Tutorials for the R/Bioconductor Package IntAssoPlot

This vignette documents usage of IntAssoPlot. IntAssoPlot was designed to plot the association, gene struture, and LD matrix in one single plot. As you read this document, you will see the input data format and the basic usage of IntAssoPlot.

1. Introduction

1.1. install compiler

The install_github(), in the R package remotes, requires that you build from source, thus make and compilers must be installed on your system -- see the R FAQ for your operating system; you may also need to install dependencies manually.

For Windows system, rtools, a compiler, is required and can be found at https://cran.r-project.org/bin/windows/Rtools/.

For Ubuntu/Linux system, compilers could be installed by the command: sudo apt-get install libcurl4-openssl-dev libssl-dev. For more information, please see https://stackoverflow.com/questions/20923209/problems-installing-the-devtools-package.

1.2. install R package devtools and remotes

install.packages(c("devtools","remotes"))

install depended packages, including ggplot2, SNPRelate, ggrepel, gdsfmt and reshape2

ggplot2, ggrepel, and reshape2 are installed from CRAN

install.packages(c("ggplot2","ggrepel","reshape2"))

SNPRelate and gdsfmt are installed from Bioconductor

if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")

BiocManager::install(c("SNPRelate","gdsfmt"))

1.3. install IntAssoPlot from Github:

library(remotes) # version 2.1.0

download, build, and install IntAssoPlot without creating vignette

install_github("whweve/IntAssoPlot")

download, build, and install IntAssoPlot with creating vignette

install_github("whweve/IntAssoPlot",build=TRUE,build_vignettes = TRUE)

1.4 example

see example("IntGenicPlot") or example("IntRegionalPlot")

2. Data formats used in IntAssoPlot

To give a quickly introduction to IntAssoPlot, we listed datasets previously published (Chia et al., 2013; Li et al., 2013; Wang et al., 2016). Below, we give detailed information for each data required.

2.1 association analysis data format

A dataset containing the association analysis file. Currently, we plot the -log10 transformed p values, which can be derived from genome scan, against the physical position of the corresponding marker. Thus, a dataframe, containing marker name and its p value, is required. The required variables are as follows: Marker (molecular marker name), Locus (the chromosome of the marker), Site (the position of the marke), p (p value of the marker) Here's a quick demo of association data format:

```
head(association)
#>     Marker Locus     Site         p
#> 1 SNP37049     9 90042244 0.2073684
#> 2 SNP37050     9 90042408 0.5028760
#> 3 SNP37052     9 90210602 0.1561271
#> 4 SNP37053     9 90255652 0.2814653
#> 5 SNP37056     9 90290939 0.6392511
#> 6 SNP44597     9 90367887 0.6582597
#the attribute of each column could be viewed as:
str(association$Marker)
#>  chr [1:2316] "SNP37049" "SNP37050" "SNP37052" "SNP37053" "SNP37056" ...
str(association$Locus)
#>  int [1:2316] 9 9 9 9 9 9 9 9 9 9 ...
str(association$Site)
#>  int [1:2316] 90042244 90042408 90210602 90255652 90290939 90367887 90377686 90385837 90560965
90759605 ...
str(association$p)
#>  num [1:2316] 0.207 0.503 0.156 0.281 0.639 ...
```

2.2 gene structure data format

A dataset containing the annotation file, usually a gtf file, WITHOUT the column name. When a genome of one species is sequenced, the gtf file can be found at genome annotation website. For the released genomes, please refer to www.ensembl.org. Because most of the gtf files are lack of colnames, the annotation file could be read in using read.table("where is your file",header=FALSE). Here's a quick demo of gtf data format:

```
head(gtf)
#>   V1             V2          V3       V4       V5 V6 V7 V8
#> 1  9 protein_coding        exon 90030791 90030847  .  -  .
#> 2  9 protein_coding        exon 90030635 90030701  .  -  .
#> 3  9 protein_coding         CDS 90030635 90030681  .  -  0
#> 4  9 protein_coding start_codon 90030679 90030681  .  -  0
#> 5  9 protein_coding        exon 90029368 90029471  .  -  .
#> 6  9 protein_coding         CDS 90029368 90029471  .  -  1
#>
V9
#> 1              gene_id "GRMZM2G416644"; transcript_id "GRMZM2G416644_T01"; exon_number "1";
seqedit "false";
#> 2              gene_id "GRMZM2G416644"; transcript_id "GRMZM2G416644_T01"; exon_number "2";
seqedit "false";
#> 3 gene_id "GRMZM2G416644"; transcript_id "GRMZM2G416644_T01"; exon_number "2"; protein_id
"GRMZM2G416644_P01";
#> 4                                gene_id "GRMZM2G416644"; transcript_id "GRMZM2G416644_T01";
exon_number "2";
#> 5              gene_id "GRMZM2G416644"; transcript_id "GRMZM2G416644_T01"; exon_number "3";
seqedit "false";
#> 6 gene_id "GRMZM2G416644"; transcript_id "GRMZM2G416644_T01"; exon_number "3"; protein_id
"GRMZM2G416644_P01";
#the attribute of each column could be viewed as:
str(gtf$V1)
#>  chr [1:4665] "9" "9" "9" "9" "9" "9" "9" "9" "9" "9" "9" "9" "9" "9" "9" ...
str(gtf$V2)
#>  chr [1:4665] "protein_coding" "protein_coding" "protein_coding" ...
str(gtf$V3)
#>  chr [1:4665] "exon" "exon" "CDS" "start_codon" "exon" "CDS" "exon" "CDS" ...
str(gtf$V4)
#>  int [1:4665] 90030791 90030635 90030635 90030679 90029368 90029368 90028806 90029227 90029224
90030697 ...
str(gtf$V5)
#>  int [1:4665] 90030847 90030701 90030681 90030681 90029471 90029471 90029282 90029282 90029226
90031013 ...
str(gtf$V6)
#>  chr [1:4665] "." "." "." "." "." "." "." "." "." "." "." "." "." "." "." ...
str(gtf$V7)
#>  chr [1:4665] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" ...
str(gtf$V8)
#>  chr [1:4665] "." "." "0" "0" "." "1" "." "2" "0" "." "0" "0" "." "0" "0" ...
str(gtf$V9)
#>  chr [1:4665] "gene_id \"GRMZM2G416644\"; transcript_id \"GRMZM2G416644_T01\"; exon_number
\"1\"; seqedit \"false\";" ...
```

2.3 genotype data format

A dataset containing the genotype file, usually a hapmap file, with the column name. Hapmap is a frequently used format to store the genotyoe information. One can high-throughput sequence important materials, align the sequences to the reference genome, extract SNPs/InDels. On the other hand, one can resequence one specific gene. To perform this, design overlapped amplication primers, amplify genomic or transcriptomic fragments, multi-align the sequence, extract SNPs/InDels. Here's a quick demo of association data format:

```
#only 20 column of the genotype markers are shown.
head(zmvpp1_hapmap[,1:20])
#>          rs allele chrom      pos strand assembly center protLSID assayLSID
#> 1 INDEL-665    -/+     9 94183611      +       NA     NA       NA        NA
#> 2   snp-663    A/T     9 94183597      +       NA     NA       NA        NA
#> 3   snp-649    A/G     9 94183594      +       NA     NA       NA        NA
#> 4   snp-646    C/T     9 94183585      +       NA     NA       NA        NA
#> 5   snp-637    G/T     9 94183584      +       NA     NA       NA        NA
#> 6   snp-636    A/G     9 94183575      +       NA     NA       NA        NA
#>   panel QCcode B73 CIMBL55 CIMBL32 CML298 CML170 CIMBL12 CIMBL95 CIMBL1 CIMBL56
#> 1    NA     NA  --      ++      ++     ++     ++      ++      ++     --      --
#> 2    NA     NA  AA      TT      TT     TT     TT      TT      TT     NN      NN
#> 3    NA     NA  AA      GG      GG     GG     GG      GG      GG     NN      NN
#> 4    NA     NA  CC      TT      TT     TT     TT      TT      TT     NN      NN
#> 5    NA     NA  GG      TT      TT     TT     TT      TT      TT     NN      NN
#> 6    NA     NA  AA      AA      AA     AA     AA      AA      AA     NN      NN
```

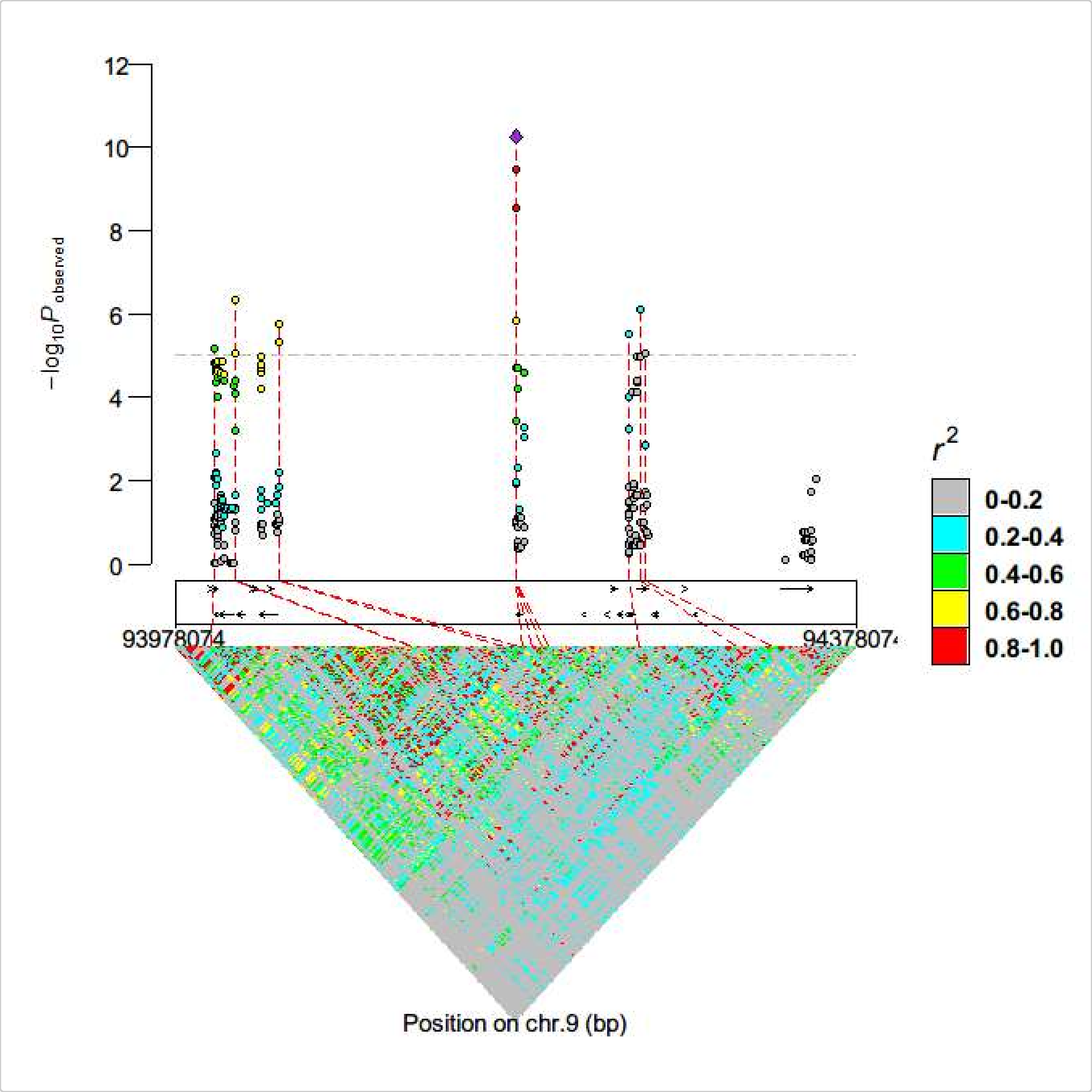
3. plot the association with annotation and LD matrix

IntAssoPlot is to automatically integrate the results of association analysis, gene structure and LD matrix into one single view. In fact, this task is difficult, because the scatter diagram plot - log10P against the physical location of SNPs, while the LD matrix has no physical location. Our work is actually to solve the above problem. Here, we present an example to show the usage of IntAssoPlot, using a previouly published data (Wang, et al., 2016).

3.1 Regional integrative plot with one set of genotype markers

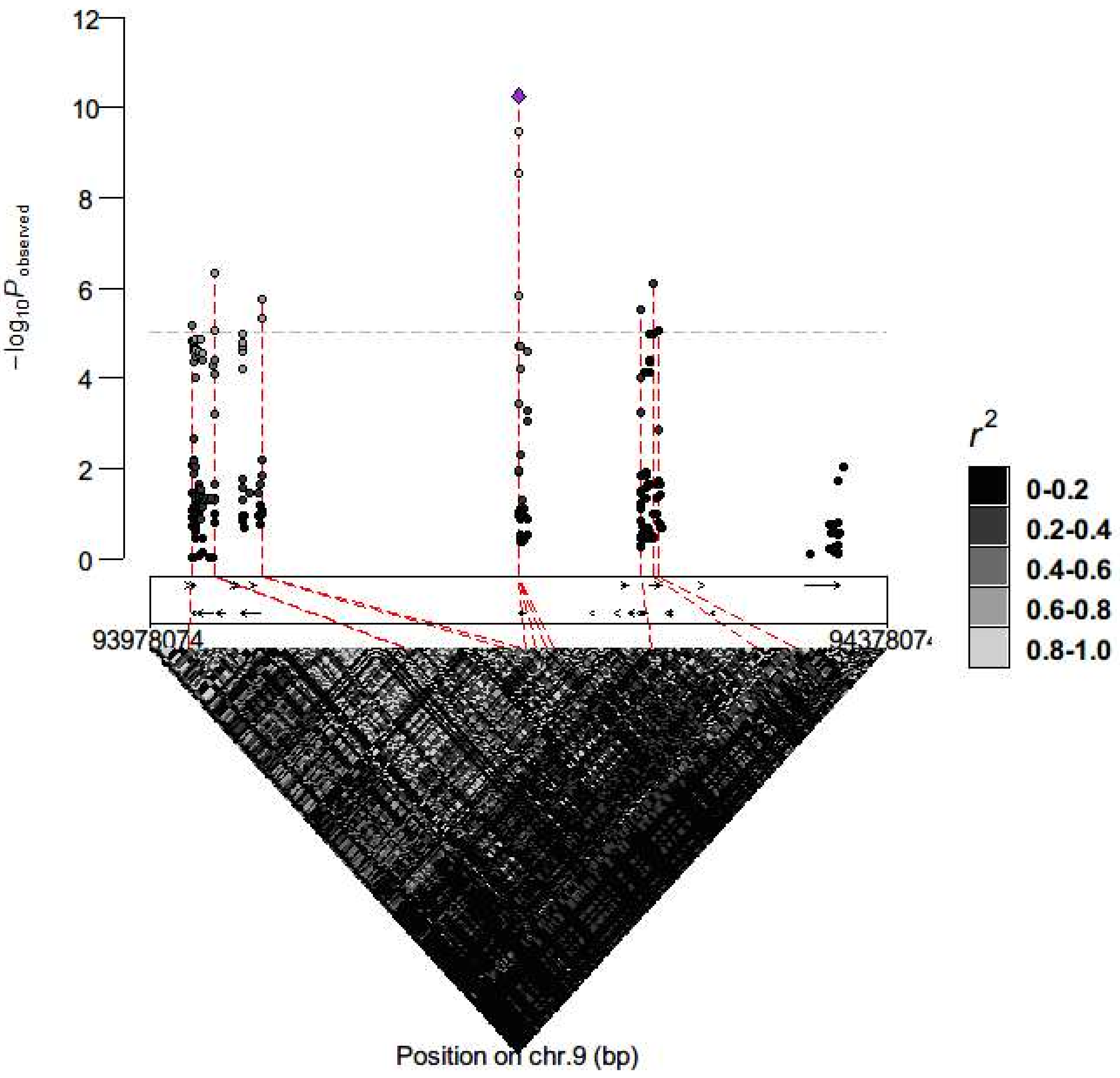
plot the association results at a region spaning a 400 kbp region, and plot the LD matrix using SNP markerers that are same as that for association mapping.

```
IntRegionalPlot(chr=9,left=94178074-
200000,right=94178074+200000,gtf=gtf,association=association,hapmap=hapmap_am368,hapmap_ld=hapmap_a

#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread

#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
```



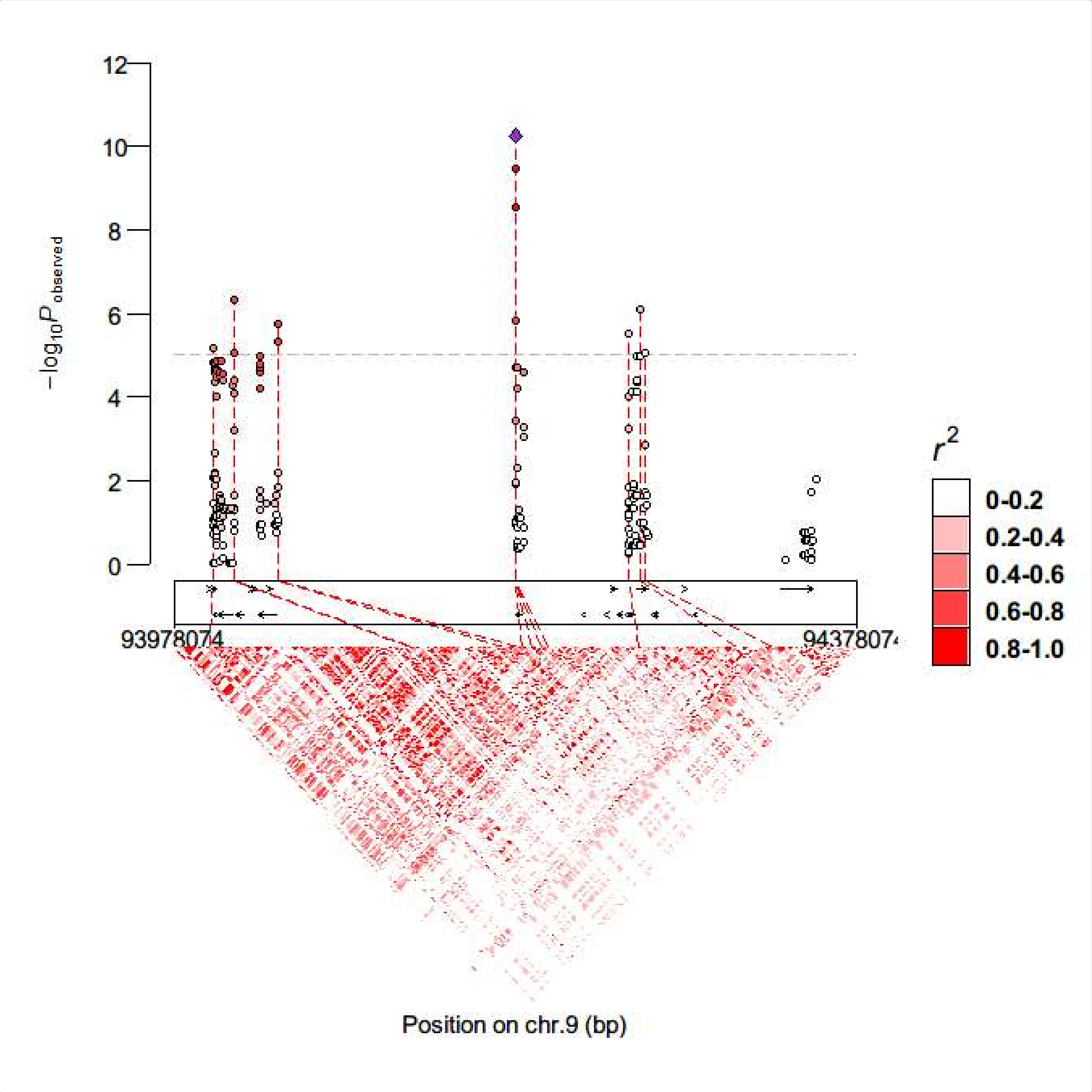
3.2 plot the LD values with colours ranging from light gray to dark gray.

```
IntRegionalPlot(chr=9,left=94178074-
200000,right=94178074+200000,gtf=gtf,association=association,hapmap=hapmap_am368,hapmap_ld=hapmap_a
 = "gray1",colour04 = "gray21",colour06 = "gray41",colour08 = "gray61",colour10 = "gray81",)
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
```



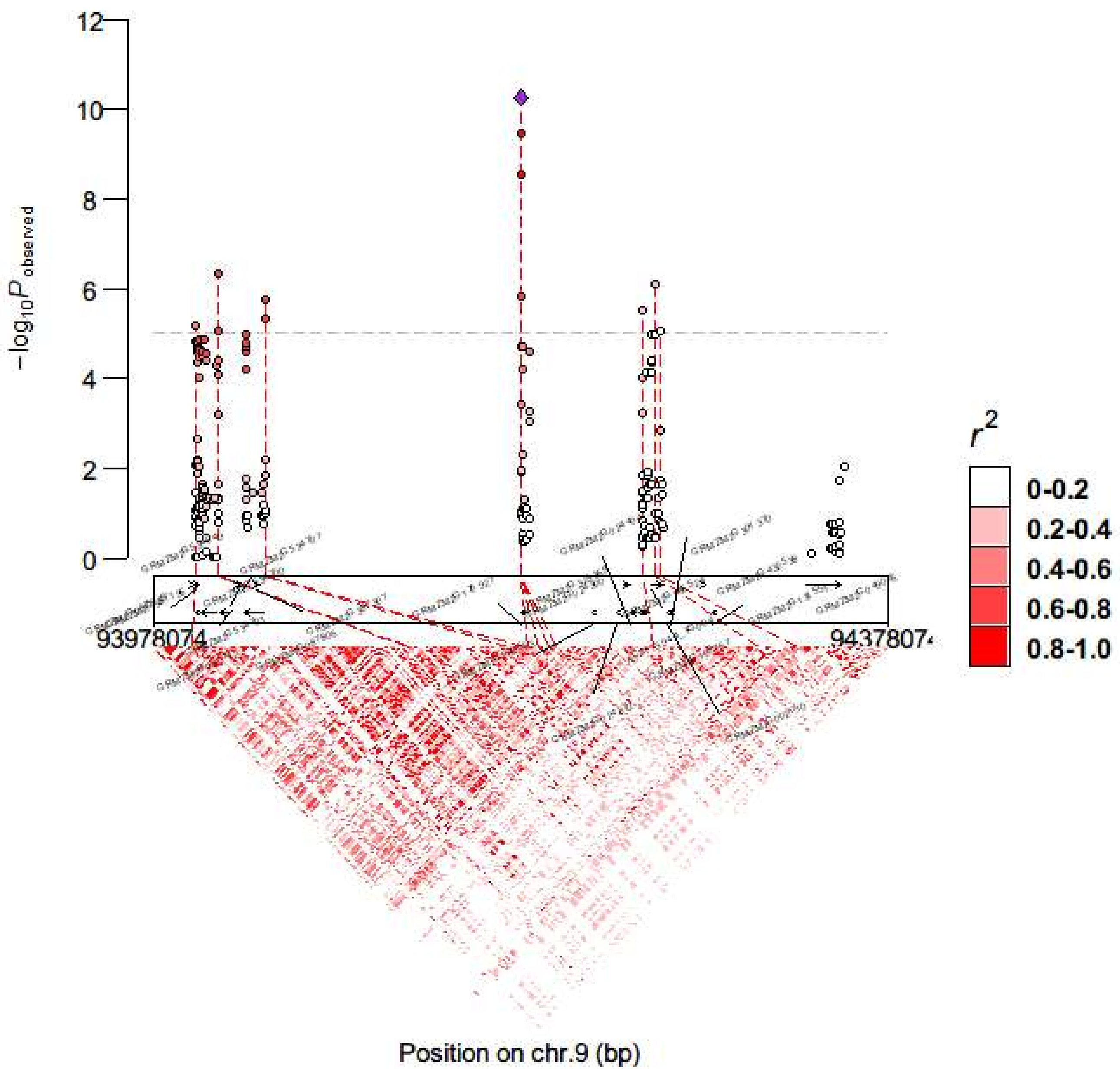
3.3 plot the LD values with colours ranging from white to red.

```
#get five colors ranging from white to red
pal <- colorRampPalette(c("white", "red"))
IntRegionalPlot(chr=9,left=94178074-
200000,right=94178074+200000,gtf=gtf,association=association,hapmap=hapmap_am368,hapmap_ld=hapmap_a
 = pal(5)[1],colour04 = pal(5)[2],colour06 = pal(5)[3],colour08 = pal(5)[4],colour10 = pal(5)[5])
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276

#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
```



3.4 plot the LD values with colours ranging from white to red and label the gene name.

```
#get five colors ranging from white to red
pal <- colorRampPalette(c("white", "red"))
IntRegionalPlot(chr=9,left=94178074-
200000,right=94178074+200000,gtf=gtf,association=association,hapmap=hapmap_am368,hapmap_ld=hapmap_a

 = pal(5)[1],colour04 = pal(5)[2],colour06 = pal(5)[3],colour08 = pal(5)[4],colour10 = pal(5)
[5],label_gene_name = TRUE)
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 270
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 143276
```

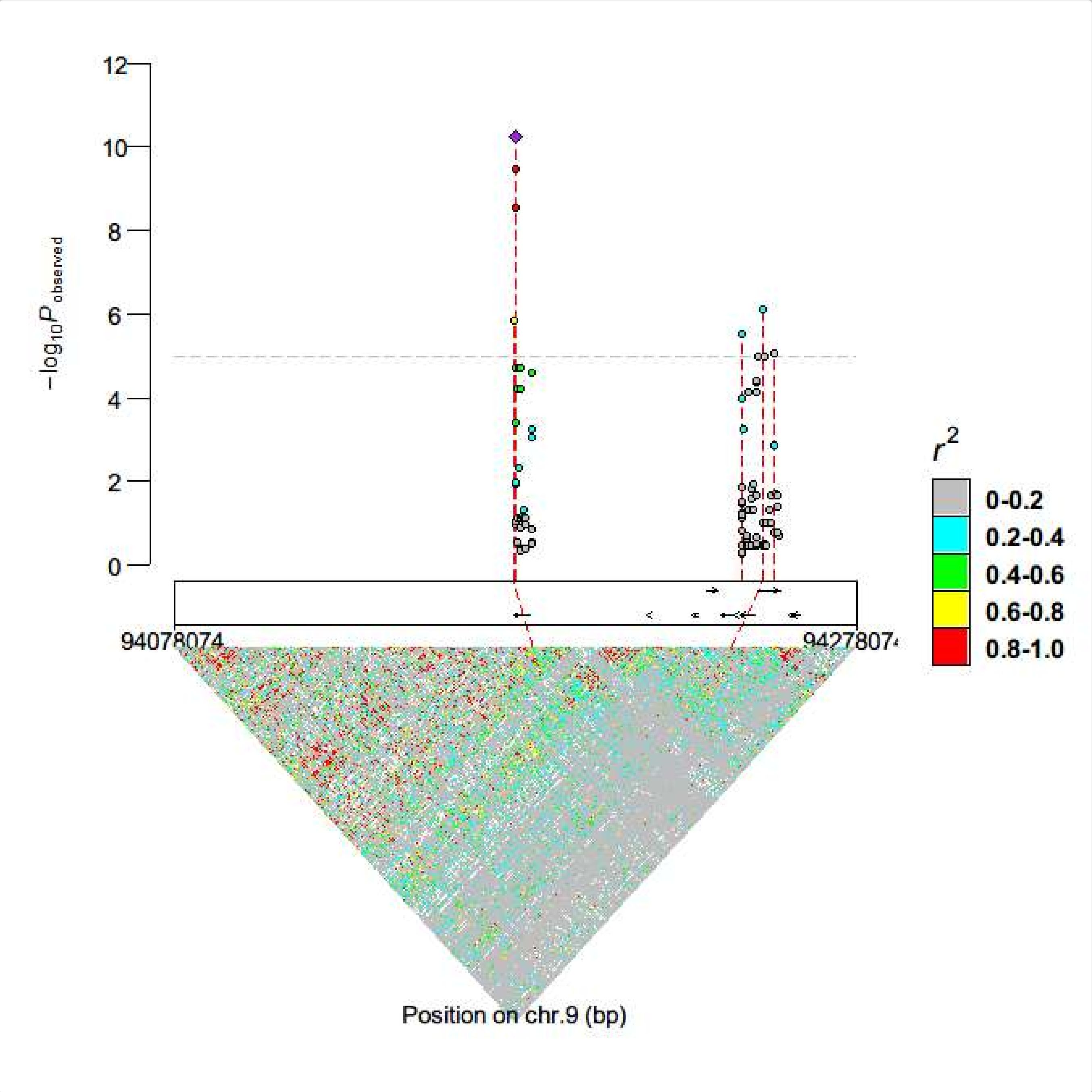


3.5 Regional integrative plot with two set of genotype markers

plot the association results at a regional spaning a 200 kbp region, and plot the LD matrix using SNP markerers that differed from that for association mapping. This feature allows reserchers investigate the LD structure at a more wide range of markers.

```
IntRegionalPlot(chr=9,left=94178074-
100000,right=94178074+100000,gtf=gtf,association=association,hapmap=hapmap_am368,hapmap_ld=hapmap2,
```

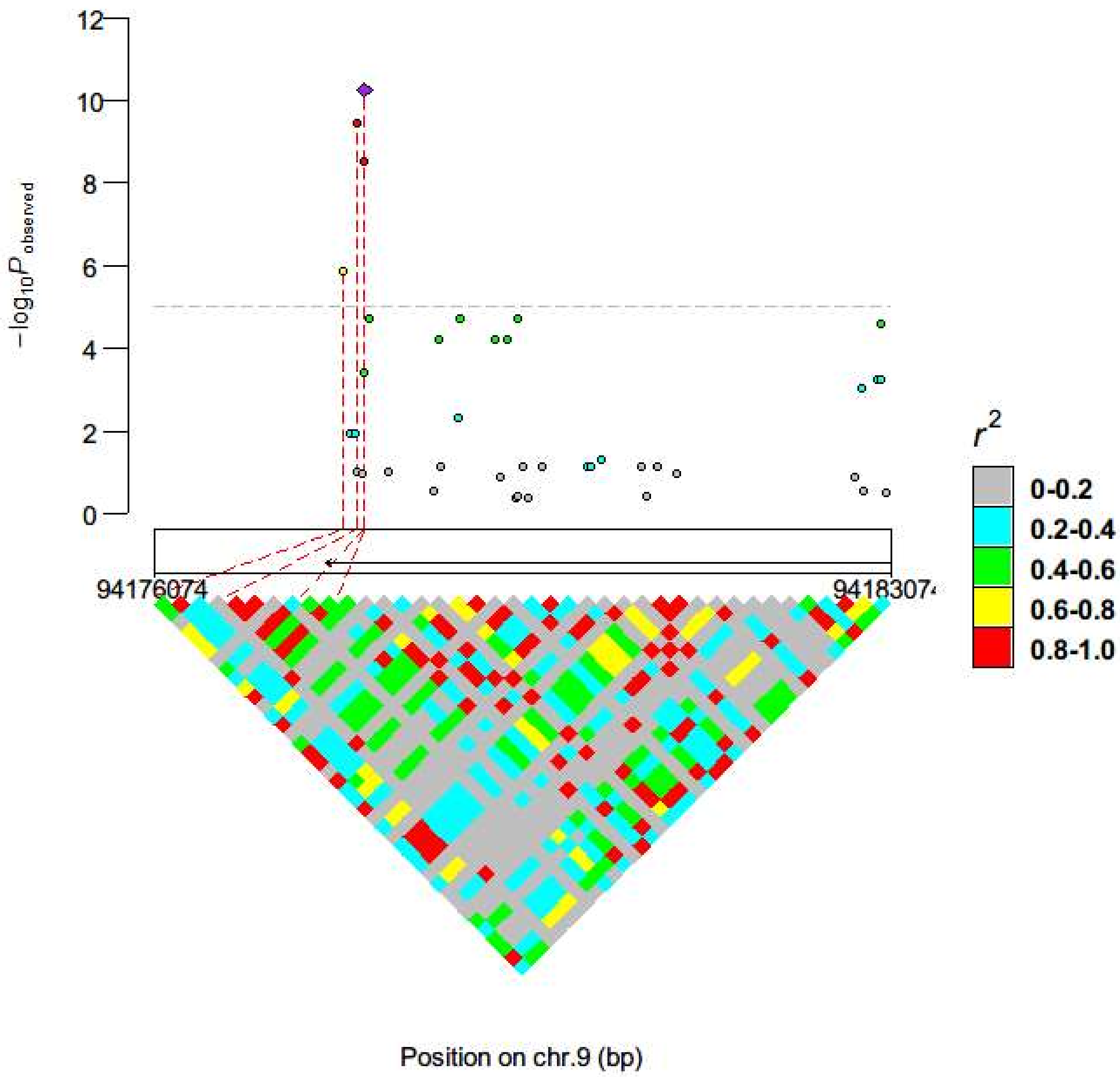
```
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 108
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 56746
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 104
#>     # of SNPs: 1,081
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 114672
```



3.6 a relative small regional integrative plot with one set of genotype markers

plot the association results at a regional covering the candidate gene, and plot the LD matrix using SNP markerers that are the same from that for association mapping.

```
IntRegionalPlot(chr=9,left=94178074-
2000,right=94178074+5000,gtf=gtf,association=association,hapmap=hapmap_am368,hapmap_ld=hapmap_am368

#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 41
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 21412
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 368
#>     # of SNPs: 41
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 21412
```



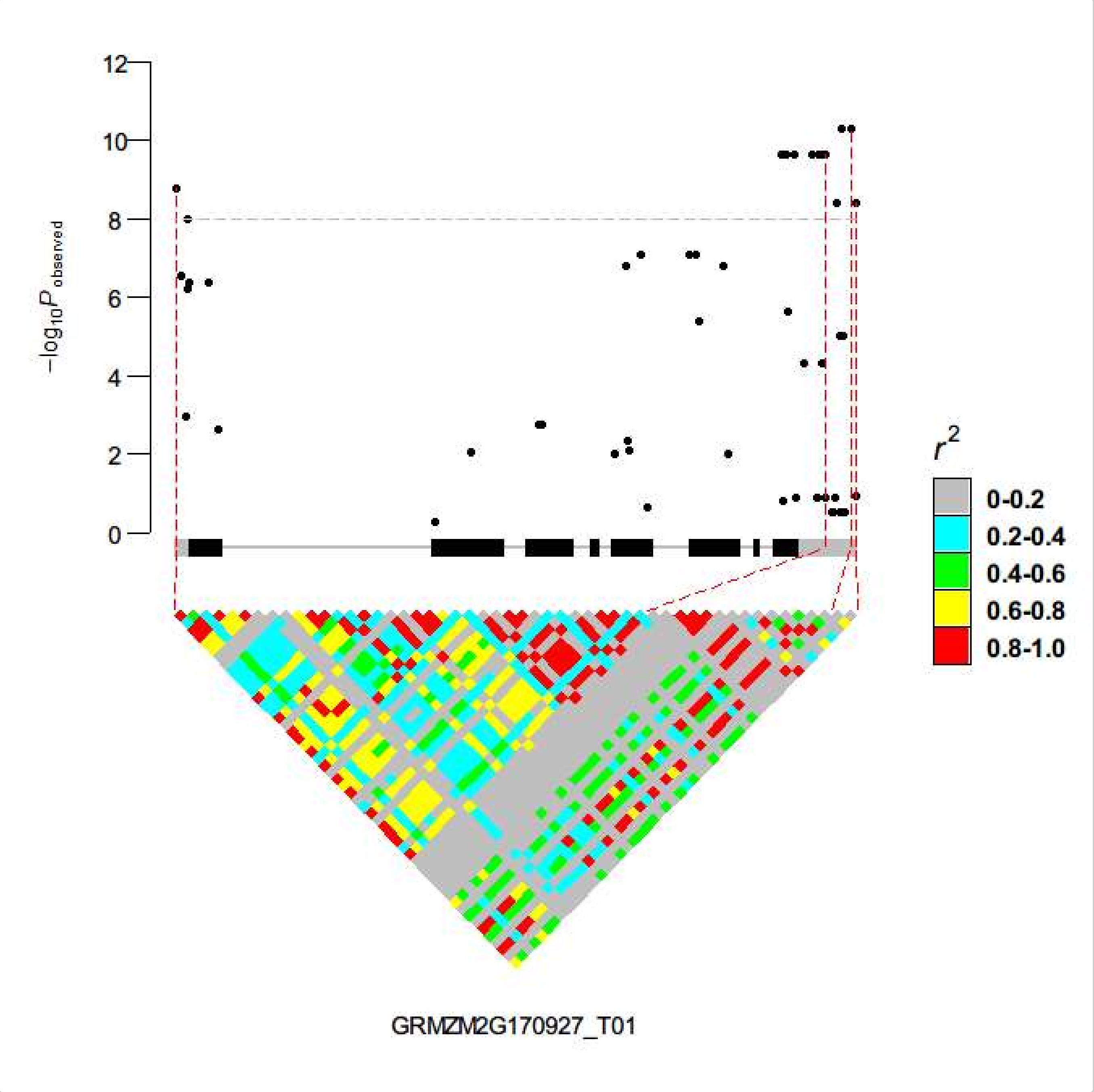
3.7 a single gene level plot

plot the association results at a given gene, and plot the LD matrix using SNP markerers that are the same from that for association mapping. Also specified markers are highlighted by various shape and colour.

3.7.1 a basic plot

```
IntGenicPlot('GRMZM2G170927_T01',gtf,association=zmvpp1_association,hapmap=zmvpp1_hapmap,hapmap_ld
 = zmvpp1_hapmap,threshold=8,leadsnpLD = FALSE)
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 141
#>     # of SNPs: 53
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 9992
#> Linkage Disequilibrium (LD) estimation on genotypes:
```

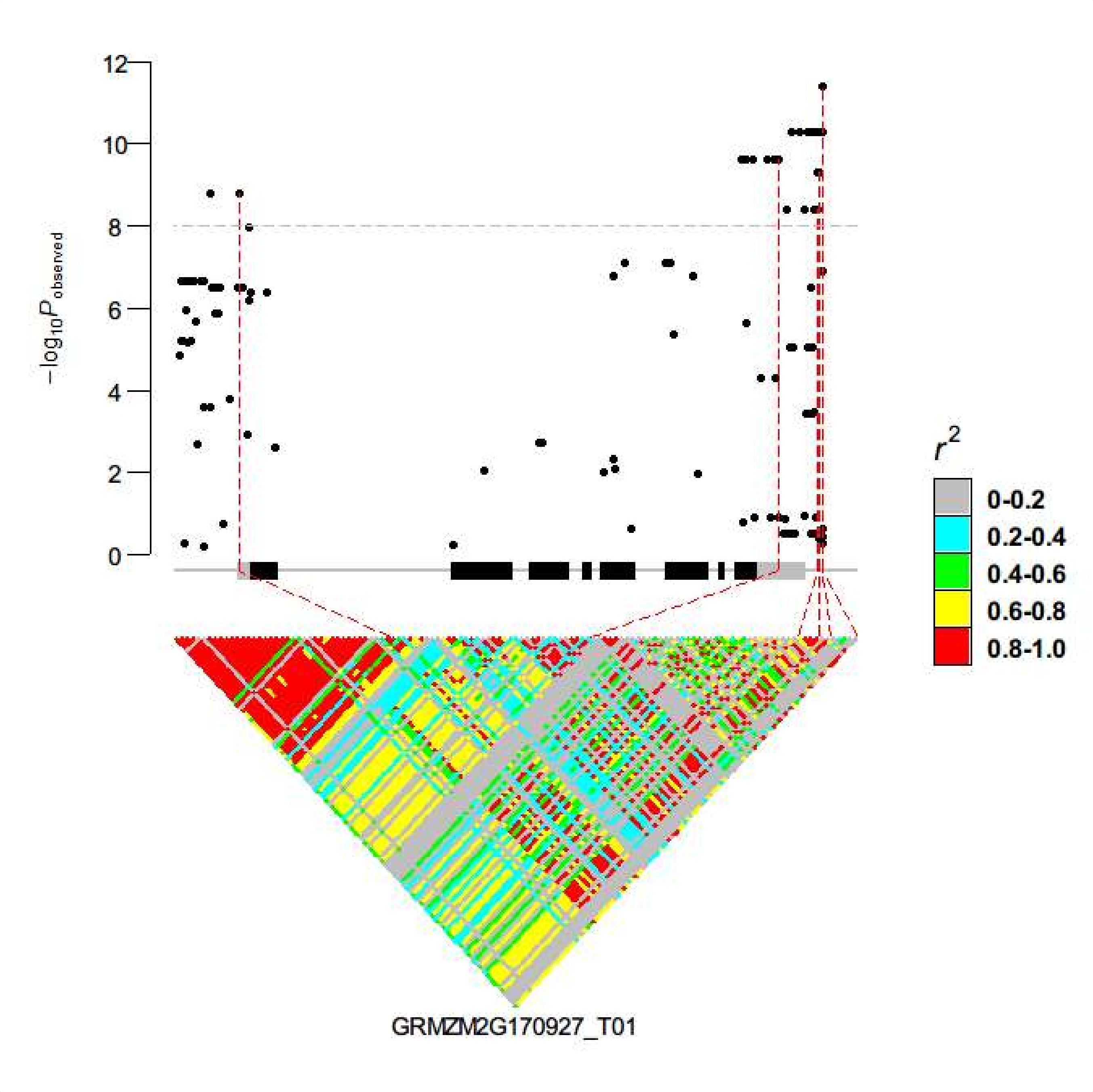
```
#>      # of samples: 141
#>      # of SNPs: 53
#>      using 1 thread
#>      method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 9992
```



3.7.2 extand region from up/down-stream of gene

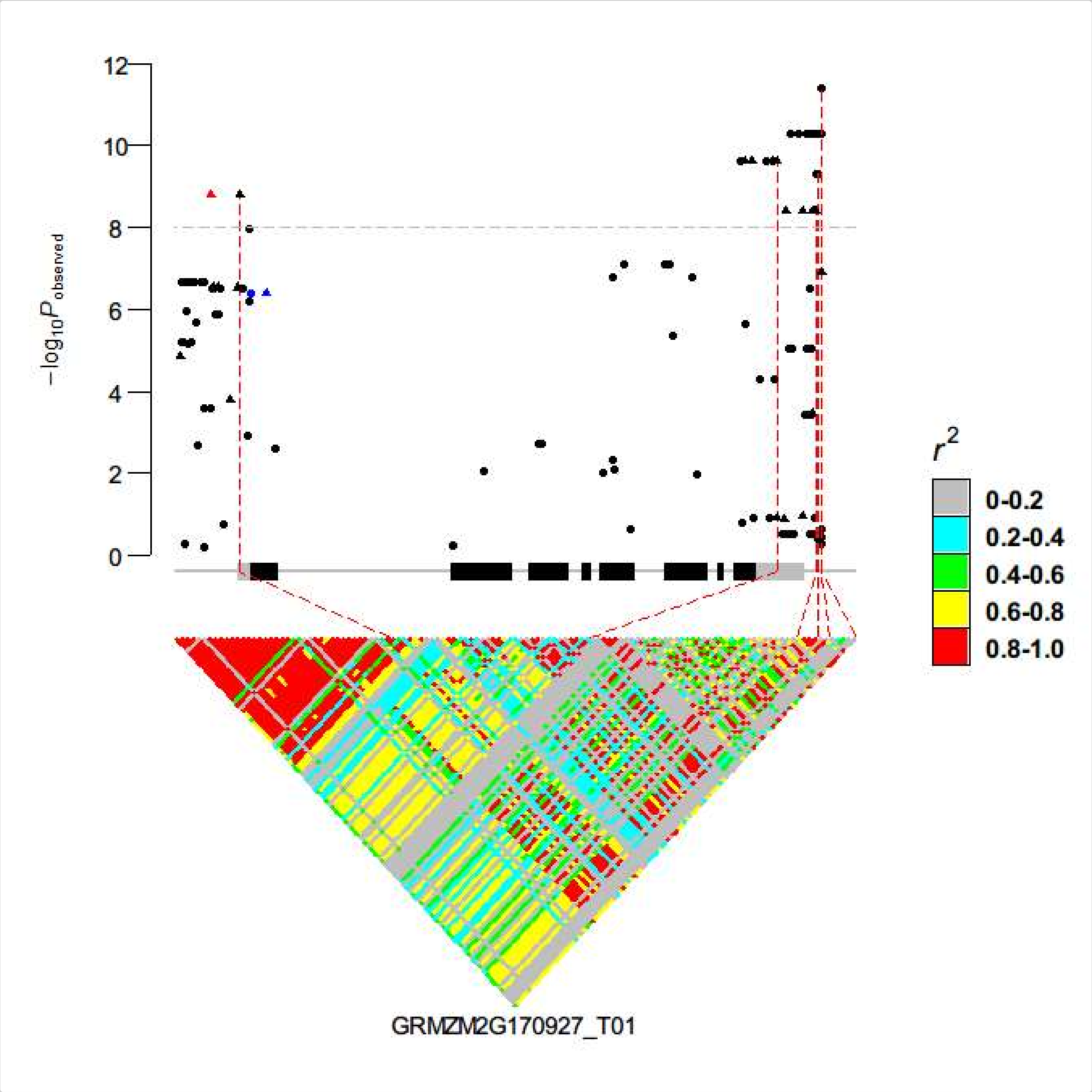
```
IntGenicPlot('GRMZM2G170927_T01',gtf,association=zmvpp1_association,hapmap=zmvpp1_hapmap,hapmap_ld
 = zmvpp1_hapmap,threshold=8,up=500,down=600,leadsnpLD = FALSE)
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>      # of samples: 141
#>      # of SNPs: 124
#>      using 1 thread
#>      method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
#> Linkage Disequilibrium (LD) estimation on genotypes:
```

```
#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
```



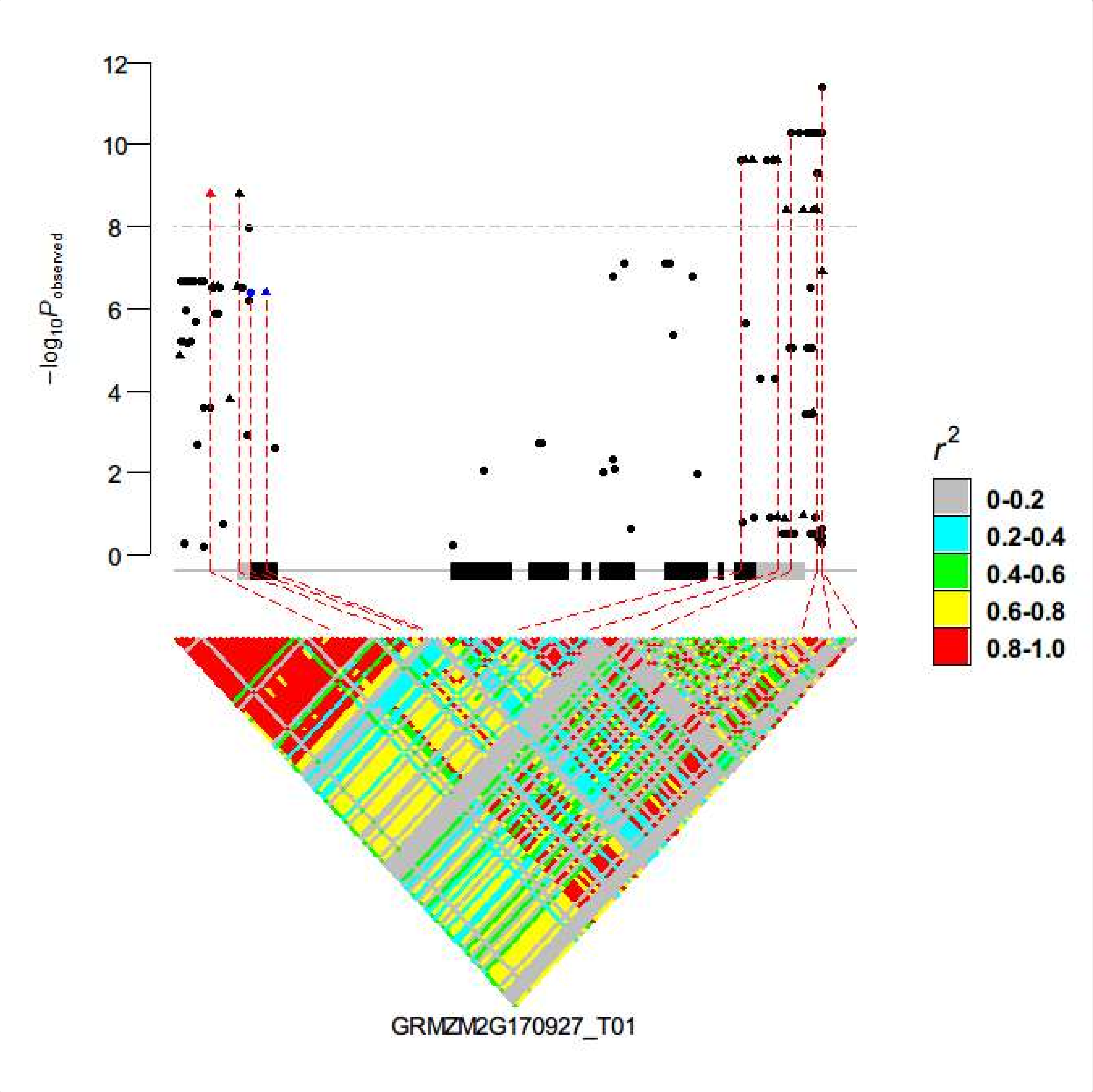
3.7.3 highlight selected marker, with colour and shape speicified in a dataframe: marker2highlight

```
IntGenicPlot('GRMZM2G170927_T01',gtf,association=zmvpp1_association,hapmap=zmvpp1_hapmap,hapmap_ld
 = zmvpp1_hapmap,threshold=8,up=500,down=600,leadsnpLD = FALSE,marker2highlight=marker2highlight)
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
```



3.7.4 add linking line

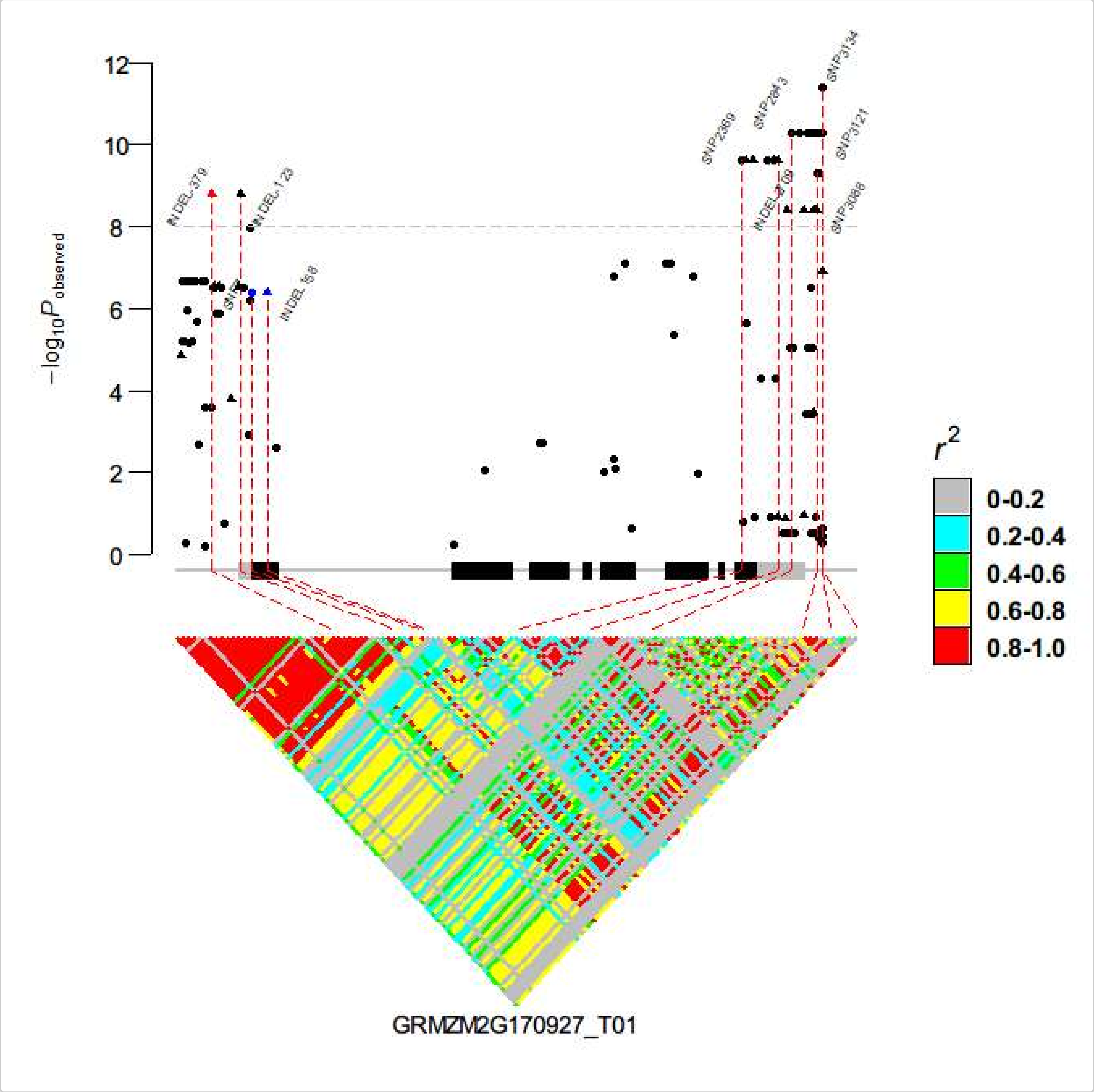
```
IntGenicPlot('GRMZM2G170927_T01',gtf,association=zmvpp1_association,hapmap=zmvpp1_hapmap,hapmap_ld
 = zmvpp1_hapmap,threshold=8,up=500,down=600,leadsnpLD =
FALSE,marker2highlight=marker2highlight,link2gene=marker2link,link2LD=marker2link)
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
```



3.7.5 add names for highlighted marker

```
IntGenicPlot('GRMZM2G170927_T01',gtf,association=zmvpp1_association,hapmap=zmvpp1_hapmap,hapmap_ld
 = zmvpp1_hapmap,threshold=8,up=500,down=600,leadsnpLD =
FALSE,marker2highlight=marker2highlight,link2gene=marker2link,link2LD=marker2link,marker2label=mark
```

```
#> Linkage Disequilibrium (LD) estimation on genotypes:

#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
#> Linkage Disequilibrium (LD) estimation on genotypes:
#>     # of samples: 141
#>     # of SNPs: 124
#>     using 1 thread
#>     method: correlation
#> LD matrix:    the sum of all selected genotypes (0,1,2) = 23488
```


$-\log_{10}P_{observed}$
INDEL379
INDEL123
SNP
INDEL58
SNP2369
SNP2843
INDEL
SNP3134
SNP3121
SNP3088
r^2
GRMZM2G170927_T01