

# Lecture 1: Case Control Association Testing & Association Testing with Quantitative Traits

Instructors: Joelle Mbatchou and Loic Yengo

Summer Institute in Statistical Genetics 2022

# Introduction

- ▶ Association mapping is now routinely being used to identify loci that are involved with complex traits.
- ▶ Technological advances have made it feasible to perform association studies on a genome-wide basis with millions of markers in a single study.
- ▶ We consider testing a genetic marker for association with a disease (e.g. 1/0, affected/unaffected, dead/alive) or a quantitative trait (e.g. height, BMI) in **a sample of unrelated subjects**.
- ▶ Vast amounts of literature on these topics!

# Case-Control Association Testing

- ▶ Allelic Association Tests
  - ▶ Allele is treated as the sampling unit
  - ▶ Typically make an assumption of Hardy-Weinberg equilibrium (HWE) - alleles within an individual are conditionally independent given the disease status
  - ▶ e.g. Pearson's  $\chi^2$
- ▶ Genotypic Association Tests
  - ▶ Individual is the sampling unit
  - ▶ Does not assume HWE
  - ▶ e.g. Logistic regression

## Pearson's $\chi^2$ Test for Allelic Association

- ▶ This test looks for deviations from independence between the trait and allele.
- ▶ Consider a single marker with 2 allelic types (e.g., a SNP) labeled "C" and "T".
- ▶ Let  $N_{ca}$  be the number of cases and  $N_{co}$  be the number of controls on which we have genotype data.

## Pearson's $\chi^2$ Test for Allelic Association

- Below is a  $2 \times 2$  contingency table for trait and allelic type

	Cases	Controls	Total
Allele C	$n_C^{ca}$	$n_C^{co}$	$N_C$
Allele T	$n_T^{ca}$	$n_T^{co}$	$N_T$
Total	$2N_{ca}$	$2N_{co}$	$2N$

- $n_C^{ca}$  is the number of C alleles in the cases and  $n_C^{ca} = 2 \times$  the number of homozygous CC cases + the number of heterozygous CT cases
- Hypotheses
  - $H_0$ : there is *no association* between the row variable and column variable
  - $H_a$ : there *is* an association between the two variables

## Pearson's $\chi^2$ Test for Allelic Association

- ▶ Can use Pearson's  $\chi^2$  test for independence. The statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed cell} - \text{Expected cell})^2}{\text{Expected cell}}$$

- ▶ What is the the expected cell number under  $H_0$ ? For each cell, we have

$$\text{Expected Cell Count} = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

- ▶ Under  $H_0$ , the  $\chi^2$  test statistic has an approximate  $\chi^2$  distribution with  $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$  degree of freedom

## LHON Example: Pearson's $\chi^2$ Test

- ▶ From Phasukijwattana et al. (2010), Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- ▶ Corresponding  $2 \times 2$  contingency table with allelic type instead of genotype

	Allele C	Allele T
Cases	20	158
Controls	86	39

- ▶ Should we reject the hypothesis that allelic type is independent of disease status?

## LHON Example: Pearson's $\chi^2$ Test

	Allele C	Allele T	Total
Cases	20	158	178
Controls	86	392	478
Total	106	550	656

- Intuition for the test: Suppose  $H_0$  is true, so allelic type and case-control status are independent, then what counts would we expect?
  - the expected number of case alleles that are of type C is:

$$\begin{aligned}n_C^{ca} &= \{\#cases\} \times P(\text{Allelic type is C} \mid \text{Allele is from a Case}) \\&= \{\#cases\} \times P(\text{Allelic type is C}) \quad \textbf{(independence)} \\&= 178 \times \left(\frac{106}{656}\right) = 28.7622\end{aligned}$$



## LHON Example: Pearson's $\chi^2$ Test

- Fill in the remaining cells for the expected counts

	Allele C	Allele T	Total
Cases	28.7622	.	178
Controls	.	.	478
Total	106	550	656

## LHON Example: Pearson's $\chi^2$ Test

- Fill in the remaining cells for the expected counts

	Allele C	Allele T	Total
Cases	28.7622	149.2378	178
Controls	77.2378	400.7622	478
Total	106	550	656

## LHON Example: Pearson's $\chi^2$ Test

- Fill in the remaining cells for the expected counts

	Allele C	Allele T	Total
Cases	28.7622	149.2378	178
Controls	77.2378	400.7622	478
Total	106	550	656

- Calculate the  $X^2$  statistic

$$\chi^2 = \frac{(20 - 28.7622)^2}{28.7622} + \dots + \frac{(392 - 400.7622)^2}{400.7622} = 4.369$$

## LHON Example: Pearson's $\chi^2$ Test

- Fill in the remaining cells for the expected counts

	Allele C	Allele T	Total
Cases	28.7622	149.2378	178
Controls	77.2378	400.7622	478
Total	106	550	656

- Calculate the  $X^2$  statistic

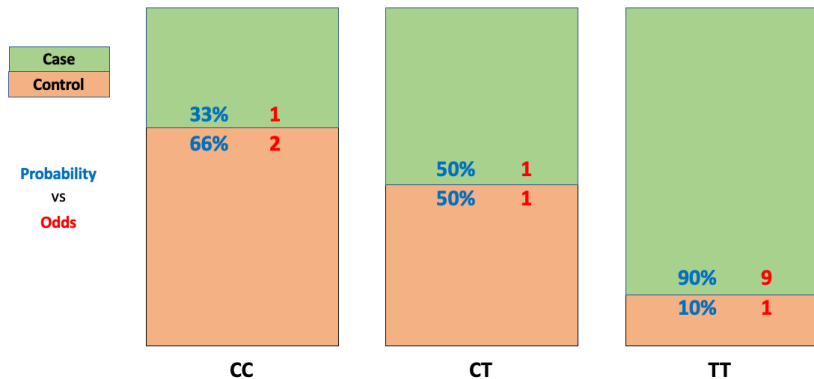
$$\chi^2 = \frac{(20 - 28.7622)^2}{28.7622} + \dots + \frac{(392 - 400.7622)^2}{400.7622} = 4.369$$

- What is the  $p$ -value?

$$P(\chi_1^2 \geq 4.369) = .037$$

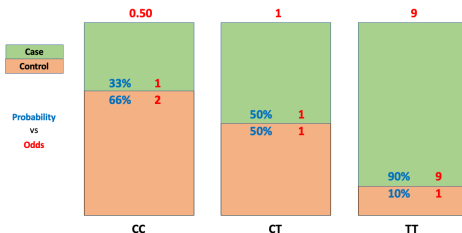
## Odds Ratios (ORs) for Genotypes

- What are **odds**? An expression of **relative probabilities**...



- Odds of disease in an individual with the CC genotype = 50%

# Odds Ratios (ORs) for Genotypes



- Typically choose a reference genotype, e.g. CC

$$OR_{TT} = \frac{\text{odds of disease with the TT genotype}}{\text{odds of disease with the CC genotype}} = \frac{9}{0.50} = 18$$

- $OR_{TT} = 1$  implies no association with disease.
- $OR_{TT} > 1$  or  $OR_{TT} < 1$  implies association with the disease.

## Genotypic Association Tests: Logistic regression

- ▶ Generally used to estimate odds ratios and get confidence intervals for genotypes.
- ▶ Let  $\pi_i$  be the probability that individual  $i$  has the disease and let  $G_i$  be the genotype at the SNP:

$$\underbrace{\log \left( \frac{\pi_i}{1 - \pi_i} \middle| G_i \right)}_{\text{odds of disease}} = \beta_0 + \beta_{CT} I\{G_i = CT\} + \beta_{TT} I\{G_i = TT\}$$

where  $I\{G_i = CT\}$  is 1 if  $G_i = CT$  and 0 otherwise, and similarly for  $I\{G_i = TT\}$ .

## Genotypic Association Tests: Logistic regression

- ▶ The coefficient estimates for  $\hat{\beta}_{CT}$  and  $\hat{\beta}_{TT}$  can be used to calculate odds ratios:

$$\begin{aligned} OR_{CT} &= \frac{\text{odds of disease with the CT genotype}}{\text{odds of disease with the CC genotype}} \\ &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_{CT})}{\exp(\hat{\beta}_0)} = \exp(\hat{\beta}_{CT}) \end{aligned}$$

Similarly,  $OR_{TT} = \exp(\hat{\beta}_{TT})$

- ▶ 95% CI for  $OR_{CT}$  is

$$\exp(\hat{\beta}_{CT} \pm 1.96 \times \text{s.e.}(\hat{\beta}_{CT}))$$



## Odds Ratios for LHON Example

- ▶ Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

- ▶ We will use the R to obtain odds ratios and confidence intervals for this data set
- ▶ We will contrast this test with an allelic test (Pearson's  $\chi^2$  test).

## Introduction to Quantitative Trait Mapping

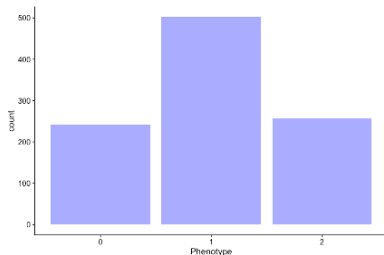
- ▶ Quantitative trait loci (QTL) mapping involves identifying genetic loci that influence the phenotypic variation of a quantitative trait.
- ▶ QTL mapping is commonly conducted with GWAS, often involving directly genotyped variants along with imputation through reference panels to result in millions of genetic variants

## Introduction to Quantitative Trait Mapping

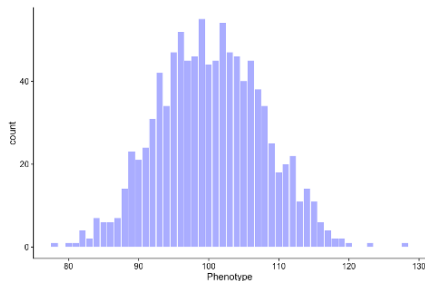
- ▶ Some quantitative traits can be largely influenced by a single gene as well as by environmental factors (e.g. monogenic or Mendelian traits)
- ▶ Influences on a quantitative trait can also be due to a number of genes → polygenicity
- ▶ Many quantitative traits of interest are complex where phenotypic variation is due to a combination of both multiple genes and environmental factors
- ▶ Examples: Blood pressure, cholesterol levels, IQ, height, weight, etc.

# Mendelian & Quantitative Genetics

**One marker**



**Many markers**



## Quantitative Genetic Model

- ▶ The classical quantitative genetics model introduced by Ronald Fisher (1918) is

$$Y = G + E$$

where  $Y$  is the phenotypic value,  $G$  is the genetic value, and  $E$  is the environmental deviation.

- ▶  $G$  is the combination of all genetic loci that influence the phenotypic value and  $E$  consists of all non-genetic factors that influence the phenotype (mean set to 0)

## Components of Genetic Variance

- ▶ Consider a single locus. Fisher modeled the genotypic value  $G$  with a linear regression model (least squares) where the genotypic value can be partitioned into an additive component ( $A$ ) and deviations from additivity as a result of dominance ( $D$ ), where

$$G = A + D,$$
$$\underbrace{Var(G)}_{\sigma_G^2} = \underbrace{Var(A)}_{\sigma_A^2} + \underbrace{Var(D)}_{\sigma_D^2}$$

- ▶  $\sigma_A^2$  is the **additive genetic variance**. It is the genetic variance associated with the average additive effects of alleles
- ▶  $\sigma_D^2$  is the **dominance genetic variance**. It is the genetic variance associated with the dominance effects.

# Heritability

- ▶ Remember

$$\begin{aligned} Y &= G + E \\ &= A + D + E, \\ \underbrace{Var(Y)}_{\sigma_Y^2} &= \underbrace{Var(A)}_{\sigma_A^2} + \underbrace{Var(D)}_{\sigma_D^2} + \underbrace{Var(E)}_{\sigma_E^2} \end{aligned}$$

- ▶ **Narrow-sense heritability** (or simply heritability) is

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}$$

- ▶  $h^2$  is the proportion of the total phenotypic variance due to additive effects.
- ▶ It can also be viewed as the extent to which phenotype is determined by the alleles transmitted from the parents.

# Heritability

- ▶ The **broad-sense heritability** is defined to be

$$H^2 = \frac{\sigma_G^2}{\sigma_Y^2} = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_Y^2}$$

- ▶  $H^2$  is the proportion of the total phenotypic variance that is due to all genetic effects (additive and dominance)
- ▶ Heritability can vary over time and with the study population as it depends also on environmental effects



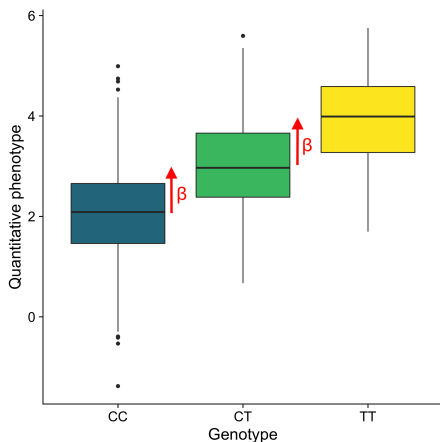
## QTL Mapping

- ▶ For traits that are heritable, i.e., traits with a non-negligible genetic component that contributes to phenotypic variability, identifying (or mapping) QTLs that influence the trait is often of interest.
- ▶ Linear regression models are commonly used for QTL mapping
  - ▶ They will often include a single genetic marker (e.g., a SNP) as predictor in the model, in addition to other relevant covariates (e.g. age, sex), with the quantitative phenotype as the response

## Linear regression with SNPs

Many analyses fit the 'additive model'

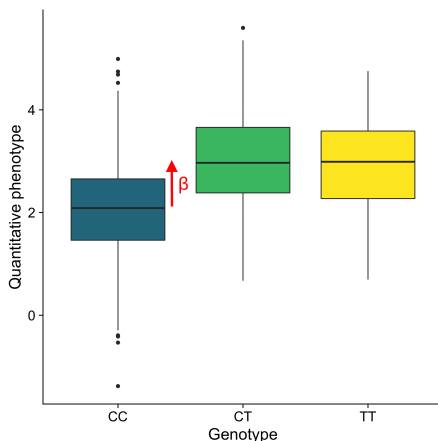
$$y = \beta_0 + \beta \times \#T \text{ alleles}$$



## Linear regression, with SNPs

An alternative is the 'dominant model';

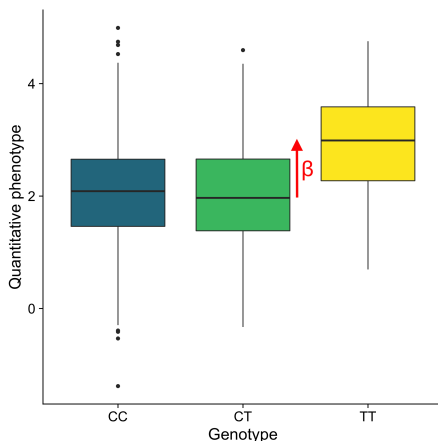
$$y = \beta_0 + \beta \times I\{G \neq CC\}$$



# Linear regression, with SNPs

or the 'recessive model';

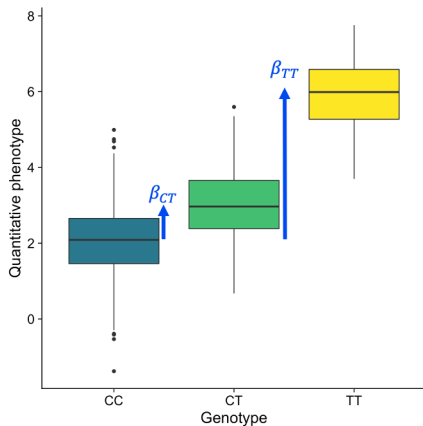
$$y = \beta_0 + \beta \times I\{G == TT\}$$



## Linear regression, with SNPs

Finally, the 'two degrees of freedom model';

$$y = \beta_0 + \beta_{CT} \times I\{G == CT\} + \beta_{TT} \times I\{G == TT\}$$



## Additive Genetic Model

- ▶ **Most GWAS perform single SNP association testing with linear regression assuming an additive model.**
- ▶ The coefficient of determination ( $r^2$ ) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the "SNP heritability"

## Additive Genetic Model

- ▶ Consider the following additive model for association testing with a quantitative trait and a SNP with alleles  $C$  and  $T$ :

$$Y = \beta_0 + \beta_1 G + \epsilon$$

where  $G$  is the number of copies of the allele  $T$ .

- ▶ How would you interpret  $\epsilon$  in this model?

## Association Testing with Additive Model

$$Y = \beta_0 + \beta_1 G + \epsilon$$

- Two test statistics for  $H_0 : \beta_1 = 0$  versus  $H_a : \beta_1 \neq 0$

$$T = \frac{\hat{\beta}_1}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim \mathbf{t}_{N-2} \approx N(0, 1) \text{ for large } N$$

$$T^2 = \frac{\hat{\beta}_1^2}{\text{var}(\hat{\beta}_1)} \sim \mathbf{F}_{1, N-2} \approx \chi_1^2 \text{ for large } N$$

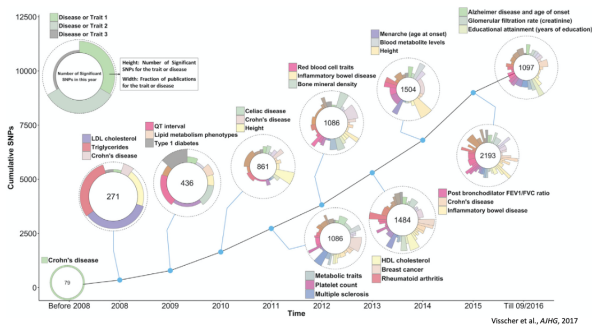
where

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{S_{GG}}$$

and  $S_{GG}$  is the corrected sum of squares for the  $G_i$ 's



# Missing Heritability



- ▶ Gap between SNP-based and pedigree-based heritability estimates
- ▶ Causal variants not well tagged? Rare variants involved?

# Time for practical exercises!

- ▶ Head to your breakout room
- ▶ Introduce yourselves:
  - ▶ Name
  - ▶ Academic background
  - ▶ Research interests
  - ▶ 1 hobby
- ▶ Work on Session 1 exercise (link on class website)  
`https://joellembatchou.github.io/SISG2022\_Association\_Mapping/Session01\_practical.html`

## References

- ▶ Phasukkijwatana, N., Kunhapan, B., Stankovich, J., Chuenkongkaew, W. L., Thomson, R., Thornton, T., Bahlo, M., Mushiroda, T., Nakamura, Y., Mahasirimongkol, S., et al. (2010). Genome-wide linkage scan and association study of PARL to the expression of LHON families in Thailand. *Human genetics*, **128**(1), 39–49.
- ▶ Visscher, P.M. et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5-22.