Genetic association studies

Genetic association studies search for alleles which occur more (or less) frequent in affecteds than in unaffecteds.

Example:

Ankylosing spondylitis and HLA-B

Group	B27 positive	B27 negative	$\sum$
Cases	72	3	75
Controls	3	72	75

Genetic association studies represent a special case of epidemiological studies evaluating the association between an exposure E (example: smoking) and a disease D (example: lung cancer).

Measures for the strength of an association

Let E^+ (and E^-) denote the event that an individual is exposed (and nonexposed). The conditional probability $P(D \mid E^+)$ that an exposed individual becomes affected is the *risk of the exposure for the disease*.

The risk ratio

$$RR_E = \frac{P(D \mid E^+)}{P(D \mid E^-)}$$

is called the *relative risk* of the exposure E for the disease D.

The *odds* of becoming affected versus nonaffected for an exposed individual is given by $P(D \mid E^+)/(1 - P(D \mid E^+))$. Therefore, the ratio

$$OR_E = \frac{P(D \mid E^+)/(1 - P(D \mid E^+))}{P(D \mid E^-)/(1 - P(D \mid E^-))}$$

is called the *odds ratio* of the exposure E for the disease D.

Measures for the strength of an association

Exercise:

Show that $RR_E \neq 1$ implies $1 < RR_E < OR_E$ or $OR_E < RR_E < 1$.

Measures of the strength of an association

For a diallelic locus $\{A, a\}$, three different genotypes AA, Aa, and aa can be distinguished (i.e., three levels of "exposure"). The *genotype specific relative risks* are

$$RR_{AA} = \frac{P(D \mid AA)}{P(D \mid aa)} \quad \text{and} \quad RR_{Aa} = \frac{P(D \mid Aa)}{P(D \mid aa)}$$

Special cases:

- $RR_{AA} > RR_{Aa} = 1$: recessive effect
- $RR_{AA} = RR_{Aa} > 1$: dominant effect
- $RR_{AA} = (RR_{Aa})^2 > 1$: multiplicative effect

Case-control studies

In a case-control study, a group of individuals who developed the disease (the cases) and a group of individuals who did not develop the disease (the controls) are examined for the presence of the risk factor. If there is no relationship between exposure and disease, then the distribution of exposure among cases should be the same as the distribution of exposure among the controls.

In genetic epidemiology, the control group often consists of individuals who have not been examined for the disease.

Estimation of RR and OR in case-control studies

$P(D \mid E^+)$ and $P(D \mid E^-)$ cannot be estimated from case-control studies.

However, it can be shown that

$$RR_E = \frac{P(E^+ \mid D)/(1 - P(E^+ \mid D))}{P(E^+)/(1 - P(E^+))}$$

and

$$OR_E = \frac{P(E^+ \mid D)/(1 - P(E^+ \mid D))}{P(E^+ \mid \bar{D})/(1 - P(E^+ \mid \bar{D}))}.$$

Therefore, OR_E or RR_E can be estimated by case-control studies. More precisely, the fraction of cases being exposed provides an estimate for $P(E^+ \mid D)$ and the fraction of controls being exposed provides an estimate for $P(E^+ \mid \bar{D})$ (if the controls are unaffected) or an estimate of $P(E^+)$ (if the controls are a random sample from the population).

Testing for association

Group	Genotype AA	Aa	aa	$\sum$
Cases	$D_2(d_2)$	$D_1(d_1)$	$D_0(d_0)$	n^D
Controls	$C_2(c_2)$	$C_1(c_1)$	$C_0(c_0)$	n^C

- diallelic marker $\{A, a\}$
- n^D cases and n^C controls
- D_i: number of cases with i alleles A
- C_i: number of controls with i alleles A
- d_i: probability that a case possesses i alleles A
- c_i: probability that a control possesses i alleles A

Comparison of genotype frequencies

Null hypothesis:

$$H_0 : (d_2, d_1, d_0) = (c_2, c_1, c_0)$$

Under H_0, the maximum likelihood estimate for $d_i (= c_i)$ is given by $\frac{D_i + C_i}{n^D + n^C}$. Therefore, $e_i^D = n^D \cdot \frac{D_i + C_i}{n^D + n^C}$ and $e_i^C = n^C \cdot \frac{D_i + C_i}{n^D + n^C}$ are the expected (under H_0) numbers of cases and controls with i alleles A.

Let

$$T_G = \sum_{i=0}^{2} \frac{(D_i - e_i^D)^2}{e_i^D} + \sum_{i=0}^{2} \frac{(C_i - e_i^C)^2}{e_i^C}$$

Under H_0, the distribution of T_G can be approximated by a χ_2^2 distribution.

Comparison of allele frequencies

Group	Allele A	Allele a	$\sum$
Cases	$2D_2 + D_1$	$D_1 + 2D_0$	$2n^D$
Controls	$2C_2 + C_1$	$C_1 + 2C_0$	$2n^C$

With $\widehat{p}_A^D = \frac{2D_2 + D_1}{2n^D}$, $\widehat{p}_A^C = \frac{2C_2 + C_1}{2n^C}$, and $\widehat{p}_A = \frac{2D_2 + D_1 + 2C_2 + C_1}{2(n^D + n^C)}$, the test statistic of the χ^2 test for 2×2 tables becomes

$$T_A = \frac{(\widehat{p}_A^D - \widehat{p}_A^C)^2}{\widehat{p}_A \cdot (1 - \widehat{p}_A) \cdot (\frac{1}{2n^D} + \frac{1}{2n^C})}.$$

Under H_0, the distribution of T_A (allele test) can be approximated by a χ_1^2 distribution.

Armitage's trend test

It can be shown that the allele test is a valid test only in case that the genotype distribution at the marker locus is in Hardy-Weinberg equilibrium (HWE). Especially in case that there is an excess of homozygous individuals, the allele test can become anti-conservative. Armitage's trend test does not require the assumption of HWE and is based on the statistic

$$T_{\text{trend}} = \frac{(\widehat{p}_A^D - \widehat{p}_A^C)^2}{\left(\widehat{p}_A \cdot (1 - \widehat{p}_A) + (\widehat{p}_{AA} - \widehat{p}_A^2)\right) \cdot (\frac{1}{2n^D} + \frac{1}{2n^C})}.$$

Under H_0, the distribution of T_{trend} can be approximated by a χ_1^2 distribution.

Gametic equilibrium

Consider two diallelic loci $\{A, a\}$ and $\{B, b\}$. If the occurrence of allele A and the occurrence of allele B in a gamete are independent events, then the probability of the joint occurrence of alleles A and B in a gamete is equal to the product of the allele frequencies, i.e.,

$$p_{AB} = p_A \cdot p_B$$

and the alleles at the two loci are said to be in *gametic equilibrium.*

Exercise:

Show that $p_{AB} = p_A \cdot p_B$ implies that $p_{Ab} = p_A \cdot p_b$, $p_{aB} = p_a \cdot p_B$, and $p_{ab} = p_a \cdot p_b$.

Gametic disequilibrium

The non-independence of the alleles in a gamete can be measured by the deviation of the probability of a haplotype from the value expected under gametic equilibrium:

haplotype	gametic probability		expected under gametic equilibrium		deviation
AB	p_{AB}	$=$	$p_A \cdot p_B$	$+$	δ
Ab	p_{Ab}	$=$	$p_A \cdot p_b$	$-$	δ
aB	p_{aB}	$=$	$p_a \cdot p_B$	$-$	δ
ab	p_{ab}	$=$	$p_a \cdot p_b$	$+$	δ

δ : *gametic disequilibrium coefficient*

Standardized gametic disequilibrium coefficient

For given p_A and p_B, the value of δ is always between

$$\delta_{\min} = \max\left(-p_A \cdot p_B, -p_a \cdot p_b\right)$$

and

$$\delta_{\max} = \min\left(p_A \cdot p_b, p_a \cdot p_B\right)$$

The *standardized gametic disequilibrium coefficient* D' is defined by

$$D' = \begin{cases} \dfrac{\delta}{\delta_{\max}} & \text{for} \quad \delta \geq 0 \\ \dfrac{\delta}{-\delta_{\min}} & \text{for} \quad \delta \leq 0 \end{cases}$$

Correlation coefficient

Another frequently used disequilibrium measure is the correlation coefficient

$$r = \frac{\delta}{\sqrt{p_A \cdot p_B \cdot p_a \cdot p_b}}$$

Exercise:

Show that $\mid r \mid \leq \mid D' \mid$

Linkage disequilibrium

Gametic disequilibrium can be due to close linkage between the two loci. Therefore, gametic disequilibrium is often named *linkage disequilibrium*. However, close linkage is not the only mechanism which can generate gametic disequilibrium. Other possible causes of gametic disequilibrium include

- selection (i.e., reproductive fitness is influenced by an individual's genotype)
- population stratification (see below)

Decay of linkage disequilibrium over time

Let $p_{AB}^{(k)}$ denote the probability of the haplotype AB in generation k. Then,

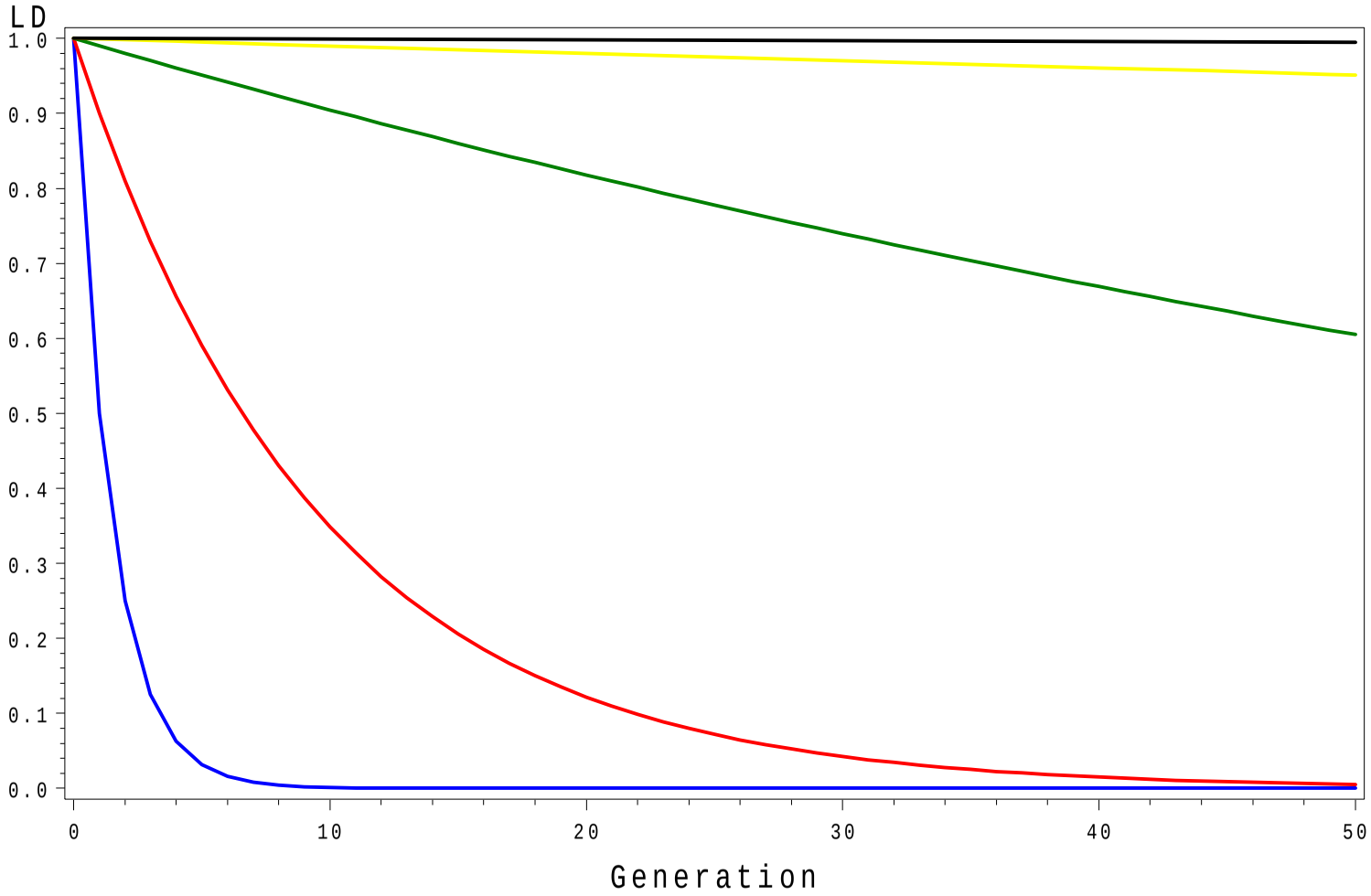
$$p_{AB}^{(k)} = (1-\theta) \cdot p_{AB}^{(k-1)} + \theta \cdot p_A \cdot p_B.$$

Therefore,

$$\begin{aligned} \delta^{(k)} &= p_{AB}^{(k)} - p_A \cdot p_B \\ &= (1-\theta) \cdot (p_{AB}^{(k-1)} - p_A \cdot p_B) = (1-\theta) \cdot \delta^{(k-1)} \end{aligned}$$

$\Rightarrow$ Gametic disequilibrium due to close linkage decreases by a factor of $(1-\theta)$ after one generation.

Decay of linkage disequilibrium over time



$\theta = 0.5, 0.1, 0.01, 0.001, 0.0001$

Population stratification

Population stratification is the phenomenon that a population consists of two or more subgroups. Population stratification can induce an increased rate of false positive association results for case-control association studies.

Example:

Study population: San Francisco population

Trait: Ability to eat with chopsticks

Marker: HLA-A locus

Result: Allele A1 at HLA-A positively associated with ability to use chopsticks

Explanation: San Francisco population consists of two subpopulations (Asians and Caucasians). The ability to eat with chopsticks **and** the allele A1 at HLA-A is more common among Asians than Caucasians.

Population stratification

Subpopulation 1:

	B	b	$\sum$
A	$p_{AB}^{(1)}$	$p_{Ab}^{(1)}$	$p_{A}^{(1)}$
a	$p_{aB}^{(1)}$	$p_{ab}^{(1)}$	$p_{a}^{(1)}$
$\sum$	$p_{B}^{(1)}$	$p_{b}^{(1)}$	1

$$\delta^{(1)} = p_{AB}^{(1)} - p_{A}^{(1)} \cdot p_{B}^{(1)}$$

Subpopulation 2:

	B	b	$\sum$
A	$p_{AB}^{(2)}$	$p_{Ab}^{(2)}$	$p_{A}^{(2)}$
a	$p_{aB}^{(2)}$	$p_{ab}^{(2)}$	$p_{a}^{(2)}$
$\sum$	$p_{B}^{(2)}$	$p_{b}^{(2)}$	1

$$\delta^{(2)} = p_{AB}^{(2)} - p_{A}^{(2)} \cdot p_{B}^{(2)}$$

Population stratification

Combined population: ($m \widehat{=}$ portion of subpopulation 1)

	B	b	$\sum$
A	$p_{AB} = m \cdot p_{AB}^{(1)} + (1-m) \cdot p_{AB}^{(2)}$	$p_{Ab} = m \cdot p_{Ab}^{(1)} + (1-m) \cdot p_{Ab}^{(2)}$	$p_A = m \cdot p_A^{(1)} + (1-m) \cdot p_A^{(2)}$
a	$p_{aB} = m \cdot p_{aB}^{(1)} + (1-m) \cdot p_{aB}^{(2)}$	$p_{ab} = m \cdot p_{ab}^{(1)} + (1-m) \cdot p_{ab}^{(2)}$	$p_a = m \cdot p_a^{(1)} + (1-m) \cdot p_a^{(2)}$
$\sum$	$p_B = m \cdot p_B^{(1)} + (1-m) \cdot p_B^{(2)}$	$p_b = m \cdot p_b^{(1)} + (1-m) \cdot p_b^{(2)}$	1

Population stratification

Linkage disequilibrium δ in the combined population:

$$\begin{aligned}
\delta &= p_{AB} - p_A \cdot p_B \\
&= m \cdot p_{AB}^{(1)} + (1-m) \cdot p_{AB}^{(2)} \\
&\quad -\left[m \cdot p_A^{(1)} + (1-m) \cdot p_A^{(2)}\right] \cdot \left[m \cdot p_B^{(1)} + (1-m) \cdot p_B^{(2)}\right] \\
&= m \cdot \delta^{(1)} + (1-m) \cdot \delta^{(2)} \\
&\quad + m \cdot (1-m) \cdot \left(p_A^{(1)} - p_A^{(2)}\right) \cdot \left(p_B^{(1)} - p_B^{(2)}\right)
\end{aligned}$$

Special case:

$$\delta^{(1)} = \delta^{(2)} = 0, p_A^{(1)} \neq p_A^{(2)}, p_B^{(1)} \neq p_B^{(2)}$$

$$\Rightarrow \delta \neq 0$$

Population stratification

- Population stratification can induce an increased rate of false positive association results only in case that the frequency of the disease *and* the marker allele frequencies differ between the subpopulations.
- Case-control association studies which sample cases and controls from different subpopulations (bad study design!) will lead to a false positive association result already in case that the marker allele frequencies differ between the subpopulations.
- An increased rate of false positive association results due to population stratification can be avoided by using family-based methods of association analysis.