Lecture 8 (part a)
Association mapping

Bruce Walsh lecture notes
Tucson Winter Institute
7 - 9 Jan 2013

Limitations of QTL mapping

- The confidence intervals for QTL are greatly large (20cM or more) unless sample is very large
  - Not suitable for fine mapping
  - *AIC* lines can be used to expand map, but a 20cM maping in an F_2 corresponds to roughly a 2 cM in an F_{10} *ACI*, not a huge improvement
- Requires a modest to large collection of relatives
  - Less problematic in plant breeding
- Relies of excess of parental gametes due to linkage to generate marker-trait associations

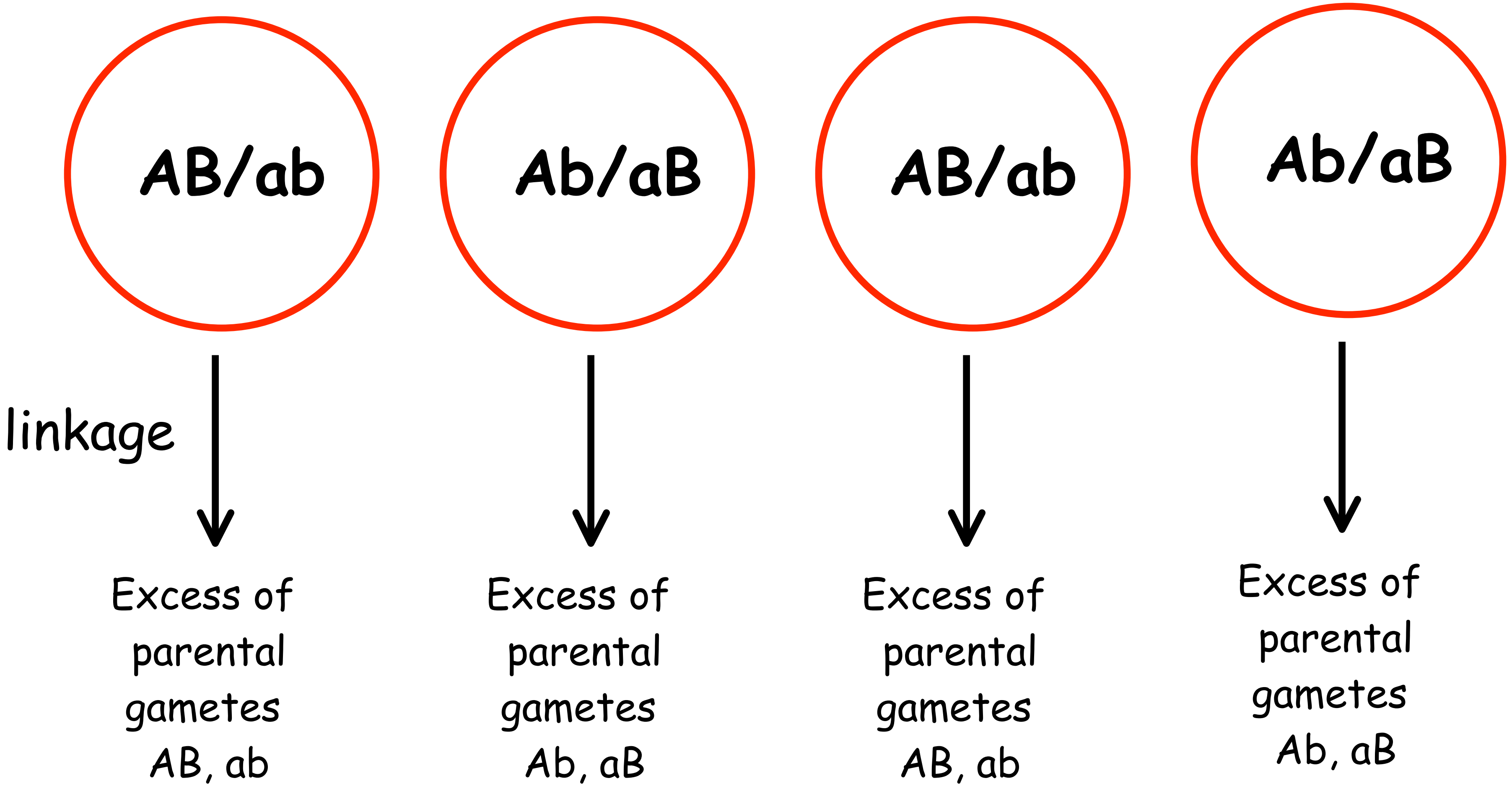
Association mapping

- Use a random collection of individuals from the population (don't need relatives)
- Needs very dense markers
- Uses linkage disequilibrium over very small regions to generate marker-trait associations
- LD over small regions means fine mapping

Linkage vs. Linkage disequilibrium

- Linkage = **excess of parental gametes** from a particular parent
- Linkage disequilibrium = **nonrandom distribution of linkage phases** in the population

No LD: random distribution of linkage phases



AB/ab

Ab/aB

AB/ab

Ab/aB

linkage

Excess of parental gametes AB, ab

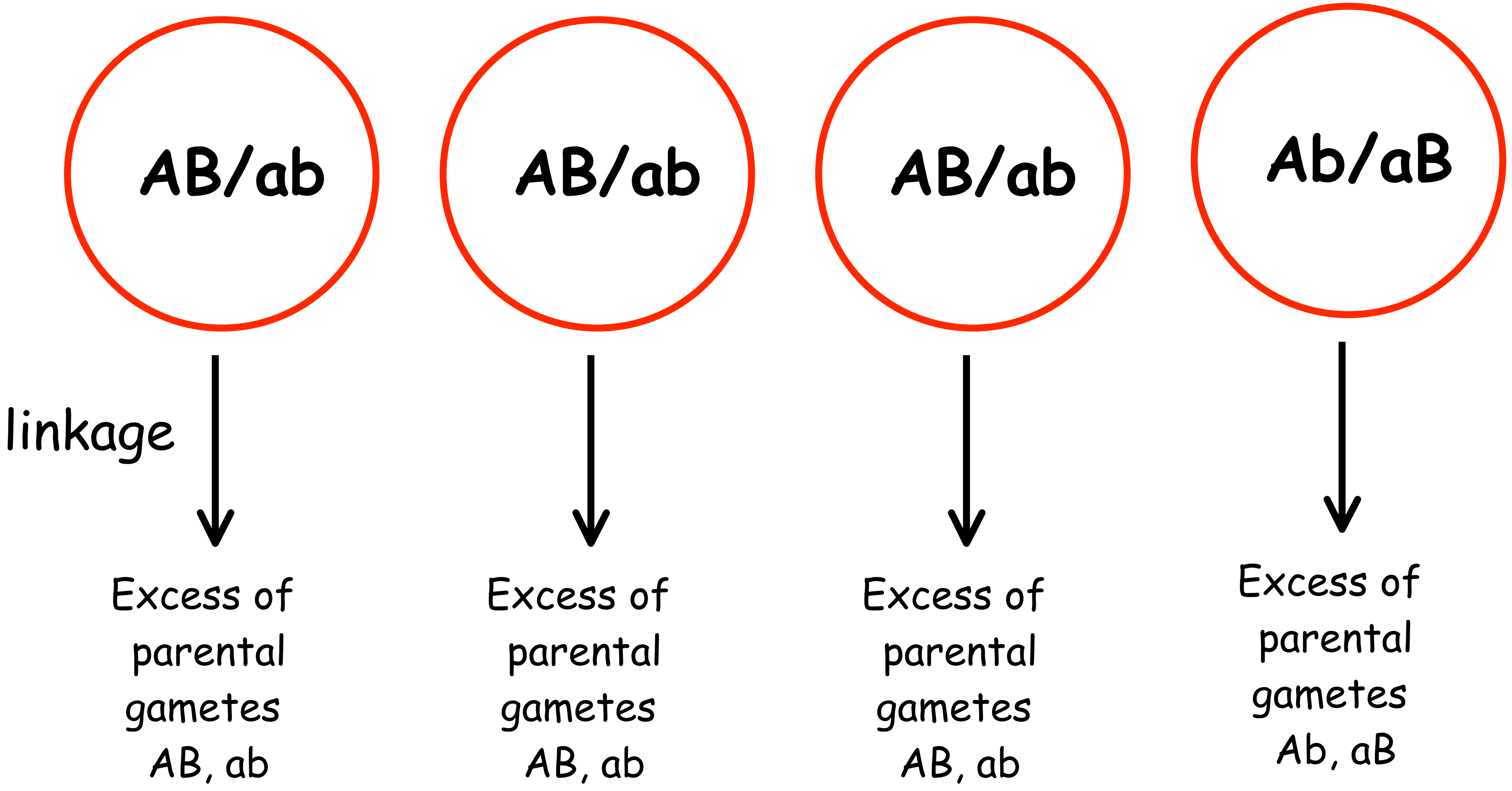
Excess of parental gametes Ab, aB

Excess of parental gametes AB, ab

Excess of parental gametes Ab, aB

Pool all gametes: AB, ab, Ab, aB equally frequent

With LD, nonrandom distribution of linkage phase



AB/ab

AB/ab

AB/ab

Ab/aB

linkage

Excess of parental gametes AB, ab

Excess of parental gametes AB, ab

Excess of parental gametes AB, ab

Excess of parental gametes Ab, aB

Pool all gametes: Excess of AB, ab due to an excess of AB/ab parents

LD: Linkage disequilibrium

$D(AB) = freq(AB) - freq(A)*freq(B)$.

LD = 0 if A and B are independent. If LD not zero, correlation between A and B in the population

If a marker and QTL are linked, then the marker and QTL alleles are in LD in close relatives, generating a marker-trait association.

The decay of D: $D(t) = (1-c)^t D(0)$

here c is the recombination rate. Tightly-linked genes (small c) initially in LD can retain LD for long periods of time

Measures of LD

- The maximum value of D is a function of allele frequencies. For two diallelic loci, let p = Freq(A), q = Freq(B)
  - D_{max} = max[-pq, -(1-p)(1-q)] for D < 0
  - D_{max} = min[p(1-q), (1-p)q] for D > 0
- Lewontin's D' (1964) defined as
  - $D' = D/|D_{max}|$
- Can also scale D by expressing it as the correlation r among alleles
  - r = D/sqrt[p(1-p)q(1-q)]
  - Under drift-mutation-recombination equilibrium, $E(r^2) \sim 1/(1+4N_e c)$

Examples

Gamete	freq
AB	0.3
Ab	0.2
aB	0.4
ab	0.1

freq(A) = p = freq(AB) + Freq(Ab) = 0.5

freq(a) = 1- p = 0.5

freq(B) = q = freq(AB) + Freq(aB) = 0.7

freq(b) = 1- q = 0.3

Linkage-equilibrium value for AB = freq(A)*freq(B) = 0.35

D_{AB} = Freq(AB) - Freq(A)*Freq(B) = -0.05

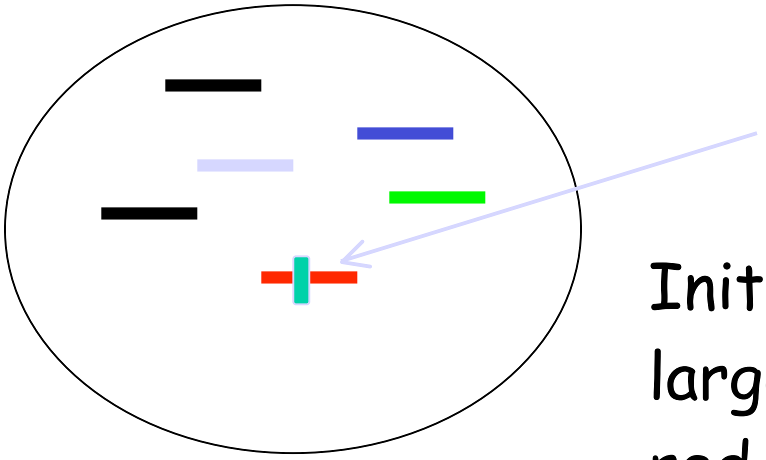
D_{max} = max[-pq, -(1-p)(1-q)] = max(-0.35, -0.15) = -0.15

D′ = D/Dmax = -0.05/0.15 = -0.33

r = D/sqrt(pq(1-p)(1-1) = -0.05/sqrt(0.35*0.15) = -0.22

Fine-mapping genes

Suppose an allele causing a large effect on the trait arose as a single mutation in a closed population



New mutation arises on red chromosome

Initially, the new mutation is largely associated with the red haplotype

Hence, markers that define the red haplotype are likely to be associated (i.e. in LD) with the mutant allele

Thus, new mutations expected to be in almost complete LD with tightly-linked sites (i.e. | D′ | ~ 1)

Dense SNP Association Mapping

Mapping genes using known sets of relatives can be problematic because of the cost and difficulty in obtaining enough relatives to have sufficient power.

By contrast, it is straightforward to gather large sets of unrelated individuals, for example a large number of cases (individuals with a particular trait/disease) and controls (those without it).

With the very dense set of SNP markers (dense = very tightly linked), it is possible to scan for markers in LD in a random mating population with QTLs, simply because c is so small that LD has not yet decayed

Population Stratification

Often try to map genes by using case/control contrasts, also called association mapping.

The frequencies of marker alleles are measured in both a
- case sample -- showing the trait (or extreme values)
- control sample -- not showing the trait

The idea is that if the marker is in tight linkage, we might expect LD between it and the particular DNA site causing the trait variation.

Problem with case-control approach: Population Stratification can given false positives.

When population being sampled actually consists of several distinct subpopulations we have lumped together, marker alleles may provide information as to which group an individual belongs. If there are other risk factors in a group, this can create a false association btw marker and trait

Example. The Gm marker was thought (for biological reasons) to be an excellent candidate gene for diabetes in the high-risk population of Pima Indians in the American Southwest. Initially a very strong association was observed:

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Problem: freq(Gm^+) in Caucasians (lower-risk diabetes Population) is 67%, Gm^+ rare in full-blooded Pima

The association was re-examined in a population of Pima that were 7/8th (or more) full heritage:

Gm^+	Total	% with diabetes
Present	17	59%
Absent	1,764	60%

F_{ST}, a measure of population structure

- One measure of population structure is given by Wright's F_{ST} statistic (also called the fixation index)
- Basically, this is the fraction of genetic variation due to between-population differences
- Consider a biallelic locus (A, a). If p denotes overall pop freq of allele A,
  - then the overall population variation is p(1-p).
  - $\mathrm{Var}(p_i)$ = variance in p over subpopulations
  - $F_{ST} = \mathrm{Var}(p_i)/[p(1-p)]$

Example

Population	Freq(A)
1	0.1
2	0.6
3	0.2
4	0.7

Assume all subpopulations contribute equally to the overall metapopulation

Overall freq(A) = p = (0.1 + 0.6 + 0.2 + 0.7)/4 = 0.4

$\mathrm{Var}(p_i) = E(p_i^2) - [E(p_i)]^2 = E(p_i^2) - p^2$

$= (0.1^2 + 0.6^2 + 0.2^2 + 0.7^2)/4 - 0.4^2 = 0.064$

Total population variance = $p(1-p) = 0.24$

Hence, $F_{ST} = \mathrm{Var}(p_i) /[p(1-p)] = 0.064/0.24 = 0.27$

More general F_{ST}

- Other more general definitions of FST.
  - If f_0 is the probability of IBD within a subpopulation and f the IBD probability for two randomly-drawn individuals from the entire population, then
    - $F_{ST} = (f_0-f)/(1-f)$
  - Alternatively, if t_0 is the average coalescent time for two individuals from the same subpopulation and t the coalescent time for two random individuals from the entire population, then
    - $F_{ST} \sim 1-t_0/t$

Linkage vs. Association

The distinction between linkage and association is subtle, yet critical

Marker allele M is associated with the trait if

$\mathrm{Cov}(M,y) \neq 0$

While such associations can arise via linkage, they can also arise via population structure.

Thus, association DOES NOT imply linkage, and linkage is not sufficient for association