



H3ABioNet

Pan African Bioinformatics Network for H3Africa

Statistical models used for Genome-Wide Association Studies (GWAS)

Cheikh LOUCOUBAR

Group for Biostatistics, Bioinformatics and Modeling
Institut Pasteur de Dakar

Outline

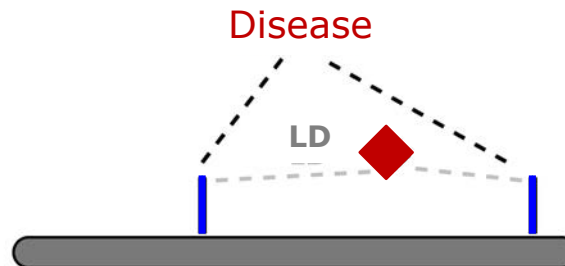
- Brief recap on types of GWAS studies
- Types of association testing models and their use
- Covariates and testing / correcting for confounding
- P values and correcting for multiple testing
- Manhattan plots and other visualizations
- Tools used for GWAS testing (e.g Plink)
- Files, formats and outputs for GWAS testing

Association Studies

Direct association



Indirect association



Statistical Association

Definition

- Any relationship between two measured quantities that shows them statistically dependent (correlation)

Main GWAS characteristics

- Large sample size requirement
- No a priori knowledge
- Raise several statistical issues

Issues & Considerations

- Management of large datasets
- Data quality control
- Controlling for confounding
 - Sex, age
 - Correlation with other variables
- Population stratification → **PCA**
- Linkage disequilibrium → LD pruning
- Cryptic relatedness → Mixed Models including **GRM**
- Statistical issues (multiple testing, e.g., **FWER should be controlled !!!**)
 - Bonferroni, **FDR**

There are several types of GWAS ...

Depending on parameters like:

Recruitment design

- Population-based, Case/Control
- Family-based (related individuals)

Nature of the measured phenotype

- Qualitative, usually binary (affected / not affected)
- Quantitative

Complexity of genetic effect tested

- Single-Marker effect: One marker analyzed at a time (univariate methods)
- Multi-Marker effect: Joint markers analysis (multivariate methods, testing combined effect of different markers from a same gene, from selected regions or at the Genome-Wide level)

Types of association testing models

Recruitment

- Population-based, Case/Control
- Family-based (related individuals)

Nature of the phenotype

- Qualitative, usually binary (affected / not affected)
- Quantitative

Complexity of genetic effect tested

- Single-Marker effect
- Multi-Marker effect

Association statistics : Chi-squared

	Cases	Controls
AA	c_0	t_0
Aa	c_1	t_1
aa	c_2	t_2

Contingency table of
observed counts (O_{ij})

Test statistic for association is X defined as:

$$X = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2 \text{ with df}=2$$

- Conclude for association if the corresponding P-value < significance threshold

Association statistics : Chi-squared

	Cases	Controls	<i>Tot_c</i>
AA	20	50	70
Aa	20	30	50
aa	60	20	80
<i>Tot_l</i>	100	100	200

Expected cell count = $(Tot_l \times Tot_c) / Total$

Expected **AA** cases = $100 \times 70 / 200 = 35$

Expected **Aa** cases = $100 \times 50 / 200 = 25$

... etc.

Association statistics : Chi-squared

	Cases	Controls
AA	20	50
Aa	20	30
aa	60	20

(O_{ij})

	Cases	Controls
AA	35	35
Aa	25	25
aa	40	40

(E_{ij})

$$X = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$X = [(20 - 35)^2/35] + [(20 - 25)^2/25] + [(60 - 40)^2/40] + [(50 - 35)^2/35] + [(30 - 25)^2/25] + [(20 - 40)^2/40]$$

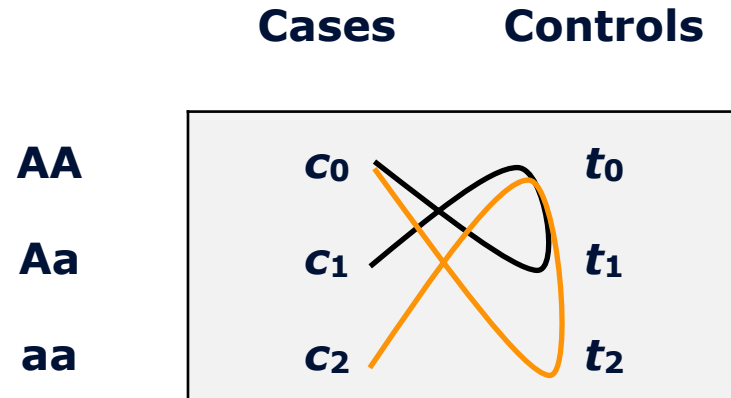
X = 34.86 , a chi-squared with **2 df**

P.value = 2.7e-08

Other association statistics

- **Used test statistics**
 - Fisher's exact test, Cochran-Armitage trend test
 - Gold Standard— **Fischer's exact test** (for case/control)

Association measure : Odds-ratio (OR)



Odds ratio for **Aa** vs. **AA** = $(c_1 \times t_0) / (c_0 \times t_1)$

Odds ratio for **aa** vs. **AA** = $(c_2 \times t_0) / (c_0 \times t_2)$

Association measure : Odds-ratio

	Cases	Controls
AA	20	50
Aa	20	30
aa	60	20

Aa vs. **AA**: $OR = (20 \times 50) / (20 \times 30) = \mathbf{1.66}$

aa vs. **AA**: $OR = (60 \times 50) / (20 \times 20) = \mathbf{7.5}$

... Optimize coding scheme

Genotypic analysis – dominance effect

"a" dominant

	Cases	Controls
AA	C_0	t_0
Aa or aa	C_1+C_2	t_1+t_2

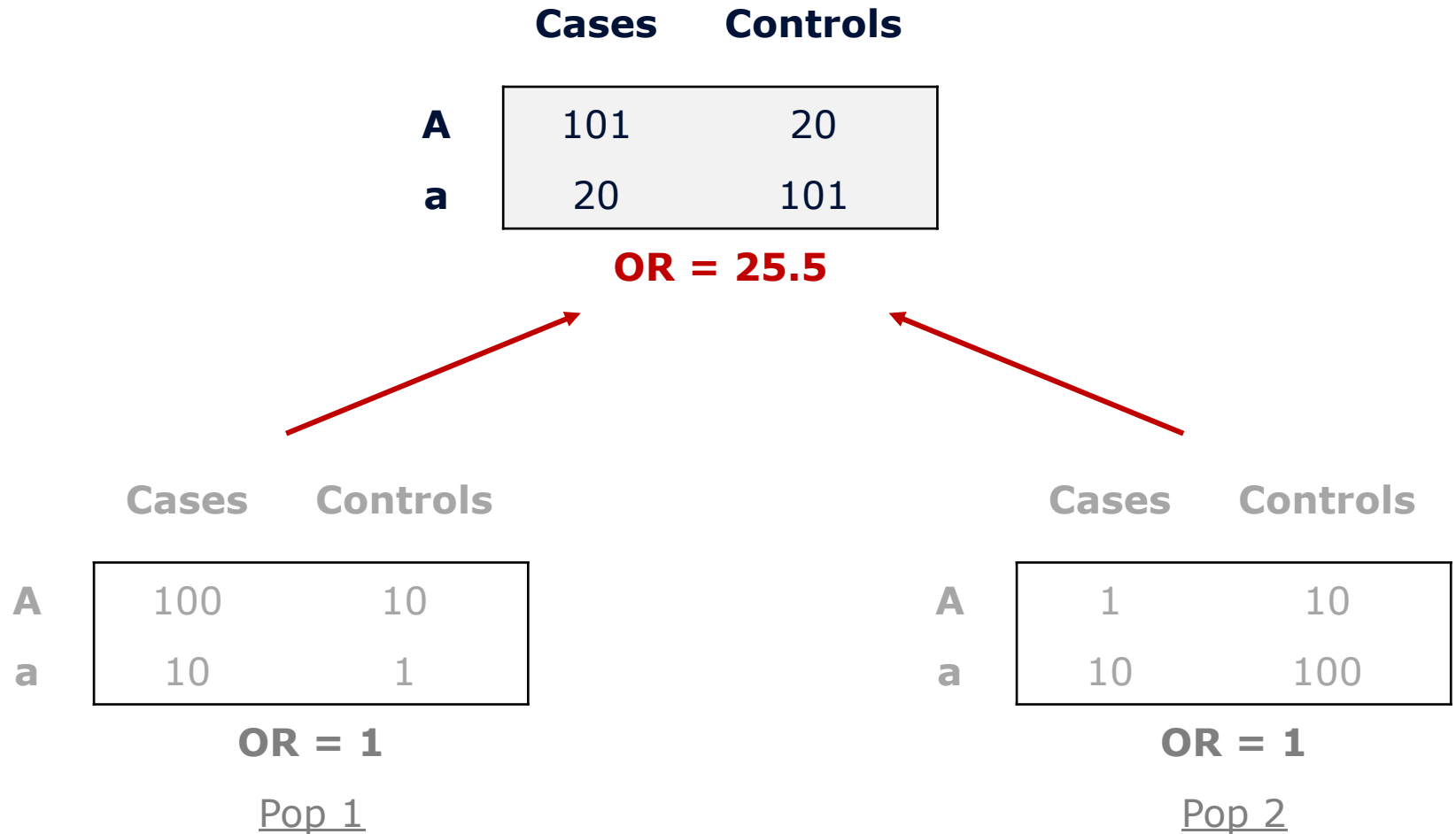
"a" recessive

	Cases	Controls
AA or Aa	C_0+C_1	t_0+t_1
aa	C_2	t_2

Test statistic for association is a χ^2 with $df=1$

	Cases	Controls	OR	P-value
AA	20	50	-	
Aa or aa	80	50	4.0	8.7e-06

Population stratification effect



To generalize: Logistic regression

$$\log(Y=1 / Y=0) = \alpha + \beta \text{ SNP} + \varepsilon$$

Convenience of regression models:

- We can adjust for covariates (to correct for confounding)
- Explanatory variables can be quantitative

Test statistic for association is:

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim \text{Standard Normal } \mathbf{N(0; 1)}$$

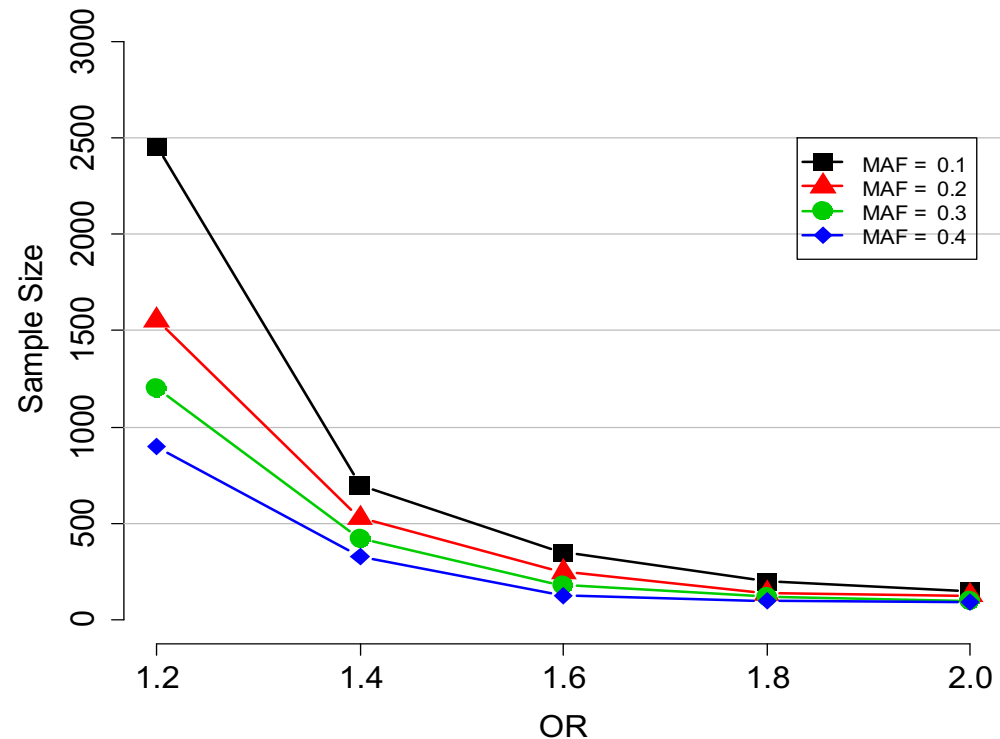
$\mathbf{Z^2 = X}$, the chi-squared statistics from the contingency table

Required sample size

Log-additive model, Power of 80%, $\alpha = 0.05$

Sample size required in a case/control study according to:

- SNP effect (Odd Ratio)
- Allele frequency



Types of association testing models

Recruitment

- Population-based, Case/Control
- Family-based (related individuals)

Nature of the phenotype

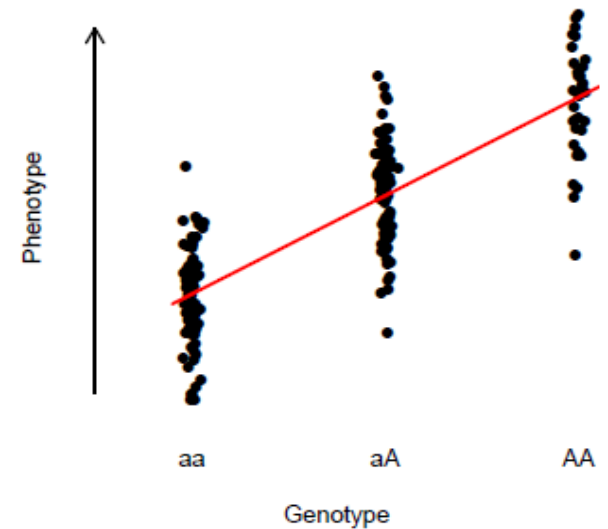
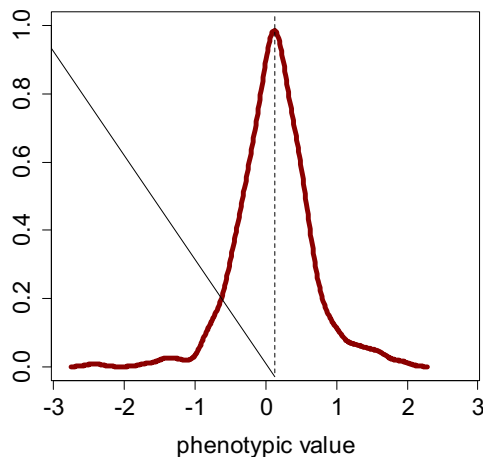
- Qualitative, usually binary (affected / not affected)
- Quantitative

Complexity of genetic effect tested

- Single-Marker effect
- Multi-Marker effect

Association statistics : Z-score from linear regression

Example of Quantitative Traits: blood glucose levels, BMI
- Usually Gaussian



e.g., the profile for a negative association between "a" allele and the trait

Univariate methods:

- Generalized linear models
- ANOVA

Classical GWAS model

Linear regressions on each genetic marker

Repetition for $i = 1, \dots$, number of SNPs analyzed

$$Y = \beta_{0i} + \beta_{1i} \text{SNP}_i + \beta_{2i} \text{PC1} + \beta_{3i} \text{PC2} + \varepsilon_i$$

Approach

- Individual test statistics
- Correction for multiple testing
 - Bonferroni
 - Gold Standard— FDR

Types of association testing models

Recruitment

Population-based, Case/Control

Family-based (related individuals)

Nature of the phenotype

Qualitative, usually binary (affected / not affected)

Quantitative

Type of genetic effect tested

Single-Marker effect

Multi-Marker effect

Combining summary statistics to test joint markers effect

Example of method combining individual test statistics

(main challenge: determine the theoretical or empirical distribution of the resulting test statistic)

REPORT

A Versatile Gene-Based Test for Genome-wide Association Studies

Jimmy Z. Liu,^{1,*} Allan F. Mcrae,¹ Dale R. Nyholt,¹ Sarah E. Medland,¹ Naomi R. Wray,¹
Kevin M. Brown,² AMFS Investigators,³ Nicholas K. Hayward,¹ Grant W. Montgomery,¹
Peter M. Visscher,¹ Nicholas G. Martin,¹ and Stuart Macgregor^{1,*}

We have derived a versatile gene-based test for genome-wide association studies (GWAS). Our approach, called VEGAS (*versatile gene-based association study*), is applicable to all GWAS designs, including family-based GWAS, meta-analyses of GWAS on the basis of summary data, and DNA-pooling-based GWAS, where existing approaches based on permutation are not possible, as well as singleton data, where they are. The test incorporates information from a full set of markers (or a defined subset) within a gene and accounts for linkage disequilibrium between markers by using simulations from the multivariate normal distribution. We show that for an associa-

Software: VEGAS (with R and Plink dependency)

Types of association testing models

Recruitment

- Population-based, Case/Control
- Family-based (related individuals)

Nature of the phenotype

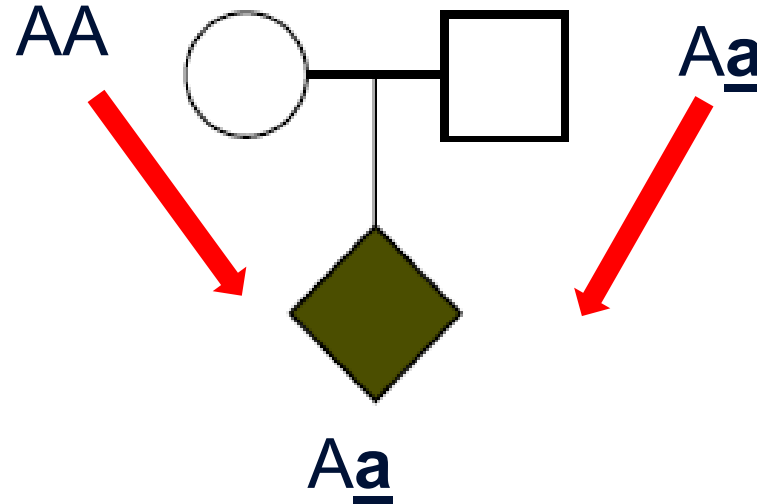
- Qualitative, usually binary (affected / not affected)
- Quantitative

Complexity of genetic effect tested

- Single-Marker effect
- Multi-Marker effect

Transmission Disequilibrium Test (TDT)

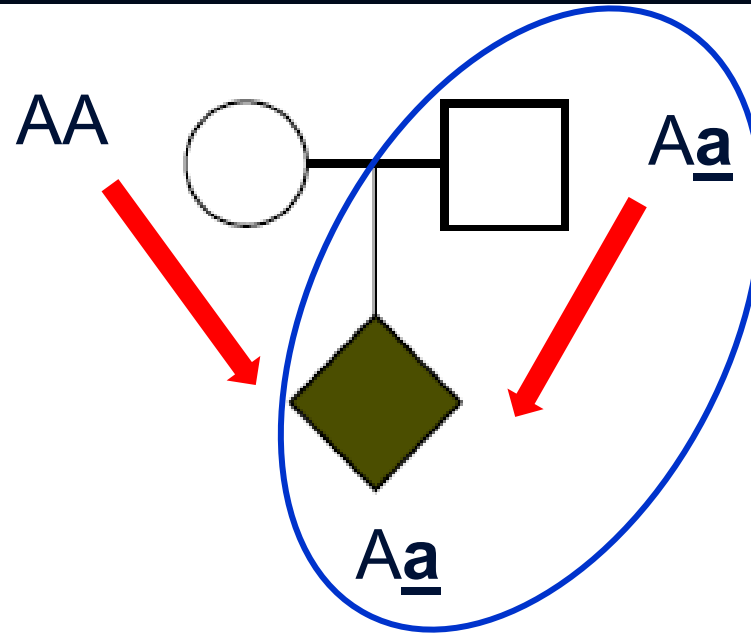
(Spielman et al. 1993)



Trios: two parents and one affected offspring

The **informative event** is the transmission of allele a by the heterozygous (Aa) parent (here the father)

TDT hypotheses



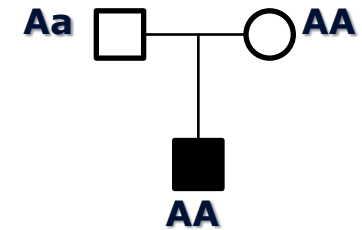
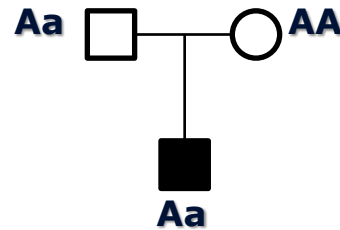
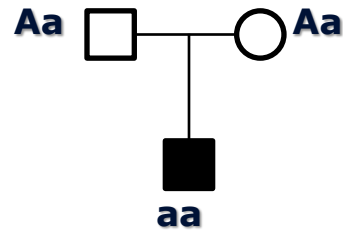
Under H_0 \underline{A} is transmitted as often as \underline{a}
The probability that Aa parents transmit $\underline{a} = 0.5$

Under H_1
The probability that Aa parents transmit $\underline{a} \neq 0.5$

Characteristics of TDT

- Offspring (case) is affected
- Genotype the case and both parents
 - test is conditional on the affected child; phenotype of parents can be ignored; but genotypes of parents are needed
- **Immune against confounding due to population stratification**

TDT test statistic (Mc Nemar's statistic)

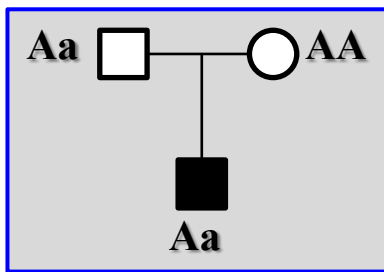


		Not transmitted	
Transmitted		A	a
	A	n_{11}	n_{12}
	a	n_{21}	n_{22}

$$\text{TDT} = (n_{12} - n_{21})^2 / (n_{12} + n_{21}) \sim \chi^2_{1\text{df}}$$

NB: n_{11} & n_{22} do not contribute to the test statistic

TDT practical



Father	Mother	Offspring
Aa	AA	Aa
Aa	Aa	Aa
Aa	Aa	aa
AA	AA	AA
Aa	AA	AA
AA	Aa	Aa
AA	Aa	AA
Aa	AA	Aa
Aa	Aa	AA

	Not transmitted	
	A	a
Transmitted		
A	7	5
a	6	0

$$\text{TDT} = (5-6)^2 / (5+6) = \mathbf{0.091} \sim \chi^2_{1\text{df}}$$

$$\rightarrow \text{P-value} = \mathbf{0.763}$$

Example: TDT within malaria resistant individual

	Not transmitted	
	A	S
Transmitted		
A	237	18
S	39	0

$$\mathbf{TDT} = (18 - 39)^2 / (18+39) = 7.74 \sim \chi^2_{1df}$$

$$\mathbf{p} = \mathbf{0.0054}$$

→ We reject H_0 and conclude that the HbS locus

or a locus in high linkage disequilibrium is involved in Malaria resistance

→ The allele S is associated with Malaria resistance

FBAT statistic, an extension of TDT

FBAT statistic, an extension of **TDT** test (Spielman et al. 1993)

Introduced by **Rabinowitz and Laird (2000)** and **Laird et al. (2000)**, builds on the original TDT method in which alleles transmitted to affected offspring are compared with the expected distribution of alleles among offspring

Robustness: Expected distribution is derived using Mendel's law of segregation and conditioning on parents genotype eliminate any potential confounding: population stratification, admixture, potential model misspecification.

FBAT statistic use a natural measure of association between two variables, covariance between the traits and the genotypes:

$$U = \sum T_{ij} (X_{ij} - E(X_{ij} | S_i))$$

i indexes family and *j* indexes nonfounders in the family

FBAT statistic, an extension of TDT

Covariance between the trait and the centered genotype

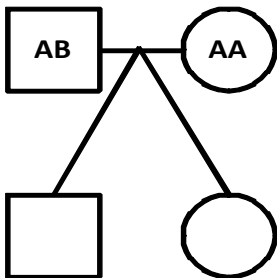
$$U = \sum T_{ij} (X_{ij} - E(X_{ij} | S_i))$$

$$Z = \frac{U}{\sqrt{\text{Var}(U)}} \sim N(0,1) \quad \text{approximately}$$

For univariate X or T:

$$\text{or equivalently, } \chi^2_{FBAT} = \frac{U^2}{\text{var}(U)}, \quad df = \text{rank}(V)$$

$(X_{ij} - E(X_{ij} | S_i))$ Can be thought of as residual of the "transmission" of parental genotype to offspring



= **0** , if both parents of ij th offspring are homozygous

- Transmissions from homozygous parents do not contribute to the test statistic (cf. Mendel Law)

Mendel law of allele's inheritance:

$$P(AA) = P(AB) = 0.5$$

TDT (likelihood version)

Design

- N independent "parents – affected offspring" trios
- 1 bi-allelic marker to test: a_1 , a_2

$\{a_1=T \mid a_1/a_2\}$: a_1 transmitted, given that parent's genotype is a_1/a_2

$p_{(1)(2)}$ = Probability ($\{a_1=T \mid a_1/a_2\}$)

$n_{(1)(2)}$ = sample frequency of the event $\{a_1=T \mid a_1/a_2\}$

Hypotheses

- $H_0: p_{(1)(2)} = p_{(2)(1)} = 0.5$ (no linkage or no association)
- $H_1: p_{(1)(2)} \neq p_{(2)(1)}$ (linkage and association)

Test Statistic (Mc Nemar)

- $X_{Mc} = \frac{(n_{(1)(2)} - n_{(2)(1)})^2}{n_{(1)(2)} + n_{(2)(1)}} \sim \chi^2(df = 1)$ under H_0

TDT (likelihood version)

Likelihood of data

- $L(\alpha) = p_{12}^{n_{12}} \times p_{21}^{n_{21}}$
 $\log L(\alpha) = n_{12} \times \log(p_{12}) + n_{21} \times \log(p_{21})$
- $\log L_0 = -\log(2) \times (n_{12} + n_{21})$ log-likelihood under the null hypothesis

Test Statistic (LRT)

- $\mathbf{X} = 2 \times \max_p \{ \log L(\alpha) - \log L_0 \} \sim \chi^2(1\text{df})$ under null hypothesis

\mathbf{X} is equivalent to the Mc Nemar statistic \mathbf{X}_{Mc}

Advantage: \mathbf{X} is more adequate for multi-marker generalisation than \mathbf{X}_{Mc}

Types of association testing models

Recruitment

- Population-based, Case/Control
- Family-based (related individuals)

Nature of the phenotype

- Qualitative, usually binary (affected / not affected)
- Quantitative

Type of genetic effect tested

- Single-Marker effect
- Multi-Marker effect

Detecting joint marker effect in family-based design

Design

- **N** independent "parents – affected offspring" trios
- **K** multi-allelic markers to test simultaneously: L^1, L^2, \dots, L^K
having respectively: l_1, l_2, \dots, l_K alleles

Detecting joint marker effect in family-based design

Log-Likelihoods

LL of the data:

$$\log L(\alpha^1, \alpha^2, \dots, \alpha^K) = \sum_{S \neq S'} n_{SS'} \times \log(p_{SS'}) + \sum_{S \neq S'} (m_{SS'} - n_{SS'}) \times \log(1 - p_{SS'})$$

- $m_{(S S')}$ = frequency of informative parents with respect to S and S'
- $n_{(S S')}$ = frequency of parents that have transmitted set S instead of S'
- $p_{(S S')} = \Pr(S \text{ Transmitted and } S' \text{ Not Transmitted})$

Under the null hypothesis of 50/50 risk:

$$(p_{SS'} = 1 - p_{SS'} = 0.5)$$

$$\log L_0 = -\log(2) \times \sum_{S \neq S'} m_{SS'}$$

Detecting multiway interaction in family data

The test statistic "X"

A likelihood Ratio based Test:

$X = 2 \times \max \{ \log L(\alpha^1, \alpha^2, \dots, \alpha^K) - \log L_0 \} \sim \chi^2(\mathbf{df})$ under Null Hypothesis

$$p_{(\mathbf{ss})} = \mathbf{f}(\alpha^1, \alpha^2, \dots, \alpha^K)$$

$$\mathbf{df} = \mathbf{h}(l_1, l_2, \dots, l_K) \quad \mathbf{f}, \mathbf{h} \text{ depend on } H_1 \text{ (tested hypothesis)}$$

Where $\alpha^i = \begin{pmatrix} \alpha_a^i \\ \vdots \\ \alpha_v^i \end{pmatrix}$ transmission intensities for each allele " a " at each locus i among heterozygous parents bearing the " a " allele.

Multiple Testing

The principal
Family-Wise Error Rate

Correction methods

- Bonferroni
- Benjamini & Hochberg method for controlling the FDR

Multiple testing

The principal

Suppose that we performed m independent tests corresponding to different alternative hypotheses H_1, H_2, \dots, H_m against the same null hypothesis H_0 .

e.g.,

H_1 : gene 1 has an effect on the trait

H_2 : gene 2 has an effect on the trait

...

H_m : gene m has an effect on the trait

H_0 : any of the genes has an effect on the trait

Multiple testing

The principal

Testing each marker for association at a given error rate α

$\Leftrightarrow \Pr(\mathbf{H}_i \mid H_0) = \text{probability to adopt hypothesis } i \text{ given that } H_0 \text{ is true} = \alpha$
(the probability to **wrongly** find gene i positive for the test)

Increasing the number of genes tested increases probability to find at least one of the genes wrongly significant, only by chance due to many trials.

A natural way to correct this increase of the risk of false findings is to set a new error rate α' for each gene such that the probability to have at least one false finding is α .
Testing m genes at error rate α' \Leftrightarrow testing a single gene at error rate α .

How to obtained α' ?

Corrected significance threshold

Family Wise Error Rate (FWER) or "Genome-wide Significance Level":

Is the probability to obtain at least one false positive (**FP**) result, and is conventionally expected to be equal to 0.05.


i.e., $FWER = 1 - \Pr(0 \text{ false positive} \mid H_0)$

If α' is the probability for each single test to be found positive wrongly,

$$\begin{aligned} \text{then, } \Pr(0 \text{ FP} \mid H_0) &= P(H_1 \text{ not FP}, H_2 \text{ not FP}, \dots, H_m \text{ not FP} \mid H_0) \\ &= (1 - \alpha') \times (1 - \alpha') \times \dots \times (1 - \alpha') \\ &= (1 - \alpha')^m \end{aligned}$$

$$\begin{aligned} \text{Then, } FWER &= 1 - (1 - \alpha')^m \\ &\leq \max(m \alpha', 1). \end{aligned}$$

$$\text{Then, } \alpha' = 1 - (1 - FWER)^{1/m}$$


 $\alpha = 0.05$
Target threshold

Bonferroni method

Bonferroni correction of p-values

α' is obtained as followed:

$$\begin{aligned}\alpha &= P(H_1 \text{ or } H_2, \text{ or } \dots \text{ or } H_m | H_0) \\ &= P(H_1 | H_0) + \dots + P(H_m | H_0) \\ &= \alpha' + \dots + \alpha' \\ &= m \times \alpha'\end{aligned}$$

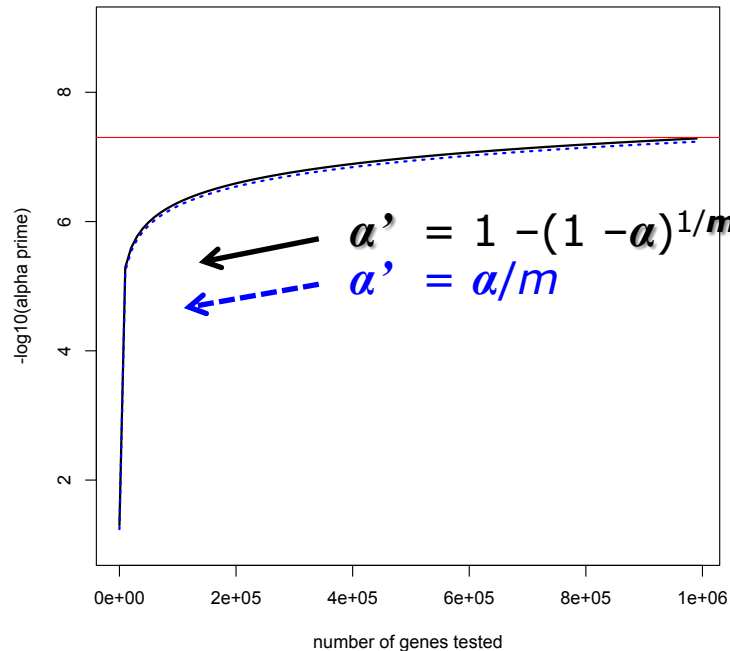
as H_1, H_2, \dots, H_m are independent

Then, $\alpha' = \alpha/m$.

Bonferroni method

Bonferroni correction of p-values

A good approximation of Family-Wise Error Rate (FWER)



$$-\log_{10}(0.05/1e6) = 7.3 ,$$

for 1,000,000 tests

(FWER)
(Bonferroni)

Benjamini & Hochberg method

False Discovery Rate (FDR)

After performing the m tests as described above, suppose that \mathbf{P} are declared positive and \mathbf{N} as negative, but in reality \mathbf{m}_1 are positive and \mathbf{m}_0 are negative as summarized here:

The Truth	Declared by the tests		Total
	Significant	Not Significant	
<i>Null is True</i>	Fp	Tn	$\mathbf{m}_0 = Fp + Tn$
<i>Alternative is True</i>	Tp	Fn	$\mathbf{m}_1 = Tp + Fn$
Total	$\mathbf{P} = Fp + Tp$	$\mathbf{N} = Tn + Fn$	\mathbf{m}

Benjamini & Hochberg method

False Discovery Rate (FDR)

The FDR method provides a control of error rate with a straightforward interpretability by setting a false discovery rate that satisfy the following condition:

$$E\left(\frac{Fp}{P} \mid P > 0\right) \leq FDR$$

i.e., given that we obtain a non null number of positive tests, the expectation of error rate which is Fp/P has to be lower than the FDR. Conventionally FDR is set to 0.05

Benjamini & Hochberg method

False Discovery Rate (FDR)

The weak control of FDR (Benjamini and Hochberg 1995) follow these three steps:

- Order the P-values from the lowest to the highest

$$P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$$

- Find the highest rank k (denoted k^*) that satisfy

$$P_{(k)} \leq k \times \alpha / m$$

- If k^* exists, adopt all hypotheses corresponding to

$$P_{(1)}, \dots, P_{(k^*)}$$

Benjamini & Hochberg method

False Discovery Rate (FDR)

Equivalently, we can calculate adjusted (or corrected) FDR's P-values (P^*).

Order the P-value: $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(m)}$

and follow these steps:

- $P^*_{(m)} = P_{(m)}$
- $P^*_{(m-1)} = \min \{P^*_{(m)} ; P_{(m-1)} \times m / (m-1)\}$
- $P^*_{(m-2)} = \min \{P^*_{(m-1)} ; P_{(m-2)} \times m / (m-2)\}$

...

- $P^*_{(1)} = \min \{P^*_{(2)} ; P_{(1)} \times m\}.$

Benjamini & Hochberg method

False Discovery Rate (FDR)

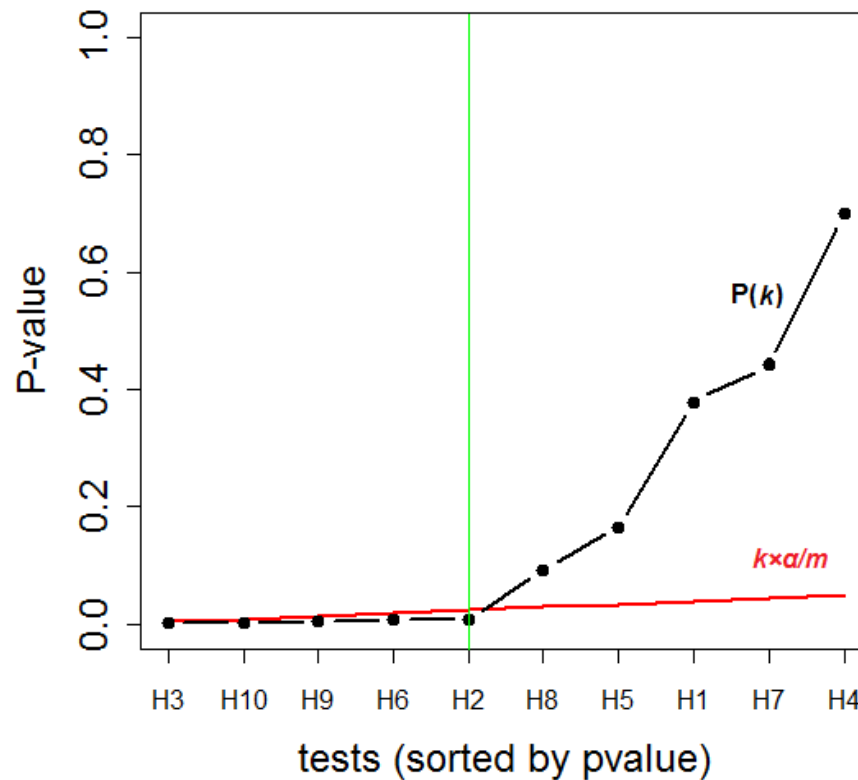
For example $m=10$ tests and $\alpha=0.05$ with following raw P-values:

Test	P-value		Ordered P-value	rank (k)	$k \times \alpha / m$	P*
gene1	0.378	g3	0.002	1	0.005	0.018
gene2	0.009	g10	0.004	2	0.010	0.018
gene3	0.002	g9	0.006	3	0.015	0.018
gene4	0.700	g6	0.008	4	0.020	0.018
gene5	0.166	g2	0.009	5	0.025	0.018
gene6	0.008	g8	0.094	6	0.030	0.157
gene7	0.443	g5	0.166	7	0.035	0.237
gene8	0.094	g1	0.378	8	0.040	0.473
gene9	0.006	g7	0.443	9	0.045	0.492
gene10	0.004	g4	0.700	10	0.050	0.700

Benjamini & Hochberg method

False Discovery Rate (FDR)

Graphically



Bonferroni vs FDR

Bonferroni:

- Simple to compute
- Less consistent if the m tests are not really independent, could be the case in genome wide studies due to linkage disequilibrium.
- Conservative (number of false positive (Fp) is evaluated with respect to the total number of tests (m))
- Not always appropriate for genetic studies where many genes are often involved

FDR:

- Acceptable way of controlling the inflation of Fp in context of genetic studies, considering the expected number of false positive among the P tests declared positive only, instead of referring to all the m tests.

Population stratification

Definition

"Population stratification" = population structure

Spurious association due to:

- Systematic difference in allele frequencies between sub-populations ...

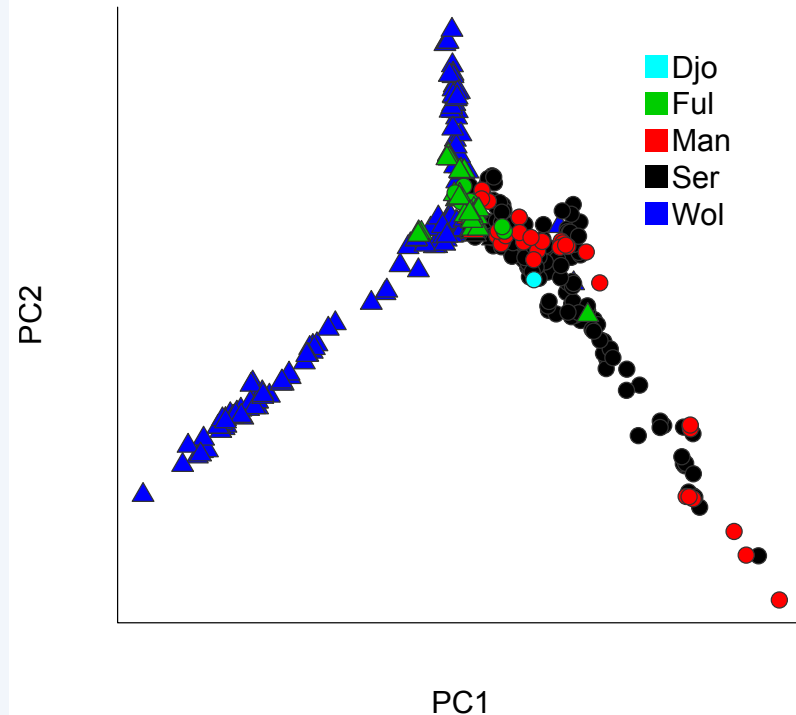
... can be due to different ancestry

To visualize:

- PCA

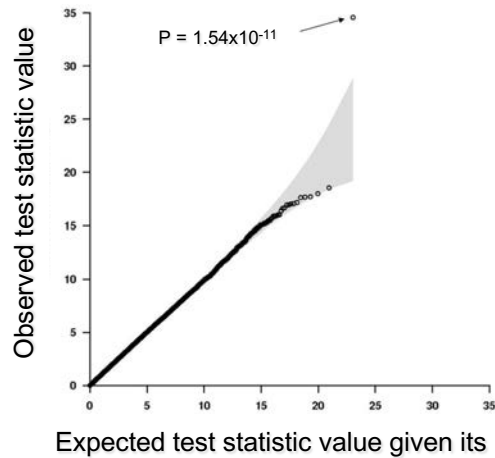
To correct:

- Genomic Control
- Regression on Principal Components of PCA

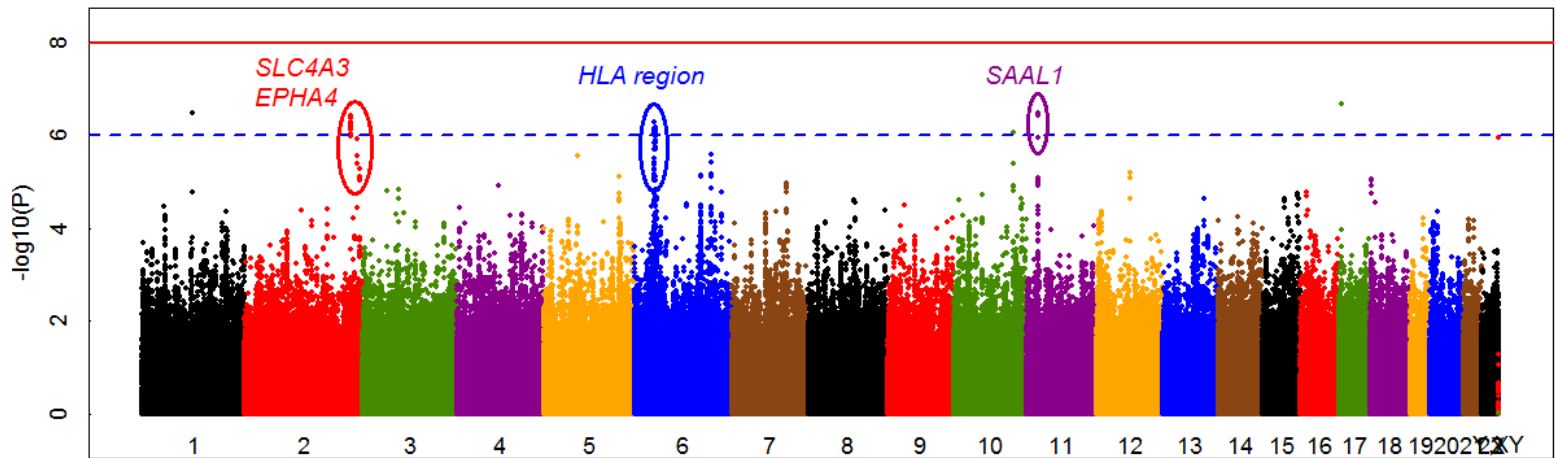


Graphical representation of results

Quantile-Quantile (QQ) Plot

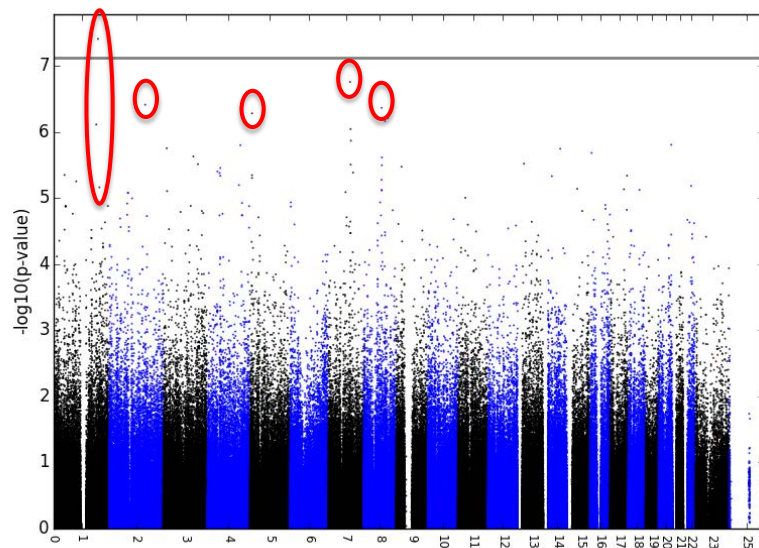


Manhattan plot



Graphical representation of results

Manhattan plot



Locus zoom plots

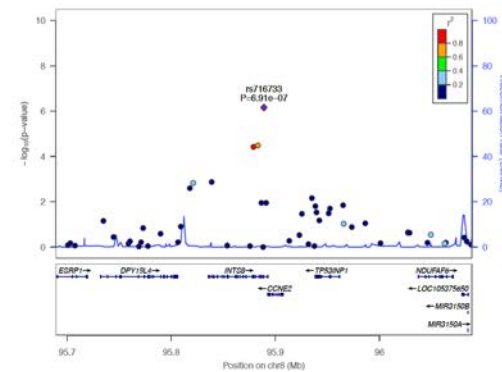
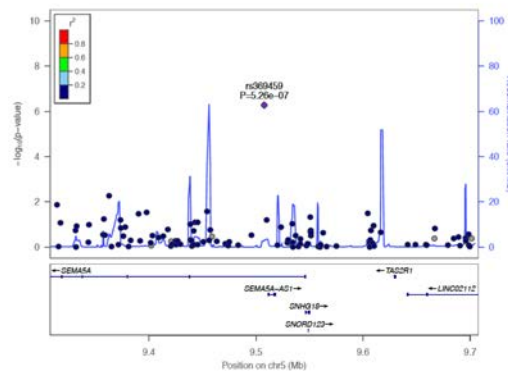
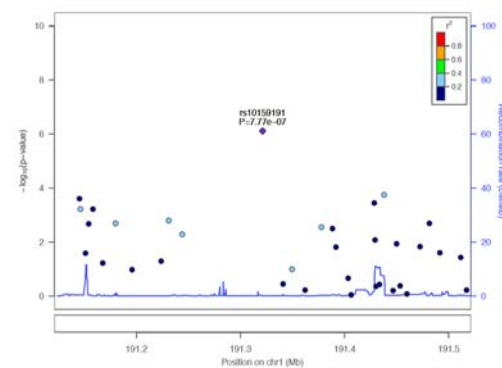
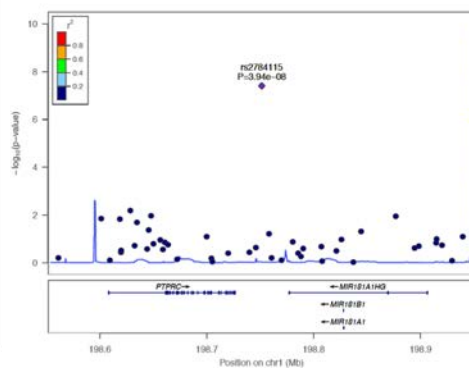


Table representation of results

SNP	CHR	Gene symbol	Position	Beta	NominalP	A1	A2	CR	MAF	HWE	FDR	Protein ID
rs10159191	1	<i>MFSD6</i>	191321186	-0.38	7.8E-07	A	G	1	0.22	0.64	0.032	Q6ZSS7
rs2784115	1	<i>PLCL1</i>	198751401	0.51	3.9E-08	G	A	1	0.15	0.01	0.006	Q15111
rs369459	5	<i>ZNF143</i>	9507624	0.46	5.3E-07	A	G	1	0.12	0.37	0.026	P52747
rs716733	8	<i>PLCE1</i>	95888981	0.33	6.9E-07	G	A	1	0.21	0.30	0.030	Q9P212

Tools used for GWAS testing

- **Plink**
- **GCTA**
- **R Bioconductor** libraries (availability of set of packages for GWAS: e.g., "GWASTools", "SNPRelate", "gdsfmt", and for graphical representation)

Files, formats and outputs for GWAS testing

GWAS results are generally exported in tabulated files, and merging with:

- **SNPs QC files (containing MAF, Call rate, HWE test, ...)**
- **Annotation files**
- Usual file formats: **.tsv, .xlsx, etc...**
- Use of programming software (**like R**) **to explore large results files**

Thanks

for your attention

QUESTIONS ?
