Footprints of Diversity in the Agricultural Landscape:
Understanding and Creating Spatial Patterns of Diversity

Quantitative Genetics for Using Genetic Diversity

Bruce Walsh

Depts of Ecology & Evol. Biology, Animal Science, Biostatistics, Plant Science

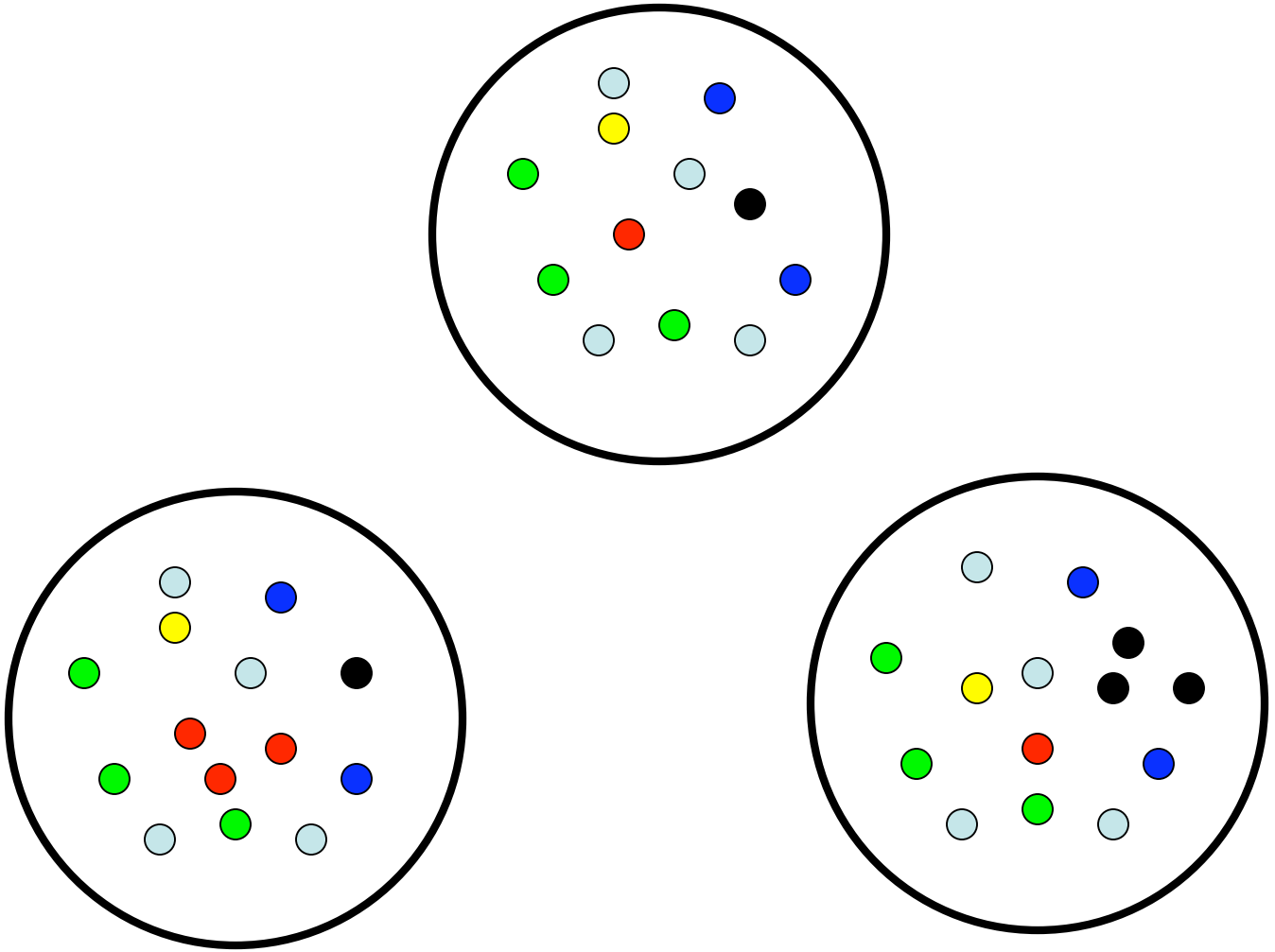
Overview

- Introductory comments
  - Processes generation spatial genetic variation
  - Molecular vs. genetic variation
  - Importance of variability
- Finding genomic locations under selection
  - Expected patterns left in the genome
  - Domestication genes as an example
  - Tools for potentially locating locally-adpative genes
- G x E tools for localizing interesting populations

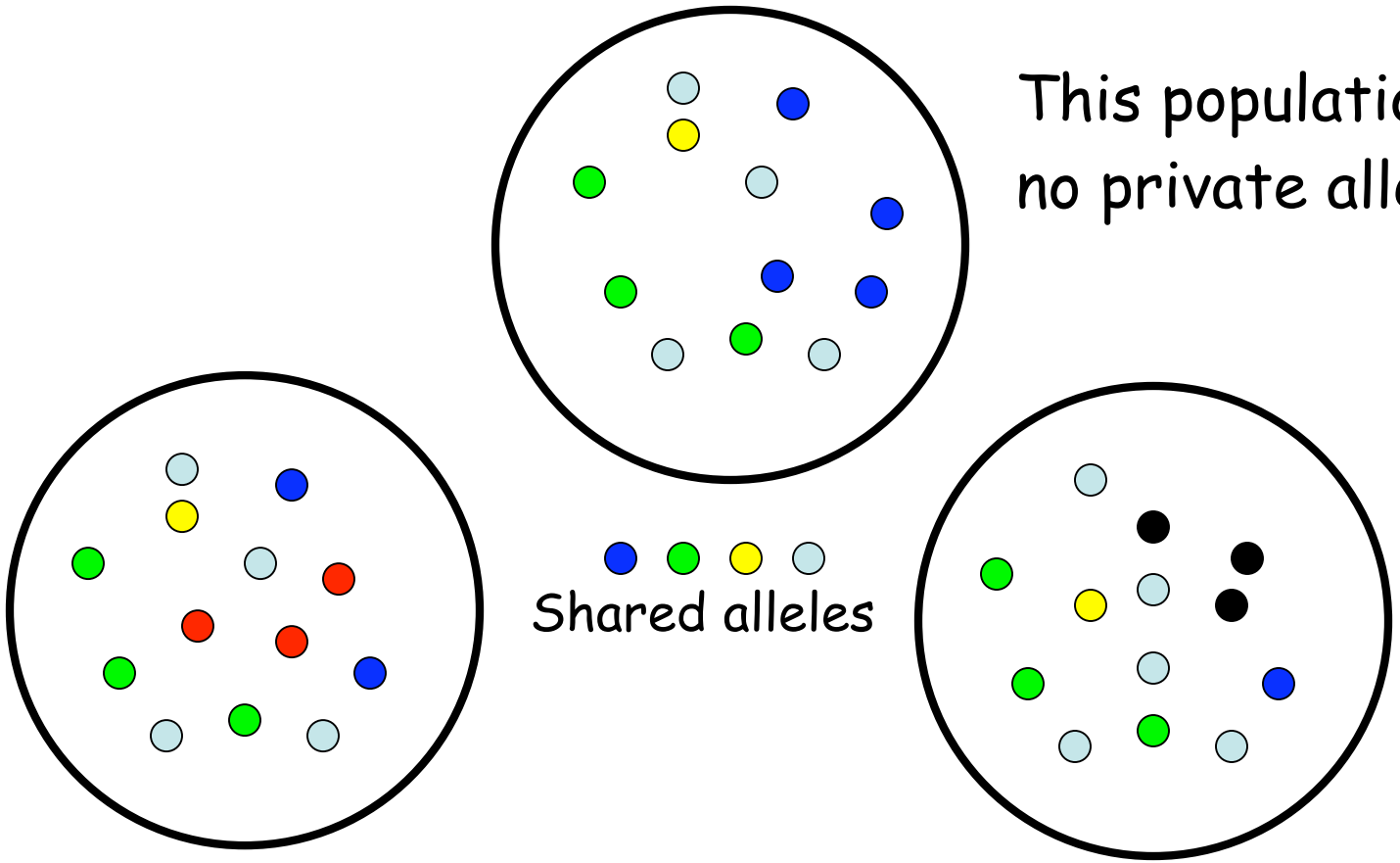
Divergences of populations over time

- The patterning of genetic variation within- and between-populations is a dynamic process
- Loss/fixation of variations via drift and creation of new genetic variation via mutation (and perhaps migration) is a constant background process
- Populations can also evolve via natural selection to be locally-adaptive

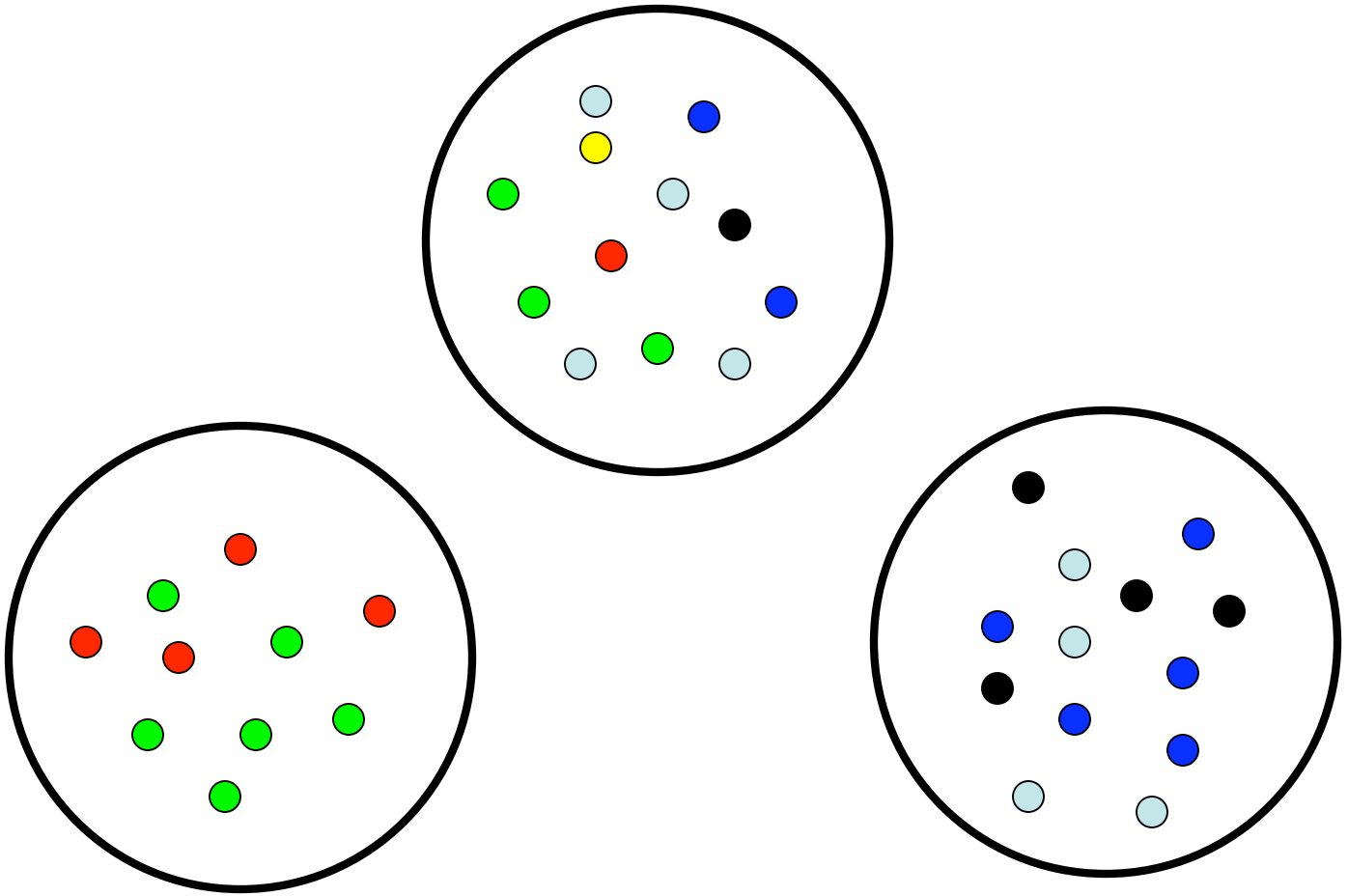
Populations show both within-population variation
As well as between-population variation (variation)



Over time, loss/fixation (via drift) of variation increases the between-population variation unless overpowered by sufficient levels of migration



Variation can also be lost (and hence between-population variation increased) in founder populations



Note reduction of within-population variation relative to founding (source) population

Quantifying levels of variation

- In an ANOVA-like framework, we can ask how much of the total variation over a series of population is in common (within-population variation) and how much is distinct (such as differences in allele frequencies)
- F_{st} = fraction of all genetic variation due to between-population divergence.

Molecular diversity

- SNPs, SSRs (STRs), and other molecular markers widely used to examine genetic variation within populations and divergence between them (such as estimating $_{st}$ and polymorphism levels).
- Much of this pattern of variation is largely shaped by the genetic drift of effectively neutral alleles (the marker alleles)
- Hence, molecular variation is a snap-shot of the neutral variation
  - All loci equally influenced by demography

Genetic divergence

- In isolated populations, drift and mutation cause allele frequencies to change between populations
- However, the breeder is usually interested in those changes from selection
  - Selective adaptation to the local environment
  - Interested in both traits that provide adaptation
  - And in the genes that under these adaptive trait values

Types of divergence

- Three sources of usable genetic variation for breeding from population divergence
  - Accumulation of new QTLs alleles for subsequent selection response
  - Divergence in allele frequencies at loci involved in heterosis
  - Fixation of locally-adaptive mutations.
- How good a predictor is divergence at neutral sites (e.g., SNP, STR data) likely to be for these three?

Accumulation of new variation

- For random quantitative traits, new variance accumulates at roughly 2t Var(M)
- The trait mutational variance/gen Var(M) is typically on the order of (1/1000) of the environmental variance
- Hence, accumulation of variation in a neutral trait tracks the accumulation of divergence at random molecular markers

- Predicts that usable variance can be generated in the cross between two divergent lines (transgressive segregation)
- Transgressive segregation is a potential example of this, the finding in many QTL mapping studies that "up" alleles for a trait are often found in populations with lower trait values (and vise-versa)
- Hence, as a rough approximation, molecular divergence can provide a guide of potentially usable quantitative trait variation

Accumulation of heterotic variation

Recall that the expected heterosis in a cross between two populations is a function of their difference (δp) in allele frequencies at loci showing dominance (d)

$$H_{F_1} = \sum_{i=1}^{n} (\delta p_i)^2 \, d_i$$

δp^2 = variance under drift = 2p(1-p)[1-exp(-t/Ne)]

Hence, δp^2 is expected to increase with divergence time, which can be predicted by levels of molecular divergence

Predicting heterosis

- While expected allele frequency differences increase with time of divergence, this does not guarantee that heterosis will increase with divergence time between populations
- Key is that strong directional dominance ($d > 0$ consistently) is required, and drift also increases the frequency differences in alleles with $d < 0$.
- Levels of marker divergence is a poor predictor of cross heterosis.

Finding genes under selection

- Overall amount of genomic molecular divergence no predictor of divergence in adaptation
- However, can use molecular markers to look for recent signatures of selection
- This, in turn, allows us to localize potentially adaptation genes

Search for Genes that experienced artificial (and natural) selection

Akin in sprit to testing candidate genes for association or using genome scans to find QTLs.

In linkage studies: Use molecular markers to look for marker-trait associations (phenotypes)

In tests for selection, use molecular markers to look for patterns of selection (patterns of within- and between-species variation)

The general approaches for using sequence data to search for signs of selection

Key: Use of features of variation at a marker locus to test for departures from strict neutrality

- Tests based on pattern and amount of within-species polymorphism (departures from neutral predictions).

- Tests based on polymorphism plus between species divergence.

Logic behind polymorphism-based tests

Key: Time to MRCA relative to drift

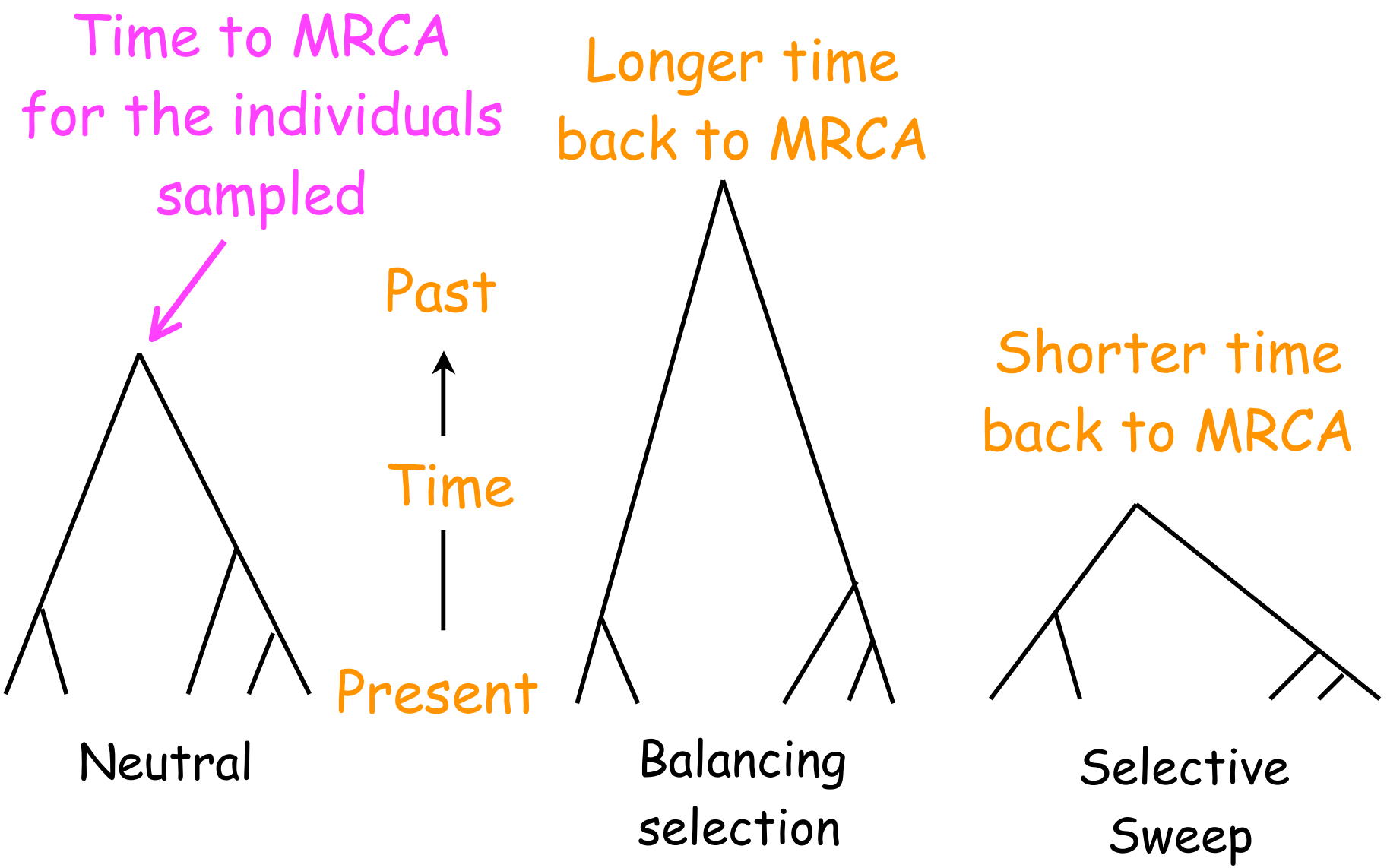
If a locus is under positive selection, more recent MRCA (shorter coalescent)

If a locus is under balancing selection, older MRCA relative to drift (deeper coalescent)

Shorter coalescent = lower levels of variation, longer blocks of disequilibrium

Deeper coalescent = higher levels of variation, shorter blocks of disequilibrium

Selection changes to coalescent times



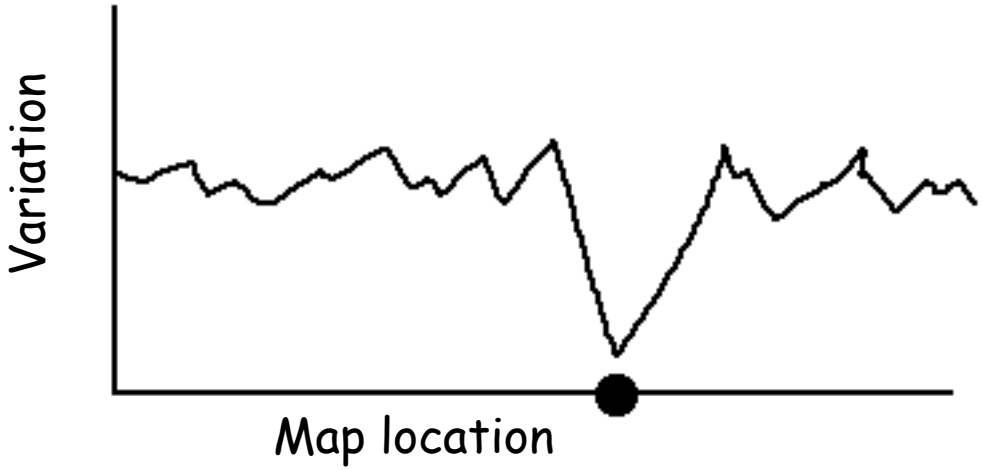
Selective sweeps result in a local decrease in the effective population size N_e around the selective site

This results in a shorter time to MRCA and a decrease in the amount of polymorphism

Note that this has *no effect on the rate of divergence of neutral sites* , as this is independent on N_e.

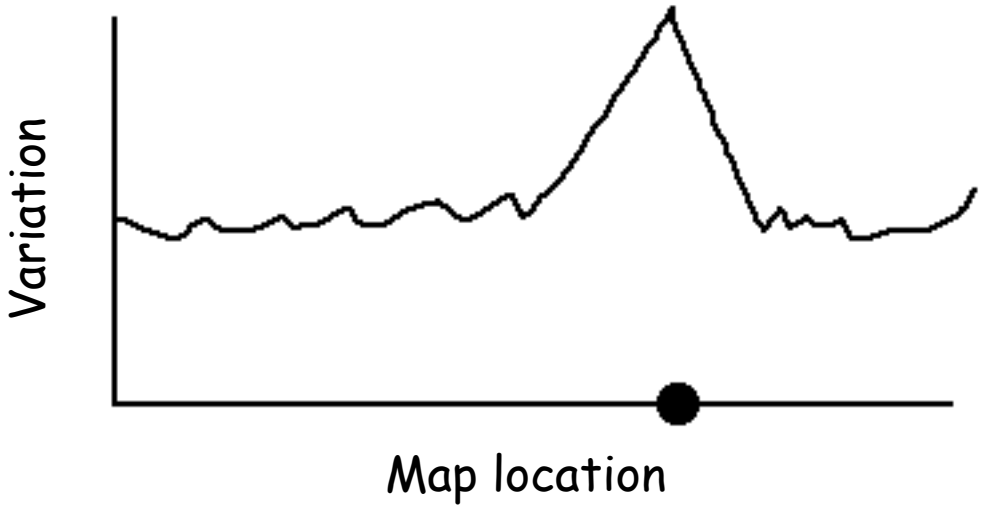
Conversely, balancing selection increases the effective population size, increasing the amount of polymorphism

A scan of levels of polymorphism can thus suggest sites under selection



Directional selection
(selective sweep)

Local region with
reduced mutation rate



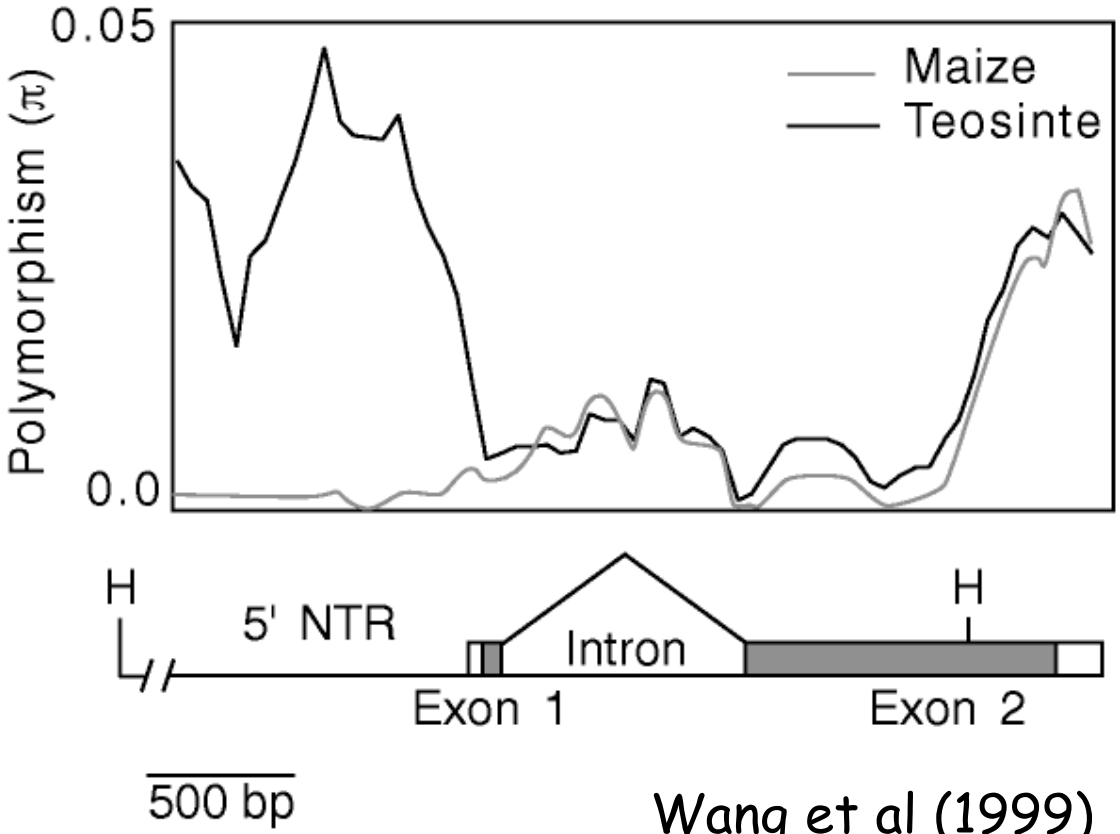
Balancing selection

Local region with
elevated mutation rate

Example: maize domestication gene *tb1*

Doebley lab identified a gene, teosinite branched 1, *tb1*, involved in many of these architectural changes

Wang et al. (1999) observed a significant decrease in genetic variation in the 5′ NTR region of *tb1*, suggesting a selective sweep influenced this region. The sweep did not influence the coding region.



Wang et al (1999) Nature 398: 236.

Polymorphism-based tests

- Given a sample of n sequences at a candidate gene, there are several different ways to measure diversity, which are related under the strict neutral model
  - number of segregating sites. $E(S) = a_n\theta$
  - number of singletons. $E(\eta) = \theta * n/(n-1)$
  - average nu. of pairwise differences, $E(k) = \theta$
- A number of tests (e.g., Tajima's D) are based on detecting departures from these expectations

Major Complication With Polymorphism-based tests

Demographic factors can also cause these departures from neutral expectations!

Too many young alleles -> recent population expansion

Too many old alleles -> population substructure

Thus, there is a composite alternative hypothesis, so that rejection of the null does not imply selection. Rather, selection is just one option.

Can we overcome this problem?

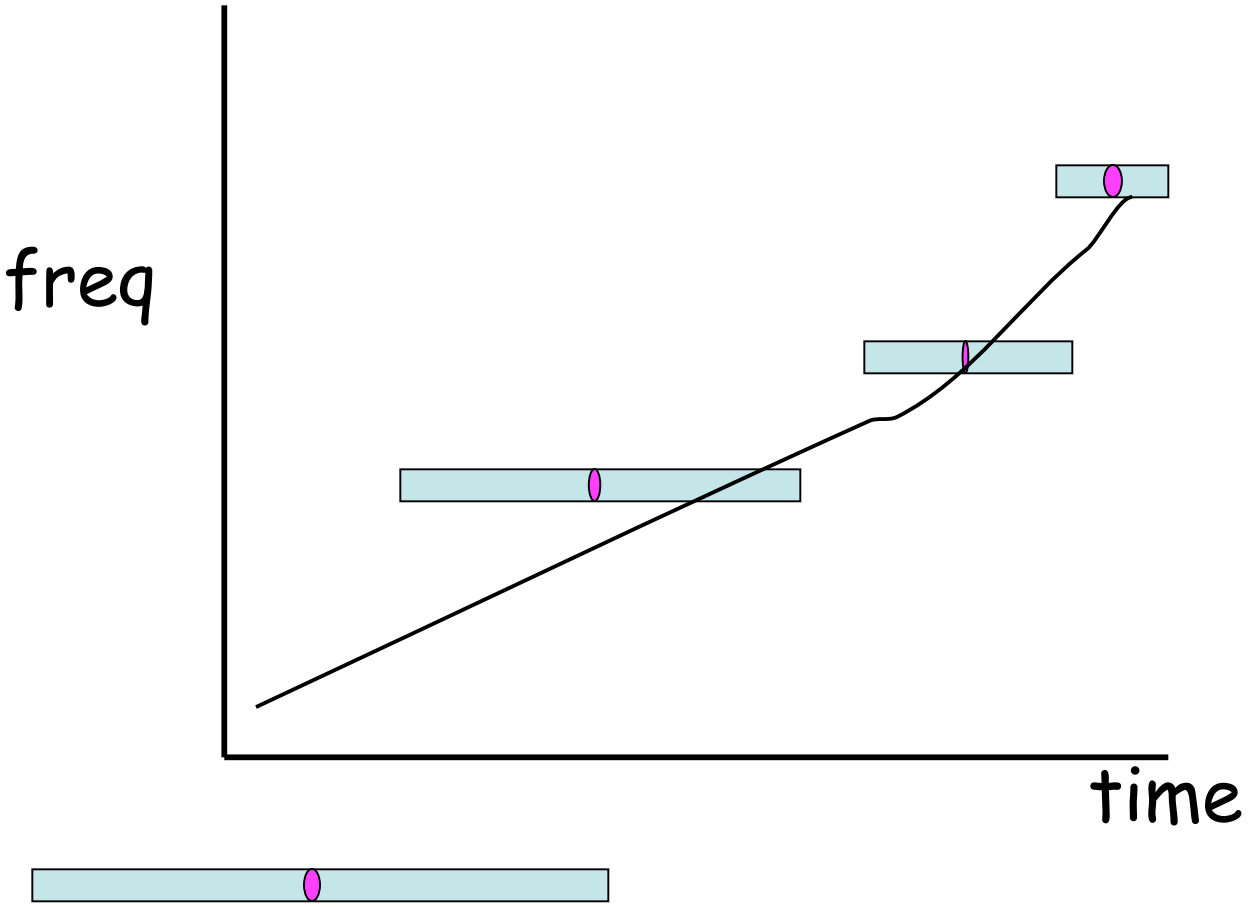
It is an important one, as only polymorphism-based tests can indicate on-going selection

Solution: demographic events should leave a constant signature across the genome

Essentially, all loci experience common demographic factors

Genome scan approach: look at a large number of markers. These generate null distribution (most not under selection), outliers = potentially selected loci (genome wide polymorphism tests)

Under pure drift, high-frequency alleles should have short haplotypes



Starting haplotype

Linkage Disequilibrium Decay

One feature of a selective sweep are derived alleles at high frequency. Under neutrality, older alleles are at higher frequencies.

Sabeti et al (2002) note that under a sweep such high frequency young alleles should (because of their recent age) have much longer regions of LD than expected.

Wang et al (2006) proposed a Linkage Disequilibrium Decay, or LDD, test looks for excessive LD for high frequency alleles

Optimal conditions for detecting selection

High levels of polymorphism at the start of selection

High effective levels of recombination gives a shorter window around the selective site

High levels of selfing reduces the effective recombination rate (eg. Maize vs. rice)

Signatures of sweeps persist for roughly N_e generations

Summary

Linkage mapping vs. detection of selected loci

Linkage: Know the target phenotype

Selection: Don't know the target phenotype

Both can suffer from low power and confounding from demographic effects

Both can significantly benefit from high-density genomic scans, but these are also not without problems.

G x E

- The flip side of molecular divergence is the direct accessment of trait values in a set of populations/lines over a series of environments
- Lines that show strong positive G x E (genotype-environment interactions) in a particular environment (or set of environments) are sources of improvement genes for a target environment (or TEP)

Basic G x E model

- Basic model is the mean value of line i in environment j is $u + G_i + E_j + GE_{ij}$
- G_i is the line average over all environments
- E_j is the environmental effect over all lines
- GE_{ij} is the G x E interaction

Looking for structure in G x E

- Often there is considerable structure in G x E, so that the ij-th term can be estimated as a simple product
  - $GE_{ij} = a_i b_j$
  - More general bilinear models can be used
  - Key: a_i can be thought of an a genotypic environmental specificity factor
- Modification is to use **factorial regression**
  - Here one used measured environmental factors (temp, rainfall, etc) to try to predict GE
  - One can also incorporate measured genes as well

$$GE_{ij} = \eta_{1i} y_{1k} + \cdots + \eta_{pi} y_{pk} + \delta_{ij}$$

- Suppose that y_1 .. y_p are p environmental factors that are measured by the breeder (e.g., degree days, rainfall,etc.), with y_{jk} the value of factor j in environment k
- The idea is to predict GE by looking at how different lines react to each environmental factor
- η_{1i} is the measure of the sensitivity to line i to environmental factor 1, η_{2i} to factor 2, etc.

- Factorial regressions allow the breeder to examine how each line reacts to a variety of environmental factors, potentially offering differential targets of selection
- Example: Epinat-Le Signor et al. (2001)
  - Looked at maize
  - A major contributor to GxE was the interaction between a line's date of flowering and water supply, with early varieties becoming more favorable as the water supply decreases

Using factorial regressions

- Specific trait-environment interactions
- Specific line-environmental factor interactions
- Specific gene-environmental factor interactions

Summary

- Level of genome-wide divergence using molecular markers
  - a weak signal for usable QTL variation
  - a very poor (at best!) signal for heterosis
  - No signal for presence of locally-adaptive genes
- Signals of adaptive genes
  - Changes in polymorphism levels around target
- Use of factorial regressions to tease out components of GXE
  - Environmental factors within E
  - Traits, genes within lines