

Association testing I

```
$ echo "Data Sciences Institute"
```

- **Objective:** establish association between a trait of interest and a genetic marker.
- Study designs: case-control, case-cohort, population-based design.
- Unrelated subjects or **population-based designs:** easy to collect so possible to achieve large sample sizes as in GWAS.
- **Family-based designs:** robust to population stratification, more difficult to collect.
Also hard to collect for late-onset diseases.

Overview

- For a quantitative trait Y : simple linear regression models to detect association between Y and G :

$$Y = \alpha + \beta G + e, e \sim N(0, \sigma^2).$$

- Coding of G : 1 d.f. additive (or dominant, recessive) model or 2 d.f. genotypic model.

$G =$	aa	Aa	AA
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1
Genotypic	"0"	"1"	"2"

Overview

- **1 d.f. model:** $Y = \alpha + \beta G + e, e \sim N(0, \sigma^2)$.
 - The test of no association is equivalent to $H_0 : \beta = 0$.
- **Genotypic (2 d.f.) model:** $Y = \alpha + \beta_1 D_1 + \beta_2 D_2 + e, e \sim N(0, \sigma^2)$.
 - $D_1 = I(G = Aa), D_2 = I(G = AA)$.
 - $H_0 : \beta_1 = \beta_2 = 0$.
- **Extended model** (includes environment, other SNPs, interactions):
$$Y = \alpha + \beta G + \gamma E + \delta GxE + e, e \sim N(0, \sigma^2).$$

Association Analysis for Binary Trait

- For a binary trait Y : $Y = 1$ denotes for being affected/cases, and $Y = 0$ for unaffected/controls
- SNP: categorical variable with three genotypes (Aa, aa and AA)
- Genotype frequency differ between cases and controls only at the DSL or markers that are in LD with the DSL
- How do we detect association between Y and G ?
 - **Compare frequency differences between cases and controls**
 - **Use regression framework** to unify analysis (QTL + binary traits, with covariates and interactions)

Association Testing (2 df test)

- Observed genotype counts at a marker under the study for r cases and s controls.

	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

- Most general test: nothing is assumed about the relationship between disease and genotype.

$$H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid Aa) = P(Y = 1 \mid aa)$$

H_A : At least one inequality holds

Contingency Table - Pearson's Homogeneity Test

- Assume we have independent observations from I multinomial distributions, each of which has J categories, e.g. $I = 2$.

	$X = 0$	$X = 1$	\dots	$X = J - 1$	Total
Cases	r_0	r_1	\dots	r_{J-1}	r
Controls	s_0	s_1	\dots	s_{J-1}	s
Total	n_0	n_1	\dots	n_{J-1}	n

- Our goal is to test whether the multinomial distributions for the two groups are identical.

$$H_0 : \pi_{1j} = \pi_{2j} \equiv \pi_j, \quad j = 0, 1, \dots, J - 2.$$

Pearson's Homogeneity Test

- Test Statistic

$$T = \sum \frac{(O - E)^2}{E} \sim \chi^2_{(I-1)(J-1)}.$$

- E: expected counts under the null hypothesis of homogeneity using pooled estimate
 $\tilde{\pi}_j = \frac{n_j}{n}.$
- df = $(I - 1)(J - 1)$: number of independent counts ($I(J - 1)$) - the number of independent parameters estimated under the null from the data ($J - 1$).

Association Testing (2 df test)

- | | 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 |

- Two df Pearson test of independence:

$$\chi^2 = \sum (O - E)^2 / E.$$

- Sum is over all six entries. e.g. $E[\text{Case \& aa}] = (r \cdot n_0) / n$.

Association Testing (1 df test)

- For example:

	aa	Aa	Total
Cases	r_0	r_1	r
Controls	s_0	s_1	s
Total	n_0	n_1	n

$$T = \frac{\left(r_0 - \frac{rn_0}{n}\right)^2}{\frac{rn_0}{n}} + \frac{\left(r_1 - \frac{rn_1}{n}\right)^2}{\frac{rn_1}{n}} + \frac{\left(s_0 - \frac{sn_0}{n}\right)^2}{\frac{sn_0}{n}} + \frac{\left(s_1 - \frac{sn_1}{n}\right)^2}{\frac{sn_1}{n}} \sim \chi^2_1.$$

Association Testing - 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 |

- Dominant model:

$$H_0 : P(Y = 1 \mid AA) = P(Y = 1 \mid Aa) = P(Y = 1 \mid aa)$$
$$H_A : P(Y = 1 \mid AA \text{ or } Aa) \neq P(Y = 1 \mid aa)$$

- Replace r_1 with $r_1 + r_2$ with 1 df chi-square test.
- Optimal when the true disease model is dominant but not for recessive.

Exercise

- 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

Association Testing - Additive Model

- We can also test for association using **allele counts**.
- $H_0 : p_{cases} = p_{controls}$

	a	A	Total
Cases	$2r_0 + r_1$	$r_1 + 2r_2$	$2r$
Controls	$2s_0 + s_1$	$s_1 + s_2$	$2s$
Total	n_a	n_A	$2n$

- Assumptions: samples are independent both within each group (cases and controls) and between groups. Under this assumption, allele count data can be modeled using two independent binomial distributions.

Estimating Effect Sizes - Risk Ratio (RR)

- A nature measure of effect size is relative risk ratio: $RR = \frac{P(\text{ disease } | \text{ exposed })}{P(\text{ disease } | \text{ unexposed })}$.
- Exposure = genotype; recessive model: AA vs (Aa, aa).
- In case-control or genotype-ascertained samples, group risks are distorted.
- Therefore RR cannot be validly estimated from those designs.

Odds Ratio

- Odds Ratio are used to approximate the relative risk (RR) for case-control or case-cohort sampling.
- OR compares odds: $\frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))}$.

- | | Disease (+) | Disease (-) | Total |
|-----------|-------------|-------------|-------|
| Exposed | a | b | a+b |
| Unexposed | c | d | c+d |

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}.$$

- When the outcome is rare, $OR \approx RR$.

Inference of ORs

- $OR = \frac{a/c}{b/d} = \frac{ad}{bc}$.
- Log(OR) approximately normal.
- Variance:

$$\text{Var}[\log(OR)] \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}.$$

- 95% CI:

$$\exp \left(\log(OR) \pm 1.96 \cdot \sqrt{\text{Var}[\log(OR)]} \right).$$

Limitations of Previous Tests

- Work well only for simple binary traits without major environmental influence
- Do not easily adjust for covariates (e.g., age, sex, ancestry)
- Population stratification can confound results if unadjusted
- Stratified analyses are one workaround, but not always sufficient

Properties of OR

- Let $\pi_1^* = P(\text{ case } | Aa)$ and $\pi_2^* = P(\text{ case } | aa)$.
- Let $\pi_1 = P(Aa | \text{ case})$ and $\pi_2 = P(Aa | \text{ control})$.
- We have the following result that links the two quantities:

$$\begin{aligned}\frac{\text{odds}(\text{ case } | Aa)}{\text{odds}(\text{ case } | aa)} &= \frac{\pi_1^*}{1 - \pi_1^*} / \frac{\pi_2^*}{1 - \pi_2^*} \\ &= \frac{\pi_1}{1 - \pi_1} / \frac{\pi_2}{1 - \pi_2} = \frac{\text{odds}(Aa | \text{ case})}{\text{odds}(Aa | \text{ control})}.\end{aligned}$$

Details of the derivation

$$\begin{aligned} \frac{\text{odds}(\text{ case } | Aa)}{\text{odds}(\text{ case } | aa)} &= \frac{\pi_1^*}{1 - \pi_1^*} / \frac{\pi_2^*}{1 - \pi_2^*} = \frac{P(\text{ case } | Aa)}{1 - P(\text{ case } | Aa)} / \frac{P(\text{ case } | aa)}{1 - P(\text{ case } | aa)} \\ &= \frac{P(\text{ case } | Aa)}{P(\text{ control } | Aa)} / \frac{P(\text{ case } | aa)}{P(\text{ control } | aa)} = \frac{P(\text{ case}, Aa)P(Aa)}{P(\text{ control}, Aa)P(Aa)} / \frac{P(\text{ case}, aa)P(aa)}{P(\text{ control}, aa)P(aa)} \\ &\quad = \frac{P(\text{ case}, Aa)}{P(\text{ control}, Aa)} / \frac{P(\text{ case}, aa)}{P(\text{ control}, aa)} \\ &= \frac{P(Aa | \text{ case})P(\text{ case})}{P(Aa | \text{ control})P(\text{ control})} / \frac{P(aa | \text{ case})P(\text{ case})}{P(aa | \text{ control})P(\text{ control})} \\ &\quad = \frac{P(Aa | \text{ case})}{P(aa | \text{ case})} / \frac{P(Aa | \text{ control})}{P(aa | \text{ control})} \\ &= \frac{P(Aa | \text{ case})}{1 - P(Aa | \text{ case})} / \frac{P(Aa | \text{ control})}{1 - P(aa | \text{ control})} = \frac{\pi_1}{1 - \pi_1} / \frac{\pi_2}{1 - \pi_2} = \frac{\text{odds}(Aa | \text{ case})}{\text{odds}(Aa | \text{ control})} \end{aligned}$$

Regression Framework: Extending to Covariates & Traits

- Use Generalized Linear Models (GLMs):

$$g(E[Y | X]) = \alpha + X\beta$$

- Link function g depends on trait type:
- Binary traits (logistic):

$$\log \frac{E[Y | X]}{1 - E[Y | X]} = \alpha + X\beta$$

- Continuous traits (linear):

$$E[Y | X] = b_0 + Xb_1$$

- X = coded genotype (additive, dominant, recessive, etc.)

Regression Framework

- Tests genetic effect:

$$H_0 : b_1 = 0$$

- Inference via likelihood ratio test, Wald test, or score test
- Key advantage: allows easy adjustment for covariates (age, sex, ancestry, etc.)
- Flexible framework for analyzing both binary and continuous traits

Remarks

- For logistic regression, the **estimated coefficient β** is equal to the $\log(\text{OR})$ (for the corresponding model)
- For continuous outcomes, the coefficient β represents differences in means by genotype group
 - For recessive model, β is the mean phenotype for the AA group - mean phenotype for Aa or aa group
 - For the additive model, b_1 is the mean increase in phenotype with each additional allele.

Interpretation of Logistic Regression

- Consider logit of μ as a linear function of X .

$$\mu(X) = E(Y | X) = P(Y = 1 | X)$$

$$\text{logit}(P(Y = 1 | X)) = \text{logit}(\mu(X)) = \log \frac{\mu(X)}{1 - \mu(X)} = \alpha + \beta X$$

$$\implies P(Y = 1 | X) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)},$$

$$P(Y = 0 | X) = \frac{1}{1 + \exp(\alpha + \beta X)}.$$

Interpretation of Logistic Regression

- What is the interpretation of the parameters, α and β ?

$$\text{logit}(\mu(X)) = \log \frac{\mu(X)}{1 - \mu(X)} = \alpha + \beta X.$$

- For individuals carry genotype aa, $X = 0$:

$$\text{logit}(\mu(X = 0)) = \alpha.$$

- Thus,

$$\alpha = \text{logit}(\mu(X = 0)) = \log \left(\frac{\mu(X = 0)}{1 - \mu(X = 0)} \right) = \log \left(\frac{P(Y = 1 | aa)}{1 - P(Y = 1 | aa)} \right)$$

is the log-odds of being affected/case for genotype aa.

Interpretation of Logistic Regression

- For individuals carry genotype Aa, $X = 1$:

$$\text{logit}(\mu(X = 1)) = \alpha + \beta.$$

- Thus,

$$\begin{aligned}\beta &= \text{logit}(\mu(X = 1)) - \text{logit}(\mu(X = 0)) = \log\left(\frac{\mu(X = 1)}{1 - \mu(X = 1)}\right) - \log\left(\frac{\mu(X = 0)}{1 - \mu(X = 0)}\right) \\ &= \log\left(\frac{P(Y = 1 | Aa)}{1 - P(Y = 1 | Aa)}\right) - \log\left(\frac{P(Y = 1 | aa)}{1 - P(Y = 1 | aa)}\right) = \log\left(\frac{\pi_1^*}{1 - \pi_1^*}\right) - \log\left(\frac{\pi_2^*}{1 - \pi_2^*}\right) \\ &= \log\left(\frac{P(Aa | Y = 1)}{1 - P(Aa | Y = 1)}\right) - \log\left(\frac{P(Aa | Y = 0)}{1 - P(Aa | Y = 0)}\right) = \log\left(\frac{\pi_1}{1 - \pi_1}\right) - \log\left(\frac{\pi_2}{1 - \pi_2}\right).\end{aligned}$$

Interpretation of Logistic Regression

- β is the log-odds ratio of being affected/case for individuals with genotype Aa ($X = 1$ copy of allele A) compared with being affected/case for individuals with genotype aa ($X = 0$ copies of allele A)
- β also equals to the log-odds ratio of having genotype Aa among cases compared with having genotype Aa among controls.

Logistic Regression Likelihood

- Probability of outcome:

$$P(Y_i = 1 \mid X_i = x_i) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)},$$

$$P(Y_i = 0 \mid X_i = x_i) = \frac{1}{1 + \exp(\alpha + \beta x_i)}.$$

- Compact form (for $y_i = 0$ or 1):

$$P(Y_i = y_i \mid X_i = x_i) = \frac{\exp((\alpha + \beta x_i)y_i)}{1 + \exp(\alpha + \beta x_i)}.$$

- Assuming independence:

$$P(Y_1 = y_1, \dots, Y_n = y_n \mid X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(Y_i = y_i \mid X_i = x_i). \quad 27$$

Logistic Regression Likelihood

- Likelihood function:

$$L(\alpha, \beta) = \prod_{i=1}^n \frac{\exp((\alpha + \beta x_i)y_i)}{1 + \exp(\alpha + \beta x_i)}.$$

- Log-likelihood:

$$l(\alpha, \beta) = \sum_{i=1}^n [(\alpha + \beta x_i)y_i - \log(1 + \exp(\alpha + \beta x_i))].$$

- Estimation:

- Parameters α, β obtained by Maximum Likelihood Estimation (MLE)
- No closed-form solution → need iterative methods (Newton-Raphson, Fisher scoring, etc.)

Logistic Regression Likelihood (More Advanced)

- | | aa | Aa | Total |
|----------|-------|-------|-------|
| Cases | r_0 | r_1 | r |
| Controls | s_0 | s_1 | s |
| Total | n_0 | n_1 | n |

- Results for β are identical to the ones derived from a 2×2 table for log OR.

$$\hat{\alpha} = \log\left(\frac{r_0}{s_0}\right), SE(\hat{\alpha}) = \sqrt{\frac{1}{r_0} + \frac{1}{s_0}},$$

$$\hat{\beta} = \log\left(\frac{r_1 s_0}{r_0 s_1}\right), SE(\hat{\beta}) = \sqrt{\frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{s_1}}.$$

MLE Derivation Details (More Advanced)

- Note that $r_0 : Y = 1, X = 0, r_1 : Y = 1, X = 1, s_0 : Y = 0, X = 0, s_1 : Y = 0, X = 1$. We can simplify the log likelihood:

$$\begin{aligned} I(\theta) &= I(\alpha, \beta) = \sum_{i=1}^n \log \left(\frac{\exp((\alpha + \beta x_i) \cdot y_i)}{1 + \exp(\alpha + \beta x_i)} \right) \\ &= r_0 \cdot \log \left(\frac{\exp(\alpha)}{1 + \exp(\alpha)} \right) + r_1 \cdot \log \left(\frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)} \right) \\ &\quad + s_0 \log \left(\frac{1}{1 + \exp(\alpha)} \right) + s_1 \cdot \log \left(\frac{1}{1 + \exp(\alpha + \beta)} \right) \\ &= r_0 \alpha - r_0 \log(1 + \exp(\alpha)) + r_1 (\alpha + \beta) - r_1 \log(1 + \exp(\alpha + \beta)) \\ &\quad - s_0 \log(1 + \exp(\alpha)) - s_1 \log(1 + \exp(\alpha + \beta)) \\ &= r\alpha + r_1\beta - n_0 \log(1 + \exp(\alpha)) - n_1 \log(1 + \exp(\alpha + \beta)). \end{aligned}$$

MLE Derivation Details (More Advanced)

- Obtain the score functions

$$\frac{\partial I(\theta)}{\partial \alpha} = r - n_0 \frac{\exp(\alpha)}{1 + \exp(\alpha)} - n_1 \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}$$

$$\frac{\partial I(\theta)}{\partial \beta} = r_1 - n_1 \frac{\exp(\alpha + \beta)}{1 + \exp(\alpha + \beta)}.$$

MLE Derivation Details (More Advanced)

- Calculate the MLE

$$\frac{\partial I(\theta)}{\partial \alpha} = 0, \frac{\partial I(\theta)}{\partial \beta} = 0 \implies$$

$$r - n_0 \frac{\exp(\hat{\alpha})}{1 + \exp(\hat{\alpha})} - r_1 = 0 \implies \exp(\hat{\alpha}) = \frac{r_0}{s_0} \implies \hat{\alpha} = \log \frac{r_0}{s_0}$$

$$\frac{\exp(\hat{\alpha} + \hat{\beta})}{1 + \exp(\hat{\alpha} + \hat{\beta})} = \frac{r_1}{n_1} \implies \exp(\hat{\alpha} + \hat{\beta}) = \exp(\hat{\alpha}) \cdot \exp(\hat{\beta}) = \frac{r_1}{n_1 - r_1} = \frac{r_1}{s_1}$$
$$\implies \exp(\hat{\beta}) = \frac{r_1}{s_1} / \frac{r_0}{s_0} = \frac{r_1 s_0}{r_0 s_1} \implies \hat{\beta} = \log \frac{r_1 s_0}{r_0 s_1}.$$

- Variance calculation involves the second derivatives and the Fisher's information.

Logistic Regression Inference

- How would we perform a formal hypothesis test of

$$H_0 : \beta = 0$$

- Wald test would be identical to the test derived for testing $\Delta = 0$

$$T = \left(\frac{\hat{\beta} - 0}{SE(\hat{\beta})} \right)^2 \sim \chi^2_1.$$

$$\hat{\beta} = \log \left(\frac{r_1 s_0}{r_0 s_1} \right),$$

$$SE(\hat{\beta}) = \sqrt{\frac{1}{r_0} + \frac{1}{r_1} + \frac{1}{s_0} + \frac{1}{s_1}}.$$

Logistic Regression Inference

- LRT

$$T = 2 \left(\log \left(L_{H_1}(\hat{\alpha}, \hat{\beta}) \right) - \log \left(L_{H_0}(\tilde{\alpha}, \beta = 0) \right) \right) \sim \chi^2_1.$$

- $\tilde{\alpha} = ?$

$$\begin{aligned} H_0 : \beta &= 0, \\ \log \left(\frac{\mu(X)}{1 - \mu(X)} \right) &= \alpha. \end{aligned}$$

$$L(\theta) = L(\alpha, 0) = \prod_{i=1}^n \frac{\exp((\alpha) \cdot y_i)}{1 + \exp(\alpha)}$$

Logistic Regression Inference

$$\begin{aligned} I(\theta) &= I(\alpha, 0) = \sum_{i=1}^n \log \left(\frac{\exp((\alpha) \cdot y_i)}{1 + \exp(\alpha)} \right) \\ &= r \cdot \log \left(\frac{\exp(\alpha)}{1 + \exp(\alpha)} \right) + s \cdot \log \left(\frac{1}{1 + \exp(\alpha)} \right) \\ &= r \cdot \alpha - n \cdot \log(1 + \exp(\alpha)) \\ \frac{\partial I(\theta)}{\partial \alpha} &= r - n \cdot \frac{\exp(\alpha)}{1 + \exp(\alpha)} \\ \frac{\partial I(\theta)}{\partial \alpha} = 0 &\implies \tilde{\alpha} = \log \frac{r/n}{(n-r)/n} = \log \frac{\mu}{1-\mu} \end{aligned}$$

- This is not surprising since r/n is a pooled estimate of $\mu = P(Y = 1)$ regardless of the value of X .

Logistic Regression Inference

- Score test involves the score function and the Fisher's information evaluated under the null hypothesis that $\beta = 0$.

$$T = S(\tilde{\alpha}, \beta = 0)' I(\tilde{\alpha}, \beta = 0)^{-1} S(\tilde{\alpha}, \beta = 0) \sim \chi_1^2.$$

- Note that the CI for OR is derived from the CI for log OR, e.g. 95%CI for OR is

$$(\exp(\hat{\beta} - 1.96SE(\hat{\beta})), (\exp(\hat{\beta} + 1.96SE(\hat{\beta})))$$

Example

- The data below come from the study by Knowler et al. (1988), on the association between IDDM type 2 and a haplotype from the GM system human immunoglobulin *G*. These data include all individuals in a sample of 4,920 Native Americans of the Pima and Papago tribes. In this example, think of the GM haplotype as just an allele at a suspected DSL.

GM haplotype	# subjects	#(%) with IDDM
Present	293	23(7.9)
Absent	4627	1343(29.0)

Example

- We can reformulate the data:

GM haplotype	affected/case (%)	unaffected/control	Total
Present, D	23(7.9)	270	293
Absent, d	1343(29.0)	3284	4627
Total	1366	3554	4920

- We are interested in comparing 7.9% with 29%.

Power and Significance

- Power $1 - \beta$: probability of detecting an effect when it truly exists.
- Significance level α : probability of false positive (rejecting H_0 when true)

- | Effect | Detect | Not Detect |
|--------|-------------|--------------|
| True | $1 - \beta$ | β |
| False | α | $1 - \alpha$ |

- Goal: minimize α and maximize $1 - \beta$.

Power Estimation

- Many genetic association tests follow a normal or chi-squared distribution
- Under H_1 : distribution becomes noncentral chi-squared with noncentrality parameter (NCP) λ
- For a test statistic T :
 - $H_0 : E(T) = df$ (central χ^2)
 - $H_1 : E(T) = df + NCP$ (noncentral χ^2)

How to Increase Power

- Example (allelic test at a SNP): true effect size $a = \log(OR)$, risk allele frequency f and proportion of cases $\phi = r/n$.
- The noncentrality parameter:

$$NCP \approx 2\phi(1 - \phi)na^2f(1 - f)$$

- Increase sample size n
- Increase effect size a (e.g., extreme case selection)
- Adjust case/control ratio ϕ

Linkage Disequilibrium (LD)

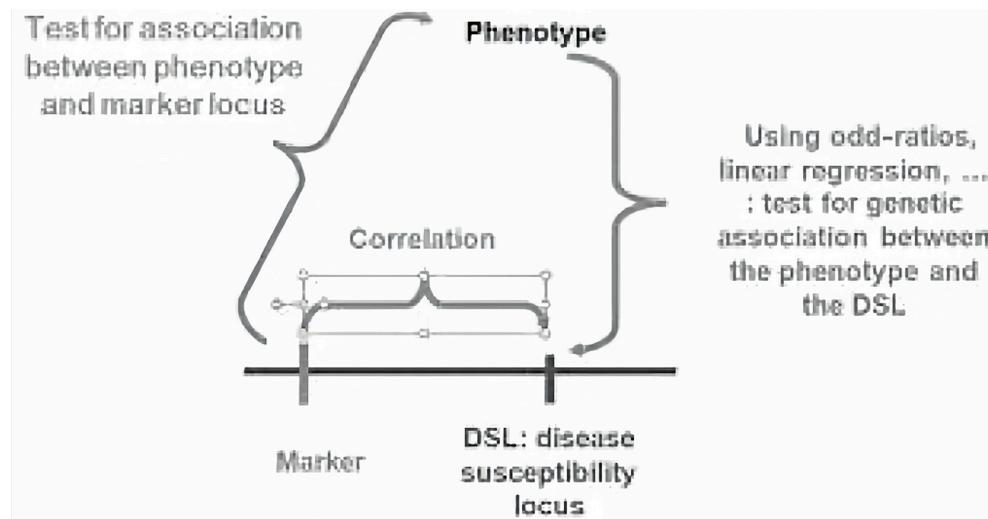
- LD is commonly measured using the Pearson correlation (ρ) between genotypes at two markers



Source: Publicly available from Google Images

Indirect Association

- We usually test genetic markers, not the actual causal mutation
- A marker may be correlated with the true disease-causing variant → indirect association
- Linkage disequilibrium (LD) between a marker and the causal variant creates an observed association with the phenotype



Source: Publicly available from Google Images

International HapMap Project

- Multinational project launched at the end of the Human Genome Project
- **Main goal:** provide data to estimate **linkage disequilibrium (LD)** across populations
- DNA samples collected from 4 groups:
 - 30 Yoruba trios (Nigeria)
 - 30 CEPH trios (European ancestry)
 - 45 Japanese (Tokyo)
 - 45 Han Chinese (Beijing)

International HapMap Project

- Data: SNP genotypes for 270 individuals + allele frequencies in each population
- Provides standard LD measures (e.g., ρ^2)
- Impact: reduced the number of SNPs needed for GWAS
 - From ~10 million SNPs → ~500,000 tag SNPs

SNP Microarrays

- SNP: single nucleotide polymorphism (usually biallelic: A/a)
- SNP arrays detect common variants ($\geq 5\%$ frequency) in a population
- Nearby SNPs are correlated → ~300K–600K tag SNPs capture most variation
- Modern platforms cover >1 million SNPs to map common variation

LD Varies by Population

- LD patterns differ across populations
- Example:
 - [LDlink tool](#)
 - SNP pair: *rs146366639* and *rs6661489*

What's next

- Population Substructure in Association Studies
- Association Analysis in Family Designs

What questions do you have about anything from today?

