Hardy-Weinberg Equilibrium

Allele Frequencies and Genotype Frequencies

- How do allele frequencies relate to genotype frequencies in a population?
- If we have genotype frequencies, we can easily get allele frequencies.

Example

Cystic Fibrosis is caused by a recessive allele. The locus for the allele is in region 7q31. Of 10,000 Caucasian births, 5 were found to have Cystic Fibrosis and 442 were found to be heterozygous carriers of the mutation that causes the disease. Denote the Cystic Fibrosis allele with cf and the normal allele with N. Based on this sample, how can we estimate the allele frequencies in the population?

- We can estimate the genotype frequencies in the population based on this sample
- $\frac{5}{10000}$ are cf, cf
- $\frac{442}{10000}$ are N, cf
- $\frac{9553}{10000}$ are N, N

Example

So we use 0.0005, 0.0442, and 0.9553 as our estimates of the genotype frequencies in the population. The only assumption we have used is that the sample is a random sample. Starting with these genotype frequencies, we can estimate the allele frequencies without making any further assumptions: Out of 20,000 alleles in the sample

- $\frac{442+10}{20000} = .0226$ are cf
- $1 - \frac{442+10}{20000} = .9774$ are N

Hardy-Weinberg Equilibrium

In contrast, going from allele frequencies to genotype frequencies requires more assumptions.

HWE Model Assumptions

- infinite population
- discrete generations
- random mating
- no selection
- no migration in or out of population
- no mutation
- equal initial genotype frequencies in the two sexes

Hardy-Weinberg Equilibrium

- Consider a locus with two alleles: A and a
- Assume in the first generation the alleles are not in HWE and the genotype frequency distribution is as follows:

1st Generation

Genotype	Frequency
AA	u
Aa	v
aa	w

where $u + v + w = 1$

- From the genotype frequencies, we can easily obtain allele frequencies:

$$P(A) = u + \frac{1}{2}v$$

$$P(a) = w + \frac{1}{2}v$$

Hardy-Weinberg Equilibrium

In the first generation: $P(A) = u + \frac{1}{2}v$ and $P(a) = w + \frac{1}{2}v$

2nd Generation

Mating Type	Mating Frequency	Expected Progeny
$AA \times AA$	u^2	AA
$AA \times Aa$	$2uv$	$\frac{1}{2}\ AA : \frac{1}{2}\ Aa$
$AA \times aa$	$2uw$	Aa
$Aa \times Aa$	v^2	$\frac{1}{4}\ AA : \frac{1}{2}\ Aa : \frac{1}{4}\ aa$
$Aa \times aa$	$2vw$	$\frac{1}{2}\ Aa : \frac{1}{2}\ aa$
$aa \times aa$	w^2	aa

$*$ Check: $u^2 + 2uv + 2uw + v^2 + 2vw + w^2 = (u + v + w)^2 = 1$

- $p \equiv P(AA) = u^2 + \frac{1}{2}(2uv) + \frac{1}{4}v^2 = \left(u + \frac{1}{2}v\right)^2$
- $q \equiv P(Aa) = uv + 2uw + \frac{1}{2}v^2 + vw = 2\left(u + \frac{1}{2}v\right)\left(\frac{1}{2}v + w\right)$
- $r \equiv P(aa) = \frac{1}{4}v^2 + \frac{1}{2}(2vw) + w^2 = \left(w + \frac{1}{2}v\right)^2$

Hardy-Weinberg Equilibrium

In the third generation:

$$P(AA) = \left(p + \frac{1}{2}q\right)^2 = \left(\left(u + \frac{1}{2}v\right)^2 + \left(\frac{1}{2}\right)2\left(u + \frac{1}{2}v\right)\left(\frac{1}{2}v + w\right)\right)^2$$

$$= \left(\left(u + \frac{1}{2}v\right)\left[\left(u + \frac{1}{2}v\right) + \left(\frac{1}{2}v + w\right)\right]\right)^2$$

$$= \left(\left(u + \frac{1}{2}v\right)[(u + v + w)]\right)^2$$

$$= \left(\left(u + \frac{1}{2}v\right)1\right)^2 = \left(\left(u + \frac{1}{2}v\right)\right)^2 = p$$

Similarly, $P(Aa) = q$ and $P(aa) = r$ for generation 3

- **Equilibrium** is reached after one generation of mating under the Hardy-Weinberg assumptions! Genotype frequencies remain the same from generation to generation.

Hardy-Weinberg Equilibrium

When a population is in Hardy-Weinberg equilibrium, the alleles that comprise a genotype can be thought of as having been chosen at random from the alleles in a population. We have the following relationship between genotype frequencies and allele frequencies for a population in Hardy-Weinberg equilibrium:

$$P(AA) = P(A)P(A)$$

$$P(Aa) = 2P(A)P(a)$$

$$P(aa) = P(a)P(a)$$

Hardy-Weinberg Equilibrium

For example, consider a diallelic locus with alleles A and B with frequencies 0.85 and 0.15, respectively. If the locus is in HWE, then the genotype frequencies are:

$$P(AA) = 0.85 * 0.85 = 0.7225$$

$$P(AB) = 0.85 * 0.15 + 0.15 * 0.85 = 0.2550$$

$$P(BB) = 0.15 * 0.15 = 0.0225$$

Hardy-Weinberg Equilibrium Example

Establishing the genetics of the ABO blood group system was one of the first breakthroughs in Mendelian genetics. The locus corresponding to the ABO blood group has three alleles, A, B and O and is located on chromosome 9q34. Alleles A and B are co-dominant, and the alleles A and B are dominant to O. This leads to the following genotypes and phenotypes:

Genotype	Blood Type
AA, AO	A
BB, BO	B
AB	AB
OO	O

Mendels first law allows us to quantify the types of gametes an individual can produce. For example, an individual with type AB produces gametes A and B with equal probability (1/2).

Hardy-Weinberg Equilibrium Example

From a sample of 21,104 individuals from the city of Berlin, allele frequencies have been estimated to be P(A)=0.2877, P(B)=0.1065 and P(O)=0.6057. If an individual has blood type B, what are the possible genotypes for this individual, what possible gametes can be produced, and what is the frequency of the genotypes and gametes if HWE is assumed?

- If a person has blood type B, then the genotype is BO or BB.
- What is P(genotype is BO|blood type is B)?
- What is P(genotype is BB|blood type is B)?
- What is P(B gamete|blood type is B)?
- What is P(O gamete|blood type is B)?

Hardy-Weinberg Equilibrium

- With HWE: allele frequencies $\Longrightarrow$ genotype frequencies.
- Violations of HWE assumption inclue:
- Small population sizes. Chance events can make a big difference.
- Deviations from random mating.
- Assortive mating. Mating between genotypically similar individuals increases homozygosity for the loci involved in mate choice without altering allele frequencies.
- Disassortive mating. Mating between dissimilar individuals increases heterozygosity without altering allele frequencies.
- Inbreeding. Mating between relatives increases homozygosity for the whole genome without affecting allele frequencies.
- Population sub-structure
- Mutation
- Migration
- Selection

Testing Hardy-Weinberg Equilibrium

- When a locus is not in HWE, then this suggests one or more of the Hardy-Weinberg assumptions is false.
- Departure from HWE has been used to infer the existence of natural selection, argue for existence of assortive (non-random) mating, and infer genotyping errors.
- It is therefore of interest to test whether a population is in HWE at a locus.
- We will discuss the two most popular ways of testing HWE:
  - Chi-Square test
  - Exact test

Chi-Square Goodness-Of-Fit Test

Compares observed genotype counts with the values expected under Hardy-Weinberg. For a locus with two alleles, we might construct a table as follows:

Genotype	Observed	Expected under HWE
AA	n_{AA}	np_A^2
Aa	n_{Aa}	$2np_A(1-p_A)$
aa	n_{aa}	$n(1-p_A)^2$

where n is the number of individuals in the sample and p_A is the probability that a random allele in the population is of type A.

- We estimate p_A with $\hat{p}_A = \frac{2n_{AA}+n_{Aa}}{2n}$

Chi-Square Goodness-Of-Fit Test

Test statistic is for Allelic Association is:

$$X^2 = \sum_{\text{genotypes}} \frac{(\text{Observed count} - \text{Expected count})^2}{\text{Expected count}}$$

$$X^2 = \frac{\left(n_{AA} - n\hat{p}_a^2\right)^2}{n\hat{p}_a^2} + \frac{\left(n_{Aa} - 2n\hat{p}_a(1-\hat{p}_a)\right)^2}{2n\hat{p}_a(1-\hat{p}_a)} + \frac{\left(n_{aa} - n(1-\hat{p}_a)^2\right)^2}{n(1-\hat{p}_a)^2}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with 1 degree of freedom
- Recall the rule of thumb for such χ^2 tests: the expected count should be at least 5 in every cell. If allele frequencies are low, and/or sample size is small, and/or there are many alleles at a locus, this may be a problem.

HWE Exact Test

- The Hardy-Weinberg exact test is based on calculating probabilities
- $P(\text{genotype counts}|\text{allele counts})$ under HWE.

HWE Exact Test Example

- Suppose we have a sample of 5 people and we observe genotypes AA, AA, AA, aa, and aa.
- If five individuals have among them 6 A alleles and 4 a alleles, what genotype configurations are possible?

HWE Exact Test Example

aa	Aa	AA	theoretical probability
2	0	3	0.048
1	2	2	0.571
0	4	1	0.381

HWE Exact Test Example

- Now suppose we have a sample of 100 individuals and we observe 21 "a" alleles and 179 "A" alleles, what genotype configurations are possible?

HWE Exact Test Example

Note that specifying the number of heterozygotes determines the number of AA and aa genotypes.

aa	Aa	AA	theoretical probability
	1		$\ll .000001$
	3		$\ll .000001$
	5		$< .000001$
	7		.000001
	9		.000047
	11		.000870
	13		.009375
	15		.059283
	17		.214465
	19		.406355
	21		.309604

Wiggington, Cutler, Abecasis (AJHG, 2005)

HWE Exact Test Example

The formula is:

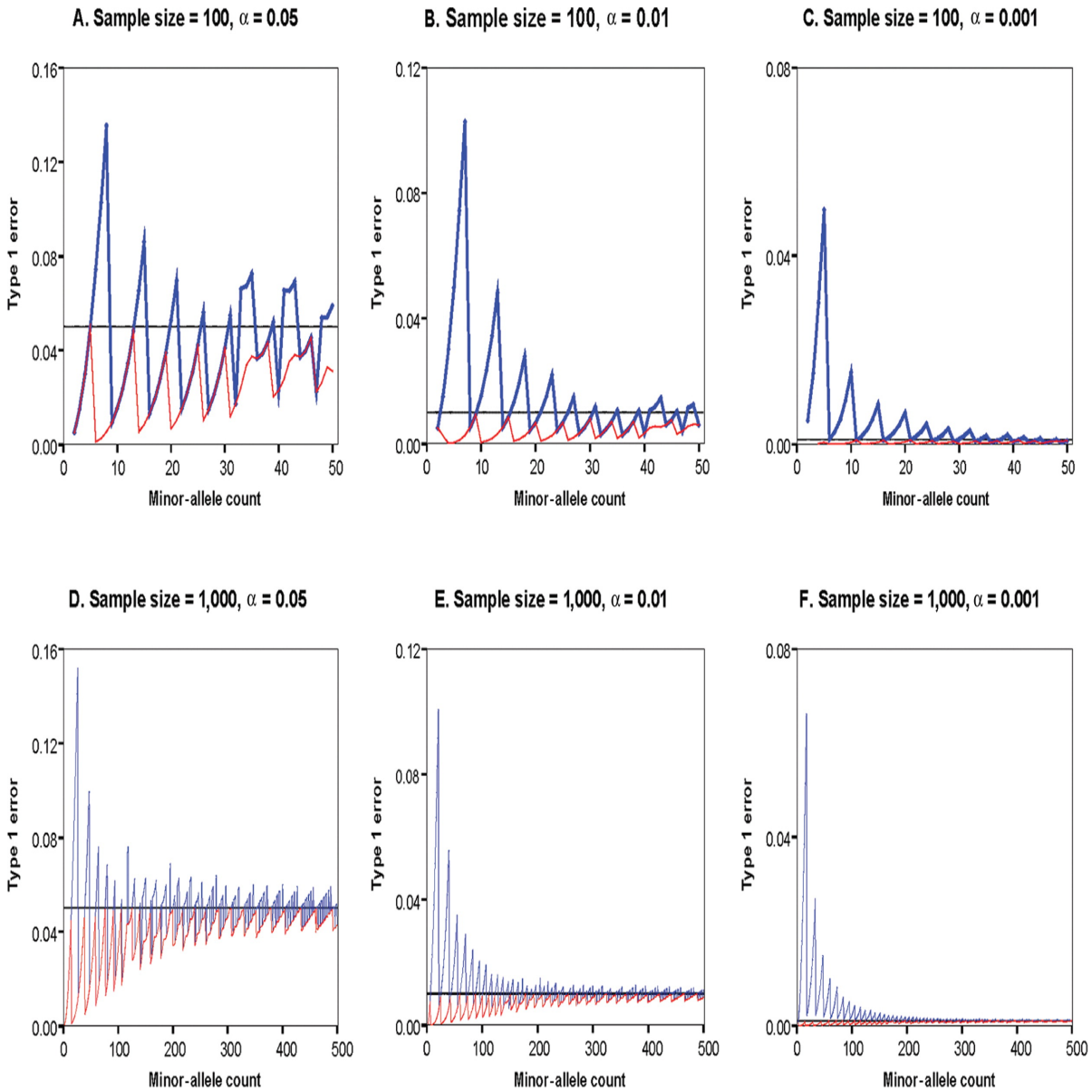
$$P(n_{Aa}|n_A, n_a, HWE) = \frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

If we had actually observed 13 heterozygotes in our sample, then the exact test p-value would be
$\approx .009375 + .000870 + .000047 + .000001 = 0.010293$ (To get the p-value, we sum the probabilities of all configurations with probability equal to or less that the observed configuration.)

Comparison of HWE χ^2 Test and Exact Test

The next slide is Figure 1 from Wigginton et al (AJHG 2005). The upper curves give the type I error rate of the chi-square test; the bottom curves give the type I error rate from the exact test. The exact test is always conservative; the chi-square test can be either conservative or anti-conservative.

HWE TYPE I ERROR



Comparison of HWE χ^2 Test and Exact Test

- The Exact Test should be preferred for smaller sample sizes and/or multiallelic loci, since the χ^2 test is not valid in these cases (rule of thumb: must expect at least 5 in each cell)
- The coarseness of Exact Test means it is conservative. In Example 4, we reject the null hypothesis that HWE holds if 13 or fewer heterozygotes are observed. But the observed p-value is actually 0.010293. Thus to reject at the 0.05 level, we actual have to see a p-value as small as 0.010293.

Comparison of HWE χ^2 Test and Exact Test

- The χ^2 test can have inflated type I error rates. Suppose we have 100 genes for which HWE holds. We conduct 100 χ^2 tests at level 0.05. We expect to reject the null hypothesis that HWE holds in 5 of the tests. However, the results of Wiggington et al (AJHG, 2005) say, on average, it can be more than 5 depending on the minor allele count. Although it is not desirable for a test to be conservative (Exact Test), an anti-conservative test is considered unacceptable.
  - Wiggington et al (AJHG, 2005) give an extreme example with a sample of 1000 individuals. At a nominal a=0.001, the true type I error rate for the χ^2 test exceeds 0.06.

Comparison of HWE χ^2 Test and Exact Test

- The χ^2 test is a two-sided test. In contrast, the Exact Test can be made one-sided, if appropriate. Specifically, one can test for a deficit of heterozygotes (if one suspects inbreeding or population stratification); test for an excess of heterozygotes (which indicate genotyping errors for some genotyping technologies).
- Exact test is more computationally intensive

Linkage Disequilibrium

Linkage Equilibrium

- Consider two linked loci
- Locus 1 has alleles $A_1, A_2, \ldots, A_m$ occurring at frequencies $p_1, p_2, \ldots, p_m$
- locus 2 has alleles $B_1, B_2, \ldots, B_n$ occurring at frequencies $q_1, q_2, \ldots, q_n$ in the population.
- How many possible haplotypes are there for the two loci?
- The possible haplotypes can be denote as $A_1B_1, A_1B_2, \ldots, A_mB_n$ with frequencies $h_{11}, h_{12}, \ldots, h_{mn}$
- The two linked loci are said to be in linkage equilibrium (LE), if the occurrence of allele A_i and the occurrence of allele B_j in a haplotype are independent events. That is, $h_{ij} = p_i q_j$ for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$.
- Two loci are said to be in linkage (or gametic) disequilibrium (LD) if their respective alleles do not associate independently
- Notice that linkage equilibrium/disequilibrium is a population-level characteristic

Linkage Disequilibrium

- Consider two bi-allelic loci.
- There are four possible haplotypes: A_1B_1, A_1B_2, A_2B_1, and A_2B_2.
- Suppose that the frequencies of these four haplotypes in the population are 0.4, 0.1, 0.2, and 0.3, respectively.
- Are the loci in linkage equilibrium?
- Which alleles on the two loci occur together on haplotypes than what would be expected under linkage equilibrium?

Measures of Linkage Disequilibrium

- The Linkage Disequilibrium Coefficient D is one measure of LD.
- For ease of notation, we define D for two biallelic loci with alleles A and a at locus 1; B and b at locus 2:

$$D_{AB} = P(AB) - P(A)P(B)$$

- What about D_{aB}? Note that

$$\begin{aligned} D_{aB} &= P(aB) - P(a)P(B) \\ &= P(aB) - (1 - P(A))P(B) \\ &= P(aB) - P(B) + P(A)P(B) \\ &= P(aB) - (P(AB) + P(aB)) + P(A)P(B) \\ &= P(aB) - P(aB) - P(AB) + P(A)P(B) \\ &= -P(AB) + P(A)P(B) = -D_{AB} \end{aligned}$$

Linkage Disequilibrium Coefficient

- Can similarly show that $D_{Ab} = -D_{AB}$ and $D_{ab} = D_{AB}$
- LD is a property of two loci, not their alleles.
- Thus, the magnitude of the coefficient is important, not the sign.
- The magnitude of D does not depend on the choice of alleles.
- The range of values the linkage disequilibrium coefficient can take on varies with allele frequencies.

Linkage Disequilibrium Coefficient

- By using the fact that $p_{AB} = P(AB)$ must be less than both $p_A = P(A)$ and $p_B = P(B)$, and that allele frequencies cannot be negative, the following relations can be obtained:
  - $0 \leqslant p_{AB} = p_A p_B + D_{AB} \leqslant p_A, p_B$
  - $0 \leqslant p_{aB} = p_a p_B - D_{AB} \leqslant p_a, p_B$
  - $0 \leqslant p_{Ab} = p_A p_b - D_{AB} \leqslant p_A, p_b$
  - $0 \leqslant p_{ab} = p_a p_b + D_{AB} \leqslant p_a, p_b$
- These inequalities lead to bounds for D_{AB} :

$$-p_A p_B, -p_a p_b \leqslant D_{AB} \leqslant p_a p_B, p_A p_b$$

Normalized Linkage Disequilibrium Coefficient

- What is the theoretical range of the linkage disequilibrium coefficient D_{AB} and its absolute value $|D_{AB}|$ under the follow scenarios?
- $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$
- $P(A) = .95$, $P(B) = .95$
- $P(A) = .95$, $P(B) = .05$
- $P(A) = \frac{1}{2}$, $P(B) = .95$?
- Under what circumstances might D_{AB} reach its theoretical maximum value? Suppose $D_{AB} = P(a)P(B)$. What does this imply? Why does this make sense?

Normalized Linkage Disequilibrium Coefficient

- We have just seen that the possible values of D depend on allele frequencies. This makes D difficult to interpret. For reporting purposes, the normalized linkage disequilibrium coefficient D' is often used.

$$D'_{AB} = \begin{cases} \frac{D_{AB}}{max(-p_A p_B, -p_a p_b)} & \text{if } D_{AB} < 0 \\ \frac{D_{AB}}{min(p_a p_B, p_A p_b)} & \text{if } D_{AB} > 0 \end{cases} \quad (1)$$

Estimating D

- Suppose we have the N haplotypes for two loci on a chromosomes that have been sampled from a population of interest. The data might be arranged in a table such as:

	B	b	Total
A	n_{AB}	n_{Ab}	n_A
a	n_{aB}	n_{ab}	n_a
	n_B	n_b	N

- We would like to estimate D_{AB} from the data. The maximum likelihood estimate of D_{AB} is

$$\hat{D}_{AB} = \hat{p}_{AB} - \hat{p}_A \hat{p}_B$$

where $\hat{p}_{AB} = \frac{n_{AB}}{N}$, $\hat{p}_A = \frac{n_A}{N}$, and $\hat{p}_B = \frac{n_B}{N}$

- So the population frequencies are estimated by the sample frequencies

Estimating D

- The MLE turns out to be slightly biased. If N gametes have been sampled, then

$$E\left(\hat{D}_{AB}\right) = \frac{N-1}{N} D_{AB}$$

- The variance of this estimate depends on both the true allele frequencies and the true level of linkage disequilibrium:
- $Var\left(\hat{D}_{AB}\right) =$
  $\frac{1}{N}\left[p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2\right]$
  Suppose we have the N haplotypes for two loci on a chromosomes that have been sampled from a population of interest. The data might be arranged in a table such as:

Testing for LD with D

- Since $D_{AB} = 0$ corresponds to the status of no linkage disequilibrium, it is often of interest to test the null hypothesis $H_0 : D_{AB} = 0$ vs. $H_a : D_{AB} \neq 0$.
- One way to do this is to use a chi-square statistic. It is constructed by squaring the asymptotically normal statistic z:

$$Z^2 = \left(\frac{\hat{D}_{AB} - E_0\left(\hat{D}_{AB}\right)}{Var_0\left(\hat{D}_{AB}\right)} \right)^2$$

  where E_0 and Var_0 are expectation and variance calculated under the assumption of no LD, i.e., $D_{AB} = 0$
- Under the null, the test statistic will follow a Chi-Squared (χ^2) distribution with one degree of freedom.

Measuring LD with R^2

- Define a random variable X_A to be 1 if the allele at the first locus is A and 0 if the allele is a.
- Define a random variable X_B to be 1 if the allele at the second locus is B and 0 if the allele is b.
- Then the correlation between these random variables is:

$$r_{AB} = \frac{COV(X_A, X_B)}{\sqrt{Var(X_A)Var(X_B)}} = \frac{D_{AB}}{\sqrt{p_A(1-p_A)p_B(1-p_B)}}$$

- It is usually more common to consider the r_{AB} value squared:

$$r^2_{AB} = \frac{D^2_{AB}}{p_A(1-p_A)p_B(1-p_B)}$$

Measuring LD with R^2

- R^2 has the same value however the alleles are labeled
- Tests for LD: A natural test statistic to consider is the contingency table test. Compute a test statistic using the Observed haplotype frequencies and the Expected frequency if there were no LD:

$$X^2 = \sum_{\text{possible haplotypes}} \frac{(\text{Observed cell} - \text{Expected cell})^2}{\text{Expected cell}}$$

- Under H_0, the X^2 test statistic has an approximate χ^2 distribution with 1 degree of freedom
- It turns out that $X^2 = N\hat{r}^2$

R^2 or D'

- If two loci both have very rare alleles but the loci are not in high LD, it is possible for D' to be 1 and r^2 to be small.
- D' is problematic to interpret with rare alleles, and r^2 is a better measure for this situation.

Linkage Disequilibrium 2

Why does linkage disequilibrium occur?

- Genetic drift: In a finite population, the gene pool of one generation can be regarded as a random sample of the gene pool of the previous generation. As such, allele and haplotypes frequencies are subject to sampling variation random chance. The smaller the population is, the larger the effects of genetic drift are.
- Mutation: If a new mutation appears in a population, alleles at loci linked with the mutant allele will maintain linkage disequilibrium for many generations. LD lasts longer when linkage is greater (that is, the recombination fraction is much smaller than $\frac{1}{2}$ - very close to 0).

Why does linkage disequilibrium occur?

- Founder effects: Applies to a population that has grown rapidly from a small group of ancestors. For example, the 5,000,000 Finns mostly descended from about 1000 people who lived about 2000 years ago. Such a population is prone to LD.
- Selection: When an individuals genotype influences his/her reproductive fitness. For example, if two alleles interact to decrease reproductive fitness, the alleles will tend to be negatively associated, i.e., they tend not to appear together on haplotypes.
- Stratification: Some populations consist of two or more subgroups that, for cultural or other reasons, have evolved more or less separately. Two loci that are in linkage equilibrium for each subpopulation may be in linkage disequilibrium for the larger population.

Linkage disequilibrium example

- Consider a population with three subpopulations.
- Consider two biallelic loci, the first locus with alleles A and a; the second locus with alleles B and b.
- Are the three subpopulations in linkage equilibrium?
- Is the population as a whole in linkage equilibrium?

N	A allele freq.	B allele freq.	AB haplotype freq.
1000	0.3	0.5	0.15
2000	0.2	0.4	0.08
10000	0.05	0.1	0.005

Linkage Disequilibrium Decay

- How is LD maintained in a population?
  - Selection
  - Non-random mating (e.g., population stratification)
  - Linkage
- Consider again two linked loci
- Locus 1 has alleles $A_1, A_2, \ldots, A_m$ occurring at frequencies $p_1, p_2, \ldots, p_m$
- locus 2 has alleles $B_1, B_2, \ldots, B_n$ occurring at frequencies $q_1, q_2, \ldots, q_n$ in the population.
- The haplotypes are $A_1B_1, A_1B_2, \ldots, A_mB_n$ with frequencies $h^0_{11}, h^0_{12}, \ldots, h^0_{mn}$ in generation 0.
- Let θ be the recombination fraction for locus 1 and locus 2.
- What is h^1_{ij}, the frequency of haplotype A_iB_j in the next generation is we assume random mating in the population?

Linkage Disequilibrium

$$h_{ij}^1 = P(haplotype^1 = A_i B_j)$$

$$= P(haplotype^1 = A_i B_j | \text{no recombination}) P(\text{no recombination})$$

$$+ P(haplotype^1 = A_i B_j | \text{recombination}) P(\text{recombination})$$

$$= P(haplotype^1 = A_i B_j | \text{no recombination})(1 - \theta)$$

$$+ P(haplotype^1 = A_i B_j | \text{recombination})\theta$$

$$= h_{ij}^0(1 - \theta) + p_i q_j \theta$$

Linkage Disquilibrium

- So $h_{ij}^1 = h_{ij}^0(1-\theta) + p_i q_j \theta$
- From this, we can obtian the difference in haplotype frequency between the two generations is:

$$h_{ij}^1 - h_{ij}^0 = \theta(p_i q_j - h_{ij}^0)$$

- When will this difference be 0? That is, when are the haplotype frequencies stable?
- Answer: $\theta = 0$ or no linkage disequilibrium.
- We can also characterize the difference between the true haplotype frequency at generation 1 and what the haplotype frequency would be under linkage equilibrium

$$h_{ij}^1 - p_i q_j = (1-\theta)(h_{ij}^0 - p_i q_j)$$

- Can extend this to the k^{th} generation

$$h_{ij}^k - p_i q_j = (1-\theta)^k(h_{ij}^0 - p_i q_j)$$

Linkage Disequilibrium

- Another way to write this is as follows

$$D_{ij}^{1} = (1-\theta)D_{ij}^{0}$$

$$D_{ij}^{k} = (1-\theta)^{k}D_{ij}^{0}$$

- On the following slide is a figure that shows the decline of linkage disequilibrium in a large, randomly mating population for various values of θ

Linkage Disequilibrium

Figure:

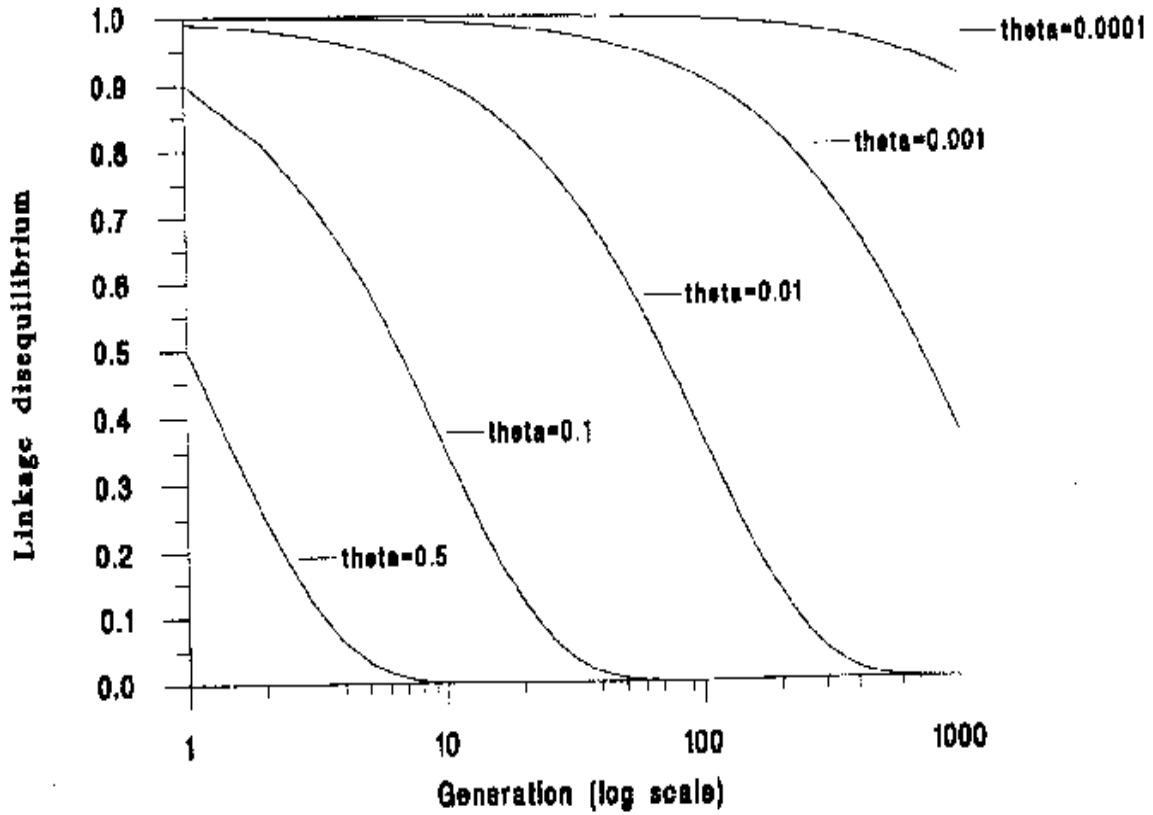


Figure 4.1 Decay of linkage disequilibrium by generation.

Linkage Disequilibrium

Figure:

- What can you say about the LD between the SNPs below?

Individual	SNP1	SNP2	SNP3	SNP4
1	A	C	A	T
	A	C	A	T
2	A	C	A	G
	G	T	A	G
3	A	C	A	T
	G	T	C	G
4	A	C	A	G
	A	C	A	G
5	G	T	C	G
	G	T	C	T

Tag SNPs using Linkage Disequilibrium Measures

- It is possible to identify genetic variation without genotyping every SNP in a haplotype block.
- By genotyping only the "Tag SNPs", it is possible to record most of the genetic variation in a haplotype block, with the fewest number of SNPs.

Tag SNPs using Linkage Disequilibrium Measures

Choosing Tag SNPs

Block 1 | Block 2

Individ:	SNP 1	2	3	4	5
1	A	A	T	A	G
2	A	A	T	A	G
3	A	A	T	A	G
4	A	A	T	A	G
5	G	T	G	A	T
6	A	A	T	C	T
7	G	T	T	A	G
8	A	A	T	A	G
9	G	T	T	C	T
10	G	T	T	C	T

Factors affecting Linkage Disequilibrium

- LD information is useful for deciding which polymorphisms to genotype.
- LD information across the whole genome can be used in a variety of ways.
- However...LD depends on population history.
- Which LD database to look at depends on which population your study individuals are from.

Population Structure

Nonrandom Mating

- HWE assumes that mating is random in the population
- Most natural populations deviate in some way from random mating
- There are various ways in which a species might deviate from random mating
- We will focus on the two most common departures from random mating:
  - inbreeding
  - population subdivision or substructure

Nonrandom Mating: Inbreeding

- Inbreeding occurs when individuals are more likely to mate with relatives than with randomly chosen individuals in the population
- Increases the probability that offspring are homozygous, and as a result the number of homozygous individuals at genetic markers in a population is increased
- Increase in homozygosity can lead to lower fitness in some species
- Increase in homozygosity can have a detrimental effect: For some species the decrease in fitness is dramatic with complete infertility or inviability after only a few generations of brother-sister mating

Nonrandom Mating: Population Subdivision

- For subdivided populations, individuals will appear to be inbred due to more homozygotes than expected under the assumption of random mating.
- Wahlund Effect: Reduction in observed heterozygosity (increased homozygosity) because of pooling discrete subpopulations with different allele frequencies that do not interbreed as a single randomly mating unit.

Wright's F Statistics

- Sewall Wright invented a set of measures called F statistics for departures from HWE for subdivided populations.
- F stands for fixation index, where fixation being increased homozygosity
- F_{IS} is also known as the inbreeding coefficient.
  - The correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (**I**ndividual within the **S**ubpopulation)
- F_{ST} is a measure of population substructure and is most useful for examining the overall genetic divergence among subpopulations
  - Is defined as the correlation of gametes within subpopulations relative to gametes drawn at random from the entire population (**S**ubpopulation within the **T**otal population).

Wright's F Statistics

- F_{IT} is not often used. It is the overall inbreeding coefficient of an individual relative to the total population (**I**ndividual within the **T**otal population).

Genotype Frequencies for Inbred Individuals

- Consider a bi-allelic genetic marker with alleles A and a. Let p be the frequency of allele A and $q = 1 - p$ the frequency of allele a in the population.
- Consider an individual with inbreeding coefficient F. What are the genotype frequencies for this individual at the marker?

Genotype	AA	Aa	aa
Frequency			

Generalized Hardy-Weinberg Deviations

- The table below gives genotype frequencies at a marker for when the HWE assumption does not hold:

Genotype	AA	Aa	aa
Frequency	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

  where $q = 1 - p$
- The F parameter describes the deviation of the genotype frequencies from the HWE frequencies.
- When $F = 0$, the genotype frequencies are in HWE.
- The parameters p and F are sufficient to describe genotype frequencies at a single locus with two alleles.

F_{st} for Subpopulations

- Example in Gillespie (2004)
- Consider a population with two equal sized subpopulations. Assume that there is random mating within each subpoulation.
- Let $p_1 = \frac{1}{4}$ and $p_2 = \frac{3}{4}$
- Below is a table with genotype frequencies

Genotype	A	AA	Aa	aa
Freq. Subpop$_1$	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop$_2$	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$

- Are the subpopulations in HWE?
- What are the genotype frequencies for the entire population?
- What should the genotypic frequencies be if the population is in HWE at the marker?

F_{st} for Subpopulations

- From the table below it is clear that there are too many homozygotes in this population.

Genotype	A	AA	Aa	aa
Freq. Subpop_1	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$
Freq. Subpop_2	$\frac{3}{4}$	$\frac{9}{16}$	$\frac{3}{8}$	$\frac{1}{16}$
Freq. Population	$\frac{1}{2}$	$\frac{5}{16}$	$\frac{3}{8}$	$\frac{5}{16}$
Hardy-Weinberg Frequencies	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

- To determine a measure of the excess in homozygosity from what we would expect under HWE, solve

$$2pq(1 - F_{ST}) = \frac{3}{8}$$

- What is F_{st}?

F_{st} for Subpopulations

- The excess homozygosity requires that $F_{ST} = \frac{1}{4}$
- For the previous example the allele frequency distribution for the two subpopulations is given.
- At the population level, it is often difficult to determine whether excess homozygosity in a population is due to inbreeding, to subpopulations, or other causes.
- European populations with relatively subtle population structure typically have an F_{st} value around .01 (e.g., ancestry from northwest and southeast Europe),
- F_{st} values that range from 0.1 to 0.3 have been observed for the most divergent populations (Cavalli-Sforza et al. 1994).

F_{st} for Subpopulations

- F_{st} can be generalized to populations with an arbitrary number of subpopulations.
- The idea is to find an expression for F_{st} in terms of the allele frequencies in the subpopulations and the relative sizes of the subpopulations.
- Consider a single population and let r be the number of subpopulations.
- Let p be the frequency of the A allele in the population, and let p_i be the frequency of A in subpopulation i, where $i = 1, \ldots, r$
- F_{st} is often defined as $F_{st} = \frac{\sigma_p^2}{p(1-p)}$, where σ_p^2 is the variance of the p_i's with $E(p_i) = p$.

F_{st} for Subpopulations

- Let the relative contribution of subpopulation i be c_i, where $\sum_{i=1}^{r} c_i = 1$.

Genotype	AA	Aa	aa
Freq. Subpop$_i$	p_i^2	$2p_iq_i$	q_i^2
Freq. Population	$\sum_{i=1}^{r} c_i p_i^2$	$\sum_{i=1}^{r} c_i 2p_iq_i$	$\sum_{i=1}^{r} c_i q_i^2$

  where $q_i = 1 - p_i$
- In the population, we want to find the value F_{st} such that $2pq(1 - F_{st}) = \sum_{i=1}^{r} c_i 2p_iq_i$
- Rearranging terms:

$$F_{st} = \frac{2pq - \sum_{i=1}^{r} c_i 2p_iq_i}{2pq}$$

- Now $2pq = 1 - p^2 - q^2$ and $\sum_{i=1}^{r} c_i 2p_iq_i = 1 - \sum_{i=1}^{r} c_i(p_i^2 + q_i^2)$

F_{st} for Subpopulations

- So can show that

$$F_{st} = \frac{\sum_{i=1}^{r} c_i(p_i^2 + q_i^2) - p^2 - q^2}{2pq}$$

$$= \frac{\left[\sum_{i=1}^{r} c_i p_i^2 - p^2\right] + \left[\sum_{i=1}^{r} c_i q_i^2 - q^2\right]}{2pq}$$

$$= \frac{Var(p_i) + Var(q_i)}{2pq}$$

$$= \frac{2Var(p_i)}{2p(1-p)}$$

$$= \frac{Var(p_i)}{p(1-p)}$$

$$= \frac{\sigma_p^2}{p(1-p)}$$

Estimating F_{st}

- Let n be the total number of sampled individuals from the population and let n_i be the number of sampled individuals from subpopulation i
- Let $\hat{p}_i$ be the allele frequency estimate of the A allele for the sample from subpopulation i
- Let $\hat{p} = \sum_i \frac{n_i}{n} \hat{p}_i$
- A simple F_{st} estimate is $\hat{F}_{ST_1} = \frac{s^2}{\hat{p}(1-\hat{p})}$, where s^2 is the sample variance of the $\hat{p}_i$'s.

Estimating F_{st}

- Weir and Cockerman (1984) developed an estimate based on the method of moments.

$$MSA = \frac{1}{r-1}\sum_{i=1}^{r} n_i(\hat{p}_i - \hat{p})^2$$

$$MSW = \frac{1}{\sum_i (n_i - 1)}\sum_{i=1}^{r} n_i \hat{p}_i(1 - \hat{p}_i)$$

- Their estimate is

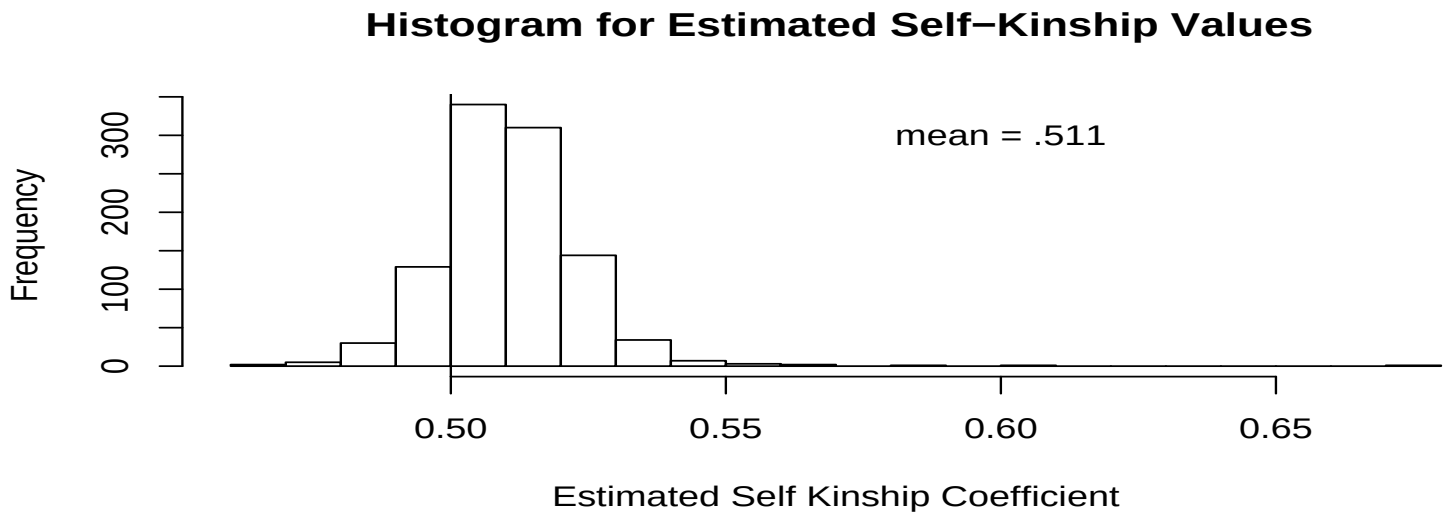
$$\hat{F}_{ST_2} = \frac{MSA - MSW}{MSA + (n_c - 1)MSW}$$

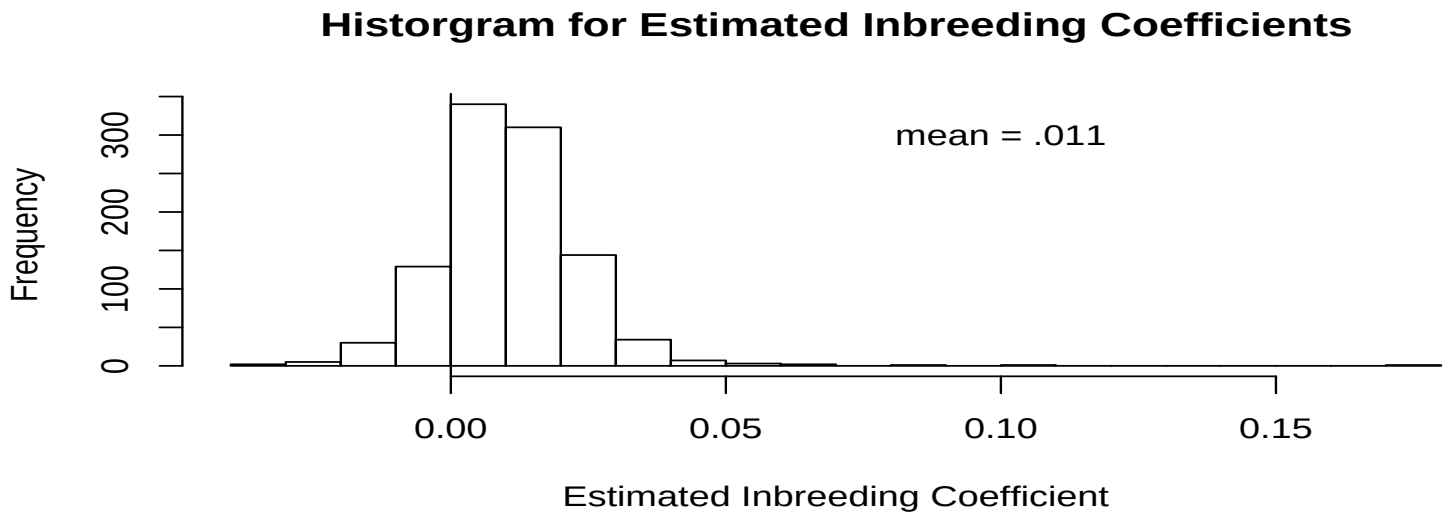
where $n_c = \sum_i n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$

GAW 14 COGA Data

- The Collaborative Study of the Genetics of Alcoholism (COGA) provided genome screen data for locating regions on the genome that influence susceptibility to alcoholism.
- There were a total of 1,009 individuals from 143 pedigrees with each pedigree containing at least 3 affected individuals.
- Individuals labeled as white, non-Hispanic were considered.
- Estimated self-kinship and inbreeding coefficients using genome-screen data

COGA Data


Histogram for Estimated Self−Kinship Values
mean = .511
Frequency
0
100
200
300
0.50
0.55
0.60
0.65
Estimated Self Kinship Coefficient


Historgram for Estimated Inbreeding Coefficients
mean = .011
Frequency
0
100
200
300
0.00
0.05
0.10
0.15
Estimated Inbreeding Coefficient