Molecular Epidemiology- Practicals

Raha Pazoki, MD, PhD

8 February 2019

Association anlysis of binary data (case/control)

Expected learning outcomes:

1- Students can cross-tabulate genotype and disease status.

2- Students can explain the relationship between genotypes and disease status in highly and non significant scenarios.

3- Student are able to run an association analysis and interpret the results.

4- Students can show the difference between additive and non- additive genetic association models.

5- Students can calculate minor allele frequency.

-A highly significant scenario

First generate your data including variables showing genotype status for a given single nucleotide polymorphism **(SNP)**, and case control status.

```r
p<-paste0("Participant",c(1:100000))
d<-c(rep("no",90000),rep("yes",10000))
g<-c(rep("GG",10000),rep("AG",30000),rep("AA",50000),rep("GG",8000),
     rep("AG",1500),rep("AA",500))
mydata<-data.frame(cbind(p,d,g));names(mydata)<-c("ID","disease","genotype")
head(mydata)
```

```
##              ID disease genotype
## 1 Participant1      no       GG
## 2 Participant2      no       GG
## 3 Participant3      no       GG
## 4 Participant4      no       GG
## 5 Participant5      no       GG
## 6 Participant6      no       GG
```

```r
attach(mydata)
table(disease,genotype)
```

```
##        genotype
## disease    AA    AG    GG
##     no  50000 30000 10000
##     yes   500  1500  8000
```

Calculate minor allele frequency (MAF)

First count the number of A alleles and G alleles in the whole sample.

```r
table(genotype)
```

```
## genotype
##    AA    AG    GG
## 50500 31500 18000
```

```
A<-(table(genotype)[1]*2 ) + (table(genotype)[2]*1)
G<-(table(genotype)[3]*2 ) + (table(genotype)[2]*1)
(min(c(A,G))/sum(c(A,G)))*100
```

```
## [1] 33.75
```

Recode for additive effect

Count number of G allele

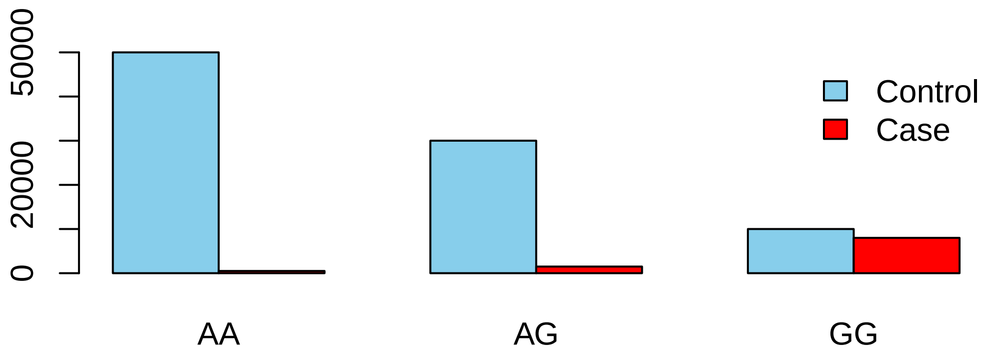
```
mydata$count_G_allele[mydata$genotype=="AA"]<-0
mydata$count_G_allele[mydata$genotype=="AG"]<-1
mydata$count_G_allele[mydata$genotype=="GG"]<-2
attach(mydata)
```

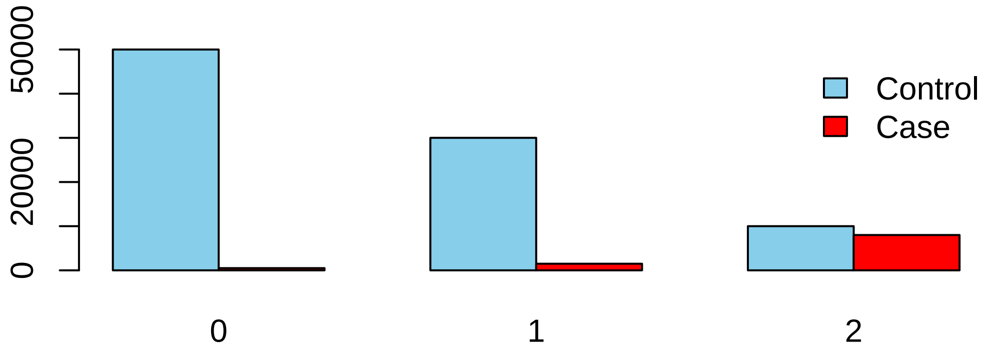
```
## The following objects are masked from mydata (pos = 3):
##
##     disease, genotype, ID
```

```
table(disease,count_G_allele)
```

```
##        count_G_allele
## disease     0     1     2
##     no  50000 30000 10000
##     yes   500  1500  8000
```

What do you infer?





Note the `G` allele ! Proportion of cases increases compare to controls when participants carry more number of G alleles.

Now run association analysis using generalized linear model (glm):

```
resultsg<-glm(disease~as.character(genotype),data=mydata,family="binomial")

summary(  resultsg)
```

```
##
## Call:
## glm(formula = disease ~ as.character(genotype), family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0842  -0.3124  -0.1411  -0.1411   3.0381
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -4.60517    0.04494 -102.47   <2e-16 ***
## as.character(genotype)AG  1.60944    0.05215   30.86   <2e-16 ***
## as.character(genotype)GG  4.38203    0.04738   92.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65017  on 99999  degrees of freedom
## Residual deviance: 42402  on 99997  degrees of freedom
## AIC: 42408
##
## Number of Fisher Scoring iterations: 7
```

Can you find effect estimate and P values?

What do you infer? Do you know what the reference group is?

Now check the results of association analysis for additive model:

```
results<-glm(disease~count_G_allele,data=mydata,family="binomial")
summary(  results)
```

```
##
## Call:
## glm(formula = disease ~ count_G_allele, family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0642  -0.3561  -0.1059  -0.1059   3.2206
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.18062    0.03800  -136.3   <2e-16 ***
## count_G_allele  2.45419    0.02197   111.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 65017  on 99999  degrees of freedom
## Residual deviance: 42662  on 99998  degrees of freedom
## AIC: 42666
##
## Number of Fisher Scoring iterations: 7
```

```
summary(results)$coefficient
```

```
##                 Estimate Std. Error   z value Pr(>|z|)
## (Intercept)    -5.180621 0.03799763 -136.3406        0
## count_G_allele  2.454192 0.02197024  111.7053        0
```

What is the difference with nonn additive model? What difference does it makes in interpration?

Now calculate Odds ratio:

```
exp(summary(results)$coefficient[2,1])
```

```
## [1] 11.63703
```

How big is this odds ratio? Do you know what it means?

-A non significant scenario (Opional) Now let's use different distribution of alleles

```
rm(list=ls())
p<-paste0("participant",c(1:10000))
d<-c(rep("no",9000),rep("yes",1000))
g<-c(rep("AA",2330),rep("AG",4330),rep("GG",3340),rep("AA",233),rep("AG",433),rep("GG",334))
mydata<-data.frame(cbind(p,d,g));names(mydata)<-c("ID","disease","genotype")
```

```
## Warning in cbind(p, d, g): number of rows of result is not a multiple of
## vector length (arg 1)
```

```
head(mydata)
```

```
##             ID disease genotype
## 1 participant1      no       AA
## 2 participant2      no       AA
## 3 participant3      no       AA
## 4 participant4      no       AA
## 5 participant5      no       AA
## 6 participant6      no       AA
```

Count number of G allele

```
mydata$count_G_allele[mydata$genotype=="AA"]<-0
mydata$count_G_allele[mydata$genotype=="AG"]<-1
mydata$count_G_allele[mydata$genotype=="GG"]<-2
attach(mydata)
```

```
## The following objects are masked from mydata (pos = 3):
##
##     count_G_allele, disease, genotype, ID
```

```
## The following objects are masked from mydata (pos = 4):
##
##     disease, genotype, ID
```

```
table(disease,genotype)
```

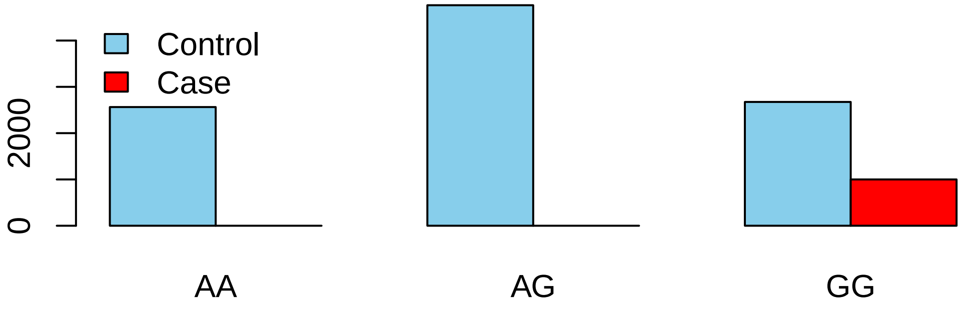
```
##        genotype
## disease   AA   AG   GG
##     no  2563 4763 2674
##     yes    0    0 1000
```

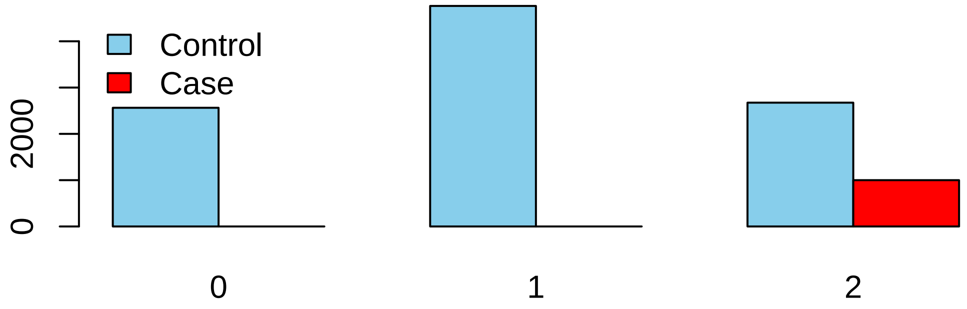
What do you infer? Waht is the difference with the highly significant scenario?

```
table(disease,count_G_allele)
```

```
##        count_G_allele
## disease    0    1    2
##     no  2563 4763 2674
##     yes    0    0 1000
```

Look at the distribution plot and think of the meaning of it.





Do you think the genotypes make any difference in disease status?

Run association analysis using generalized linear model (glm)

```
resultsg<-glm(disease~as.character(genotype),data=mydata,family="binomial")
summary(  resultsg)
```

```
##
## Call:
## glm(formula = disease ~ as.character(genotype), family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.79713  -0.00005  -0.00005  -0.00005   1.61325
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.057e+01  3.502e+02  -0.059    0.953
## as.character(genotype)AG  7.568e-11  4.343e+02   0.000    1.000
## as.character(genotype)GG  1.958e+01  3.502e+02   0.056    0.955
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6702.0  on 10999  degrees of freedom
## Residual deviance: 4301.7  on 10997  degrees of freedom
## AIC: 4307.7
##
## Number of Fisher Scoring iterations: 19
```

```
results<-glm(disease~count_G_allele,data=mydata,family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(  results)
```

```
##
## Call:
```

```
## glm(formula = disease ~ count_G_allele, family = "binomial",
##     data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.79713  -0.00005  -0.00005   0.00000   1.61325
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -40.22     523.07  -0.077    0.939
## count_G_allele    19.62     261.53   0.075    0.940
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6702.0  on 10999  degrees of freedom
## Residual deviance: 4301.7  on 10998  degrees of freedom
## AIC: 4305.7
##
## Number of Fisher Scoring iterations: 20
```

```
summary(results)$coefficient
```

```
##                 Estimate Std. Error     z value Pr(>|z|)
## (Intercept)    -40.21994   523.0667 -0.07689256 0.938709
## count_G_allele  19.61818   261.5333  0.07501216 0.940205
```

Can you explain why the results are not statistically significant?