Genetic Association Studies with R (part I)

in

Master in Omic data analysis

Universitat Vic (Uvic)

Juan R Gonzalez

(jrgonzalez@creal.cat)

BRGE - Bioinformatics Research Group in Epidemiology

http://www.creal.cat/brge

Barcelona Institute for Global Health (ISGlobal)

Departament of Mathematics, Universidad Autonoma de Barcelona (UAB)

Genetic Association analysis using R

- **Part I:** Single association analysis: `SNPassoc` package
  - Descriptive analysis
  - HWE test
  - Association analysis
  - Haplotype Analysis
- **Part II:** GWAS `snpStats` package
  - Quality control
  - Association analysis
  - Population stratification
  - Multiple comparisons
  - Manhantan plots

Single association analysis

Case-control study in asthma:

```
data.frame:   1578 obs. of  57 variables:
 $ country    : Factor w/ 10 levels "Australia","Belgium",..: 5
 $ gender     : Factor w/ 2 levels "Females","Males": 2 2 2 1 1
 $ age        : num  42.8 50.2 46.7 47.9 48.4 ...
 $ bmi        : num  20.1 24.7 27.7 33.3 25.2 ...
 $ smoke      : int  2 0 1 1 0 2 1 0 0 0 ...
 $ casecontrol: int  0 0 0 0 1 0 0 0 0 0 ...
 $ rs4490198  : Factor w/ 3 levels "A/A","A/G","G/G": 3 3 3 2 2
 $ rs4849332  : Factor w/ 3 levels "G/G","G/T","T/T": 3 2 3 2 1
 $ rs1367179  : Factor w/ 3 levels "C/C","G/C","G/G": 2 2 2 3 3
  [list output truncated]
```

Genetic Association analysis using R

Required library that is available on CRAN

```
install.packages("SNPassoc")
```

```
library(SNPassoc)
```

Let us load the data

```
asthma <- read.table("datasets/asthma.txt", header=TRUE)
asthma[1:5, 1:10]
```

```
##   country  gender      age      bmi smoke casecontrol rs4490198 rs4849332
## 1 Germany   Males 42.80630 20.14797     2           0       G/G       T/T
## 2 Germany   Males 50.22861 24.69136     0           0       G/G       G/T
## 3 Germany   Males 46.68857 27.73230     1           0       G/G       T/T
## 4 Germany Females 47.86311 33.33187     1           0       A/G       G/T
## 5 Germany Females 48.44079 25.23634     0           1       A/G       G/G
##   rs1367179 rs11123242
## 1       G/C        C/T
## 2       G/C        C/T
## 3       G/C        C/T
## 4       G/G        C/C
## 5       G/G        C/C
```

Let us prepare the data

```
asthma.s <- setupSNP(asthma, 7:ncol(asthma), sep="/")
args(setupSNP)

## function (data, colSNPs, sort = FALSE, info, sep = "/", ...)
## NULL

args(snp)

## function (x, sep = "/", name.genotypes, reorder = "common", remove.spaces =
##     allow.partial.missing = FALSE)
## NULL
```

Genetic Association analysis using R

Columns containing information about SNPs are now objects of class `snp`

```
head(asthma$rs1422993)

## [1] G/G G/T G/G G/T G/T G/G
## Levels: G/G G/T T/T

head(asthma.s$rs1422993)

## [1] G/G G/T G/G G/T G/T G/G
## Genotypes: G/G G/T T/T
## Alleles:  G T
```

Genetic Association analysis using R

Descriptive analysis can be performed using several generic functions

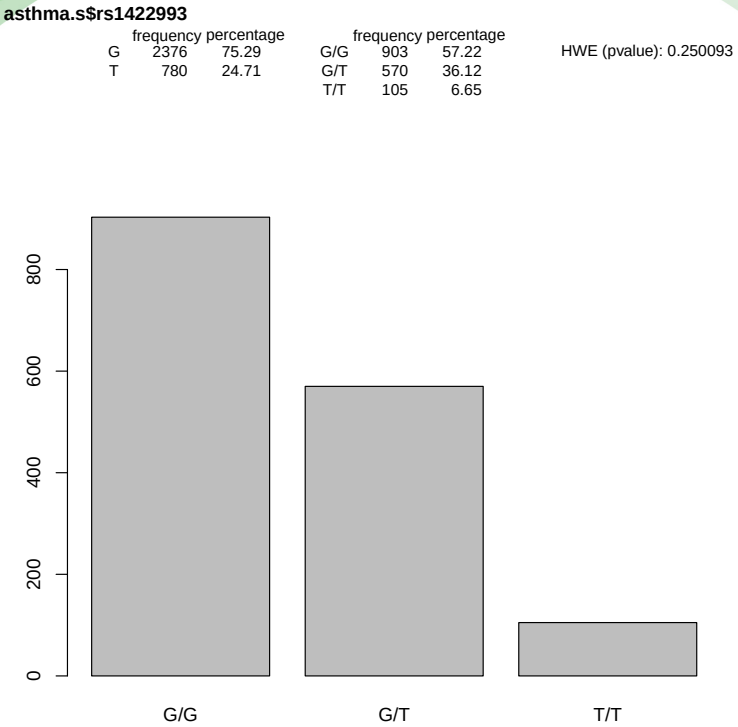
```
summary(asthma.s$rs1422993)

## Genotypes:
##     frequency percentage
## G/G       903  57.224335
## G/T       570  36.121673
## T/T       105   6.653992
##
## Alleles:
##   frequency percentage
## G      2376   75.28517
## T       780   24.71483
##
## HWE (p value): 0.250093
```

Genetic Association analysis using R

Descriptive analysis can be performed using several generic functions

```
plot(asthma.s$rs1422993)
```



Genetic Association analysis using R

Descriptive analysis can be performed using several generic functions

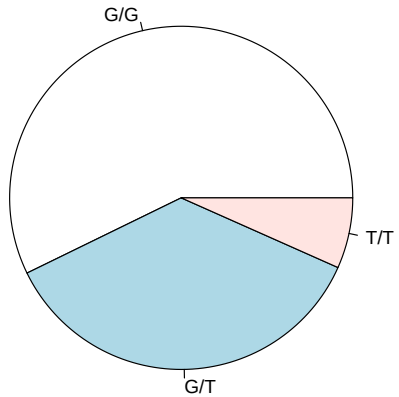
```
plot(asthma.s$rs1422993, type=pie)
```

asthma.s$rs1422993

	frequency	percentage
G	2376	75.29
T	780	24.71

	frequency	percentage
G/G	903	57.22
G/T	570	36.12
T/T	105	6.65

HWE (pvalue): 0.250093



Genetic Association analysis using R

Descriptive analysis can be performed using several generic functions

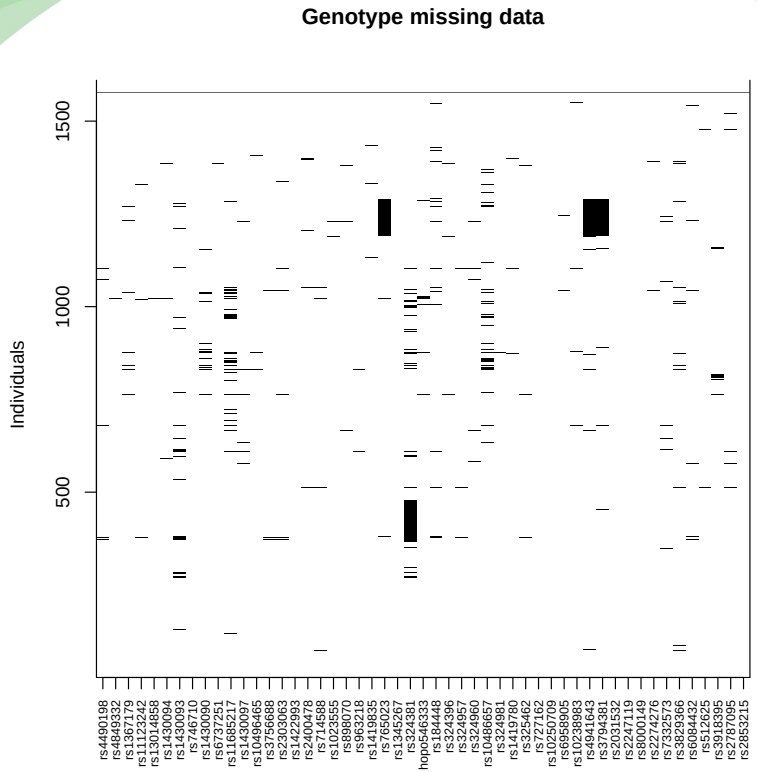
```
summary(asthma.s)
```

```
##            alleles major.allele.freq HWE      missing (%)
## rs4490198  A/G     59.2              0.174133  0.6
## rs4849332  G/T     61.8              0.522060  0.1
## rs1367179  G/C     81.4              0.738153  1.0
## rs11123242 C/T     81.7              0.932898  0.6
## rs13014858 G/A     58.3              0.351116  0.1
## rs1430094  G/A     66.9              0.305509  0.4
## rs1430093  C/A     66.6              0.817701  3.5
## rs746710   G/C     51.5              0.614368  0.0
## rs1430090  T/G     70.0              0.025180  1.6
## rs6737251  C/T     69.3              0.235996  0.3
## rs11685217 C/T     80.1              0.009462  4.5
## rs1430097  C/A     65.1              0.738166  1.0
## rs10496465 A/G     85.8              0.917997  0.6
## rs3756688  T/C     63.9              0.154632  0.6
## rs2303063  A/G     53.0              0.722069  1.1
## rs1422993  G/T     75.3              0.250093  0.0
## rs2400478  G/A     62.6              0.256786  0.9
## rs714588   A/G     54.9              0.838329  0.8
## rs1023555  T/A     76.8              0.943443  0.5
## rs898070   G/A     62.6              1.000000  0.6
```

Genetic Association analysis using R

Descriptive analysis can be performed using several generic functions

```
plotMissing(asthma.s)
```



Genetic Association analysis using R

Hardy-Weinberg equililbrium can also be tested using

```
tableHWE(asthma.s)

##            HWE (p value) flag
## rs4490198  0.1741
## rs4849332  0.5221
## rs1367179  0.7382
## rs11123242 0.9329
## rs13014858 0.3511
## rs1430094  0.3055
## rs1430093  0.8177
## rs746710   0.6144
## rs1430090  0.0252        <-
## rs6737251  0.2360
## rs11685217 0.0095        <-
## rs1430097  0.7382
## rs10496465 0.9180
## rs3756688  0.1546
## rs2303063  0.7221
## rs1422993  0.2501
## rs2400478  0.2568
## rs714588   0.8383
## rs1023555  0.9434
## rs898070   1.0000
```

Only in controls

```
tableHWE(asthma.s, casecontrol)
```

```
##            all.groups     X0     X1
## rs4490198      0.1741 0.0980 0.9116
## rs4849332      0.5221 0.6303 0.6459
## rs1367179      0.7382 1.0000 0.4920
## rs11123242     0.9329 0.9233 0.5990
## rs13014858     0.3511 0.2407 0.9119
## rs1430094      0.3055 0.0808 0.3362
## rs1430093      0.8177 1.0000 0.6260
## rs746710       0.6144 0.6490 0.8287
## rs1430090      0.0252 0.0166 0.8967
## rs6737251      0.2360 0.3141 0.5293
## rs11685217     0.0095 0.0123 0.4043
## rs1430097      0.7382 0.6156 0.9038
## rs10496465     0.9180 0.9073 1.0000
## rs3756688      0.1546 0.1093 1.0000
## rs2303063      0.7221 0.4912 0.6616
## rs1422993      0.2501 0.0509 0.2816
## rs2400478      0.2568 0.1770 0.9094
## rs714588       0.8383 0.8625 0.4441
## rs1023555      0.9434 0.9359 1.0000
## rs898070       1.0000 0.5414 0.2508
```

Genetic Association analysis using R

Association analysis for one SNP can be performed by executing

```
association(casecontrol ~ rs1422993, asthma.s)

##
## SNP: rs1422993  adjusted by:
##                  0    %   1    %   OR lower upper  p-value  AIC
## Codominant
## G/G            730 59.0 173 50.9 1.00                0.017768 1642
## G/T            425 34.3 145 42.6 1.44  1.12  1.85
## T/T             83  6.7  22  6.5 1.12  0.68  1.84
## Dominant
## G/G            730 59.0 173 50.9 1.00                0.007826 1642
## G/T-T/T        508 41.0 167 49.1 1.39  1.09  1.77
## Recessive
## G/G-G/T       1155 93.3 318 93.5 1.00                0.877863 1649
## T/T             83  6.7  22  6.5 0.96  0.59  1.57
## Overdominant
## G/G-T/T        813 65.7 195 57.4 1.00                0.005026 1641
## G/T            425 34.3 145 42.6 1.42  1.11  1.82
## log-Additive
## 0,1,2         1238 78.5 340 21.5 1.22  1.01  1.47 0.040151 1644
```

Only one mode of inheritance

```
association(casecontrol ~ rs1422993, asthma.s, model="dominant")

##
## SNP: rs1422993  adjusted by:
##             0  %   1    %   OR lower upper  p-value  AIC
## Dominant
## G/G       730 59 173 50.9 1.00             0.007826 1642
## G/T-T/T   508 41 167 49.1 1.39  1.09  1.77
```

Adjusted analysis

```
association(casecontrol ~ rs1422993 + country + smoke, asthma.s)

##
## SNP: rs1422993  adjusted by: country smoke
##                  0    %   1    %   OR lower upper p-value  AIC
## Codominant
## G/G            728 59.1 173 51.0 1.00                0.06957 1407
## G/T            423 34.3 144 42.5 1.38  1.05  1.82
## T/T             81  6.6  22  6.5 1.07  0.62  1.85
## Dominant
## G/G            728 59.1 173 51.0 1.00                0.03380 1406
## G/T-T/T        504 40.9 166 49.0 1.33  1.02  1.73
## Recessive
## G/G-G/T       1151 93.4 317 93.5 1.00                0.80821 1411
## T/T             81  6.6  22  6.5 0.94  0.55  1.60
## Overdominant
## G/G-T/T        809 65.7 195 57.5 1.00                0.02163 1406
## G/T            423 34.3 144 42.5 1.37  1.05  1.79
## log-Additive
## 0,1,2         1232 78.4 339 21.6 1.19  0.96  1.46 0.10926 1408
```

Genetic Association analysis using R

Stratified analysis

```
association(casecontrol ~ rs1422993 + strata(gender), asthma.s)
```

```
##
##        strata: Females
## SNP: rs1422993  adjusted by:
##                 0    %   1    %   OR lower upper p-value   AIC
## Codominant
## G/G           340 57.4  96 48.7 1.00             0.09208 888.0
## G/T           209 35.3  86 43.7 1.46  1.04  2.04
## T/T            43  7.3  15  7.6 1.24  0.66  2.32
## Dominant
## G/G           340 57.4  96 48.7 1.00             0.03371 886.3
## G/T-T/T       252 42.6 101 51.3 1.42  1.03  1.96
## Recessive
## G/G-G/T       549 92.7 182 92.4 1.00             0.87068 890.8
## T/T            43  7.3  15  7.6 1.05  0.57  1.94
## Overdominant
## G/G-T/T       383 64.7 111 56.3 1.00             0.03704 886.5
## G/T           209 35.3  86 43.7 1.42  1.02  1.97
## log-Additive
## 0,1,2         592 75.0 197 25.0 1.25  0.97  1.61 0.08311 887.8
##
##        strata: Males
## SNP:  adjusted by:
##                 0    %   1    %   OR lower upper p-value   AIC
## Codominant
## G/G           390 60.4  77 53.8 1.00             0.20428 749.6
## G/T           216 33.4  59 41.3 1.38  0.95  2.02
## T/T            40  6.2   7  4.9 0.89  0.38  2.05
## Dominant
## G/G           390 60.4  77 53.8 1.00             0.15263 748.8
## G/T-T/T       256 39.6  66 46.2 1.31  0.91  1.88
## Recessive
## G/G-G/T       606 93.8 136 95.1 1.00             0.54390 750.5
## T/T            40  6.2   7  4.9 0.78  0.34  1.78
## Overdominant
## G/G-T/T       430 66.6  84 58.7 1.00             0.07852 747.7
```

Subset analysis

```
association(casecontrol ~ rs1422993, asthma.s,
                subset=country=="Spain")

##
## SNP: rs1422993  adjusted by:
##                  0    %  1    %   OR lower upper p-value   AIC
## Codominant
## G/G            179 54.6 22 44.9 1.00                0.3550 295.2
## G/T            125 38.1 24 49.0 1.56  0.84  2.91
## T/T             24  7.3  3  6.1 1.02  0.28  3.66
## Dominant
## G/G            179 54.6 22 44.9 1.00                0.2059 293.7
## G/T-T/T        149 45.4 27 55.1 1.47  0.81  2.70
## Recessive
## G/G-G/T        304 92.7 46 93.9 1.00                0.7576 295.2
## T/T             24  7.3  3  6.1 0.83  0.24  2.85
## Overdominant
## G/G-T/T        203 61.9 25 51.0 1.00                0.1502 293.2
## G/T            125 38.1 24 49.0 1.56  0.85  2.85
## log-Additive
## 0,1,2          328 87.0 49 13.0 1.23  0.77  1.96  0.3816 294.5
```

Genetic Association analysis using R

Analysis of quantitative traits (stratification, adjusting, subsetting, ... also works)

```
association(bmi ~ rs1422993, asthma.s)

##
## SNP: rs1422993  adjusted by:
##                    n    me     se       dif   lower  upper p-value  AIC
## Codominant
## G/G              896 25.53 0.1446  0.000000                  0.9069 9069
## G/T              565 25.50 0.1834 -0.027059 -0.4874 0.4332
## T/T              105 25.71 0.4676  0.178076 -0.7057 1.0619
## Dominant
## G/G              896 25.53 0.1446  0.000000                  0.9818 9067
## G/T-T/T          670 25.54 0.1710  0.005089 -0.4324 0.4426
## Recessive
## G/G-G/T         1461 25.52 0.1135  0.000000                  0.6694 9067
## T/T              105 25.71 0.4676  0.188540 -0.6769 1.0540
## Overdominant
## G/G-T/T         1001 25.55 0.1383  0.000000                  0.8424 9067
## G/T              565 25.50 0.1834 -0.045739 -0.4965 0.4050
## log-Additive
## 0,1,2                             0.033951 -0.3153 0.3832  0.8489 9067
```

Genetic Association analysis using R

Analysis of multiple SNPs can be performed using

```
ans <- WGassociation(casecontrol, asthma.s)
```

Adjusted analysis can also be performed unsing `WGassociation` function

```
ans.adj <- WGassociation(casecontrol ~ country + smoke, asthma.s)
```

Fast version (only compute p-values)

```
ans.fast <- scanWGassociation(casecontrol, asthma.s)
```

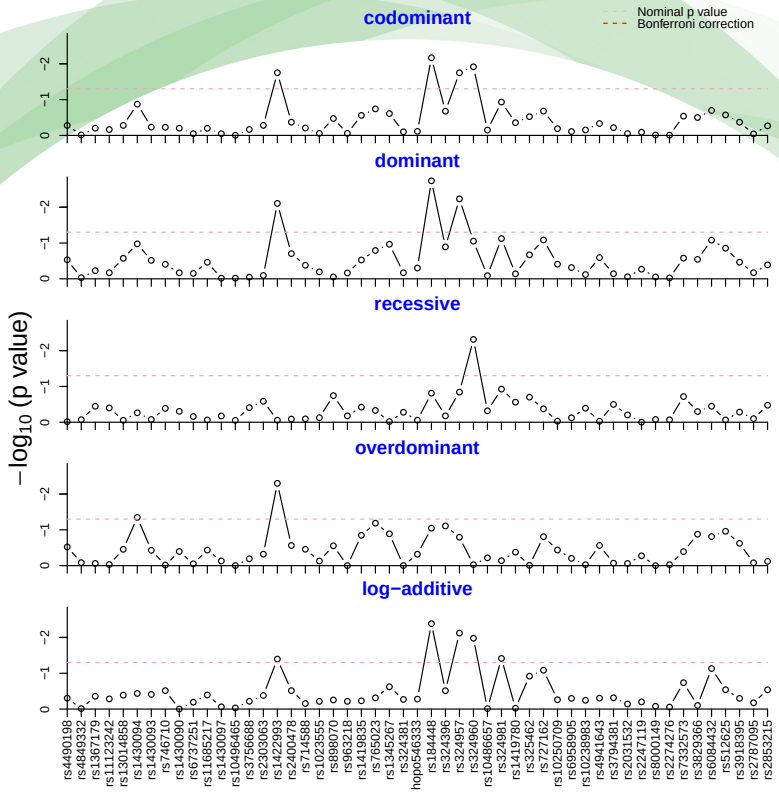
Genetic Association analysis using R

Analysis of multiple SNPs can be visualized by

```
plot(ans)
```

```
## Warning: No SNP is statistically significant after
##          Bonferroni Correction under codominant model
## Warning: No SNP is statistically significant after
##          Bonferroni Correction under dominant model
## Warning: No SNP is statistically significant after
##          Bonferroni Correction under recessive model
## Warning: No SNP is statistically significant after
##          Bonferroni Correction under overdominant model
## Warning: No SNP is statistically significant after
##          Bonferroni Correction under log-additive model
```

Genetic Association analysis using R



Genetic Association analysis using R

Complete tables can be obtained by executing

```
infoTable <- WGstats(ans)
```

The information for a given SNP is

```
infoTable$rs1422993
```

```
##
## SNP: rs1422993  adjusted by:
##                  0    %   1    %   OR lower upper  p-value  AIC
## Codominant
## G/G            730 59.0 173 50.9 1.00                0.017768 1642
## G/T            425 34.3 145 42.6 1.44  1.12  1.85
## T/T             83  6.7  22  6.5 1.12  0.68  1.84
## Dominant
## G/G            730 59.0 173 50.9 1.00                0.007826 1642
## G/T-T/T        508 41.0 167 49.1 1.39  1.09  1.77
## Recessive
## G/G-G/T       1155 93.3 318 93.5 1.00                0.877863 1649
## T/T             83  6.7  22  6.5 0.96  0.59  1.57
## Overdominant
## G/G-T/T        813 65.7 195 57.4 1.00                0.005026 1641
## G/T            425 34.3 145 42.6 1.42  1.11  1.82
## log-Additive
## 0,1,2         1238 78.5 340 21.5 1.22  1.01  1.47 0.040151 1644
```

Genetic Association analysis using R

Max-statistic can be computed by executing

```
maxstat(asthma.s, casecontrol)
```

```
##            dominant recessive log-additive MAX-statistic  Pr(>z)
## rs4490198     1.097     0.002        0.466          1.097 0.50224
## rs4849332     0.008     0.037        0.001          0.037 0.97613
## rs1367179     0.287     0.845        0.602          0.845 0.58871
## rs11123242    0.175     0.714        0.417          0.714 0.63831
## rs13014858    1.230     0.023        0.683          1.230 0.46452
## rs1430094     2.617     0.368        0.821          2.617 0.20814
## rs1430093     1.051     0.042        0.743          1.051 0.51960
## rs746710      0.728     0.679        1.051          1.051 0.51594
## rs1430090     0.172     0.463        0.000          0.463 0.74267
## rs6737251     0.143     0.156        0.217          0.217 0.86880
## rs11685217    0.894     0.030        0.705          0.894 0.56930
## rs1430097     0.003     0.183        0.029          0.183 0.88857
## rs10496465    0.003     0.020        0.008          0.020 0.98741
## rs3756688     0.016     0.738        0.266          0.738 0.62575
## rs2303063     0.060     1.271        0.658          1.271 0.45316
## rs1422993     7.073     0.024        4.291          7.073 0.01780
## rs2400478     1.662     0.056        1.055          1.662 0.35957
## rs714588      0.659     0.061        0.150          0.659 0.65880
## rs1023555     0.221     0.104        0.261          0.261 0.84650
## rs898070      0.020     1.794        0.346          1.794 0.33191
```

Haplotype Analysis

First, let's have a look at Haplotype Blocks

```
require(LDheatmap)
require(SNPassoc)
require(genetics)
data(SNPs)
ls()
```

```
##  [1] "ans"           "asthma"        "asthma.s"      "datos.s"
##  [5] "em"            "geno"          "genoH"         "haplo.score"
##  [9] "i"             "infoTable"     "mod"           "MyHeatmap"
## [13] "name.snps"     "sel"           "snpGeno"       "SNPpos"
## [17] "SNPs"          "SNPs.info.pos" "SNPs.sel"      "snpsH"
```

Genetic Association analysis using R

```
head(SNPs.info.pos)

##         snp  chr     pos
## 1 snp10001 Chr1 2987398
## 2 snp10002 Chr1 1913558
## 3 snp10003 Chr1 1982067
## 4 snp10004 Chr1  447403
## 5 snp10005 Chr1 2212031
## 6 snp10006 Chr1 2515720
```

Let's select only SNPs at Chr1

```
sel <- SNPs.info.pos$snp[SNPs.info.pos$chr=="Chr1"]
SNPs.sel <- SNPs[ , c("casco", as.character(sel))]
datos.s <- setupSNP(SNPs.sel, 2:ncol(SNPs.sel), sep="")
```

Get Physical map position (required for LD plot)

```
SNPpos <- SNPs.info.pos$pos[SNPs.info.pos$snp%in%sel]
```

Genetic Association analysis using R

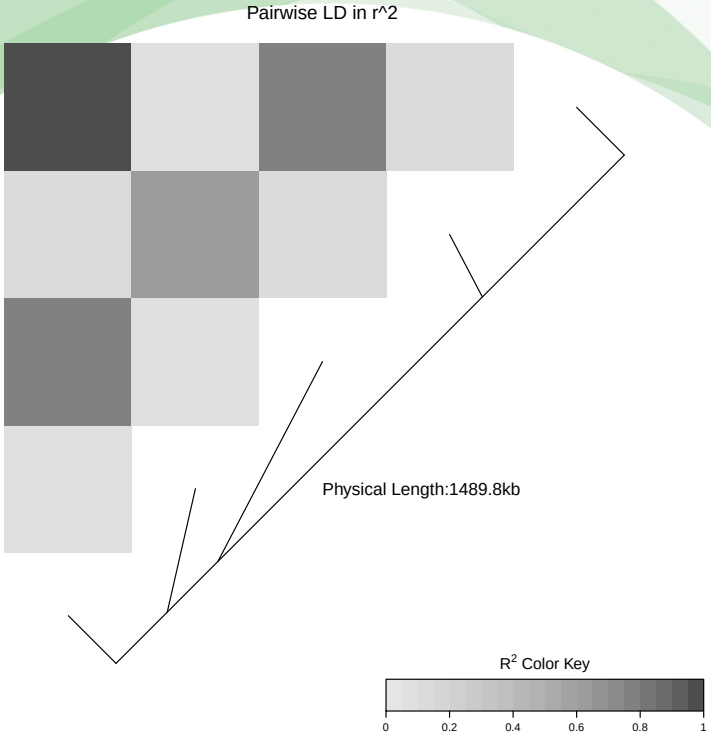
Create a data.frame with SNPs of class genotype

```
name.snps <- labels(datos.s)
snpGeno <- data.frame(lapply(datos.s[, name.snps], genotype))
```

Create LD heatmap plot

```
MyHeatmap <- LDheatmap(snpGeno, SNPpos, LDmeasure = "r",
 title = "Pairwise LD in r^2", add.map = TRUE,
 color = grey.colors(20), name = "myLDgrob", add.key = TRUE)
```

Genetic Association analysis using R



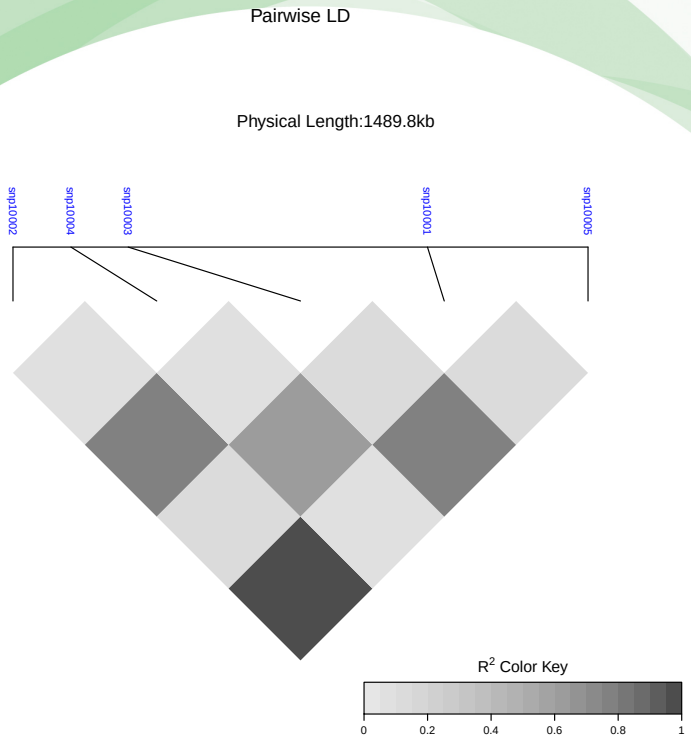
Genetic Association analysis using R

```
args(LDheatmap)

## function (gdat, genetic.distances = NULL, distances = "physical",
##     LDmeasure = "r", title = "Pairwise LD", add.map = TRUE, add.key = TRUE,
##     geneMapLocation = 0.15, geneMapLabelX = NULL, geneMapLabelY = NULL,
##     SNP.name = NULL, color = NULL, newpage = TRUE, name = "ldheatmap",
##     vp.name = NULL, pop = FALSE, flip = NULL, text = FALSE)
## NULL
```

```
LDheatmap(MyHeatmap, flip=TRUE, SNP.name=name.snps)
```

Genetic Association analysis using R



Genetic Association analysis using R

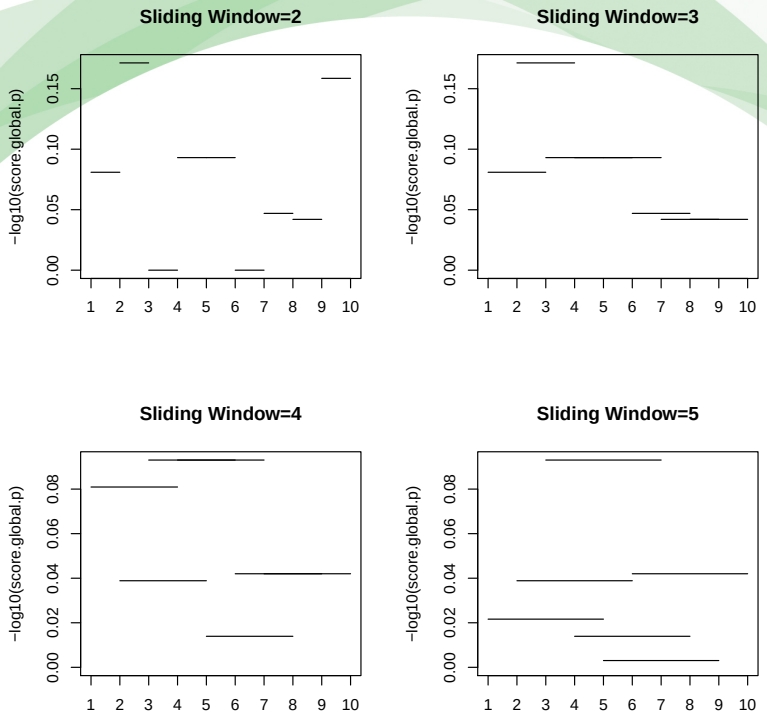
Haplotype analysis: selection of best haplotype (sliding window)

```
geno <- make.geno(datos.s, name.snps)
haplo.score <- list()
for (i in 2:5) {
   haplo.score[[i-1]] <- haplo.score.slide(datos.s$casco, geno,
                          trait.type="binomial",
                          n.slide=i,
                          simulate=TRUE,
                          sim.control=score.sim.control(min.sim=100,
                                      max.sim=200))
 }
```

Haplotype analysis: selection of best haplotype (sliding window)

```
par(mfrow=c(2,2))
for (i in 2:5) {
   plot(haplo.score[[i-1]])
   title(paste("Sliding Window=", i, sep=""))
 }
```

Genetic Association analysis using R



Haplotype analysis: OR estimates

Let's assume that the best SNP combination is:

```
snpsH <- c("snp10008", "snp10009")
genoH<-make.geno(datos.s, snpsH)
```

Haplotype estimation

```
em <- haplo.em(genoH, locus.label = snpsH, miss.val = c(0, NA))
em

## ============================================================
##                          Haplotypes
## ============================================================
##   snp10008 snp10009 hap.freq
## 1        1        1  0.51798
## 2        1        2  0.28457
## 3        2        1  0.19745
## 4        2        2  0.00000
## ============================================================
##                            Details
## ============================================================
## lnlike =  -247.8262
## lr stat for no LD =  18.91331 , df =  0 , p-val =  NA
```

Haplotype analysis: OR estimates

```
mod <- haplo.glm(datos.s$casco~genoH,
                 family="binomial",
                 locus.label=snpsH,
                 allele.lev=attributes(genoH)$unique.alleles,
                 control = haplo.glm.control(haplo.freq.min=0.05))
```

Haplotype analysis: OR estimates

```
intervals(mod)
```

```
##               freq   or   95%   C.I.   P-val
##          CA 0.5180   1.00 Reference haplotype
##          CG 0.2845   1.15 (   0.60 -   2.20 )  0.6740
##          GA 0.1975   1.00 (   0.55 -   1.81 )  0.9989
```