

CARMA Tutorial

Zikun Yang

4/04/2022

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 being causal. In this document, we will illustrate typical fine-mapping studies with two types of datasets:

- Summary statistics based on individual level phenotype and genotype data, and in-sample linkage disequilibrium (LD) matrix.
- Summary statistics generated by meta-analysis, and LD matrix extracted from reference panels.

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

Also, open terminal and download the example datasets from GitHub repository ZikunY/CARMA.

```
git clone https://github.com/ZikunY/CARMA.git
```

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 a Gaussian regression model with the simulated genotypes \mathbf{X} :

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

where $\boldsymbol{\beta}$ is a sparse coefficient vector such as $\beta_i \neq 0$ if the i th SNP is a causal SNP, and $\boldsymbol{\epsilon}$ is the standard Gaussian error. The probability of a variant being causal is decided through a logistic regression model such that

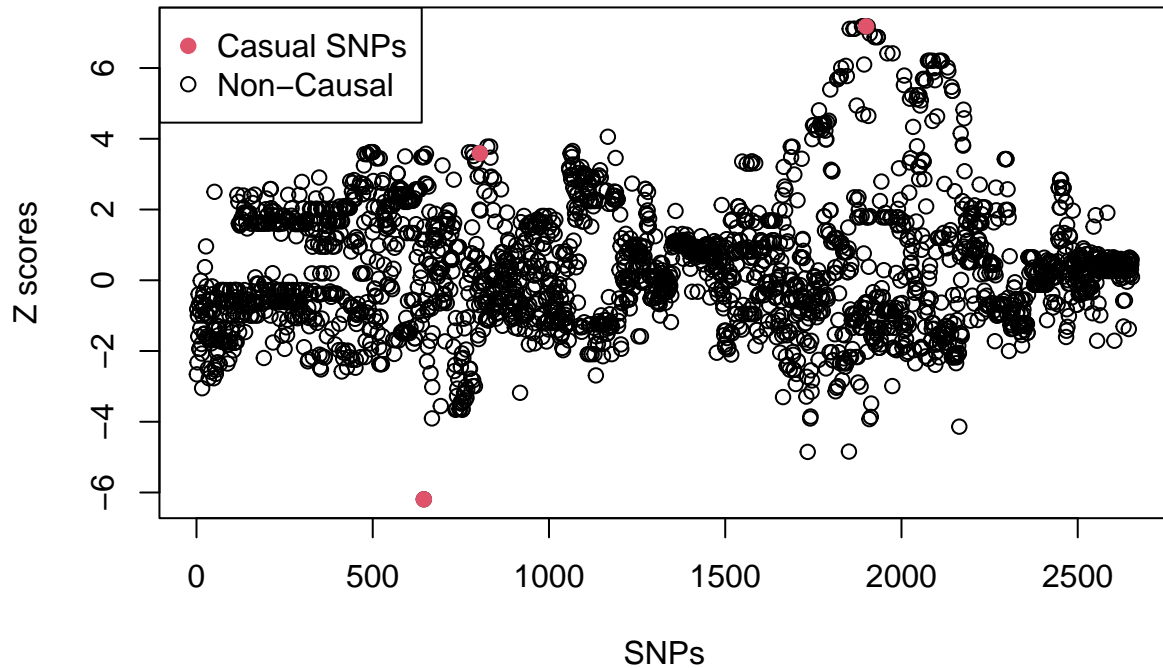
$$\Pr\{\beta_i \neq 0 | \mathbf{w}_i, \boldsymbol{\theta}\} = \frac{1}{1 + \exp\{-\mathbf{w}_i' \boldsymbol{\theta}\}},$$

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 of locus 128952507-129961171 on chromosome 11

In this section, we use the simulated data based on the locus chr11:128952507-129961171. We computed the summary statistics (Z-scores) and the LD matrix. The pre-determined causal SNPs are the **645**th, **804**th, and **1900**th SNPs at the locus.

Locus: chr11 128952507–129961171



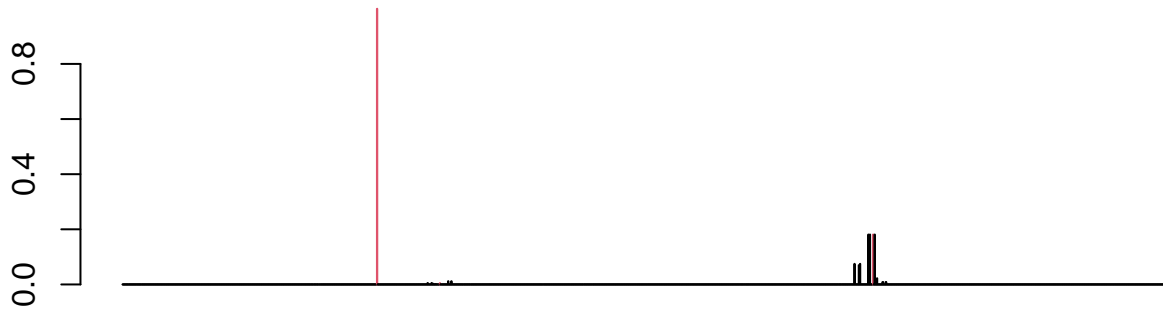
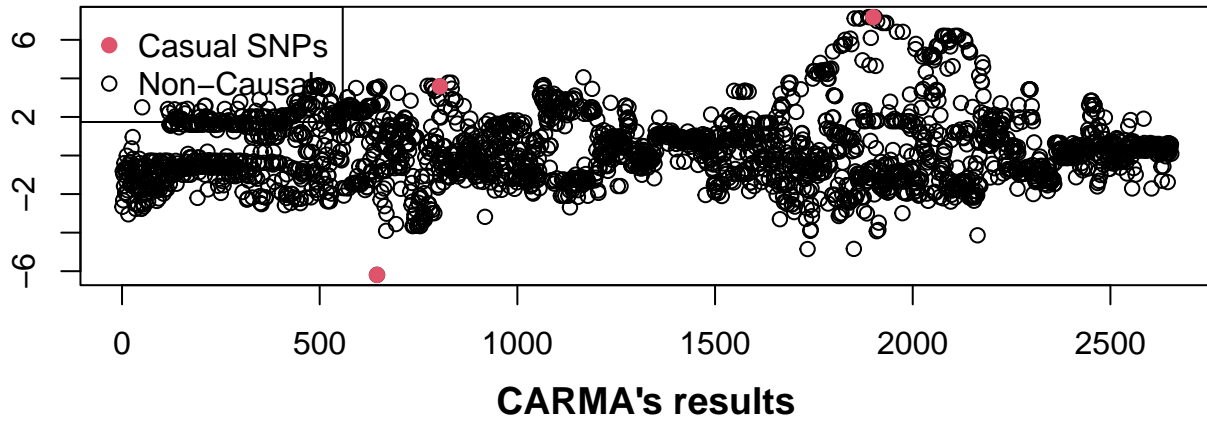
As shown in the figure, one of the causal SNPs is relatively independent of the other SNPs, whereas the other two SNPs are highly correlated to the surrounding SNPs with similar values of Z-scores.

Running CARMA without annotations We run CARMA without annotations first. The input format of CARMA is the list class. We use the “CARMA” function in the package, which is designed to run in-sample data. As recommended in the paper, we choose the dimensional hyperparameter η as $1/\sqrt{p}$, where p is the total number of SNPs at the locus.

```
setwd('Your Working Dir.') ### setting up the working directory
data<-readRDS('in-sample_data.RData')
z.list<-list()
ld.list<-list()
lambda.list<-list()
z.list[[1]]<-data$Z
ld.list[[1]]<-data$LD
lambda.list[[1]]<-1/sqrt(nrow(ld.list[[1]]))
CARMA.results<-CARMA(z.list,ld.list,lambda.list=lambda.list)
```

We can check the results.

Locus: chr11 128952507–129961171



##	SNPs.index	Causal.status	Z.scores	PIPs
## 645	645	True causal	-6.191607	0.999998111
## 1888	1888	Non-causal	7.178537	0.181027464
## 1892	1892	Non-causal	7.178537	0.181027464
## 1900	1900	True causal	7.178537	0.181027464
## 1904	1904	Non-causal	7.178537	0.181027464
## 1853	1853	Non-causal	7.108896	0.074692442
## 1867	1867	Non-causal	7.108896	0.074692442
## 1864	1864	Non-causal	7.108896	0.069958505
## 1910	1910	Non-causal	6.972467	0.022387332
## 825	825	Non-causal	3.778803	0.011750509
## 833	833	Non-causal	3.778803	0.011750509
## 1924	1924	Non-causal	6.874909	0.008507082
## 1925	1925	Non-causal	6.874909	0.008507082
## 1932	1932	Non-causal	6.874909	0.008507082
## 1933	1933	Non-causal	6.874909	0.008507082
## 773	773	Non-causal	3.616948	0.005337156
## 783	783	Non-causal	3.616948	0.005337156
## 804	804	True causal	3.590660	0.004749275
## 786	786	Non-causal	3.513737	0.002994474
## 834	834	Non-causal	3.465383	0.001743837

We can observe that the 645th SNP, which is a true causal SNP and fairly independent of other SNPs, received a large PIP value. On the other hand, the 1900th SNP, which is also a true causal SNP and highly correlated to other surrounding SNPs, shared the PIP with the highly correlated SNPs. We can also check the credible sets and credible models.

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

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

```
## [1] 645
##
## [[2]]
## [1] 1888 1892 1900 1904 1853 1867 1864 1910 1924 1925 1932
```

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

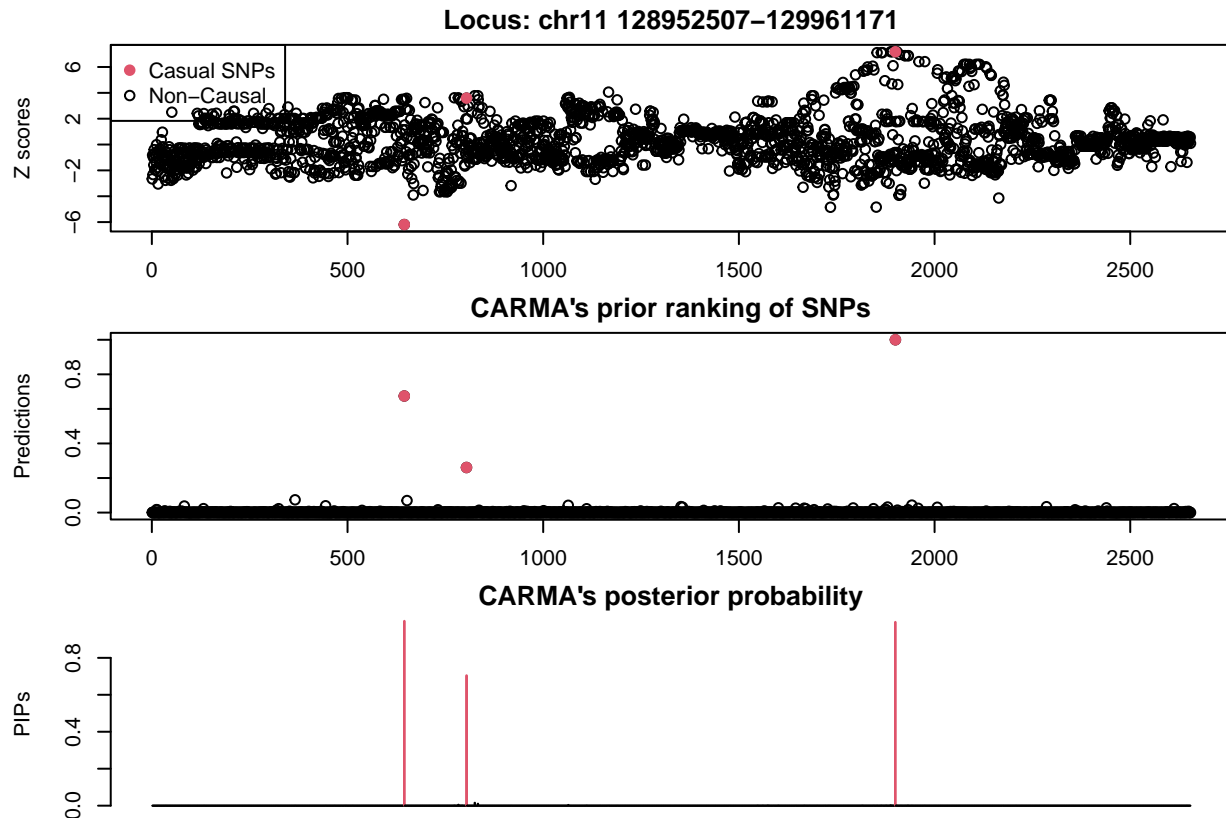
```
## [1] 645 1892 1888 1900 1904 1853 1864 1867 1910
```

Given the threshold for the credible set $\rho = 0.99$, the 645th SNP formulate a credible set with single SNP. The second credible set, which include the true causal SNP (the 1900th SNP), includes 11 SNPs with a minimum LD 0.894 among the SNPs. The number of SNPs identified by the credible model is 9, which is smaller than for the credible set.

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

```
data<-readRDS('in-sample_data.RData')
z.list<-list()
ld.list<-list()
lambda.list<-list()
annot.list<-list()
z.list[[1]]<-data$Z
ld.list[[1]]<-data$LD
lambda.list[[1]]<-1/sqrt(nrow(ld.list[[1]]))
annot.list[[1]]<-data$Annotations
CARMA.results<-CARMA(z.list,ld.list,lambda.list=lambda.list,w.list=annot.list)
```

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



##	SNPs.index	Causal.status	Z.scores	PIPs
## 645	645	True causal	-6.191607	0.9999999961
## 1900	1900	True causal	7.178537	0.9943640877
## 804	804	True causal	3.590660	0.7056123753
## 825	825	Non-causal	3.778803	0.0159377924
## 833	833	Non-causal	3.778803	0.0106795049
## 783	783	Non-causal	3.616948	0.0051777324
## 1064	1064	Non-causal	3.594629	0.0040586959
## 1888	1888	Non-causal	7.178537	0.0023769107
## 786	786	Non-causal	3.513737	0.0011411282
## 834	834	Non-causal	3.465383	0.0010816345
## 1574	1574	Non-causal	3.356648	0.0010565376
## 773	773	Non-causal	3.616948	0.0010400024
## 1932	1932	Non-causal	6.874909	0.0008463874
## 1548	1548	Non-causal	3.356648	0.0007822398
## 1877	1877	Non-causal	-2.890220	0.0007598126
## 1111	1111	Non-causal	2.535731	0.0006876151
## 1892	1892	Non-causal	7.178537	0.0006706814
## 1601	1601	Non-causal	1.133419	0.0006540633
## 793	793	Non-causal	3.389752	0.0006520183
## 1904	1904	Non-causal	7.178537	0.0005534825

As observed from the figure above, the prior helps CARMA distinguish the true causal variants from the highly correlated SNPs, such as the 1900th SNP which in the absence of functional annotations cannot be distinguished from other highly correlated SNPs. Also, the 804th SNP which was missed before receives a high PIP this time. We can also examine the credible sets and credible models of CARMA.

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

```
## [[1]]
## [1] 645
##
## [[2]]
## [1] 1900
```

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

```
## [1] 645 804 1900
```

The numbers of included SNPs in both credible sets and credible models have been reduced significantly, with the top candidate model successfully identifying the three causal SNPs.

Summary statistics and LD matrix extracted from reference panels

Usually, individual level data are not available in large 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 924 functional annotations including DeepSEA (Zhou and Troyanskaya 2015), CADD (Kircher et al. 2014), PO-EN (Yang et al. 2021), and PolyFun (Weissbrod et al. 2020).

Demonstration with the loci ADAMTS4 and CR1

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

Data of the locus ADAMTS4

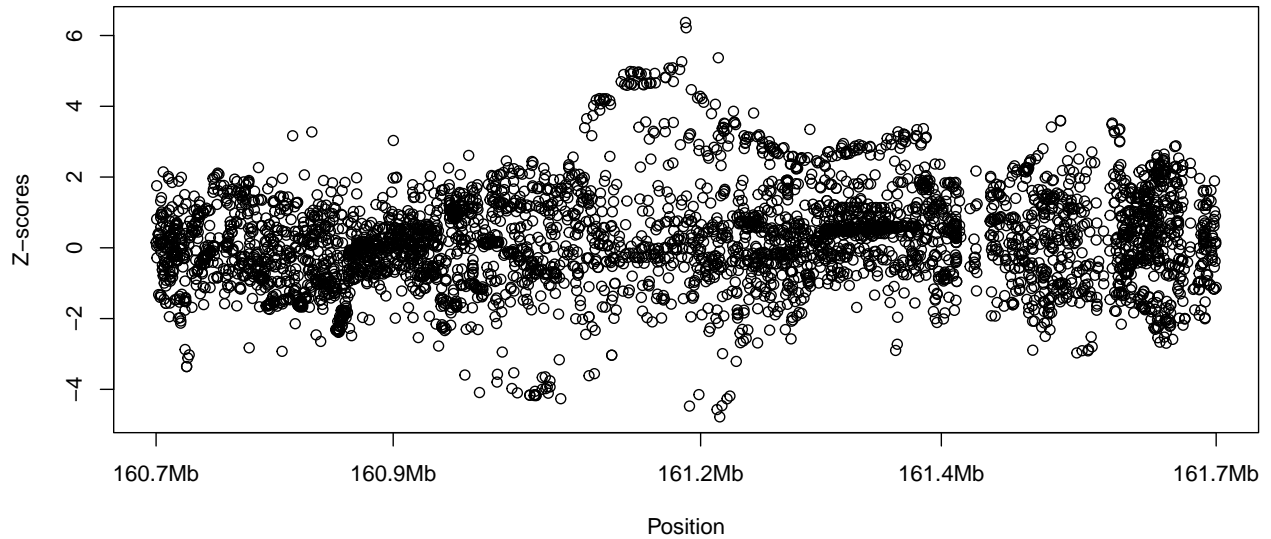
##	uniqID.a1a2	CHR	BP	A1	A2	SNP	Z	P	Nsum
## 1	1:160656603_A_T	1	160656603	A	T	rs6702441	0.14921734	0.8813821	429975
## 2	1:160657127_T_C	1	160657127	T	C	rs143426473	0.81166395	0.4169845	435185
## 3	1:160657137_G_A	1	160657137	G	A	rs11589131	0.04962457	0.9604216	429757
## 4	1:160657197_C_G	1	160657197	C	G	rs10908797	1.75450624	0.0793438	377075
## 5	1:160657356_C_T	1	160657356	C	T	rs145169682	-0.28401355	0.7764000	17477
## 6	1:160658364_G_A	1	160658364	G	A	rs7539434	0.23514690	0.8140947	380902
##	Neff	dir	EAF	BETA	SE				
## 1	423496.9	?-++	0.3695570	0.0003359043	0.002251107				
## 2	428659.9	?+++	0.0337234	0.0048561053	0.005982901				
## 3	423280.9	?-++	0.3693410	0.0001117523	0.002251954				
## 4	375814.5	?++?	0.0967510	0.0068457320	0.003901800				
## 5	17477.0	???	0.0117955	-0.0140704510	0.049541479				
## 6	379634.8	?++?	0.0280567	0.0016341896	0.006949654				

Data of the locus CR1

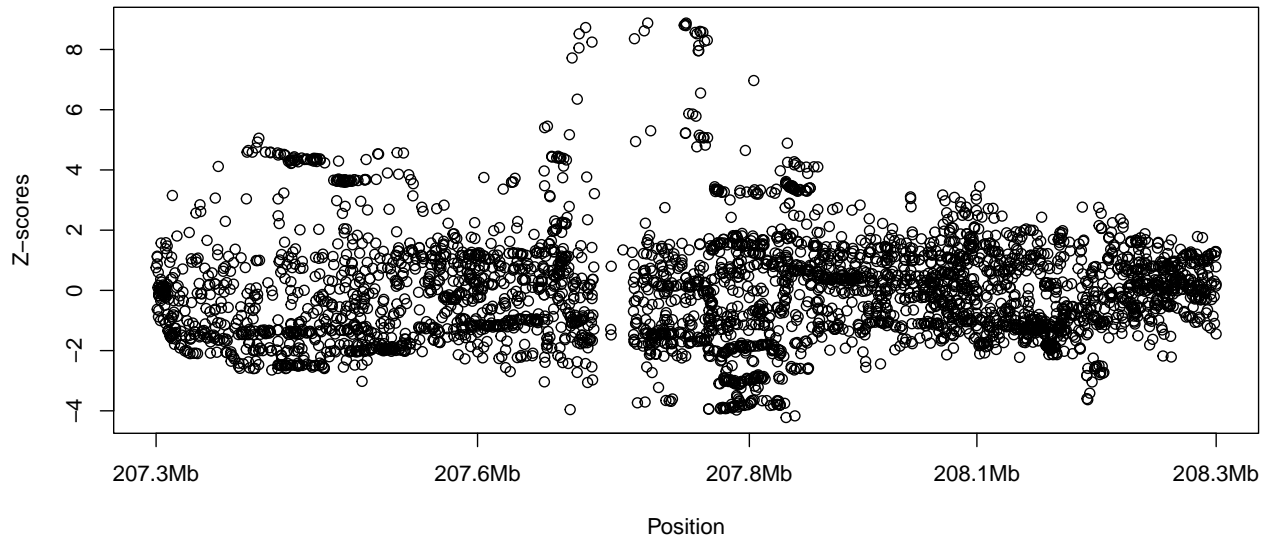
##	uniqID.a1a2	CHR	BP	A1	A2	SNP	Z	P	Nsum
## 1	1:207287187_T_C	1	207287187	T	C	rs2808470	0.75499216	0.4502537	433909
## 2	1:207288258_C_T	1	207288258	C	T	rs147553990	-0.62678527	0.5308000	364527
## 3	1:207288297_T_C	1	207288297	T	C	rs17020983	1.13570436	0.2560803	434723
## 4	1:207288309_T_G	1	207288309	T	G	rs79498904	0.87863355	0.3796000	364051
## 5	1:207288392_G_A	1	207288392	G	A	rs17020993	1.10788037	0.2679135	436498
## 6	1:207288897_T_C	1	207288897	T	C	rs12031629	0.09394084	0.9251562	71639
##	Neff	dir	EAF	BETA	SE				
## 1	427395.4	?-++	0.19396400	0.0020652587	0.002735470				
## 2	364527.0	??-?	0.00332615	-0.0127494442	0.020341008				
## 3	428202.0	?-++	0.09601170	0.0041656405	0.003667892				
## 4	364051.0	?++?	0.01279640	0.0091614442	0.010426923				
## 5	429961.1	?+++	0.15706800	0.0032833906	0.002963669				
## 6	71639.0	?-??	0.42993500	0.0005013039	0.005336379				

From the AD data we use Z-scores. Notice that the sample size values in the column “Nsum” can vary from 9,703 to 444,006 depending on which datasets are included in the meta-analyses of the AD study.

ADAMTS4



CR1

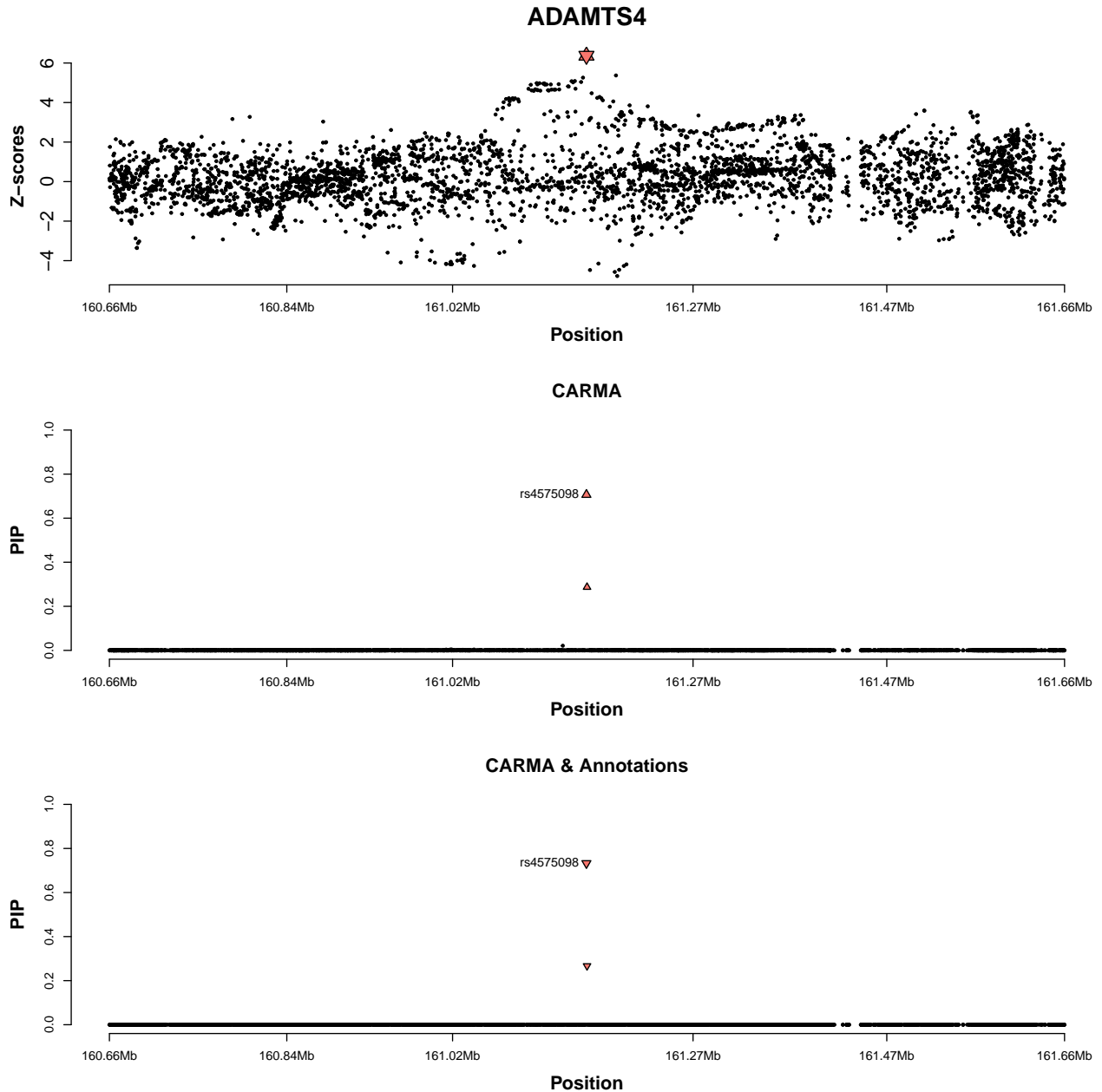


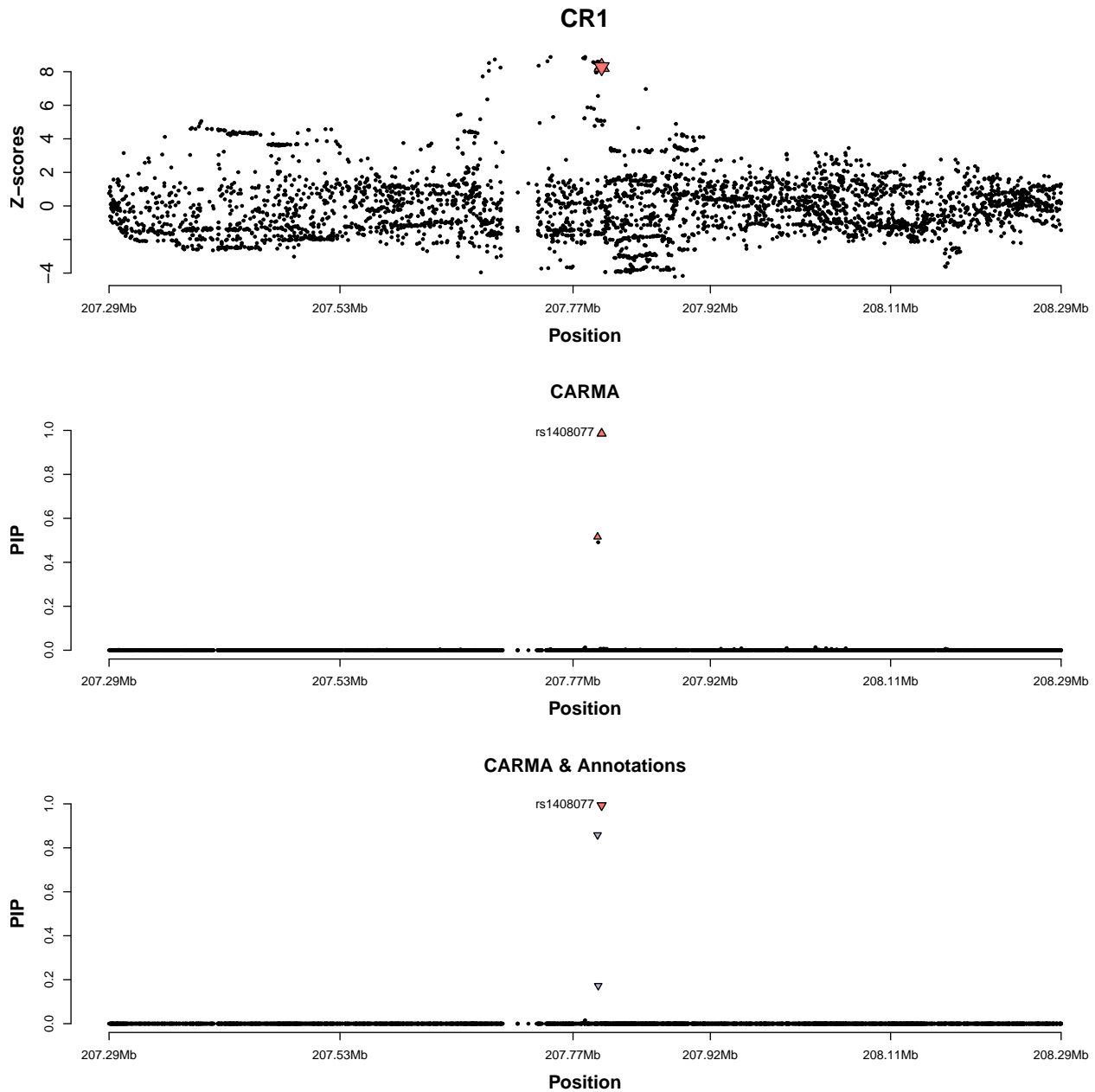
Next we run CARMA with two settings: 1. without annotations, and 2. with annotations as described above. We use the function “CARMA_fixed_sigma” to run the meta-analysis and the external LD. Also, the hyperparameter η is chosen by an adaptive procedure to control the false positives caused by possible discrepancies between Z-scores and external LD values.

```
Data_ADAMTS4<-readRDS('ADAMTS4.RData')
Data_CR1<-readRDS('CR1.RData')
z.list<-list()
ld.list<-list()
z.list[[1]]<-Data_ADAMTS4$`Meta-data`$Z
z.list[[2]]<-Data_CR1$`Meta-data`$Z
ld.list[[1]]<-Data_ADAMTS4$LD
ld.list[[2]]<-Data_CR1$LD
CARMA.results_no_annot<-CARMA_fixed_sigma(z.list,ld.list)
```

```
#####With annotations
#####The first 6 column of annotation file include location information etc.
annot.list<-list()
annot.list[[1]]<-as.matrix(cbind(1,Data_ADAMTS4$Annotations[,-(1:6)])
annot.list[[2]]<-as.matrix(cbind(1,Data_CR1$Annotations[,-(1:6)])
CARMA.results_annot<-CARMA_fixed_sigma(z.list,ld.list,w.list=annot.list)
```

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.

```
## [1] "This is the first credible set of the locus ADAMTS4"
##      CHR      BP A1 A2      SNP      Z      PIPs
## 2184    1 161155392  A  G  rs4575098 6.369505 0.7327455
## 2186    1 161156033  A  C  rs11585858 6.217633 0.2670102

## [1] "This is the first credible set of the locus CR1"
##      CHR      BP A1 A2      SNP      Z      PIPs
## 1639    1 207804141  A  C  rs1408077 8.276385 0.9931106
## [1] "This is the second credible set of the locus CR1"
##      CHR      BP A1 A2      SNP      Z      PIPs
## 1621    1 207799874  A  C  rs10863420 8.601581 0.8584831
```

```
## 1626 1 207800555 T C rs1408078 8.582403 0.1727934
```

We can also examine the credible models.

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

```
##      CHR      BP A1 A2      SNP      Z      PIPs
## 2184 1 161155392 A G rs4575098 6.369505 0.7327455
## 2186 1 161156033 A C rs11585858 6.217633 0.2670102
```

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

```
##      CHR      BP A1 A2      SNP      Z      PIPs
## 1621 1 207799874 A C rs10863420 8.601581 0.8584831
## 1639 1 207804141 A C rs1408077 8.276385 0.9931106
```

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.
- Kircher, Martin, Daniela M Witten, Preti Jain, Brian J O'roak, Gregory M Cooper, and Jay Shendure. 2014. "A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants." *Nature Genetics* 46 (3): 310–15.
- 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.
- Yang, Zikun, Chen Wang, Stephanie Erjavec, Lynn Petukhova, Angela Christiano, and Iuliana Ionita-Laza. 2021. "A Semisupervised Model to Predict Regulatory Effects of Genetic Variants at Single Nucleotide Resolution Using Massively Parallel Reporter Assays." *Bioinformatics* 37 (14): 1953–62.
- Zhou, Jian, and Olga G Troyanskaya. 2015. "Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model." *Nature Methods* 12 (10): 931–34.