Linkage Disequilibrium

a **statsTeachR** resource

These slides were written for statsTeachR by Eric Reed.

Introduction

Linkage Disequilibrium refers a number of measures of association between **alleles**[1] at different **sites**[2] along the same **chromosome**[3].

- It occurs as a result of genetic **recombination** across the chromosome.
  - Recombination is crossover between two **homologous**[4] chromosomes during meiosis (*see next slide*).
  - Sites that are further apart, and have higher recombination rates between them have lower LD values.
- Linkage disequilibrium is not relevant between sites on different chromosomes, which segregate independently.

---

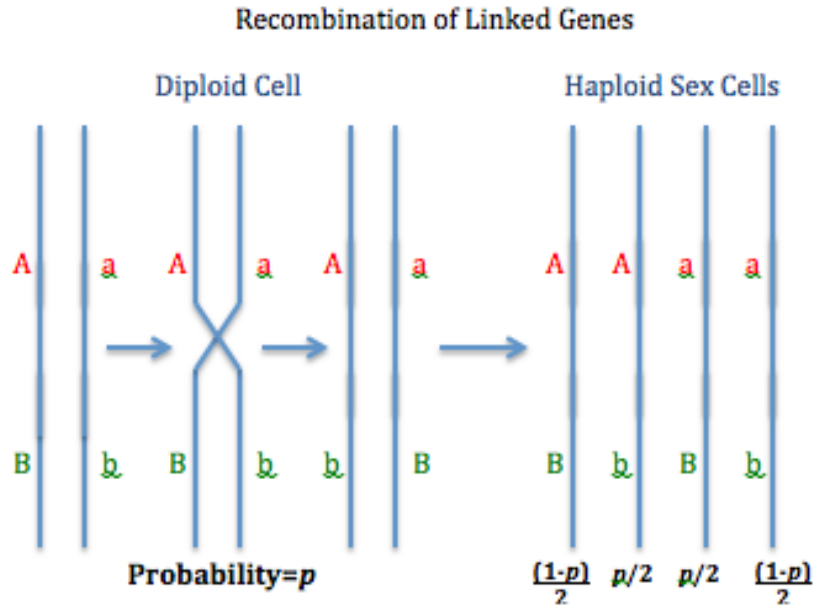
[1] The identity of a genetic information at a specific location on the chromosome

[2] A specific location on the chromosome

[3] A structure of DNA and protein that contains genetic information

[4] Refers to either of the maternal or paternally inherited chromosomes that contain genetic information for the same set of sites

Recombination



Ploidy: The number of sets of homologous chromosomes in a particular cell

- Haploid: Consisting of one set of homologous chromosomes
- Diploid: Consisting of two sets of homologous chromosomes.
- Triploid: Seedless watermelon and bananas!

Haplotype Frequencies Under Independence

Under independence the distribution of **haplotypes**[5]. counts at two sites are as follows:

		Site 2		
		B	b	
Site 1	A	$P_A P_B$	$P_A P_b$	P_A
	a	$P_a P_B$	$P_a P_b$.	P_a
		P_B	P_b	1

[5] The sequence of alleles located across a single chromosome

Haplotype Counts Under Independence

Since each individual n in a population, has two sets of homologous chromosomes, the total counts of haplotypes are equal to $2n$.

		Site 2		
		B	b	
Site 1	A	NP_AP_B	NP_AP_b	NP_A
	a	NP_aP_B	NP_aP_b.	NP_a
		NP_B	NP_b	$N = 2n$

Haplotype Frequencies Under Linkage Disequilibrium

Under linkage disequilibrium observed haplotype counts depart from expected haplotype counts under independence by frequency D.

		Site 2		
		B	b	
Site 1	A	$N(P_A P_B + D)$	$N(P_A P_b - D)$	NP_A
	a	$N(P_a P_B - D)$	$N(P_a P_b + D)$.	NP_a
		NP_B	NP_b	$N = 2n$

Measuring LD with r^2

r^2 is a commonly used measure of LD. It is the square of the correlation between two sites, and can be calculated via Pearson's χ^2-statistic, with the form:

$$r^2 = \chi_1^2 / N,$$

where

$$\chi_1^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})}{E_{ij}}.$$

i and$j$ refer to the rows and columns respectively on the contingency tables for the observed haplotype counts O, and the haplotype counts expected under independence E.

- In cases where haplotype information on study participants is available this is straight-forward to calculate.
- More often **genotype**[6] data is only available for study participants, in which case additional haplotype estimation steps are required.

[6] Genetic information pertaining to the combination of alleles on both homologous chromosomes at a specific site

Now You Try!

In the accompanied labs we will calculate r^2 values directly using haplotype data, as well as employ the "genetics" package in R to estimate r^2 values from genotype data.