Linkage Disequilibrium (LD) Vocabulary

- Allelic phase: alignment of nucleotides on a single homolog (one of the two homologous chromosomes)
- Haplotype: combination of alleles (sequence of bases) on a single homolog
  - At times this will be referred to as a gamete
- Diplotype: the pair of haplotypes that an individual has

e.g. for genotype (Aa,Bb), the possible pairs of haplotypes (i.e., diplotypes) are (AB,ab) and (Ab, aB).

Linkage Disequilibrium (LD) vs. Linkage Analysis

- *Linkage Analysis* is a set of statistical methods to identify the chromosomal location of a gene
  - Typically involves Family-based studies
  - Family-based studies require methods for clustered data
- LD is an association among the alleles found at each of two genomic sites
  - Applicable to Population-based and Family-based studies
- Population-based studies involve unrelated individuals.
  - Allelic phase is unobservable in population-based investigations.

Reasons for Linkage Disequilibrium (LD)

- Recent origin of a mutation
- Selection for certain alleles or haplotypes
- Migration and admixture
- Genetic drift

Linkage Disequilibrium (LD)

A measure of the non-random association of alleles at two loci.

An example: (European Population)

- Observed haplotype frequency (HF):

$$P_{A*0201:B*0704} = .036$$

- Observed allele frequencies:

$$P_{A*0201} = 0.246$$
$$P_{B*0704} = 0.069$$

- LD (Individual Coefficient):

$$D_{ij} = P_{A*0201:B*0704} - P_{A*0201} \cdot P_{B*0704}$$
$$= 0.036 - \underbrace{0.246 \cdot 0.069}_{Expected\ HF\ with\ no\ LD} = 0.019$$

Linkage Disequilibrium (LD)

Two primary pairwise measures based on $D = p_{AB} - p_A p_B$

1. D' is rescaled to take into account constraints on cell counts/frequencies:

$$D' = |D| / D_{max}$$

$$D_{max} = \min(p_A p_b, p_a p_B) \text{ if } D > 0$$
$$\min(p_A p_B, p_a p_b) \text{ if } D < 0$$

2. $r^2 = D^2 / p_A p_B p_a p_b = \chi_1^2 / N$

- The difference between D' and r^2 is in the type of normalization made to D
- r^2 is often preferred due to relationship with χ^2 statistic.

Linkage disequilibrium (LD)

- Note that:

$$(O_{ij} - E_{ij}) = ND$$

- So we can write:

$$\chi_1^2 = \sum_{i,j} \frac{(ND)^2}{(n_i\, n_j)/N}$$
$$= (ND)^2 \left(\frac{1}{Np_A p_B} + \frac{1}{Np_A p_b} + \frac{1}{Np_a p_B} + \frac{1}{Np_a p_b} \right)$$
$$= ND^2 \left(\frac{p_a p_b + p_a p_B + p_A p_b + p_A p_B}{p_A p_B p_a p_b} \right) = \frac{ND^2}{p_A p_B p_a p_b}$$

Expected Allele Distributions Under Independence (No LD)

		Site 2		
		B	b	
Site 1	A	$n_{AB} = N \bullet p_A p_B$	$n_{Ab} = N \bullet p_A p_b$	$n_{A.} = N \bullet p_A$
	a	$n_{aB} = N \bullet p_a p_B$	$n_{AB} = N \bullet p_a p_b$	$n_{a.} = N \bullet p_a$
		$n_{.B} = N \bullet p_B$	$n_{.b} = N \bullet p_b$	N = 2n

Observed Allele Distributions Under LD

		Site 2		
		B	b	
Site 1	A	$n_{AB} = N(p_A p_B + D)$	$n_{Ab} = N(p_A p_b + D)$	$n_{A.} = N\, p_A$
	a	$n_{aB} = N(p_a p_B + D)$	$n_{AB} = N(p_a p_b + D)$	$n_{a.} = N\, p_a$
		$n_{.B} = N\, p_B$	$n_{.b} = N\, p_b$	N = 2n

Observed Haplotype Distributions at 2 Drosophila Loci (*Bam*HI & *Xho*I)

		Site 2 (*Xho*I)		
		+	-	
Site 1 (*Bam*HI)	+	5	6	11
	-	6	0	6
		11	6	17

- Haplotypes are assumed to be known
- How can we test for the significance of LD?

Exact Test for Linkage Disequilibrium (*Bam*HI & *Xho*I)

Gamete						
+ +	+ -	- +	- -	Prob	Cum.Prob	Chi-square
11	0	0	6	0.0001	0.0001	17.00
10	1	1	5	0.0053	0.0054	9.37
5	**6**	**6**	**0**	**0.0373**	**0.0427**	**5.06**
9	2	2	4	0.0667	0.1094	4.00
6	5	5	1	0.2240	0.3334	1.41
8	3	3	3	0.2666	0.6000	0.88
7	4	4	2	0.4000	1.0000	0.02

Under Ho (No LD) the likelihood is a function of the allele frequencies

		Site 2		
		B	b	
Site 1	A	$n_{AB} = N \bullet p_A p_B$	$n_{Ab} = N \bullet p_A p_b$	
	a	$n_{aB} = N \bullet p_a p_B$	$n_{AB} = N \bullet p_a p_b$	
				N = 2n

Under H_1 (LD) the likelihood is a function of the haplotype frequencies

		Site 2		
		B	b	
Site 1	A	$n_{AB} = N \bullet p_{AB}$	$n_{Ab} = N \bullet p_{Ab}$	
	a	$n_{aB} = N \bullet p_{aB}$	$n_{AB} = N \bullet p_{ab}$	
				N = 2n

Observed multi-locus genotypes (Under LD)

		Site 2		
		BB	Bb	bb
Site 1	AA	n_{11}	n_{12}	n_{13}
	Aa	n_{21}	$n_{22} = n_{AaBb}$	n_{23}
	aa	n_{31}	n_{32}	n_{33}

- *Phase* can be determined for all but one category
- For $n_{22} = n_{AaBb}$ we can not distinguish between (AB; ab) and (Ab; aB)

- Recall that $D = p_{AB} - p_A p_B$
- Haplotypes need to be inferred (via the E-M algorithm or an alternative) since the number $N \bullet p_{AB}$ of AB haplotypes (homologs with A and B alleles) is not observed.

LD & Population Stratification

- Recall that population stratification (or population substructure) is the presence of multiple subgroups between which there is minimal mating or gene transfer

- Population stratification can lead to erroneous conclusions about the presence of LD between markers/genes/SNPs ("Simpson's paradox").