Lecture 6: Population-Based Association Analysis

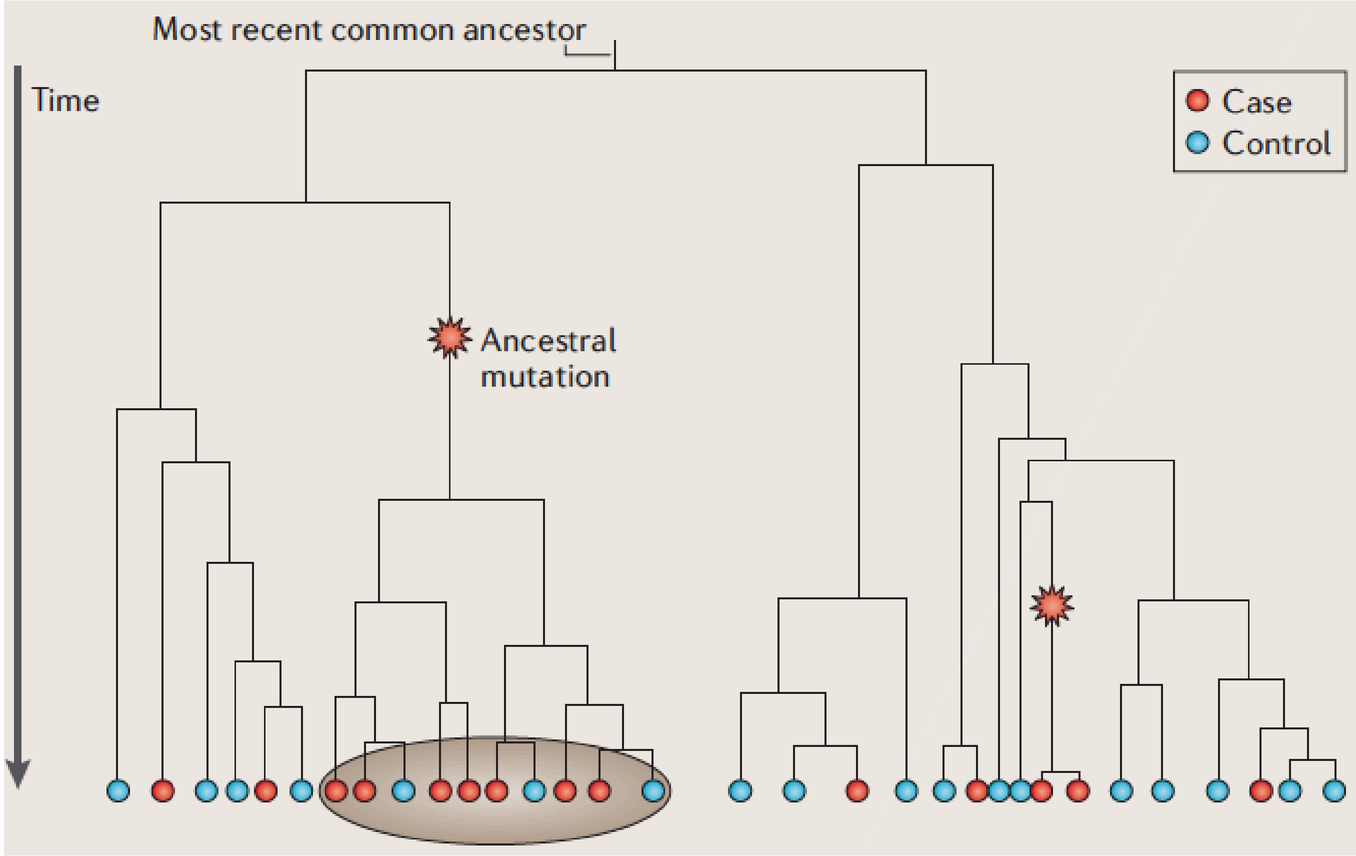
Population-Based Association Analysis

- **Key Goals of Association Analysis**
  - Test associations between each locus and the interested trait
  - Understand the biological function of these associated loci (Challenging)
- **Association Analysis**
  - Dichotomous traits (i.e., Case-control studies)
  - Quantitative traits (i.e., height, BMI, Age-to-onset)
- **Population-based association analysis**: study unrelated individuals (not relatives).

Rationale for Association Studies

Trace transmissions of phenotype over generations is no longer possible. Thus the association study must rely on the **correlations** of current phenotype with current marker alleles.

Such a correlation exists when one or more groups of cases share a relatively recent common ancestor (share a mutated allele) at a causal locus.

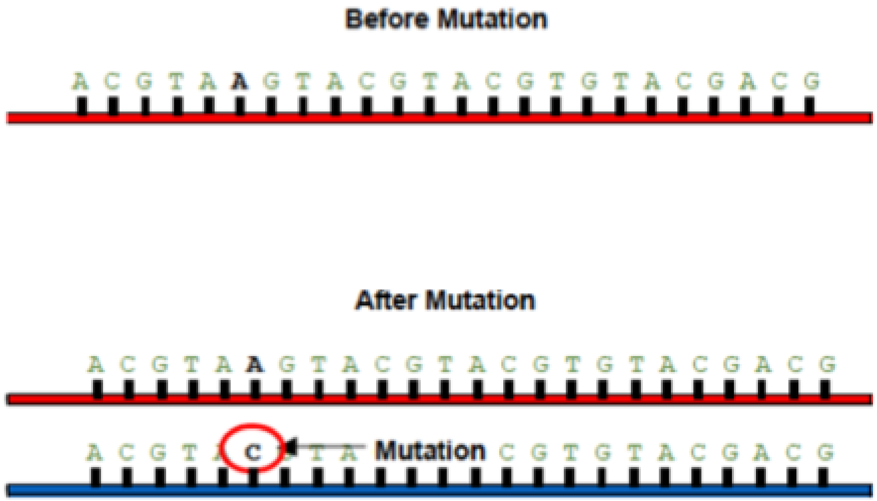


Linkage Disequilibrium (LD)

- **Linkage Disequilibrium (LD)** is the non-random association of alleles at different loci in a given population..
- Nearby markers are likely to be correlated, why?
- Origin of LD?

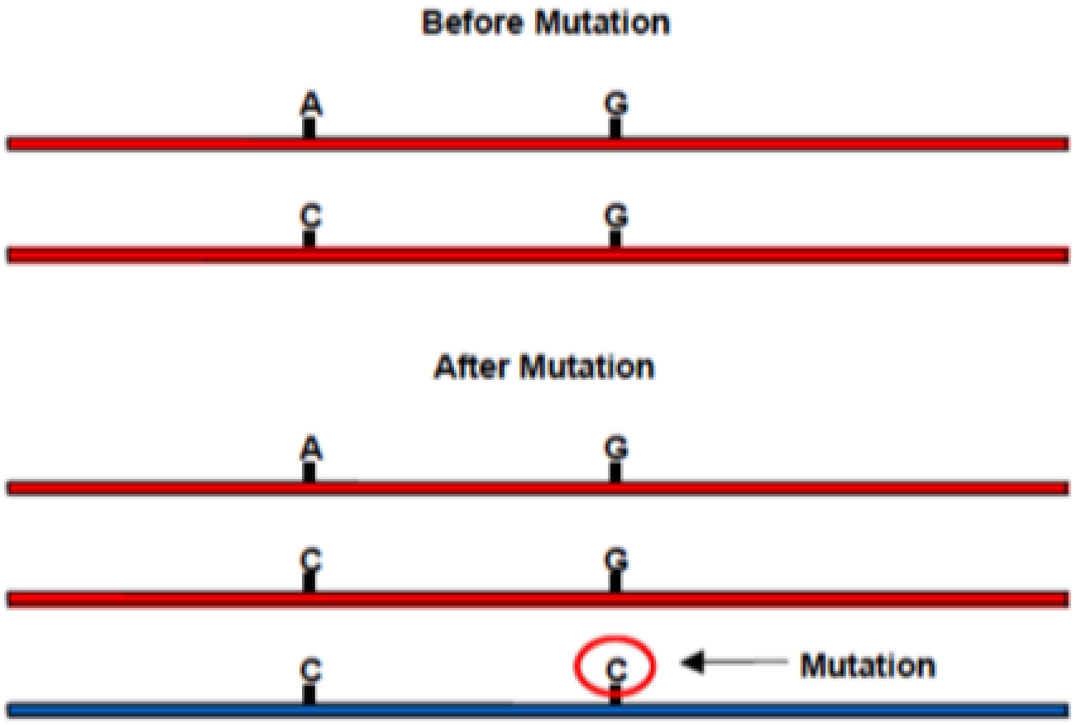
Linkage Disequilibrium (LD)

- Consider the history of two neighboring single nucleotide polymorphism (SNP)
- SNPs exist today arose through ancient mutation events…


Before Mutation
A C G T A A G T A C G T A C G T G T A C G A C G
After Mutation
A C G T A A G T A C G T A C G T G T A C G A C G
A C G T A C G T A C G T G T A C G A C G
Mutation

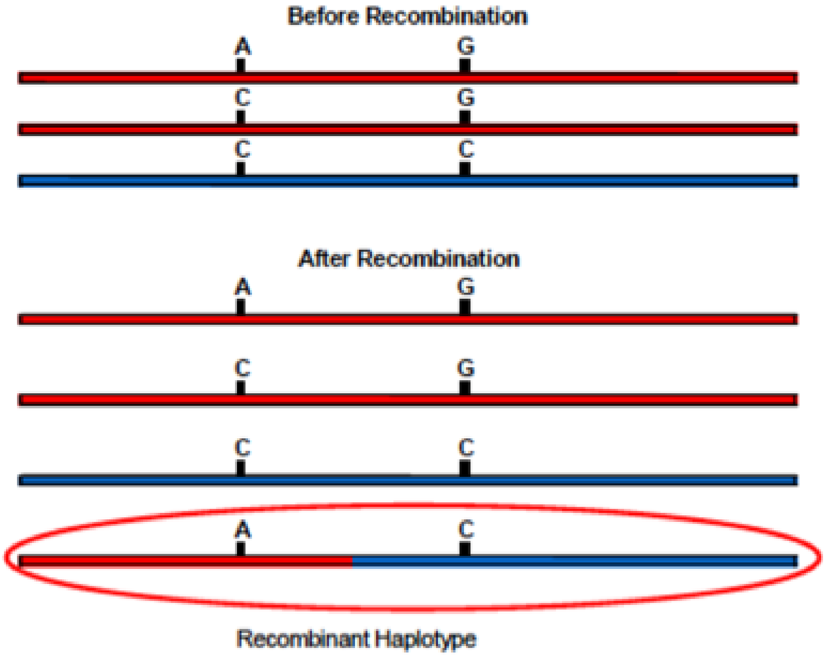
Linkage Disequilibrium (LD)

- One SNP arose first and then the other …



Linkage Disequilibrium (LD)

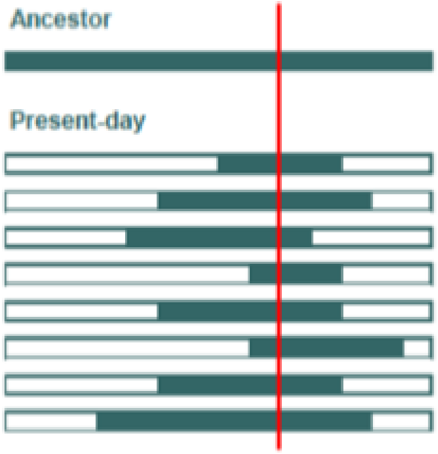
- Recombination generates new arrangements for the ancestral alleles



Linkage Disequilibrium (LD)

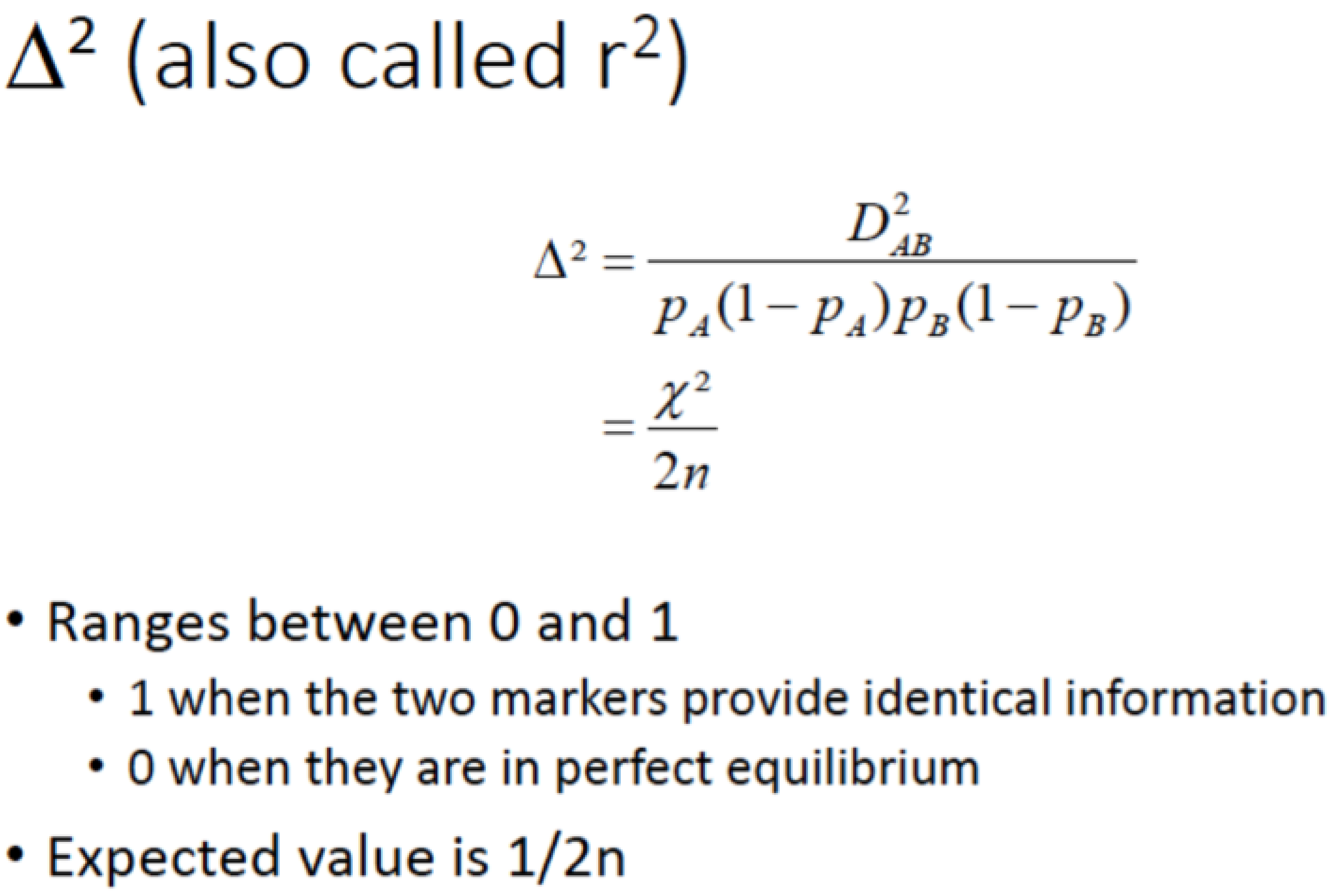
- Chromosomes are mosaics

- Extent and conservation of mosaic pieces depends on
  - Recombination rate
  - Mutation rate
  - Population size
  - Natural selection



- Combinations of alleles at very close markers reflect ancestral haplotypes

Linkage Disequilibrium (LD)



Δ² (also called r²)

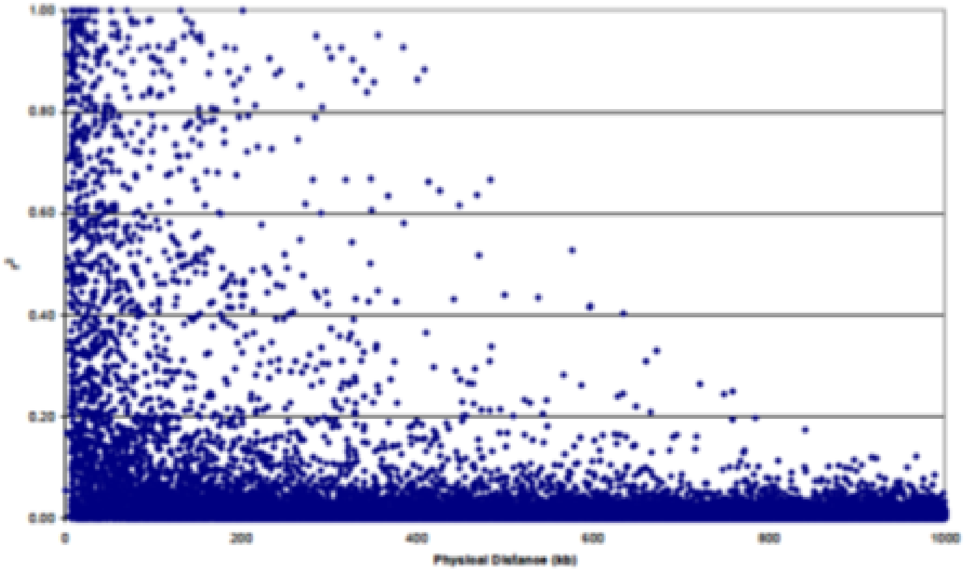
$$\begin{aligned}\Delta^2 &= \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}\\ &= \frac{\chi^2}{2n}\end{aligned}$$

- Ranges between 0 and 1
  - 1 when the two markers provide identical information
  - 0 when they are in perfect equilibrium
- Expected value is 1/2n

Linkage Disequilibrium (LD)

Genotype data for multiple samples from a population

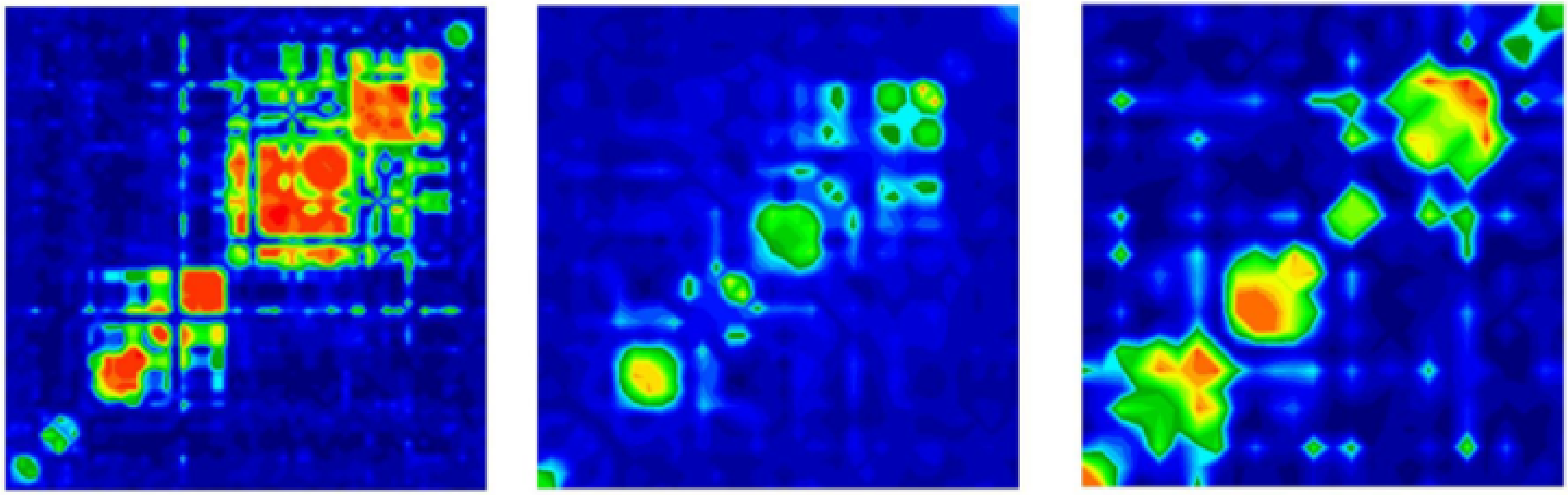
- SNP1: x1 = (0, 1, 2, 1, 0, 0, ...)
- SNP2: x2 = (1, 1, 2, 0, 0, 0, ...)
- r^2 = (correlation(x1, x2) $)^2$
- Raw r^2 from CHR22



Dawson et al, *Nature*, 2002

Linkage Disequilibrium (LD)

Linkage Disequilibrium in Three Regions



2q13
(63 markers)

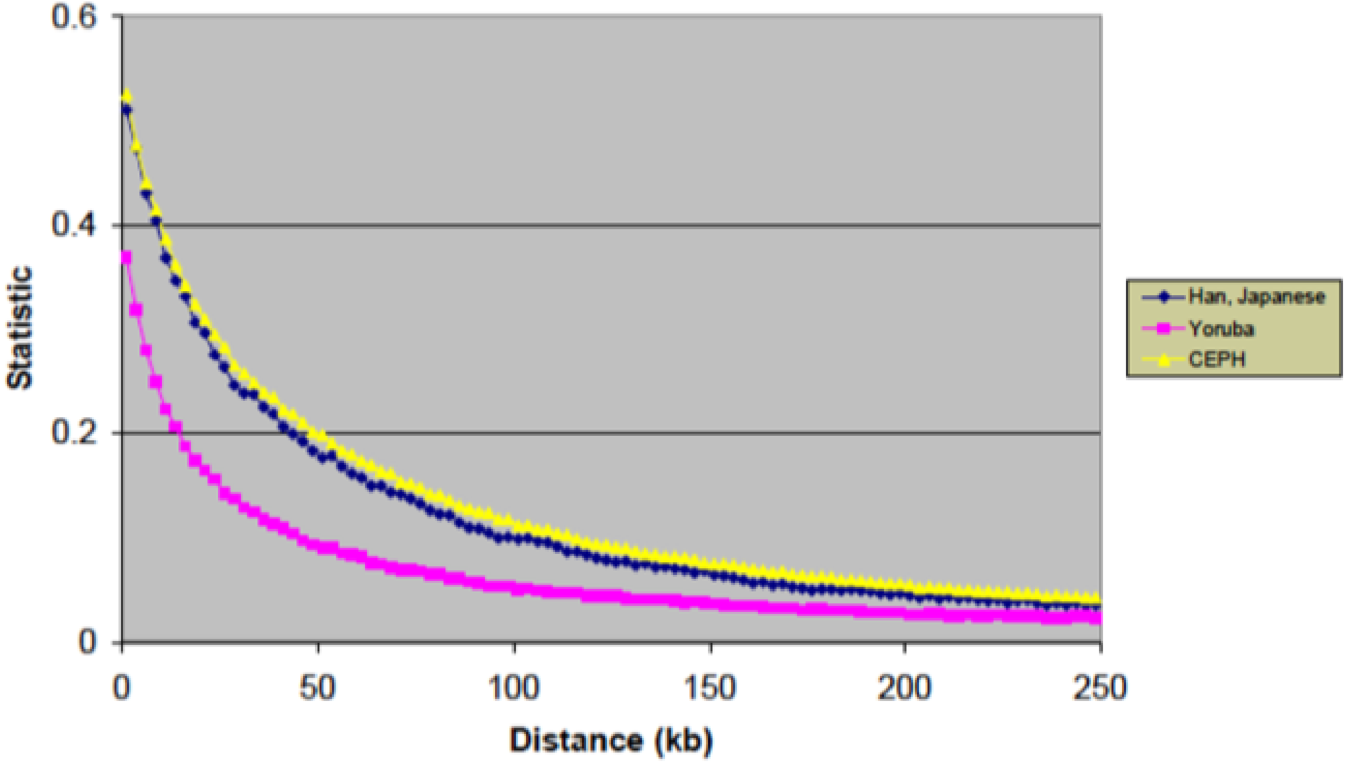
13q13
(38 markers)

14q11
(26 markers)

Abecasis et al, *Am J Hum Genet*, 2001

Linkage Disequilibrium (LD)

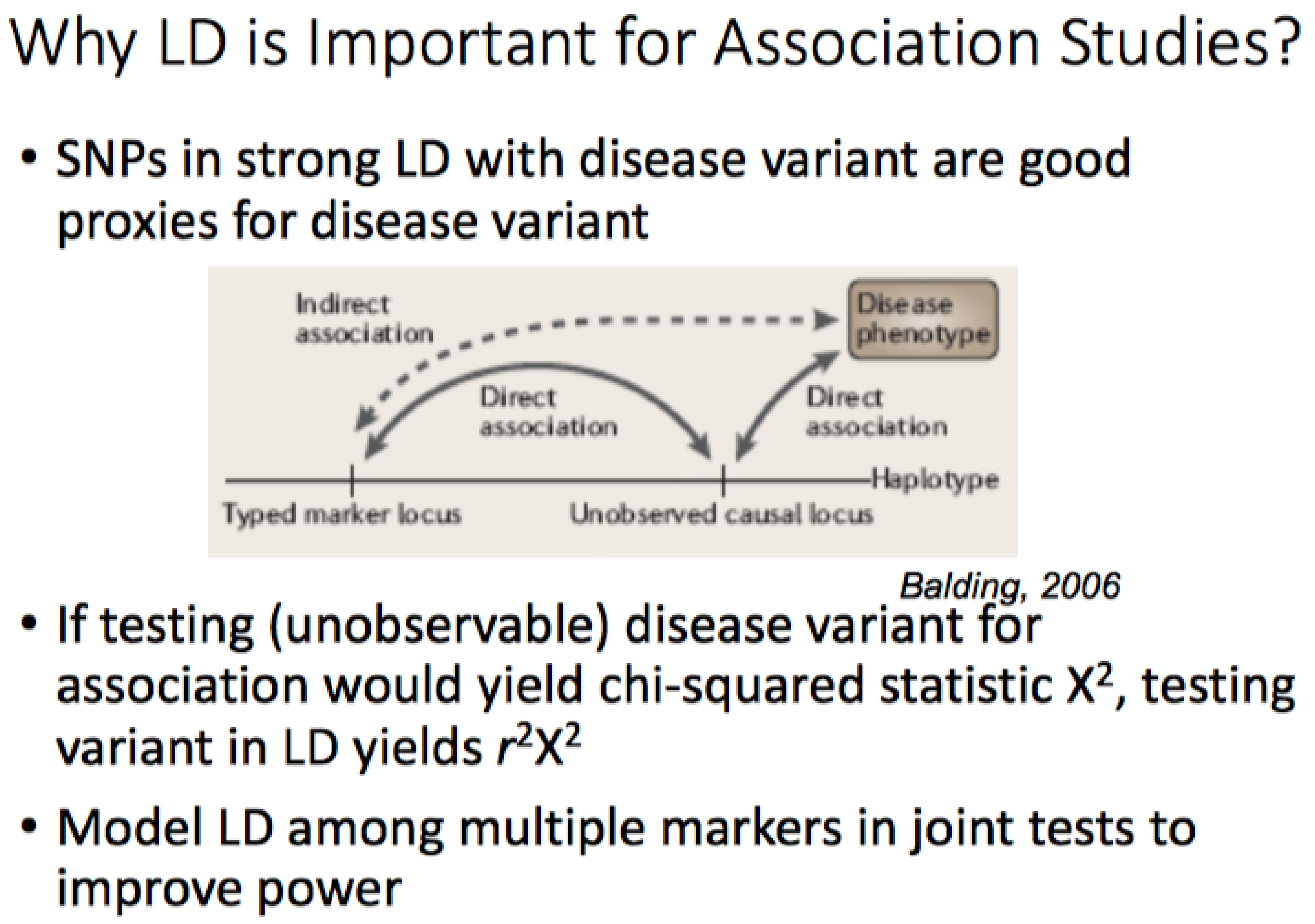
Comparing Populations ...



LD extends further in CEPH and the Han/Japanese than in the Yoruba

International HapMap Consortium, *Nature*, 2005

Linkage Disequilibrium (LD)



Why LD is Important for Association Studies?

- SNPs in strong LD with disease variant are good proxies for disease variant

Indirect association
Disease phenotype
Direct association
Direct association
Haplotype
Typed marker locus
Unobserved causal locus

Balding, 2006

- If testing (unobservable) disease variant for association would yield chi-squared statistic X^2, testing variant in LD yields r^2X^2
- Model LD among multiple markers in joint tests to improve power

Types of Population-Based Association Studies

Candidate polymorphism (rs12255372)
Focus on an individual polymorphism that is a disease susceptibility locus, or in LD with the disease susceptibility locus.

Candidate gene (TCF7L2 for Type 2 Diabetes)
Typing/sequencing a genetic region around the candidate gene (often designed to include coding sequence and flanking regions, and perhaps including splice or regulatory sites).

The gene can be a positional candidate from prior linkage analysis.

Fine mapping
The candidate region might have been identified by linkage analysis and contain perhaps 5–50 genes, 1-10 Mb length, hundreds or thousands of SNPs.

Genome-wide Association Studies (GWAS)
$\geq 0.5M$ well-chosen SNP markers throughout the genome (often imputed to higher resolution with $\sim 10M$ SNPs), or $\geq 10M$ SNPs from whole genome sequencing data. Without prior knowledge

Study Designs

Use standard epidemiological designs for studying the relationship between general risk factors and disease.

Case-Control Study

Ascertain subjects on the basis of dichotomous disease outcome

- informative
- efficient
- low cost
- selection bias, recall bias
- cannot estimate disease prevalence

Cohort Study

Follow subjects over time for development of disease and/or risk factors

- no selection and recall bias
- reliable pre-disease exposure information
- a full range of diseases and traits
- many years of follow-up

Association Test for Dichotomous Traits

Standard contingency table based methods:

- Chi-square or likelihood ratio test
- Large-sample Z-test comparing two proportions
- Fisher's exact test

Frequently-used tests:

1. Genotypic Association test (2-*df* test)
2. Genotypic Association test with dominant/recessive disease models
3. Allelic Association test
4. Cochran-Armitage tend test
5. Logistic regression

Genotypic Association Test

- Compare genotype frequencies in cases and controls in a 2×3 table
- Not assuming any specific disease model

	AA	Aa	aa	Total
Case	n_{10}	n_{11}	n_{12}	$n_{1.}$
Control	n_{00}	n_{01}	n_{02}	$n_{0.}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	n

The genotype/codominant test: D – disease status; G – genotype

$$H_0 : \Pr(D = 1|Geno = \text{AA}) = \Pr(D = 1|Geno = \text{Aa}) = \Pr(D = 1|Geno = \text{aa})$$
$$H_1 : \text{At least one inequality holds}$$

The standard $2\ df$ Pearson χ^2 test of independence for a 2×3 table is:

$$X_G^2 = \sum_{i=0,1} \sum_{j=0,1,2} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 2$$

– $O_{ij} = n_{ij}$: observed count in the cell

– $E_{ij} = n_{i.}n_{.j}/n$: expected count under independence: $np_{D=i}p_{G=j} = n(n_{i.}/n)(n_{.j}/n)$

Genotypic Association Test: Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G

	GG	GT	TT	Total
Case	661	255	20	936
Control	724	354	50	1128
Total	1385	609	70	2064

$$X_G^2 = (661 - 628.08)^2/628.08 + \ldots \approx 14.08 \quad \sim \quad \chi^2,\ df = 2$$

$$p = .0009$$

$\Pr(Geno|D)$

	GG	GT	TT	Total
Case	0.71	0.27	0.02	1.0
Control	0.64	0.31	0.05	1.0

$\Pr(D|Geno)$

	GG	GT	TT	Total
Case	0.48	0.42	0.29	0.45
Control	0.52	0.58	0.71	0.55

Dominant or Recessive Disease Models

- Compare frequencies of AA or Aa with aa in cases and controls in a 2×2 table
- Assume dominant or recessive Mendelian disease model
- More powerful than genotype test if the disease model is true

With dominant disease model:

	AA or Aa	aa	Total
Case	$n_{10} + n_{11}$	n_{12}	$n_{1.}$
Control	$n_{00} + n_{01}$	n_{02}	$n_{0.}$
Total	$n_{.0} + n_{.1}$	$n_{.2}$	n

$$H_0: \ \Pr(D = 1|\text{AA}) = \Pr(D = 1|\text{Aa}) = \Pr(D = 1|\text{aa})$$

$$H_1: \ \Pr(D = 1|\text{AA or Aa}) \neq \Pr(D = 1|\text{aa})$$

The standard $1 \ df$ Pearson χ^2 test of independence for a 2×2 table is:

$$X_D^2 = \sum_{i=0,1} \sum_{j=0,1} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2, \ df = 1$$

How to obtain E_{ij}?

Dominant or Recessive Disease Models: Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G
- Allele T is dominant to G

	GG	GT+TT	Total
Case	661	255+20=275	936
Control	724	354+50=404	1128
Total	1385	609+70=679	2064

$$X_D^2 \approx 9.60 \quad \sim \quad \chi^2,\ df = 1$$

$$p = .0019$$

Allelic Association Test

- Compare frequencies of alleles A and a in cases and controls in a 2×2 table
- **Assume additive disease model**: the risk associated with the heterozygote genotype is intermediate between the two homozygotes. (mostly used model)
- Assume HWE: allele frequencies in a population will remain constant from generation to generation, with random mating and in the absence of other evolutionary influences (selection, mutation, genetic drift)
- The allele test is the most powerful test for additive model.

	A	a	Total
Case	$n_{1A} = 2n_{10} + n_{11}$	$n_{1a} = n_{11} + 2n_{12}$	$2n_{1.}$
Control	$n_{0A} = 2n_{00} + n_{01}$	$n_{0a} = n_{01} + 2n_{02}$	$2n_{0.}$
Total	$n_{.A} = 2n_{.0} + n_{.1}$	$n_{.a} = n_{.1} + 2n_{.2}$	$2n$

The allele test:

$$H_0 : \Pr(\mathrm{A}|D = 1) = \Pr(\mathrm{A}|D = 0)$$

The standard $1\ df$ Pearson χ^2 test of independence for a 2×2 table is:

$$X_L^2 = \sum_{i=0,1} \sum_{j=0,1} (O_{ij} - E_{ij})^2 / E_{ij} \quad \sim \quad \chi^2,\ df = 1$$

Allele Test with Allelic Frequencies

It can also be derived as a test of the difference in allelic frequencies. Let

$$\bar{p}_{\text{Case}} \equiv \Pr(\mathsf{A}|D = 1) = n_{1A}/2n_{1.}$$

$$\bar{p}_{\text{Control}} \equiv \Pr(\mathsf{A}|D = 0) = n_{0A}/2n_{0.}$$

$$\bar{p} \equiv \Pr(\mathsf{A}) = n_{.A}/2n$$

Under H_0,

$$\mathrm{E}(\bar{p}_{\text{Case}} - \bar{p}_{\text{Control}}) = 0$$

Under HWE,

$$\widehat{\mathrm{Var}}(\bar{p}_{\text{Case}} - \bar{p}_{\text{Control}}) = \bar{p}(1 - \bar{p})\left(\frac{1}{2n_{0.}} + \frac{1}{2n_{1.}}\right) = \bar{p}(1 - \bar{p})\frac{n}{2n_{0.}n_{1.}}$$

Hence,

$$Z_L = 2\sqrt{n_{0.}n_{1.}}(\bar{p}_{\text{Case}} - \bar{p}_{\text{Control}})/\sqrt{2n\bar{p}(1 - \bar{p})} \quad \sim \quad N(0, 1)$$

Allele Test: Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G

	G	T	Total
Case	1577	295	1872
Control	1802	454	2256
Total	3379	749	4128

$$X_L^2 \approx 13.13 \quad \sim \quad \chi^2,\ df = 1$$

$$p = .0003$$

$$Z_L = \ 3.63$$

$$p = \ .0003$$

Cochran-Armitage Trend Test

Define X as the number of A allele in an individual and compare the means of X in the case and control groups:

	AA	Aa	aa	Total	$\bar{p}_.$	$\bar{X}_.$
Case	n_{10}	n_{11}	n_{12}	$n_{1.}$	$\bar{p}_{\text{Case}} = (2n_{10}+n_{11})/2n_{1.}$	$\bar{X}_{\text{Case}} = 2\bar{p}_{\text{Case}}$
Control	n_{00}	n_{01}	n_{02}	$n_{0.}$	$\bar{p}_{\text{Control}} = (2n_{00}+n_{01})/2n_{0.}$	$\bar{X}_{\text{Control}} = 2\bar{p}_{\text{Control}}$
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	n	$\bar{p} = (2n_{.0}+n_{.1})/2n$	$\bar{X} = 2\bar{p}$

Under $H_0 : \text{E}(X|\text{Case}) = \text{E}(X|\text{Control})$,

$$\text{E}(\bar{X}_{\text{Case}} - \bar{X}_{\text{Control}}) = 0$$

$$\text{Var}(\bar{X}_{\text{Case}} - \bar{X}_{\text{Control}}) = \text{Var}(X)\left(\frac{1}{n_{0.}} + \frac{1}{n_{1.}}\right)$$

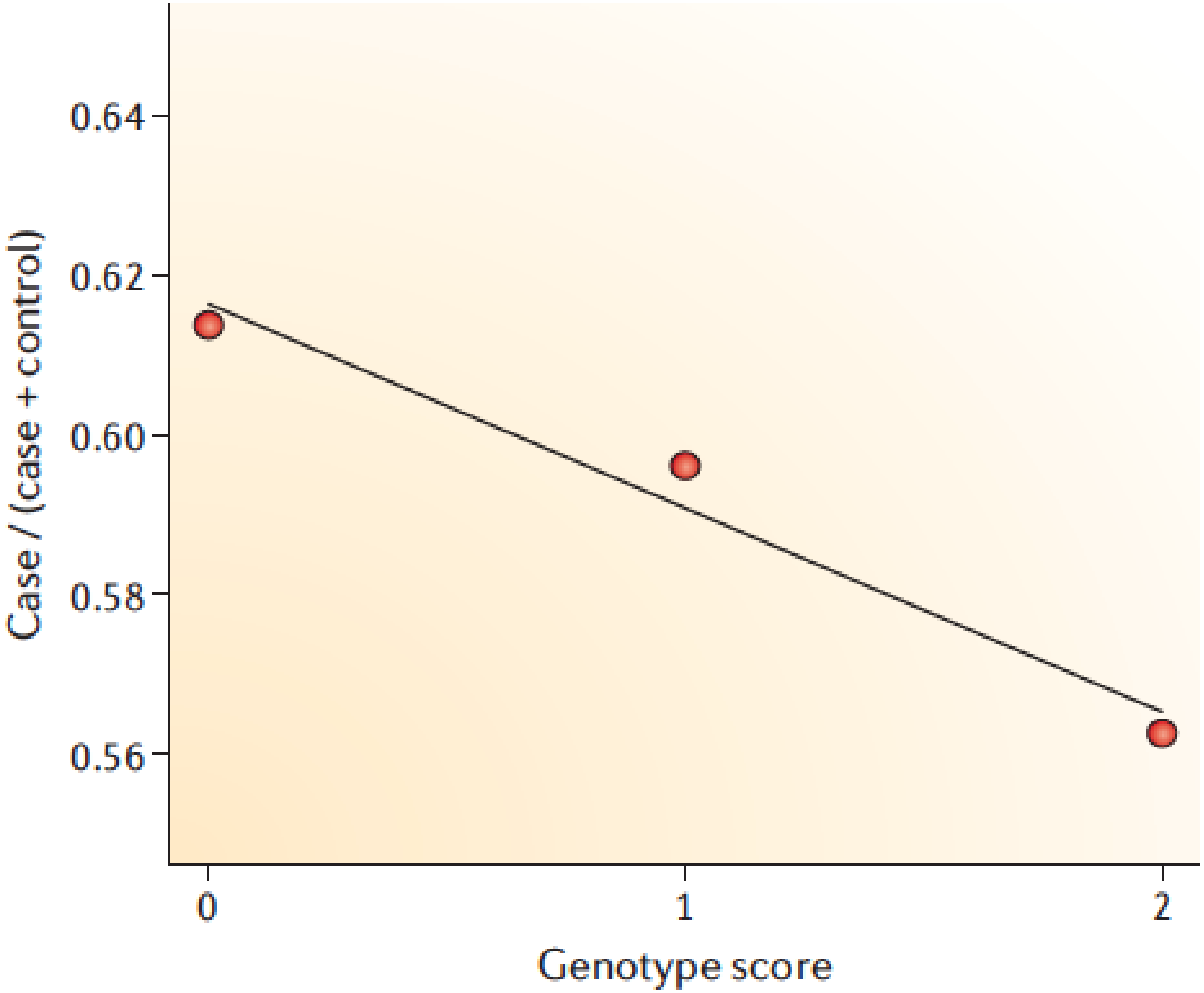
in which $\text{Var}(X)$ can be estimated, without assuming HWE, by

$$\widehat{\text{Var}}(X) = \frac{4n_{.0}+n_{.1}}{n} - \bar{X}^2 \qquad \text{(Why?)}$$

Hence,

$$Z_T = (\bar{X}_{\text{Case}} - \bar{X}_{\text{Control}}) / \sqrt{\frac{4n_{.0}+n_{.1}-n\bar{X}^2}{n_{0.}n_{1.}}} \quad \sim \quad N(0,1)$$

Cochran-Armitage Trend Test (continued)



Cochran-Armitage Trend Test vs. Allele Test

$$Z_T^2 = Z_L^2 \frac{2\bar{p}(1-\bar{p})}{(4n_{.0} + n_{.1} - n\bar{X}^2)/n}$$

Commonality:

- Same null hypothesis: $H_0: \; p_{\text{Case}} = p_{\text{Control}}$
- Both tests use the $\bar{p}_{\text{Case}} - \bar{p}_{\text{Control}}$ (in the numerator)
- Assume additive disease model

Difference: how the variance of the estimated allele frequencies is calculated.

- Allele test requires that HWE holds under H_0
- Trend test does not require HWE under H_0

Cochran-Armitage Trend Test vs. Allele Test (continued)

Sasieni(1997) showed:

- Allele test has inflated type I error if HWE fails
- Trend test is robust to violation of HWE
- The two tests are asymptotically equivalent if HWE holds

Observations:

- Power for trend and allele tests are similar even for small samples
- For complex diseases, it is rare to see departure from HWE
- Trend test is generally preferred for being robust with similar computation cost

Cochran-Armitage Trend Test: Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G

	GG	GT	TT	Total
Case	661	255	20	936
Control	724	354	50	1128
Total	1385	609	70	2064

$$Z_T^2 = 13.04 \quad \sim \quad \chi^2,\ df = 1$$

$$p = .0003$$

Summary of Association Tests for the Example

Test	X^2	df	p-value
Genotype	14.08	2	.0009
Dominant	9.60	1	.0019
Allele (Additive)	13.13	1	.0003
Trend (Additive)	13.04	1	.0003

Measure of Association Strength: Odds Ratio

	Exposed (E)	Not Exposed ($\bar{E}$)
Case (D)	a	b
Control ($\bar{D}$)	c	d

Odds ratio:

$$\begin{aligned} OR &= \frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(\bar{D}|\bar{E})} \\ &= \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})} \\ &= ad/bc \end{aligned}$$

– Exposed = carry certain genotype

– Counts pertain to individuals, not alleles.

Measure of Association Strength: Odds Ratio (continued)

Genotype Model ($\bar{E}$=aa)

	AA	Aa	aa
Case	n_{10}	n_{11}	n_{12}
Control	n_{00}	n_{01}	n_{02}

$$OR_{het} = (n_{11}n_{02})/(n_{01}n_{12})$$
$$OR_{hom} = (n_{10}n_{02})/(n_{00}n_{12})$$

Dominant Model ($\bar{E}$=aa)

	AA or Aa	aa
Case	$n_{10} + n_{11}$	n_{12}
Control	$n_{00} + n_{01}$	n_{02}

$$OR_D = [(n_{10} + n_{11})n_{02}]/[(n_{00} + n_{01})n_{12}]$$

Allele Model ($\bar{E}$=a)

	A	a
Case	$2n_{10} + n_{11}$	$n_{11} + 2n_{12}$
Control	$2n_{00} + n_{01}$	$n_{01} + 2n_{02}$

$$OR_L = [(2n_{10} + n_{11})(n_{01} + 2n_{02})]/[(2n_{00} + n_{01})(n_{11} + 2n_{12})]$$

Trend Model

estimate OR by maximum likelihood

OR_T: logistic regression

Summary of OR's for the Example

- TCF7L2 for Type 2 Diabetes in Finns
- SNP rs12255372 has alleles T and G

Comparison	OR
GT vs. GG	$OR_{het} = 1.27$
TT vs. GG	$OR_{hom} = 2.28$
T- vs. GG	$OR_D = 1.34$
Allele T vs. G	$OR_L = 1.35$
Trend	$OR_T = 1.36$

Calculating Confidence Interval for Odds Ratio

In large samples and when OR is estimated from the contingent table, $\log(\widehat{OR})$ is approximately normally distributed, with estimated variance

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

where a, b, c, d are the cells contributing to the estimation of OR.

A $(1 - \alpha)100$th confidence interval for the population OR :

$$\exp^{\log(\widehat{OR}) \pm z_{(1-\alpha/2)} \sqrt{\widehat{\text{Var}}[\log(OR)]}}$$

where $z_{(1-\alpha/2)}$ is the $(1 - \alpha/2)100$th percentile of the standard normal.

Logistic Regression

- Y = dichotomous phenotype
- X = a coding for the genotype

Genotype	Codominant	Dominant	Recessive	Additive
AA	$X = (0, 1)^{\mathrm{T}}$	$X = 1$	$X = 1$	$X = 2$
Aa	$X = (1, 0)^{\mathrm{T}}$	$X = 1$	$X = 0$	$X = 1$
aa	$X = (0, 0)^{\mathrm{T}}$	$X = 0$	$X = 0$	$X = 0$

Assume a logistic regression model:

$$\log\left[\frac{\Pr(Y = 1|X)}{\Pr(Y = 0|X)}\right] = \beta_0 + \alpha C + \beta_1 X$$

where β_0 is the intercept, α is the coefficient for covariates C, and β_1 is the genetic effect-size (i.e., log(Odds-Ratio)).

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Logistic Regression vs. Contingency Table

- Likelihood ratio test of logistic regression $\approx$ chi-square tests for appropriate contingency tables.
- The estimated coefficients $=$ log of the corresponding odds ratios.
- For the additive model, the trend test $\approx$ likelihood ratio test from logistic regression with additive coding for X.
- Because the logistic regression operate on variables defined for individuals, not chromosomes, there is no underlying assumption about HWE.

Extension to other phenotypes:

- The phenotype Y can be a count or a continuous outcome.
- The generalized linear model is given by

  $$g[\mathrm{E}(Y|X)] = \beta_0 + \alpha C + \beta_1 X$$

  where $g(.)$ is a link function.
- $$H_0 : \beta_1 = 0$$

  $$H_a : \beta_1 \neq 0$$

CDCV Hypothesis

Hypothesis underlying association studies in this lecture:

Common-Disease Common-Variant (CDCV)

- Single-variant association studies are powerful only for common causal variants ($MAF > 5\%$)
- Common diseases tend to be late-onset (e.g., Type 2 Diabetes, Alzheimer's disease)

⇒ Selection pressure is expected to be weak on late-onset diseases and on variants that contribute only a small risk

⇒ Causal variants tend to become common in the population