Linkage Disequilibrium: Part 1

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?

Linkage Equilibrium

- The possible haplotypes can be denote as $A_1B_1, A_1B_2, \dots, A_mB_n$ with frequencies $h_{11}, h_{12}, \dots, h_{mn}$
- 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 $1 \leq i \leq m$ and $1 \leq j \leq n$.
- Remember that Hardy Weinberg Equilibrium (HWE) requires independent assortment of alleles at a single locus. Under HWE, we can obtain genotype frequencies at a locus based on the allele frequencies
- Linkage equilibrium requires independent assortment of the alleles at two linked loci. We can obtain haplotype frequencies for two loci based on the allele frequencies at the two loci

Linkage Disequilibrium

- Two loci are said to be in linkage (or gametic) disequilibrium (LD) if their respective alleles do not associate independently
- 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}?

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:
    - $0 \leq p_{AB} = p_A p_B + D_{AB} \leq p_A, p_B$
    - $0 \leq p_{aB} = p_a p_B - D_{AB} \leq p_a, p_B$
    - $0 \leq p_{Ab} = p_A p_b - D_{AB} \leq p_A, p_b$
    - $0 \leq p_{ab} = p_a p_b + D_{AB} \leq p_a, p_b$
- These inequalities lead to bounds for D_{AB} :

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

Linkage Disequilibrium Coefficient

- bounds for D_{AB} :

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

- What is the theoretical range of the linkage disequilibrium coefficient D_{AB} and its absolute value $|D_{AB}|$ under the follow scenario: $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$

Normalized Linkage Disequilibrium Coefficient

- 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.

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

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:
- $Var\left(\hat{D}_{AB}\right) =$ $\frac{1}{N}\left[p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{AB} - D_{AB}^2\right]$

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:

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

  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 $X^2 = N\hat{r}^2$

D' and r^2

- The case when $D' = 1$ is referred to as **Complete LD**
  - In this case, there are at most 3 of the 4 possible haplotypes present in the populations. The intuition behind complete LD is that the two loci are not being separated by a recombination in this population since at least one of the haplotypes does not occur in the population.
- The case when $r^2 = 1$ is referred to as **Perfect LD**
  - The case of perfect LD occurs when there are exactly 2 of the 4 possible haplotypes present in the population, and as a result, the two loci also have the same allele frequencies.
- Loci that are in perfect LD are necessarily in complete LD

D' and r^2

- Consider two bi-allelic loci with the following 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 0.9, 0.04, 0.05, and 0.01, respectively.
- Calculate D' and r^2 for these two loci.
- Now suppose that the frequencies of these four haplotypes in the population are 0.9, 0.05, 0.05, and 0, respectively. Calculate D' and r^2 for these two loci.

D' and r^2

- If the two loci both have very rare alleles and the rare alleles do not occur together on a haplotype, for example, it is possible for D' to be 1 (since 1 of the haplotypes does not occur in the populations) and for r^2 to be small (when the alleles at the two loci for the 3 remaining haplotypes are not correlated).
- For this and other reasons, it is often useful to calculate both r^2 and D' for loci