Introduction to genetic analisys using R

Yurii Aulchenko, Najaf Amin

March 19, 2007

In the first part, you will be guided step by step through simple genetic analysis exercise using a small example data set. In the second part, you will investigate a bigger data set as based on the knowledge obtained in the first part, and will answer the questions.

Start R by double-click on the file ge03d1p1.RData. Load library genetics, which we will need for testing Hardy-Weinberg equilibrium (HWE) and computations of Linkage Disequilibrium (LD) and library dgc.genetics, which we will need for association analysis by typing

```
> library(dgc.genetics)
```

1 Example session

The file you have loaded contains two data frames. A data frame is an R term for a data table. In such tables, it is usually assumed that rows correspond to subjects (observations) and columns correspond to variables.

You can see the names of the loaded objects by using the command ls():

```
> ls()

[1] "example"
```

You can see that there is a single object loaded, which is a data frames, as could be seeing from

```
> class(example)

[1] "data.frame"
```

We will investigate the data presented in the example data frame. To see what variables are measured, use command names():

```
> names(example)

[1] "subj" "sex"  "aff"  "qt"   "snp4" "snp5" "snp6"
```

The 7 variables correspond to the personal ID, sex, affection status, quantitative trait qt and several SNPs.

You can explore the raw data contained in a data frame by using fix() command (e.g. fix(example)). However, normally this is not necessary.

First, let us check how many cases and controls are presented in the data set. To access some variable var in a data frame frame, you can use syntax frame$var:

```
> example$aff

  [1] 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0
 [38] 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0
 [75] 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0
[112] 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1
[149] 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1
[186] 1 1 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0
[223] 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0
```

which shows the vector of values of `aff`.

The function `table(x)` produces a frequency table for the variable `x`. Thus, we can use

```
> table(example$aff)

  0   1
194  56
```

which tells us that there are 56 cases and 194 controls in this data set.

A more convenient way to access data presented in a data frame is through "attaching" it to the `R` search path by

```
> attach(example)
```

After that, the variables can be accessed directly, e.g.

```
> table(aff)

aff
  0   1
194  56
```

The summary statistics on the distribution of a variable can be obtained by `summary()` function. For example, for the quantitative traits `qt`

```
> summary(qt)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.7240 -0.7503 -0.1447 -0.1192  0.4819  2.8660
```

Tip: `summary()` is quite useful function working with a range of data objects. Try `summary(example)`.

You can also draw a histogram of the distribution by

```
> hist(qt)
```

The resulting graph is presented in figure 1.

To see the allelic frequencies and other summary statistics for a SNP, you can use

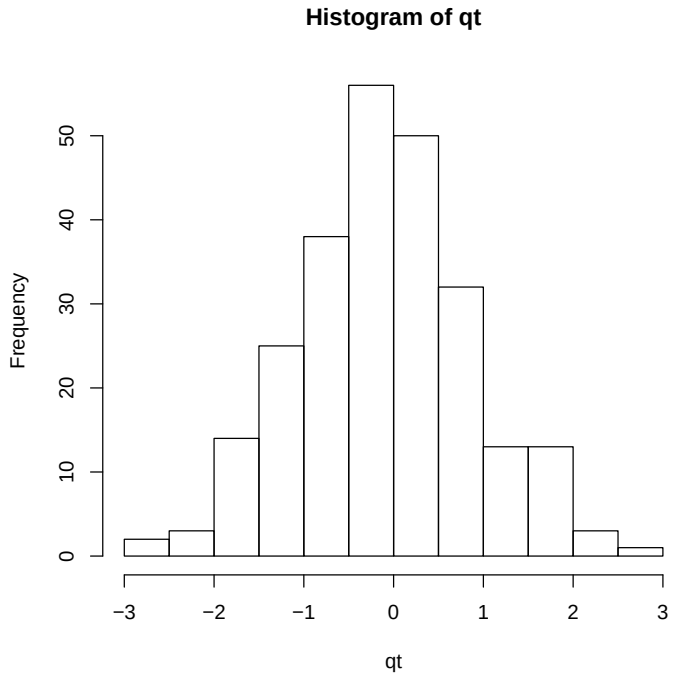

Figure 1: Histogram of the variable `qt`

```
> summary(snp4)

Number of samples typed: 243 (97.2%)

Allele Frequency: (2 alleles)
   Count Proportion
A    323       0.66
B    163       0.34
NA    14         NA


Genotype Frequency:
    Count Proportion
A/A   109       0.45
A/B   105       0.43
B/B    29       0.12
NA      7         NA

Heterozygosity (Hu)  = 0.4467269
Poly. Inf. Content   = 0.3464355
```

Tip: on R command line pressing the "up-arrow" button makes the last typed command re-appear (pressing it one more time will bring you to the one before the last, so on). This is very handy when you have to repeat the same analysis of different variables

To check these characteristics in controls and cases separately, you can use

```
> summary(snp4[aff == 0])

Number of samples typed: 190 (97.9%)

Allele Frequency: (2 alleles)
   Count Proportion
A    255       0.67
B    125       0.33
NA     8         NA


Genotype Frequency:
    Count Proportion
A/A    87       0.46
A/B    81       0.43
B/B    22       0.12
NA      4         NA

Heterozygosity (Hu)  = 0.4426469
Poly. Inf. Content   = 0.3440288

> summary(snp4[aff == 1])

Number of samples typed: 53 (94.6%)

Allele Frequency: (2 alleles)
   Count Proportion
A     68       0.64
B     38       0.36
NA     6         NA


Genotype Frequency:
    Count Proportion
A/A    22       0.42
A/B    24       0.45
B/B     7       0.13
NA      3         NA

Heterozygosity (Hu)  = 0.4643306
Poly. Inf. Content   = 0.3541731
```

Let us check if HWE holds for the SNPs described in this data frame. We can do exact test for HWE by

```
> HWE.exact(snp4)

        Exact Test for Hardy-Weinberg Equilibrium

data:  snp4
N11 = 109, N12 = 105, N22 = 29, N1 = 323, N2 = 163, p-value = 0.666
```

If you want to check HWE using controls only, you can do it by

```
> HWE.exact(snp4[aff == 0])

        Exact Test for Hardy-Weinberg Equilibrium

data:  snp4[aff == 0]
N11 = 87, N12 = 81, N22 = 22, N1 = 255, N2 = 125, p-value = 0.6244
```

Let us check if the there is LD between snp4 and snp5:

```
> LD(snp4, snp5)

Pairwise LD
-----------
                   D        D'      Corr
Estimates: 0.2009042 0.9997352 0.8683117

              X^2 P-value   N
LD Test: 354.3636       0 235
```

The output shows results of the test for significance of LD, and estimates of the magnitude of LD (D' and correlation, r). To obtain r^2, you can either square the correlation manually

```
> 0.8683117 * 0.8683117

[1] 0.7539652
```

or simply ask `LD()` to report it by

```
> LD(snp4, snp5)$"R^2"

[1] 0.7539652
```

Tip: the latter command is possible because the `LD()` function actually computes more things than it reports. This is quite common for `R` functions. You can apply `names()` function to the analysis objects to see (at least part of) what was actually computed. Try

```
> ld45 <- LD(snp4, snp5)
```

and check what are the sub-objects contained in this analysis object

```
> names(ld45)
```

```
[1] "call"    "D"       "D'"      "r"       "R^2"     "n"       "X^2"
[8] "P-value"
```

Any of these variables can be accessed through `object$var` syntax, e.g. to check D' we can use

```
> ld45$"D'"

[1] 0.9997352
```

To check LD for more that two SNPs, we can compute an LD analysis object by

```
> ldall <- LD(data.frame(snp4, snp5, snp6))
```

and later check

```
> ldall$"P-value"
```

```
     snp4 snp5 snp6
snp4   NA    0    0
snp5   NA   NA    0
snp6   NA   NA   NA
```

to see significance,

```
> ldall$"D'"
```

```
     snp4      snp5      snp6
snp4   NA 0.9997352 0.8039577
snp5   NA        NA 0.9997231
snp6   NA        NA        NA
```

for D' and

```
> ldall$"R^2"
```

```
     snp4      snp5      snp6
snp4   NA 0.7539652 0.5886602
snp5   NA        NA 0.8278328
snp6   NA        NA        NA
```

for r^2.

You can also present e.g. r^2 matrix as a plot by

```
> image(ldall$"R^2")
```

A more neat way to present it requires specification of the set of threshold (break points) and colors to be used (you do not need to try this example if you do not want):

```
> image(ldall$"R^2", breaks = c(0.5, 0.6, 0.7, 0.8, 0.9, 1), col = heat.colors(5))
```

Resulting plot is shown at figure 2.

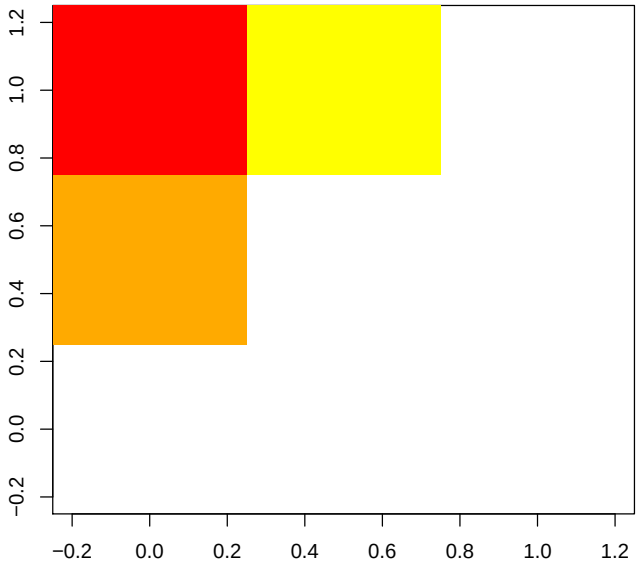

Figure 2: r^2 plot for snp4, snp5 and snp6

Tip: for any `R` command, you can get help by typing `help(command)`. Try `help(image)` if you are interested to understand what are "breaks" and "col"; or try `help(heat.colors)` to figure this color schema out.

Similar to our HWE checks, we may want to compute (and compare) LD in cases and controls separately:

```
> ldcases <- LD(data.frame(snp4, snp5, snp6)[aff == 1, ])
> ldcases$"R^2"

     snp4      snp5      snp6
snp4   NA 0.7615923 0.6891558
snp5   NA        NA 0.8943495
snp6   NA        NA        NA

> ldcontr <- LD(data.frame(snp4, snp5, snp6)[aff == 0, ])
> ldcontr$"R^2"

     snp4      snp5      snp6
snp4   NA 0.7512458 0.5616395
snp5   NA        NA 0.8075894
snp6   NA        NA        NA
```

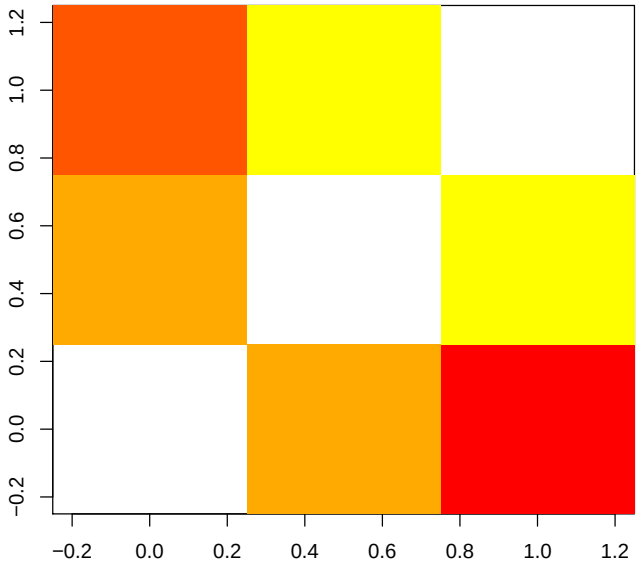

Figure 3: r^2 plot for snp4, snp5 and snp6. Above diagonal: LD in cases; below: controls

and even present it results for cases and controls on the same graph (you do not need to produce this graph, which is presented at the figure 3):

```
> image(ldcases$"R^2", breaks = c(0.5, 0.6, 0.7, 0.8, 0.9, 1),
+     col = heat.colors(5))
> image(t(ldcontr$"R^2"), breaks = c(0.5, 0.6, 0.7, 0.8, 0.9, 1),
+     col = heat.colors(5), add = T)
```

Now, after we have described genetic and phenotypic data separately, we are ready to test association between these two. First, we will investigate relation between the quantitative trait `qt` and the SNPs by using linear regression

```
> mg <- lm(qt ~ snp4)
> summary(mg)

Call:
lm(formula = qt ~ snp4)

Residuals:
     Min       1Q   Median       3Q      Max
-2.63700 -0.62291 -0.01225  0.58922  3.05561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.081114   0.092517  -0.877    0.382
snp4A/B     -0.108366   0.132079  -0.820    0.413
snp4B/B     -0.006041   0.201820  -0.030    0.976

Residual standard error: 0.9659 on 240 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-Squared: 0.003049,        Adjusted R-squared: -0.005259
F-statistic: 0.367 on 2 and 240 DF,  p-value: 0.6932
```

It is clear that the model assumes arbitrary (estimated) effects of the genotypes AA, AB and BB. Neither effect of AB nor BB is significant in this case. The global test on two degrees of freedom (bottom of the output) is also not significant.

If you want to include some covariate into your model, e.g. sex, you can easily do that by adding the term to the formula:

```
> summary(lm(qt ~ sex + snp4))

Call:
lm(formula = qt ~ sex + snp4)

Residuals:
      Min        1Q    Median        3Q       Max
-2.645225 -0.618989 -0.001171  0.587321  3.076356

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -1.485e-01  1.427e-01   -1.040    0.299
sex          1.307e-01  2.107e-01    0.620    0.536
snp4A/B     -1.042e-01  1.324e-01   -0.787    0.432
snp4B/B      6.436e-05  2.023e-01 0.000318    1.000

Residual standard error: 0.9671 on 239 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-Squared: 0.004651,        Adjusted R-squared: -0.007843
F-statistic: 0.3723 on 3 and 239 DF,  p-value: 0.7731
```

You can also allow for interaction by using the "*" operator

```
> summary(lm(qt ~ sex * snp4))

Call:
lm(formula = qt ~ sex * snp4)

Residuals:
      Min        1Q    Median        3Q       Max
-2.633100 -0.628752  0.008546  0.614107  3.042126

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2771     0.1856  -1.493    0.137
```

```
sex           0.3803     0.3119   1.219    0.224
snp4A/B       0.1286     0.2599   0.495    0.621
snp4B/B       0.2189     0.3929   0.557    0.578
sex:snp4A/B  -0.4652     0.4481  -1.038    0.300
sex:snp4B/B  -0.4422     0.7043  -0.628    0.531

Residual standard error: 0.9688 on 237 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-Squared: 0.009593,        Adjusted R-squared: -0.0113
F-statistic: 0.4591 on 5 and 237 DF,  p-value: 0.8064
```

Note that both main effects of sex and snp4, and also effects of interaction are estimated in this model.

We can also test the additive model, which assumes that the deviation from AA (reference) to BB is twice the deviation to AB. In other words, the mean value of the trait for heterozygous genotypes is right in between the two homozygotes. To test the additive model you first need to specify "additive" contrasts for the SNP:

```
> gcontrasts(snp4) <- "additive"
```

Now, running `logit()` produces a test for additive effect: ma <- lm(qt snp4) summary(ma)

You can revert to the original contrasts model for the snp4 by

```
> gcontrasts(snp4) <- "genotype"
```

To test association with a binary outcome, we will use `logit` function from `dgc.genetics`:

```
> logit(aff ~ snp4)

Logistic regression:  aff ~ snp4

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

              OR     Lower    Upper    z-test  P-value
snp4A/B 1.171717 0.6099236 2.250972 0.4757324 0.634265
snp4B/B 1.258264 0.4766694 3.321441 0.4638853 0.642730
```

To make a test of global significance of the SNP effect, you can use

```
> anova(logit(aff ~ snp4), test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: aff

Terms added sequentially (first to last)
```

```
      Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    242    254.908
snp4   2    0.329       240    254.579     0.848
```

To test the additive model, use

```
> gcontrasts(snp4) <- "additive"
> logit(aff ~ snp4)

Logistic regression:  aff ~ snp4

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

               OR    Lower    Upper    z-test   P-value
snp4:a:B 1.135596 0.728091 1.771177 0.5607114 0.5749943
```

When using the `logit()` function, you can allow for additional covariates and interactions in the same way as you did with linear regression using `lm()` function.

Now you have learned all commands necessary to answer the questions of the next section.

Exit R by typing `q()` command (do not save image) and and proceed to the self exercise.

2 Exercise

Start `R` by double-click over the file `ge03d1p2.RData`. Explore the data frame present and answer the questions.

Question 1 *How many SNPs are described in this data frame?*

Question 2 *What is the frequency (proportion) of snp1 allele A? What is its frequency in these affected (aff==1)?*

Question 3 *How many cases and controls are present?*

Question 4 *If all subjects are used to test HWE, are there any SNPs out of HWE at nominal $P \leq 0.05$? Which ones?*

Question 5 *If only controls are used to test the SNPs which are out of HWE in total sample, are these still out of HWE?*

Question 6 *Which SNP pairs are in strong LD ($r^2 \geq 0.8$)?*

Question 7 *For SNPs in strong LD, what is r^2 for separate samples of cases and controls?*

Question 8 *Is there significant association between affection status and sex? What is P-value for association?*

Question 9 *Is association between the disease and `qt` significant?*

Question 10 *Which SNPs are associated with the quantitative trait `qt` at nominal $P \leq 0.05$? Use 2 d.f. test.*

Question 11 *Test each SNP for association with the affection status, using 2 d.f. test. Which SNPs are significantly associated at nominal $P \leq 0.05$? How can you describe the model of action of the significant SNPs?*

Question 12 *For the SNPs selected in previous question, test association using additive model. Which SNPs are still associated?*

Question 13 *If you adjust the analysis under additive model (question 12) for significant covariates which you discovered in questions 8 and 9, are these findings still significant?*

Question 14 *Test association between `aff` and snp5 and snp10, allowing for the SNPs interaction effect. Use arbitrary (not an additive) model. Do you observe significant interaction? How can you describe the model of concert action of snp5 and snp10?*

3 Answers

Q.1 : How many SNPs are described in this data frame?

```
> attach(popdat)

        The following object(s) are masked from example :

         aff qt sex snp4 snp5 snp6 subj

> names(popdat)

 [1] "subj"  "sex"   "aff"   "qt"    "snp1"  "snp2"  "snp3"  "snp4"  "snp5"
[10] "snp6"  "snp7"  "snp8"  "snp9"  "snp10"
```

The answer is 10 snps

Q.2 : What is the frequency (proportion) of snp1 allele A? What is its frequency in these affected (aff==1)?

```
> summary(snp1)

Number of samples typed: 2374 (95%)

Allele Frequency: (2 alleles)
   Count Proportion
A   3462       0.73
B   1286       0.27
NA   252         NA


Genotype Frequency:
    Count Proportion
A/A  1287       0.54
A/B   888       0.37
B/B   199       0.08
NA    126         NA

Heterozygosity (Hu)  = 0.3950646
Poly. Inf. Content   = 0.3169762
```

The frequency of A in all subjects is 0.73.

```
> summary(snp1[aff == 1])

Number of samples typed: 519 (94.5%)

Allele Frequency: (2 alleles)
   Count Proportion
A    729        0.7
B    309        0.3
NA    60         NA
```

```
Genotype Frequency:
    Count Proportion
A/A   258       0.50
A/B   213       0.41
B/B    48       0.09
NA     30         NA

Heterozygosity (Hu)  = 0.4185428
Poly. Inf. Content   = 0.3307192
```

The frequency of A in affected subjects is 0.7.

Q.3 : How many cases and controls are present?

```
> table(aff)

aff
   0    1
1951  549
```

There are 549 cases and 1951 controls.

Q.4 : If all subjects are used to test HWE, are there any SNPs out of HWE at nominal $P \leq 0.05$? Which ones?

```
> HWE.exact(snp1)

        Exact Test for Hardy-Weinberg Equilibrium

data:  snp1
N11 = 1287, N12 = 888, N22 = 199, N1 = 3462, N2 = 1286, p-value =
0.01083
```

...

```
> HWE.exact(snp10)

        Exact Test for Hardy-Weinberg Equilibrium

data:  snp10
N11 = 1792, N12 = 552, N22 = 40, N1 = 4136, N2 = 632, p-value = 0.7897
```

Only SNP 1 is out of HWE in the total sample.

Q.5 : If only controls are used to test the SNPs which are out of HWE in total sample, are these still out of HWE?

```
> HWE.exact(snp1[aff == 0])

        Exact Test for Hardy-Weinberg Equilibrium

data:  snp1[aff == 0]
N11 = 1029, N12 = 675, N22 = 151, N1 = 2733, N2 = 977, p-value =
0.008393
```

Yes, SNP 1 is out of HWE also in controls.

Q.6 : Which SNP pairs are in strong LD ($r^2 \geq 0.8$)?

```
> LD(popdat[, 5:14])$"R^2"
```

```
      snp1  snp2  snp3  snp4  snp5  snp6  snp7  snp8  snp9 snp10
snp1    NA 0.016 0.235 0.206 0.258 0.227 0.152 0.117 0.090 0.000
snp2    NA    NA 0.004 0.004 0.005 0.004 0.000 0.000 0.000 0.000
snp3    NA    NA    NA 0.602 0.457 0.346 0.641 0.031 0.042 0.001
snp4    NA    NA    NA    NA 0.803 0.650 0.729 0.027 0.037 0.002
snp5    NA    NA    NA    NA    NA 0.874 0.586 0.034 0.046 0.002
snp6    NA    NA    NA    NA    NA    NA 0.670 0.030 0.040 0.002
snp7    NA    NA    NA    NA    NA    NA    NA 0.020 0.027 0.003
snp8    NA    NA    NA    NA    NA    NA    NA    NA 0.002 0.000
snp9    NA    NA    NA    NA    NA    NA    NA    NA    NA 0.001
snp10   NA    NA    NA    NA    NA    NA    NA    NA    NA    NA
```

SNP pairs 4-5 and 5-6 have $r^2 \geq 0.8$.

Q.7 : For SNPs in strong LD, what is r^2 for separate samples of cases and controls?

For controls,

```
> LD(data.frame(snp4, snp5, snp6)[aff == 0, ])$"R^2"
```

```
     snp4     snp5      snp6
snp4   NA 0.806591 0.6419715
snp5   NA       NA 0.8661005
snp6   NA       NA        NA
```

For cases,

```
> LD(data.frame(snp4, snp5, snp6)[aff == 1, ])$"R^2"
```

```
     snp4      snp5      snp6
snp4   NA 0.7951475 0.6773275
snp5   NA        NA 0.9083237
snp6   NA        NA        NA
```

Note that the fact that LD is higher in cases may mean nothing because the estimates of LD are biased upwards with smaller sample sizes. For example in a small sample (5 people) of controls we expect even higher LD because of strong upward bias:

```
> LD(popdat[which(aff == 0)[1:5], 8:10])$"R^2"
```

```
     snp4      snp5      snp6
snp4   NA 0.9995876 0.9995876
snp5   NA        NA 0.9995876
snp6   NA        NA        NA
```

More elaborate methods, such as that by Zaykin, are required to contrast LD between sample of unequal size.

Q.8 : Is there significant association between affection status and sex? What is P-value for association?

```
> logit(aff ~ sex)

Logistic regression:  aff ~ sex

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

          OR    Lower    Upper   z-test    P-value
sex 1.444688 1.045393 1.996497 2.228963 0.02581635
```

There is significant ($P = 0.03$) association.

Q.9 : Is association between the disease and `qt` significant?

```
> logit(aff ~ qt)

Logistic regression:  aff ~ qt

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

          OR     Lower    Upper     z-test   P-value
qt 0.9751773 0.8865446 1.072671 -0.5170283 0.6051364
```

There is no significant ($P = 0.6$) association.

Q.10 : Which SNPs are associated with the quantitative trait `qt` at nominal $P \leq 0.05$? Use 2 d.f. test.

```
> summary(lm(qt ~ snp1))

Call:
lm(formula = qt ~ snp1)

Residuals:
     Min       1Q   Median       3Q      Max
-3.52609 -0.66427 -0.01110  0.67648  3.54622

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02846    0.02758  -1.032   0.3022
snp1A/B      0.08200    0.04316   1.900   0.0575 .
snp1B/B      0.18644    0.07536   2.474   0.0134 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9893 on 2371 degrees of freedom
  (126 observations deleted due to missingness)
Multiple R-Squared: 0.00335,        Adjusted R-squared: 0.002509
F-statistic: 3.985 on 2 and 2371 DF,  p-value: 0.01873

...
```

```
> summary(lm(qt ~ snp10))

Call:
lm(formula = qt ~ snp10)

Residuals:
      Min        1Q    Median        3Q       Max
-3.586464 -0.677484  0.001935  0.673270  3.412527

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01915    0.02344   0.817    0.414
snp10A/B     0.01277    0.04829   0.264    0.792
snp10B/B     0.17178    0.15860   1.083    0.279

Residual standard error: 0.9921 on 2381 degrees of freedom
  (116 observations deleted due to missingness)
Multiple R-Squared: 0.0005072,        Adjusted R-squared: -0.0003324
F-statistic: 0.6041 on 2 and 2381 DF,  p-value: 0.5467
```

SNPs 1, 4, 5 an 9 are significantly associated at nominal $P \leq 0.05$.

Q.11 : Test each SNP for association with the affection status, using 2 d.f. test. Which SNPs are significantly associated at nominal $P \leq 0.05$? How can you describe the model of action of the significant SNPs?

```
> x <- logit(aff ~ snp5)
> x

Logistic regression:  aff ~ snp5

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

              OR    Lower    Upper   z-test     P-value
snp5A/A 1.235176 0.940558 1.622080 1.519212 0.128709107
snp5B/B 1.403072 1.124687 1.750364 3.001367 0.002687707

> anova(x, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: aff

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    2382    2440.40
snp5    2     9.24      2380    2431.16      0.01
```

...

```
> x <- logit(aff ~ snp10)
> x

Logistic regression:  aff ~ snp10

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

                OR     Lower    Upper     z-test    P-value
snp10A/B 1.3376929 1.0695740 1.673023  2.5493546 0.01079225
snp10B/B 0.8350447 0.3664215 1.902999 -0.4289534 0.66795715

> anova(x, test = "Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: aff

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     2383    2475.13
snp10    2     6.73      2381    2468.39      0.03
```

The SNPs 5 an 10 are significantly associated at $P \leq 0.05$. The model of action of SNP5 can be ddescribed as recessive (while the risk for AA and AB is not significantly different, there is 1.4 times increased risk for these homozygous for BB). The SNP 10 demonstrates somewhat weird action with the risk increased in heterozygous AB individuals. However, the confidence interval for BB is large and therefore we can not claim that BB is not increasing the risk.

Q.12 : For the SNPs selected in previous question, test association using additive model. Which SNPs are still associated?

```
> gcontrasts(snp5) <- "additive"
> logit(aff ~ snp5)

Logistic regression:  aff ~ snp5

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

                OR     Lower    Upper    z-test   P-value
snp5:a:A 0.8964111 0.7765197 1.034813 -1.492841 0.1354787

> gcontrasts(snp10) <- "additive"
> logit(aff ~ snp10)
```

```
Logistic regression:  aff ~ snp10

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

                OR   Lower    Upper   z-test    P-value
snp10:a:B 1.218450 1.00014 1.484412 1.961389 0.04983367
```

Only SNP 10 is significantly associated under the additive model.

Q.13 : If you adjust the analysis under additive model (question 12) for significant covariates which you discovered in questions 8 and 9, are these findings still significant?

```
> logit(aff ~ sex + snp10)

Logistic regression:  aff ~ sex + snp10

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

                OR    Lower    Upper   z-test    P-value
sex       1.453497 1.040060 2.031281 2.190016 0.02852308
snp10:a:B 1.222662 1.003450 1.489764 1.994160 0.04613457
```

Yes, SNP 10 becomes even a bit more significantly associated after adjusting for sex.

Q.14 : Test association between `aff` and snp5 and snp10, allowing for the SNPs interaction effect. Use arbitrary (not an additive) model. Do you observe significant interaction? How can you describe the model of concert action of snp5 and snp10?

```
> gcontrasts(snp5) <- "genotype"
> gcontrasts(snp10) <- "genotype"
> logit(aff ~ snp5 * snp10)

Logistic regression:  aff ~ snp5 * snp10

Odds ratios (1 unit change), lower and upper confidence limits, and tests:

                         OR      Lower      Upper      z-test      P-value
snp5A/A           0.6583495 0.44728351  0.9690143 -2.11960143 3.403967e-02
snp5B/B           1.3971228 1.07526418  1.8153233  2.50317688 1.230840e-02
snp10A/B          0.9860685 0.68953030  1.4101353 -0.07687059 9.387265e-01
snp10B/B          0.8608534 0.29134395  2.5436212 -0.27105727 7.863470e-01
snp5A/A:snp10A/B 4.4091743 2.32051700  8.3777958  4.53036148 5.888285e-06
snp5B/B:snp10A/B 1.1387038 0.66501059  1.9498132  0.47334571 6.359666e-01
snp5A/A:snp10B/B 2.2784250 0.32752451 15.8498682  0.83211257 4.053454e-01
snp5B/B:snp10B/B 0.7515445 0.06730847  8.3915024 -0.23201853 8.165236e-01
```

It appears that SNP10 genotype is only relevant in these who are homozygous for the low-risk A allele at the SNP5; in such cases SNP 10 allele B is

risk increasing. In these homozygous for SNP 5 A, we observe highly significant increase in risk for heterozygotes for SNP10 and increased (though not significantly) risk for SNP 10 BB.