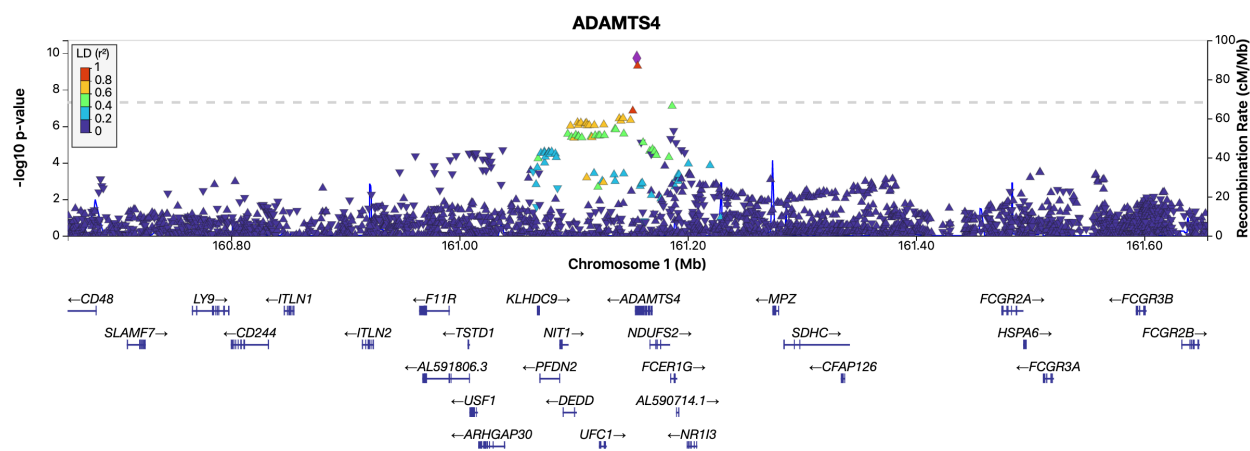


Figure 2: LocusZoom plot for ADAMTS4.

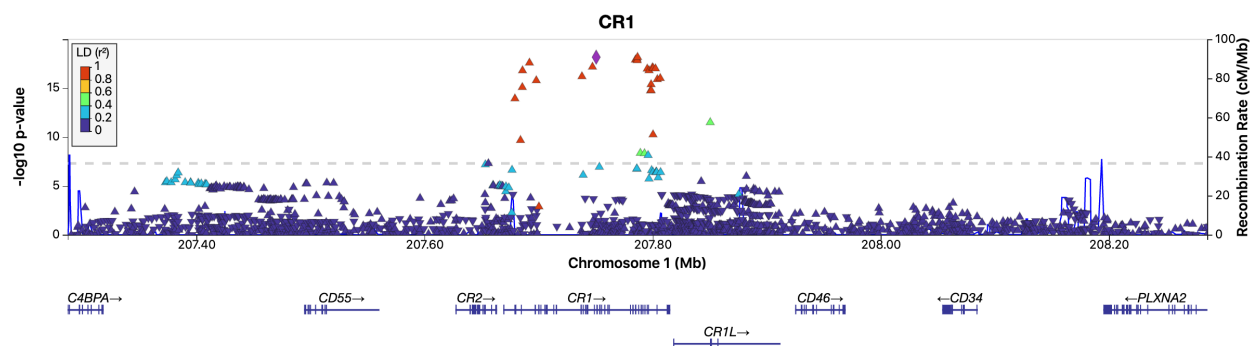


Figure 3: LocusZoom plot for CR1.

Next we run CARMA with two settings: 1. without annotations, and 2. with annotations as described above. We use the function “CARMA_fixed_sigma” to run CARMA with the external LD and $\eta = 1$ as the default setting. Given by that the LD matrix is extracted from reference panel (UKBB) instead of in-sample LD, we turn on the outlier detection setting ‘outlier.switch=TRUE’.


```

##### load the GWAS summary statistics (part of AD GWAS sumstats from Jansen et al., 2019)
sumstat.1 = fread(file = "Sample_data/ADAMTS4_sumstats.txt.gz",
                  sep = "\t", header = T, check.names = F, data.table = F,
                  stringsAsFactors = F)
sumstat.2 = fread(file = "Sample_data/CR1_sumstats.txt.gz",
                  sep = "\t", header = T, check.names = F, data.table = F,
                  stringsAsFactors = F)

##### load the functional annotations for the variants included in
##### GWAS summary statistics (assuming the variants are sorted in
##### the same order as the variants in sumstat file)
annot.1 = fread(file = "Sample_data/ADAMTS4_annotations.txt.gz",
                sep = "\t", header = T, check.names = F, data.table = F,
                stringsAsFactors = F)
annot.2 = fread(file = "Sample_data/CR1_annotations.txt.gz",
                sep = "\t", header = T, check.names = F, data.table = F,
                stringsAsFactors = F)

##### load the pair-wise LD matrix (assuming the variants are sorted in
##### the same order as the variants in sumstat file)
ld.1 = fread(file = "Sample_data/ADAMTS4_ld.txt.gz",
             sep = "\t", header = F, check.names = F, data.table = F,
             stringsAsFactors = F)
ld.2 = fread(file = "Sample_data/CR1_ld.txt.gz",
             sep = "\t", header = F, check.names = F, data.table = F,
             stringsAsFactors = F)

z.list<-list()
ld.list<-list()
lambda.list<-list()
z.list[[1]]<-sumstat.1$Z
z.list[[2]]<-sumstat.2$Z
ld.list[[1]]<-as.matrix(ld.1)
ld.list[[2]]<-as.matrix(ld.2)
lambda.list[[1]]<-1
lambda.list[[2]]<-1
##### Without annotations
CARMA.results_no_annot<-CARMA_fixed_sigma(z.list,ld.list,lambda.list = lambda.list,
                                          outlier.switch=T)

##### With annotations
##### Exclude the variant information columns in annotation file
##### such as positions and REF/ALT alleles.
annot.list<-list()
annot.list[[1]]<-as.matrix(cbind(1, annot.1 %>% select(-(uniqID.a1a2:SNP))))
annot.list[[2]]<-as.matrix(cbind(1, annot.2 %>% select(-(uniqID.a1a2:SNP))))

CARMA.results_annot<-CARMA_fixed_sigma(z.list,ld.list,w.list=annot.list,
                                       lambda.list = lambda.list,
                                       input.alpha=0, outlier.switch=T)

##### Posterior inclusion probability (PIP) and credible set (CS)
sumstat.1 = sumstat.1 %>% mutate(PIP = CARMA.results_annot[[1]]$PIPs, CS = 0)

```

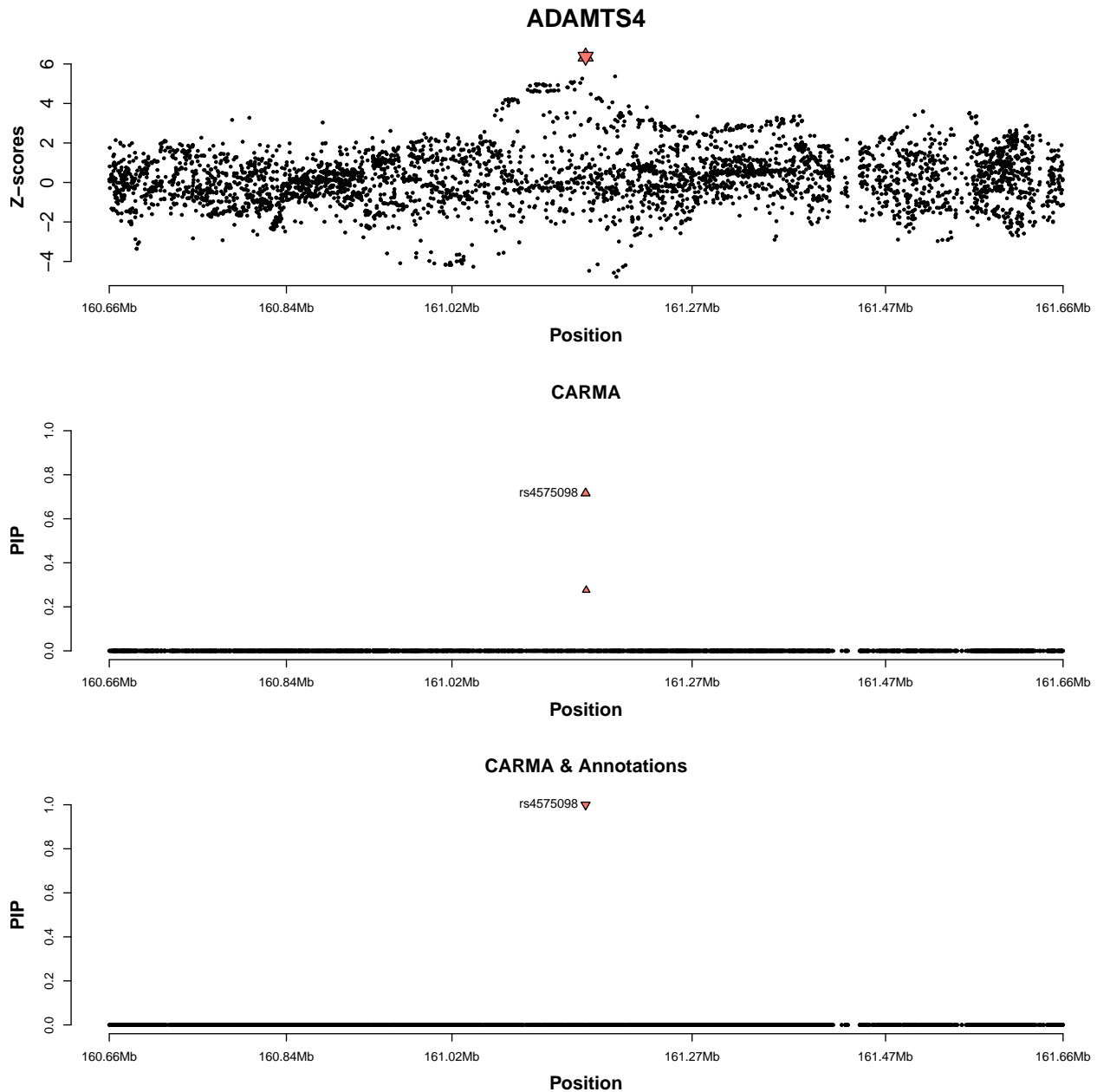
```

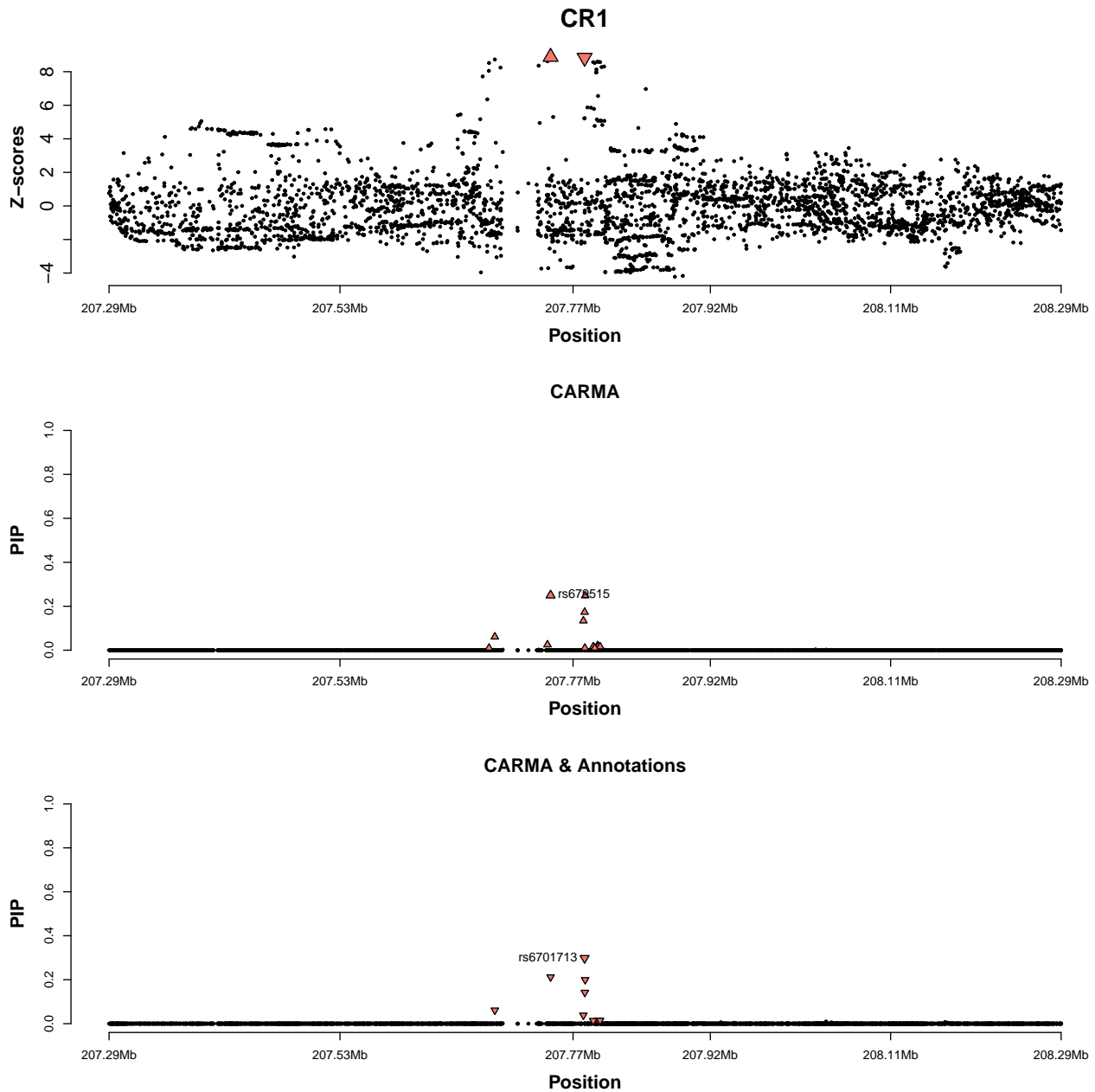
sumstat.1$CS[CARMA.results_annot[[1]]$`Credibile set`[[2]][[1]]] = 1
sumstat.2 = sumstat.2 %>% mutate(PIP = CARMA.results_annot[[2]]$PIPs, CS = 0)
sumstat.2$CS[CARMA.results_annot[[2]]$`Credibile set`[[2]][[1]]] = 1

##### write the GWAS summary statistics with PIP and CS
fwrite(x = sumstat.1, file = "Sample_data/ADAMTS4_carma.txt.gz",
       sep = "\t", quote = F, na = "NA", row.names = F, col.names = T, compress = "gzip")
fwrite(x = sumstat.2, file = "Sample_data/CR1_carma.txt.gz",
       sep = "\t", quote = F, na = "NA", row.names = F, col.names = T, compress = "gzip")

```

First, we examine the PIPs estimated by CARMA.





In the figure above, the credible sets are highlighted by colored shapes. Next, we can examine the SNPs included in the credible sets. For simplicity we only show the credible sets when including functional annotations.

```
carma_annot[[1]]$`Credibile set` #the first element of the list
                                #is the result of the first locus ADAMTS4
```

```
## [1] "This is the first credible set of the locus ADAMTS4"
##      CHR      BP A1 A2      SNP    Z PIPs
## 2184    1 161155392  A  G rs4575098 6.4    1
```

```
carma_annot[[2]]$`Credibile set` #the second element of the list
                                #is the result of the second locus CR1
```

```
## [1] "This is the first credible set of the locus CR1"
```

##	CHR	BP	A1	A2	SNP	Z	PIPs
## 1563	1	207786289	A	G	rs6701713	8.8	0.298
## 1426	1	207750568	T	C	rs679515	8.9	0.212
## 1570	1	207786828	A	G	rs2093760	8.9	0.200
## 1567	1	207786542	A	G	rs2093761	8.8	0.142
## 1343	1	207692049	A	G	rs6656401	8.7	0.061
## 1559	1	207784968	A	G	rs3818361	8.8	0.038
## 1626	1	207800555	T	C	rs1408078	8.6	0.016
## 1629	1	207802552	A	C	rs4844610	8.6	0.015
## 1598	1	207795320	A	G	rs2296160	8.6	0.014

We can also examine the credible models.

```
carma_annot[[1]]$`Credible model`#the credible model of locus ADAMTS4
```

```
## [1] "This is the credible model of the locus ADAMTS4"
```

##	CHR	BP	A1	A2	SNP	Z	PIPs
## 2184	1	161155392	A	G	rs4575098	6.4	1

```
carma_annot[[2]]$`Credible model`#the credible model of locus CR1
```

```
## [1] "This is the credible model of the locus CR1"
```

##	CHR	BP	A1	A2	SNP	Z	PIPs
## 1563	1	207786289	A	G	rs6701713	8.8	0.298
## 1426	1	207750568	T	C	rs679515	8.9	0.212
## 1570	1	207786828	A	G	rs2093760	8.9	0.200
## 1567	1	207786542	A	G	rs2093761	8.8	0.142
## 1343	1	207692049	A	G	rs6656401	8.7	0.061

References

- Dimitromanolakis, Apostolos, Jingxiong Xu, Agnieszka Krol, and Laurent Briollais. 2019. "sim1000G: A User-Friendly Genetic Variant Simulator in r for Unrelated Individuals and Family-Based Designs." *BMC Bioinformatics* 20 (1): 26.
- Jansen, Iris E, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, et al. 2019. "Genome-Wide Meta-Analysis Identifies New Loci and Functional Pathways Influencing Alzheimer's Disease Risk." *Nature Genetics* 51 (3): 404–13.
- Weissbrod, Omer, Farhad Hormozdiari, Christian Benner, Ran Cui, Jacob Ulirsch, Steven Gazal, Armin P Schoech, et al. 2020. "Functionally Informed Fine-Mapping and Polygenic Localization of Complex Trait Heritability." *Nature Genetics*, 1–9.