Linkage Disequilibrium

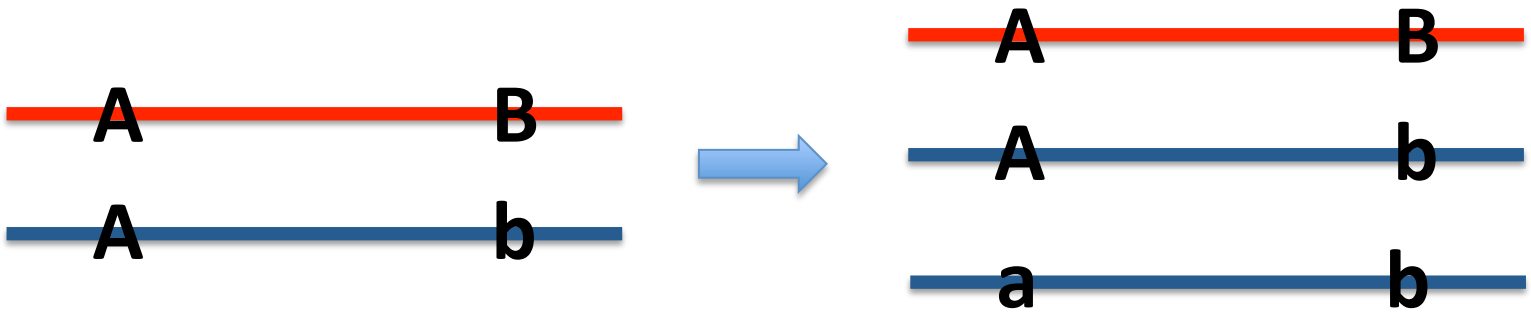
Typed up notes by Derek Aguiar

Linkage Disequilibrium

- Abbreviated LD, represents a statistical linkage between alleles on the same chromosome
- Recombination breaks linkage disequilibrium
- For example, we have one ancestral haplotype that gains a mutation, so now two haplotypes exist in the population.

Linkage disequilibrium

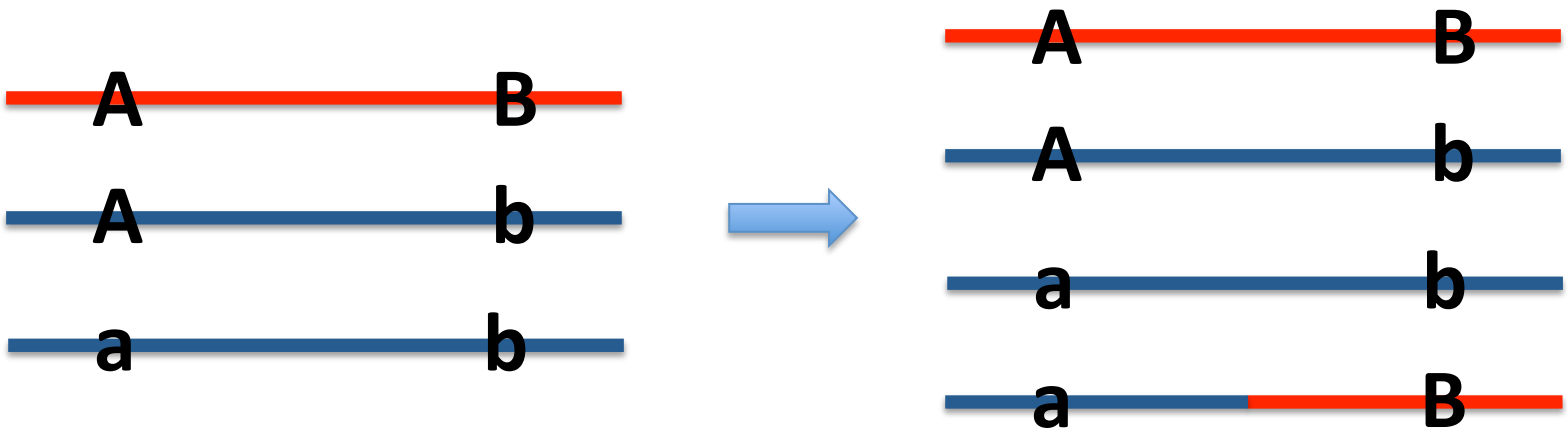
If another site is mutated, three haplotypes exist



But the *a* allele completely determines the *b* allele so the sites are still in strong linkage disequilibrium

Linkage disequilibrium

- Because the probability or mutating either the A/a or B/b variant is very low, linkage disequilibrium is more often broken by recombination



An example in drosophila

Three alleles (genes or variants)

Haplotype	Count
ABC	264
ABc	13
AbC	29
Abc	8
aBC	152
aBc	7
abC	15
abc	1

In the case of two alleles (pairwise measurements)

Haplotype	Frequency
AB	p1q1
Ab	p1q2
aB	p2q1
ab	p2q2

Allele	Frequency
A	p1
a	p2
B	q1
b	q2

Linkage equilibrium

Allele	Frequency
A	p1
a	p2
B	q1
b	q2

Haplotype	Frequency
AB	p1q1
Ab	p1q2
aB	p2q1
ab	p2q2

Real data frequency
P11=frequency of haplotype AB in population
P12=frequency of haplotype Ab in population
P21=frequency of haplotype aB in population
P22=frequency of haplotype ab in population

So we can measure the frequency of haplotypes and subtract of the expected frequency given from the allele frequencies.

This is the fundamental concept for linkage disequilibrium, D

D=P11-p1q1
If D=0, we have linkage equilibrium,
If LD!=0, we have some level of linkage disequilibrium

D

Linkage disequilibrium D
P11=p1q1+D
P12=p1q2-D
P21=p2q1-D
P22=p2q2+D

Lemma

D=P11P22-P12P21

Proof

P11P22=(p1q1+D)(p2q2+D)=p1q1p2q2+p1q1D+p2q2D+D^2

P12P21=(p1q2-D)(p2q1-D)=p1p2q1q2-Dp2q1-Dp1q2+D^2

P11P22-P12P21=D

So, D is indeed a measure that takes into account all of the haplotype frequencies.

D’

D’= D/Dmax when D>0 and
D/Dmin when D<0

where

Dmin=Max{-p1q1,-p2q2}

Dmax=Min{p1q2,p2q1}

D’

- We know P11,P12,P21,P22 >= 0
- If D<0 then Dmin is negative and D’ is positive.
- If D>0 then Dmax is positive and D’ is positive.

- If D<0 then P12P21>P11P22
- If D>0 then P11P22>P12P21

r^2

- r^2=D^2/p1p2q1q2
- r=D/sqrt(p1p2q1q2) is the correlation coefficient between pairs of variants

Statistical testing for LD

- Given a pair of sites, are they significantly associated?
- Hypothesis testing
  - LD detected is statistically significant
- Null hypothesis
  - Linkage equilibrium, Pij=piqj, the frequency of the haplotype is the product of the allele frequencies
- Alternative hypothesis

Hypothesis testing

- 2x2 contingency table, nij= the number of alleles with i in the first position and j in the second
- Sum of the nij = n over all i's and j's
- Apply Fisher's exact test or X^2 test
  - If n is large Fisher is difficult to compute but X^2 is a good approximation
  - X=Sum[(nij-piqj)^2/piqj] where this is approximately X^2 with 1 degree of freedom distributed for large n

Example in drosophila

- Computing X^2 for two sites (ignore C/c allele)

Haplotype	Count	Frequencies
AB	277	0.5656
Ab	37	0.0757
aB	159	0.3251
ab	16	0.03272

  - D=0.0061 and r^2=0.001659

Example in drosophila

- X^2=r^2*n=0.81
- X^2 with 1 degree of freedom gives a probability of 0.37 which is not significant so we cannot reject the null hypothesis

Looking at three sites

- X=13.0
- X^2 prob of 0.0003
- Dmax={0.053,0.102}
- D’=0.012/0.053=.226