Genetic association analysis

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

jan.graffelman@upc.edu

December 10, 2019

Contents

1. Introduction
2. Allele based tests
3. Genotype based tests
4. Multiple polymorphisms
5. Computer exercise

Genetic association studies

Goal:

- Investigate associations between markers and a trait (disease).

Designs:

- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

We will focus on population-based association studies

Preliminaries

- The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous).
- $Y_i = 1$ individual i has the trait, $Y_i = 0$, individual i does not have the trait.
- The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

Levels of analysis

- Tests of association at the level of alleles
  - We are sampling alleles
  - Alleles assumed to be independent
  - Rely on the Hardy-Weinberg equilibrium assumption
  - Chi-square test of the alleles by trait cross table
  - Fisher exact test of the alleles by trait cross table
  - Test on the odds ratio of the alleles by trait cross table
- Tests of association at the level of the genotypes
  - We are sampling individuals
  - Hardy-Weinberg equilibrium assumption is not needed
  - Co-dominant, dominant and recessive Chi-square tests
  - Cochran-Armitage trend test
  - Logistic regression

The data table

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

The alleles test

- Let p be the allele frequency of the A allele.

$$\begin{cases} H_0 : p_{cases} = p_{controls} \\ H_1 : p_{cases} \neq p_{controls} \end{cases}$$

- The test assumes Hardy-Weinberg equilibrium
- The test is a χ^2 test for independence in a 2×2 table of alleles.

	a	A	Total	$\hat{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/(2r)$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/(2s)$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/(2n)$

The alleles test

- Chi-square test for independence

$$X^2 = \sum_{i,j}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Expected count e_{ij} = total row $i \times$ total colum $j/$total of table
- If H_0 is true, then $X^2 \sim \chi^2_1$
- p-value $= P\left(\chi^2_1 \geq X^2\right)$

Example alleles test

A polymorphism in the Dopamine receptor is supposed to be involved in Schizophrenia. In a case-control study, the following data were obtained:

	11	12	22	Total
Cases	7	69	57	133
Controls	20	56	33	109
Total	27	125	90	242

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

Example alleles test

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

	1	2	Total
Cases	98.38	167.62	266
Controls	80.62	137.38	218
Total	179	305	484

$$X^2 = \frac{(83 - 98.38)^2}{98.38} + \cdots + \frac{(122 - 137.38)^2}{137.38} = 8.4671$$

$$\text{p-value} = P\left(\chi^2_1 \leq 8.4671\right) = 0.0036$$

R code alleles test

```
> X <- matrix(c(7,69,57,20,56,33),byrow=TRUE,ncol=3)
> colnames(X) <- c("11","12","22")
> rownames(X) <- c("Cases","Controls")
> X
         11 12 22
Cases     7 69 57
Controls 20 56 33
> Y <- cbind(2*X[,1]+X[,2],2*X[,3]+X[,2])
> colnames(Y) <- c("1","2")
> Y
          1   2
Cases    83 183
Controls 96 122
> chisq.test(Y,correct=FALSE)

Pearson’s Chi-squared test

data:  Y
X-squared = 8.4671, df = 1, p-value = 0.003616
```

Fisher's Exact test

- Often used for cross tables with low counts in the margin, or when $e_{ij} < 5$.
- If the margins are considered fixed, the probability of the table can be calculated, using the hypergeometric distribution.
- The exact p-value is the sum of the probabilities of all possible tables with the same margins that have a probability that is less or equal than the observed table.

For the same data:

```
> Y
          1   2
cases    83 183
controls 96 122
> fisher.test(Y)

Fisher's Exact Test for Count Data

data:  Y
p-value = 0.00448
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3903016 0.8512261
sample estimates:
odds ratio
 0.5770451
```

Odds ratio test for alleles

- Definition of odds

$$\text{Odds} = \frac{P\left(\text{success}\right)}{P\left(\text{failure}\right)} = \frac{P\left(\text{disease}\right)}{P\left(\text{no disease}\right)} = \frac{p}{1-p}$$

- The Odds ratio (OR) compares the odds of the disease for the two alleles:

$$OR = \frac{\text{Odds of disease with A allele}}{\text{Odds of disease with B allele}}$$

	A	B
Cases	n_{11}	n_{12}
Controls	n_{21}	n_{22}

$$OR = \frac{\left(n_{11}/n_{21}\right)}{\left(n_{12}/n_{22}\right)} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

- An odds ratio based test assumes an additive model: AA doubles the risk of AB.
- OR $= 1$ corresponds to independence; OR > 1 or OR < 1 implies association.
- Known result:

$$V\left(\ln\left(OR\right)\right) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

- Allows calculation of confidence intervals for the OR.

Odds ratio test for alleles

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

$$OR = \frac{83 \cdot 122}{96 \cdot 183} = 0.5764$$

$$se_{ln(OR)} = \sqrt{\frac{1}{83} + \frac{1}{183} + \frac{1}{96} + \frac{1}{122}} = 0.1900$$

$$CI(\text{True ln(OR)}) = \ln(OR) \pm z_{\alpha/2} se_{ln(OR)}$$

$$CI(\text{True OR}) = e^{\ln(OR) \pm z_{\alpha/2} se_{ln(OR)}}$$

$$CI(\text{True OR}) = e^{\ln(0.5764) \pm 1.96 \cdot 0.1900} = (0.397; 0.837)$$

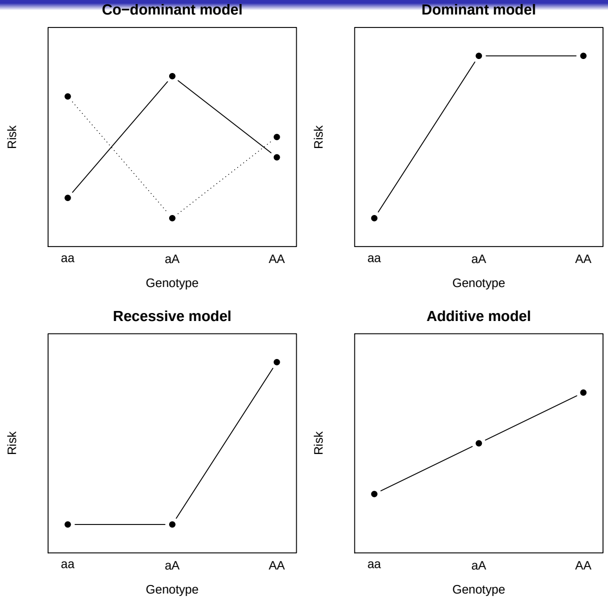
The data table

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

We can test for association using different genetic models:

- A codominant model
- A dominant model
- A recessive model
- An additive model

Genetic association models


Co−dominant model

Dominant model

Recessive model

Additive model

Codominant test

- We test the null hypothesis of no effect of the marker on the trait.
- Formally:

$$\begin{cases} H_0 : P(Y=1|AA) = P(Y=1|Aa) = P(Y=1|aa) \\ H_1 : \text{At least one pair different} \end{cases}$$

- Test statistic

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_2$
- The test makes no assumptions about the relationship between genotype and trait.
- Under H_1, each genotype can have a different disease rate.
- The test can reject the null if the data support heterozygote advantage (overdominance).

Example codominant test

TNF genotype (G/A polymorphism) in a study on acne patients and controls

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

R code codominant test

```
> X <- matrix(c(66,43,4,99,15,0),byrow=TRUE,ncol=3)
> colnames(X) <- c("GG","GA","AA")
> rownames(X) <- c("Acne","Contro")
> X
       GG GA AA
Acne   66 43  4
Contro 99 15  0
> results <- chisq.test(X)
Warning message:
In chisq.test(X) : Chi-squared approximation may be incorrect
> print(results)

Pearson's Chi-squared test

data:  X
X-squared = 24.1133, df = 2, p-value = 5.806e-06

> results$expected
             GG       GA       AA
Acne   82.13656 28.87225 1.991189
Contro 82.86344 29.12775 2.008811

> fisher.test(X)

Fisher's Exact Test for Count Data

data:  X
p-value = 1.97e-06
alternative hypothesis: two.sided
```

Dominant test

- Columns in the original table are combined to produce 2 × 2 tables.
- Dominant model:

	aa	aA or AA	Total
Cases	r_0	$r_1 + r_2$	r
Controls	s_0	$s_1 + s_2$	s
Total	n_0	$n_1 + n_2$	n

- Test:

$$\begin{cases} H_0 : \text{Disease does not depend on presence of A} \\ H_1 : \text{Disease does depend on the presence of A} \end{cases}$$

- Statistic:

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_1$

R code dominant test

```
> Y <- cbind(X[,1],X[,2]+X[,3])
> colnames(Y) <- c("GG","GA or AA")
> rownames(Y) <- c("Acne","Control")
> Y
        GG GA or AA
Acne    66       47
Control 99       15
> results <- chisq.test(Y)
> print(results)

Pearson's Chi-squared test with Yates' continuity correction

data:  Y
X-squared = 21.7021, df = 1, p-value = 3.184e-06

> results <- chisq.test(Y,correct=FALSE)
> print(results)

Pearson's Chi-squared test

data:  Y
X-squared = 23.1122, df = 1, p-value = 1.528e-06

>
```

Recessive test

- Recessive model:

	aa or aA	AA	Total
Cases	$r_0 + r_1$	r_2	r
Controls	$s_0 + s_1$	s_2	s
Total	$n_0 + n_1$	n_2	n

- Test:

$$\begin{cases} H_0 : \text{Disease does not depend on being homozygote AA} \\ H_1 : \text{Disease does depend on being homozygote AA} \end{cases}$$

- Statistic:

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_1$

The additive genetic model

- Basic idea: disease risk increases as a function of the number of alleles (0,1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - Cochran-Armitage trend test

Armitage trend test

- The trend test is based on the linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

- X is the disease status (0 or 1)
- Y is the number of A alleles (0, 1 or 2)
- Tests $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$
- Armitage trend test statistic

$$A = \frac{\hat{\beta}_1^2}{V\left(\hat{\beta}_1\right)} = n \cdot r_{xy}^2$$

- Under H_0, $A \sim \chi_1^2$.

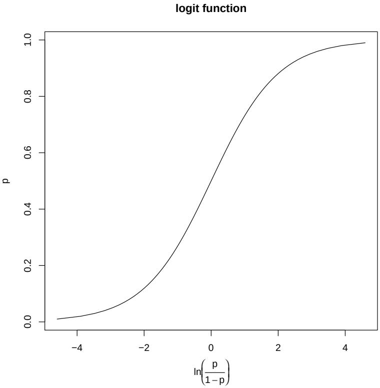
Example Armitage trend test

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

$$A = 227 \cdot (0.3253)^2 = 24.02$$

$$P\left(\chi_1^2 \geq 24.02\right) = 9.49e - 07$$

Logistic regression



Logit (or logistic) function:

$$logit(\pi) = \frac{\pi}{1-\pi}$$

Inverse of the logit function

$$logit^{-1}(\pi) = \frac{e^{\pi}}{e^{\pi}+1}$$

Using logit(π) as the response is the basis of logistic regression

The logistic regression model

$$\pi(x) = E\left(Y|x\right) = P\left(Y = 1|x\right)$$

$$y = \pi(x) + \varepsilon \quad y \sim Bin(n, \pi(x))$$

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x$$

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{e^{\beta_0 + \beta_1 x} + 1}$$

Note that

- $0 \leq \pi(x) \leq 1$
- $-\infty \leq g(x) \leq +\infty$

Model and likelihood

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right]\}$$

We maximize $l(\beta_0, \beta_1)$ by numerical methods

Odds ratios and logistic regression with genetic predictors

- One genotype is the reference genotype (e.g. AA)
- Of interest are the odds ratios

$$OR_{BB} = \frac{\text{Odss disease for a BB person}}{\text{Odds disease AA person}}$$

$$OR_{AB} = \frac{\text{Odss disease for a AB person}}{\text{Odds disease AA person}}$$

- These OR are estimated by logistic regression.
- Logistic regression is attractive as it allows to adjust for covariates.
- Model

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_{AB} I_{AB} + \beta_{BB} I_{BB}$$

- $OR_{AB} = e^{\hat{\beta}_{AB}}$ and $OR_{BB} = e^{\hat{\beta}_{BB}}$

Example logistic regression in R

```
Cases    <- c(MM=149,Mm=269,mm=91)
Controls <- c(MM=153,Mm=325,mm=180)

cas <- rep(c("MM","Mm","mm"),Cases)
con <- rep(c("MM","Mm","mm"),Controls)

ncas <- length(cas)
ncon <- length(con)

y <- c(rep(1,ncas),rep(0,ncon))
x <- factor(c(cas,con))

out.lm <- glm(y˜x, family = binomial(link = "logit"))
summary(out.lm)

or <- exp(coefficients(out.lm))
```

Example logistic regression in R

```
> out.lm <- glm(y~x, family = binomial(link = "logit"))
> summary(out.lm)

Call:
glm(formula = y ~ x, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1662  -1.0982  -0.9046   1.2587   1.4773

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6821     0.1286  -5.303 1.14e-07 ***
xMm           0.4930     0.1528   3.227 0.001251 **
xMM           0.6556     0.1726   3.798 0.000146 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1598.7  on 1166  degrees of freedom
Residual deviance: 1582.7  on 1164  degrees of freedom
AIC: 1588.7

Number of Fisher Scoring iterations: 4

> or <- exp(coefficients(out.lm))
> or
(Intercept)         xMm         xMM
  0.5055556   1.6371936   1.9263090
```

Multiple polymorphisms

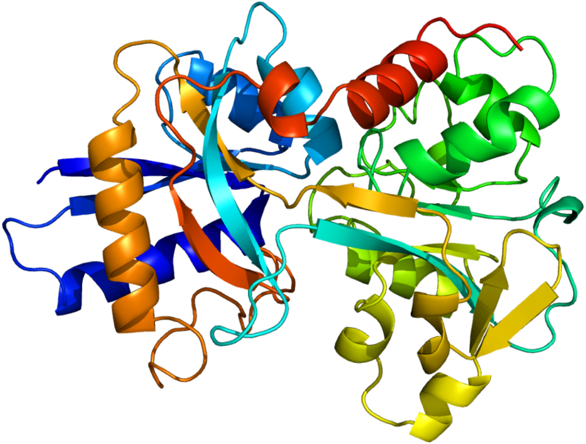
Presented methods so far have focused on a single genetic variant

How to deal with multiple SNPs?

- Multiple regression models (for a small amount of SNPs)
- Regression with haplotypes
- Test all variants: Genome wide association studies (GWAS)
- ....

In the following, we illustrate a GWAS for transferrin

Transferrin

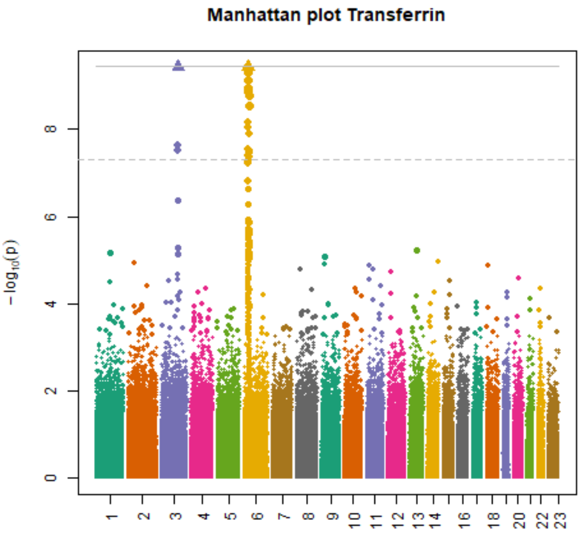


Protein responsible for iron transport

Analysis

- 2,362 individuals for which (adjusted) transferrin serum levels are available.
- 281,313 SNPs from all chromosomes.
- Filters: missing rate < 0.01; MAF > 0.05; HWE p-value > 0.001.
- We use an additive model for each SNP and fit this model with PLINK.

Results

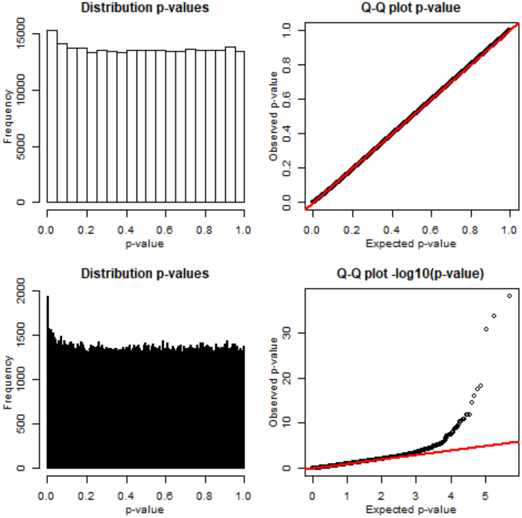

Manhattan plot Transferrin
$-\log_{10}(p)$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
23

The top 25

```
          SNP CHR        BP       B      SE      R2       T         P -log10(P)
1   rs3811647   3 134966719  0.3832 0.02889 0.06936  13.260 8.965e-39 38.047450
2   rs6794945   3 135001153  0.3652 0.02940 0.06136  12.420 2.324e-34 33.633764
3   rs1800562   6  26201120 -0.5884 0.04988 0.05572 -11.800 2.968e-31 30.527536
4  rs13214703   6  28049366 -0.4378 0.04886 0.03292  -8.961 6.390e-19 18.194499
5   rs1358024   3 134966878  0.3290 0.03745 0.03168   8.785 2.941e-18 17.531505
6   rs2274089   6  25596562 -0.3791 0.04551 0.02856  -8.330 1.352e-16 15.869023
7   rs4525863   3 134918826  0.2399 0.03017 0.02609   7.951 2.845e-15 14.545918
8   rs1867503   3 134893338  0.2039 0.02864 0.02103   7.120 1.428e-12 11.845272
9   rs1867504   3 134893351  0.2039 0.02864 0.02103   7.120 1.428e-12 11.845272
10  rs9853615   3 135002671 -0.2083 0.02929 0.02098  -7.111 1.523e-12 11.817300
11 rs12216125   6  26105437 -0.1974 0.02891 0.01936  -6.826 1.107e-11 10.955852
12  rs9379818   6  26131185 -0.1931 0.02838 0.01925  -6.805 1.276e-11 10.894149
13 rs13194984   6  26608542 -0.2719 0.04060 0.01865  -6.698 2.638e-11 10.578725
14   rs932316   6  25749179 -0.2371 0.03557 0.01849  -6.664 3.309e-11 10.480303
15 rs17270561   6  25928418 -0.2183 0.03292 0.01829  -6.631 4.108e-11 10.386370
16  rs2013063   6  26102077 -0.1798 0.02823 0.01690  -6.369 2.285e-10  9.641114
17  rs1543680   6  26211156 -0.2036 0.03259 0.01627  -6.247 4.944e-10  9.305922
18 rs10484432   6  26116855 -0.2013 0.03256 0.01595  -6.183 7.390e-10  9.131356
19  rs2009610   6  26075047 -0.1959 0.03205 0.01558  -6.111 1.158e-09  8.936291
20   rs707889   6  26203910 -0.1969 0.03238 0.01543  -6.082 1.383e-09  8.859178
21  rs1029328   6  28555894 -0.2509 0.04150 0.01526  -6.047 1.709e-09  8.767258
22 rs11757000   6  28592848 -0.2307 0.03868 0.01486  -5.966 2.806e-09  8.551912
23   rs169219   6  26065371  0.1669 0.02870 0.01413   5.816 6.861e-09  8.163613
24  rs7748771   6  25463078 -0.2678 0.04645 0.01389  -5.765 9.249e-09  8.033905
25  rs3130253   6  29741991 -0.2769 0.04845 0.01365  -5.715 1.238e-08  7.907279
```

Significant polymorphisms in the TF gene on CHR 3

Distribution of the p-values



Some statistical concerns

- Effect of filters applied?
- Multiple testing problem?
- X-chromosome adequately dealt with?
- Family structure accounted for?
- Adjustment for covariates?
- Power?

Computer exercise

- A particular SNP is supposed to be involved in Alzheimer's disease. A case control study has been performed, obtaining the following results:

	MM	Mm	mm
Cases	112	278	150
Controls	206	348	150

- Perform the alleles test for this data set.
- Perform Armitage trend test for this data set.
- Plot the risk of disease as a function of the number of *m* alleles. Fit a linear model and add the regression line to the plot. Test the null hypothesis $\beta_1 = 0$.
- Is there evidence for association of this marker with the disease?
- Also test for association using a codominant, a dominant and a recessive model.
- Which model seems most appropriate?
- Estimate odds ratios using logistic regression

Bibliography

- Laird, N.M. & Lange, C. (2011) The fundamentals of modern statistical genetics. Springer.