

# Molecular Population Genetics Analysis

## Marker data can be useful

- Indicate the haplotype state of an individual

## Haplotype

From the term  
**HAPLOid genoTYPE**

- The specific combination of nucleotides across a locus shared by multiple individuals in a population

- Whole genome level

Requires whole-genome resequencing of a **population** of individuals that vary across the species

- Sequence the best indicator
- Develops “hapmaps” ← **Key Term**
- Species-wide effort to describe variation across the whole genome of a species
  - Human Hapmap
  - Medicago Hapmap

- Collection of markers

- Selected marker loci distributed “across” the genome
  - Why “across” in parenthesis?

\$25-\$50 per

Requires genotyping a variable **population** with a SNP chip

- Most markers are from the euchromatic region of the genome.
- Deep sequencing using Next Generation Sequencing provides more coverage
  - But a reference genome needed for mapping SNPs to a location

- Gene sequences

- The combination of the various SNPs in a gene or a gene region

Requires sequence data for a single gene from members of a variable **population**

**20X - 40X**  
Genome = 500Mb  
1X = 500Mb data  
10X = 5000Mb (5Gb) data  
20X = 10,000Mb (10Gb) data

## Haplotype structure example

- Intron of chalcone isomerase intron 3 of common bean
- Sample of 67 individuals (landraces and cultivars)
  - 10 haplotypes observed

Haplotype of a single region of a single gene

H

a

**Position of variable SNPs in sample**

p

l

o

1111111111111111111111111122222222333334445566  
 783333334444459999999990003446912671290703  
 681567890123491234567890121367517566156812

---

**1 GCTTTTTTTTGTGATACGAACACAGAAGTTCACTGTTTCGAC**  
 2 .....A..  
 3 T.....  
 4 .A.....- - - - .T.....GGC.CTGTC A.-.....  
 5 .A.....- - - - .T.....GGCCCTGTCA.-.....  
 6 AA.....- - - - .T.....GGC.CTGTC A.-.....  
 7 ..A- - - - - - - - - - .- - - - - - - - - -GGC.....-T.G.  
 8 ..A- - - - - - - - - - .- - - - - - - - - -GGC.....A-T.G.  
 9 ..A- - - - - - - - - - .G- - - - - - - - - -GGC.....-T...

Reference

The haplotype is shared by ancestry.

# A whole genome HapMap

## Medicago HapMap GBrowser

www.medicagohapmap.org

The screenshot displays the Medicago HapMap GBrowser interface. At the top, the browser title is "Medicago truncatula 3.0: chr2:364000\_370000 - Mozilla Firefox". The address bar shows the URL "www.medicagohapmap.org/cgi-bin/gbrowse/mhapmap/". The main header is green and contains "Mt3.0" and "Medicago truncatula HAPMAP PROJECT". Below the header, it indicates "Showing 6.001 kbp from chr2, positions 364,000 to 370,000".

The interface includes several sections:

- Instructions:** Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed. Navigation: Click one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position. Examples: chr2:364000\_370000, chr5:13115000\_13125000, chr5:210000\_215000.
- Search:** Landmark or Region: chr2:364000\_370000. Data Source: Medicago truncatula 3.0.
- Reports & Analysis:** Annotate Coverage Ratio Plot. Scroll/Zoom: Show 6.001 kbp.
- Overview:** A genomic ruler showing the location of the region on chromosome 2.
- Region:** A detailed view of the region on chromosome 2, with a yellow highlight indicating the current view.
- Details:** A track showing Gene Models (Genes) and SNPs found within all lines and individual lines. The gene model track shows the gene Medtr2g060620 (Putative ion channel [K1-1], identical). The SNP track shows SNPs found within all lines (All Lines Combined (Variations)) and SNPs found within individual lines (HM001 - SA22322 (Variations), HM002 - SA28064 (Variations), HM003 - ESP105-1 (Variations), HM004 - DZA045-6 (Variations), HM005 - DZA315-16 (Variations), HM006 - F83005-5 (Variations)).

Annotations on the screenshot include:

- "Location in genome" pointing to the genomic ruler in the Region section.
- "Specific gene" pointing to the gene model track in the Details section.
- "SNPs found within all lines" pointing to the All Lines Combined (Variations) track in the Details section.
- "SNPs found within individual lines" pointing to the individual variation tracks in the Details section.

The bottom of the interface shows the Tracks section with "Overview" selected and "Centromere" checked. The Windows taskbar at the bottom shows the Start button and various application icons, with the system clock displaying 9:41 AM.

# Phylogenetics

- Definition
  - The study of the evolutionary relationship between a collection of genotypes
- Can be based on
  - Phenotype
  - Molecular markers
  - Sequence data
- Creates a branching pattern that
  - Depicts the relationship of the members of the population

**Important Result of Phylogenetics!!!**



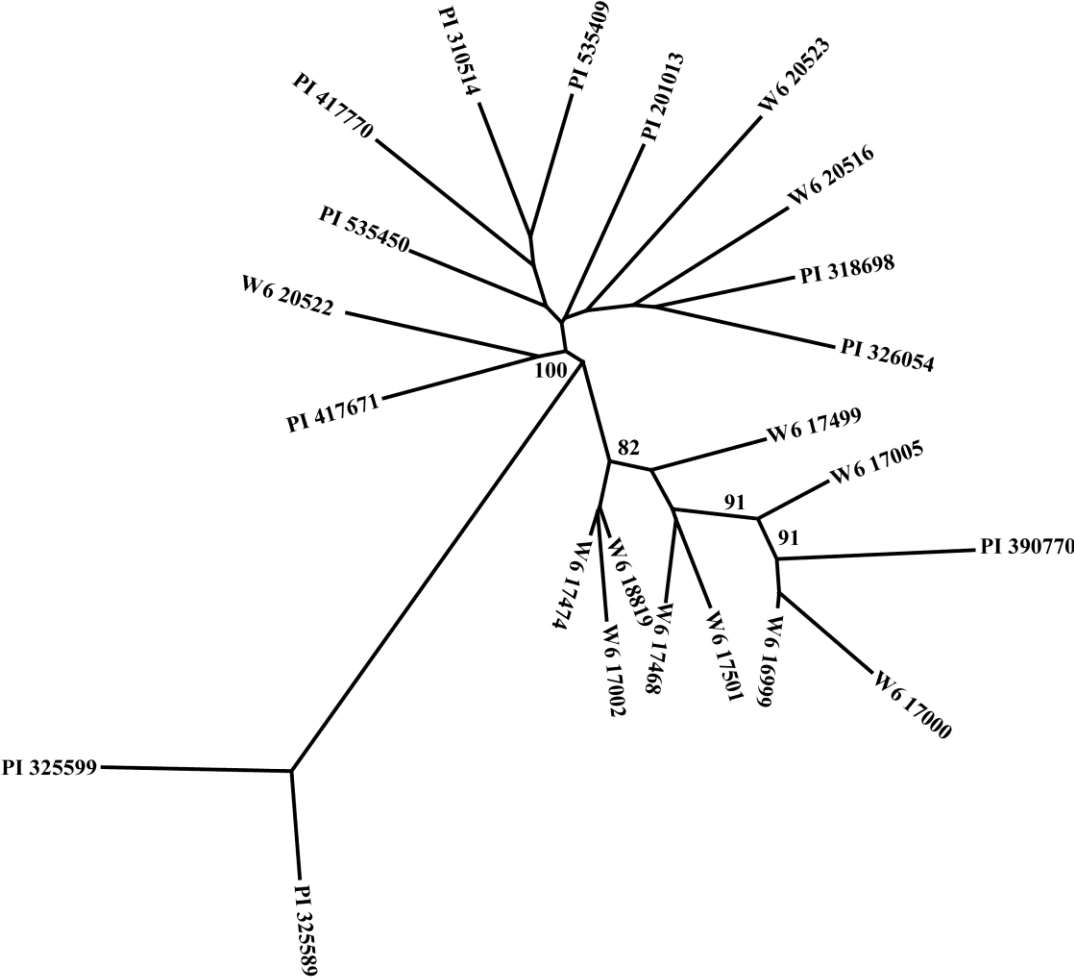
## Two Major Approaches Used in Applied Crop Phylogenetics

### Neighbor joining

- Popular approach
- Based on distance between individuals
  - Distance measure based on marker or sequence data
- Theory
  - Minimum evolutionary steps approach
    - Evolution proceeds by the fewest possible steps

**Two individuals with the fewest differences are the most closely related individuals.!!!**

# NJ Phylogenetic Tree Example – Common Bean



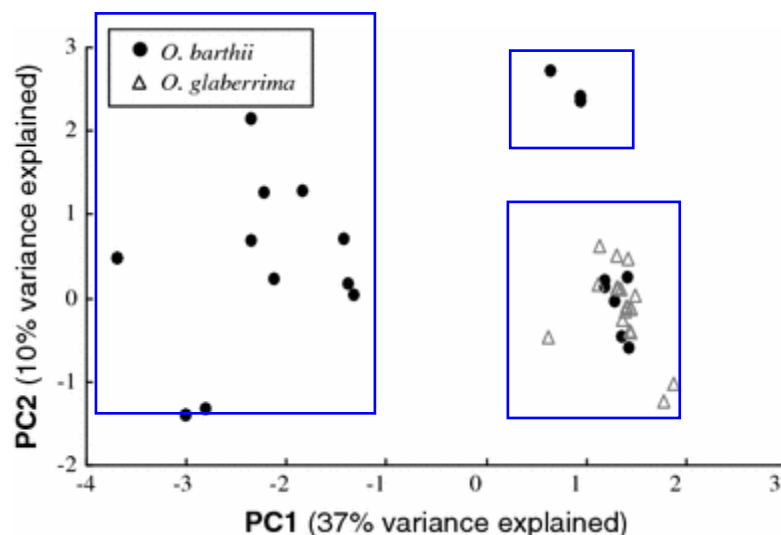
0.1



## Principle component analysis

- “A mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components.” (from Wikipedia)
  - Result
    - Successive components that each account for a decreasing amount of the variation of the data
  - PCA and Molecular Phylogenetics
    - PC 1
      - Related to an important feature of the population
    - Other PCs
      - Show the relationship among individuals for other features
      - Can show transition between individuals in the populations
    - Should confirm the tree building approach

**Fig. 4** Principal component analysis (PCA) of 40 *O. glaberrima* and *O. barthii* samples based on sequences of 14 nuclear loci. The first eigenvector (PC1) explained 37% of variation and the second (PC2) explained 10% of variation



**Li et al. 2011. Genetic diversity and domestication history of African rice... *Theor Appl Genet* 123:21-31.**

## STRUCTURE Software

- Popular software often used in studies that define the organization of a population of genotypes
- Defines the number of subpopulations that “best” define a population
- Describes the ancestry of an individual relative to the subpopulations
  - Ancestry expressed as a percentage ( $q_{kn}$ ) of each subpopulation

	Subpopulations			
	Subpop 1 ( $q_{k1}$ )	Subpop 2 ( $q_{k2}$ )	Subpop 3 ( $q_{k3}$ )	Subpop 4 ( $q_{k4}$ )
Individual A	90%	5%	0%	5%
Individual B	10%	75%	5%	10%
Individual C	35%	5%	15%	45%

### Interpreting results

**Individual A:** Subpopulation 1 membership

**Individual B:** Subpopulation 2 membership

**Individual C:** Admixed individual; subpopulation membership not assigned

### General approach of STRUCTURE

- Bayesian-model based approach
  - Model is
    - The number of subpopulations

- Individuals are assigned to a subpopulation based on genotype

## Principle

- Attempts to account for Hardy-Weinberg and linkage disequilibrium by ***imposing population substructure*** on the data

## Assumptions

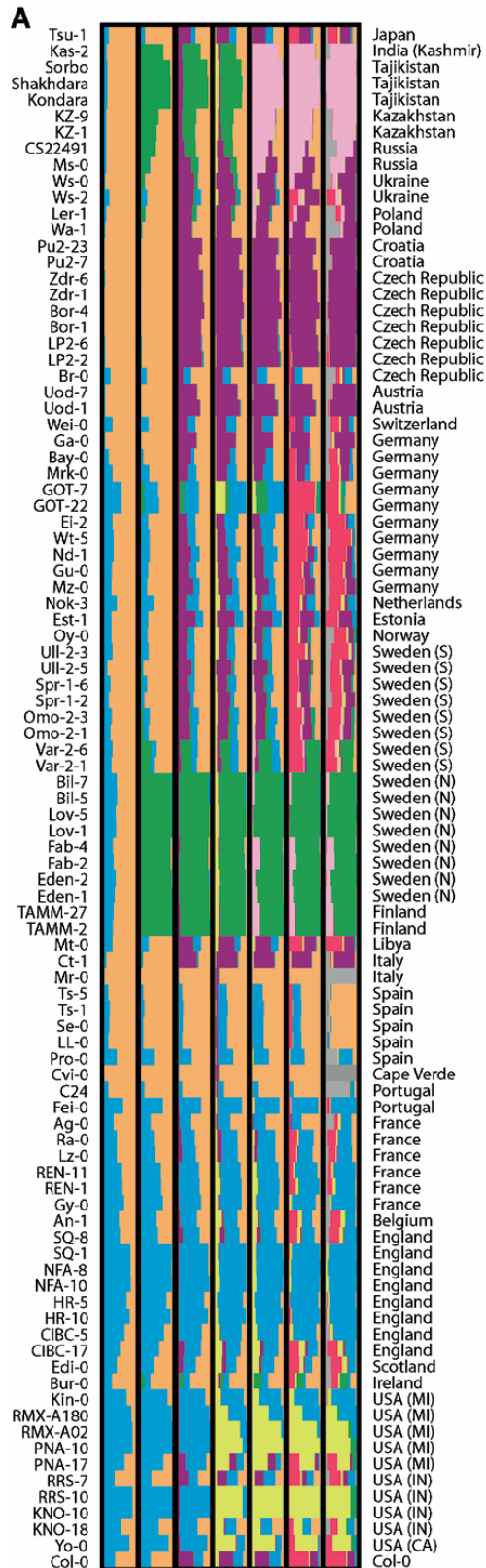
- All individuals in the full population are members of a specific  $k$  subpopulation ( $k_n$ )
- All loci within a subpopulation are in ***Hardy-Weinberg equilibrium***
- All loci within a subpopulation are in ***linkage equilibrium***
- A genotype can be defined relative to the percentage of each subpopulation in its ancestry ( $q_{kn}$ )
- Admixture among subpopulations has not occurred
  - Admixture definition
    - Intermating among previously separated populations
      - Current version of STRUCTURE allows for admixture
- Loci are unlinked
  - Original feature
    - Linked loci are now allowed in current version

Assumptions not usually met with populations!! Be cautious when interpreting results

Primary paper: Pritchard et al (2000) Genetics 155:945

Software: <http://pritch.bsd.uchicago.edu/software.html>





## STRUCTURE Example

from: Nordborg et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biology 3:e196

96 individuals

SNP data for 876 loci

$k=2$

- population split along an East-West gradient

$k=3$

- Sweden/Finland cluster separates

$k=3-8$

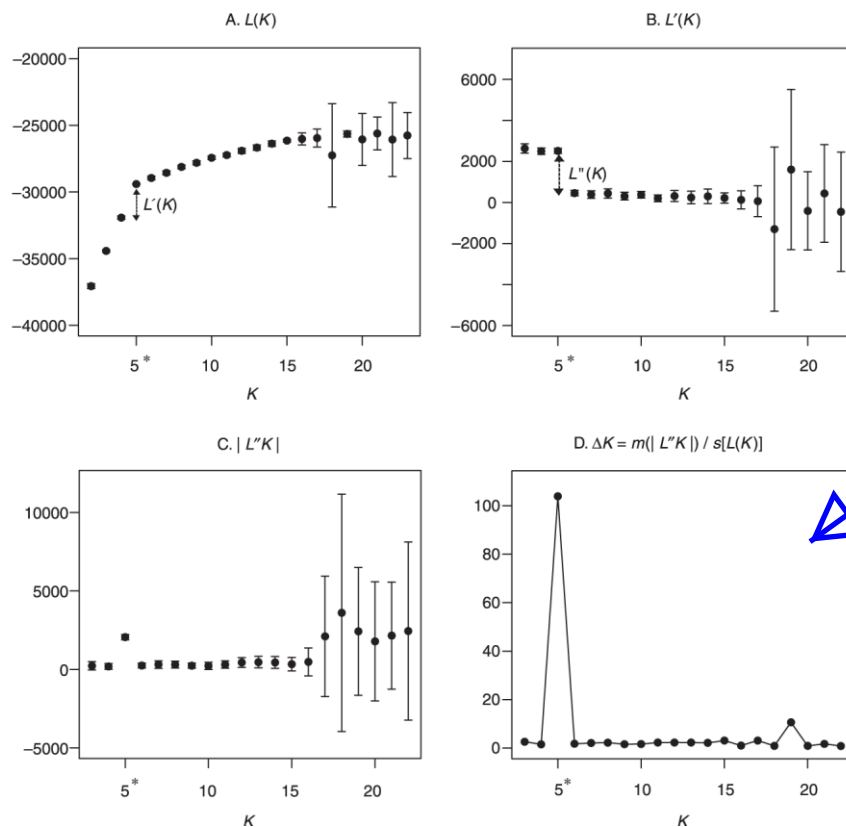
- cluster split further along geographic borders

$K = 2$   
 $K = 3$   
 $K = 4$   
 $K = 5$   
 $K = 6$   
 $K = 7$   
 $K = 8$

# Determining the number of subpopulations in the sample

Evanno et al. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* 14:2611

**Fig. 2** Description of the four steps for the graphical method allowing detection of the true number of groups  $K^*$ . (A) Mean  $L(K)$  ( $\pm$  SD) over 20 runs for each  $K$  value. The model considered here is a hierarchical island model using all 100 individuals per population and 50 AFLP loci. (B) Rate of change of the likelihood distribution (mean  $\pm$  SD) calculated as  $L'(K) = L(K) - L(K - 1)$ . (C) Absolute values of the second order rate of change of the likelihood distribution (mean  $\pm$  SD) calculated according to the formula:  $|L''(K)| = |L'(K + 1) - L'(K)|$ . (D)  $\Delta K$  calculated as  $\Delta K = m|L''(K)| / s[L(K)]$ . The modal value of this distribution is the true  $K^*$  or the uppermost level of structure, here five clusters.



- Good for determining the most basic structural features of the data
- **Not good at measuring fine-structure population features**

## Uses of Molecular Phylogenetics

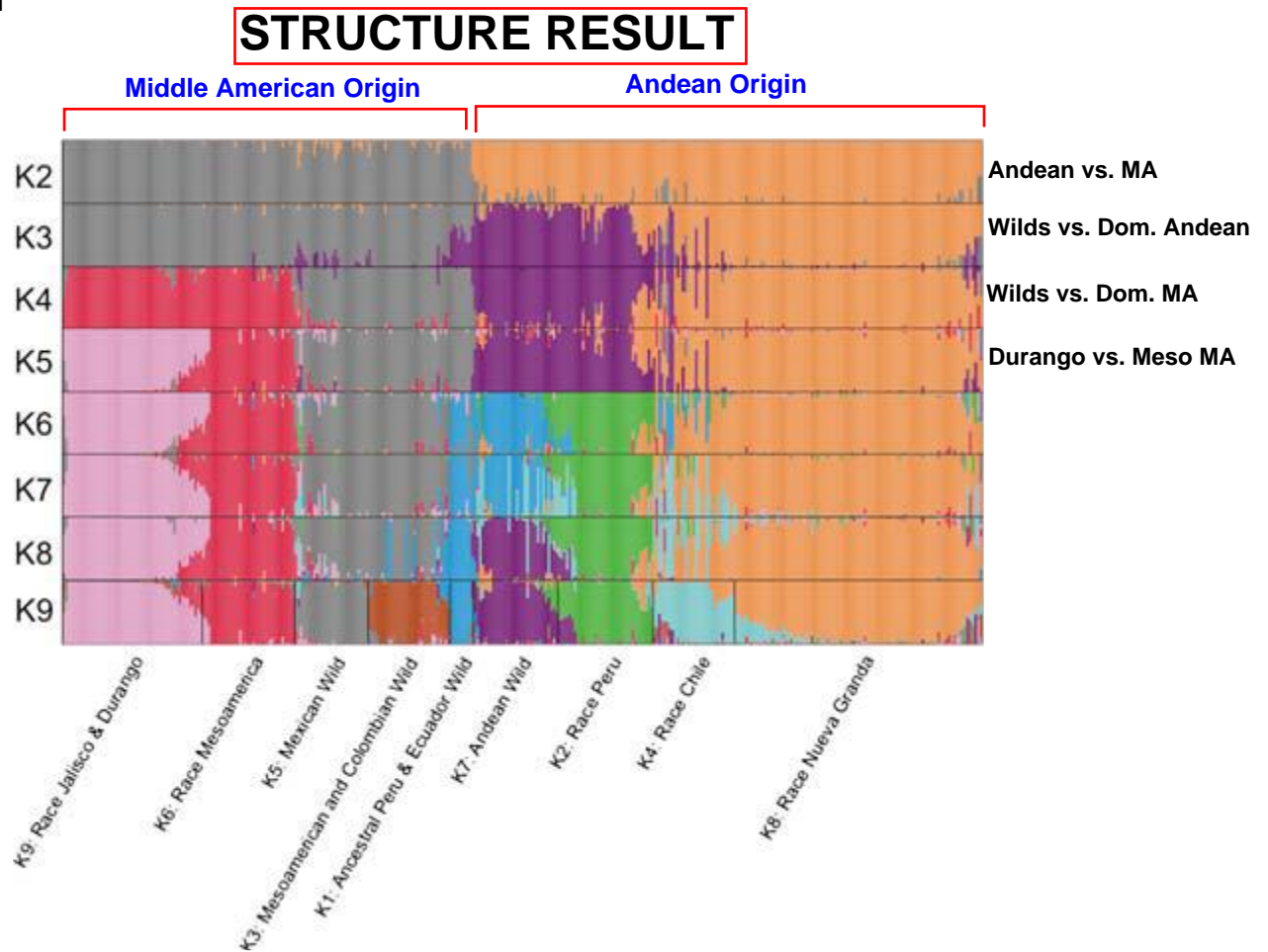
- Describes trees of life
  - At any taxonomical level
- Follows the relationships of haplotypes
- Evaluates the origin of a genotypic group
  - Defines ancestral materials that are progenitors of a current population

# Complete Phylogenetic Example Using All the Current Analytic Tools

Kwak and Gepts (2009) Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae) Theoretical and Applied Genetics 118:979-992.

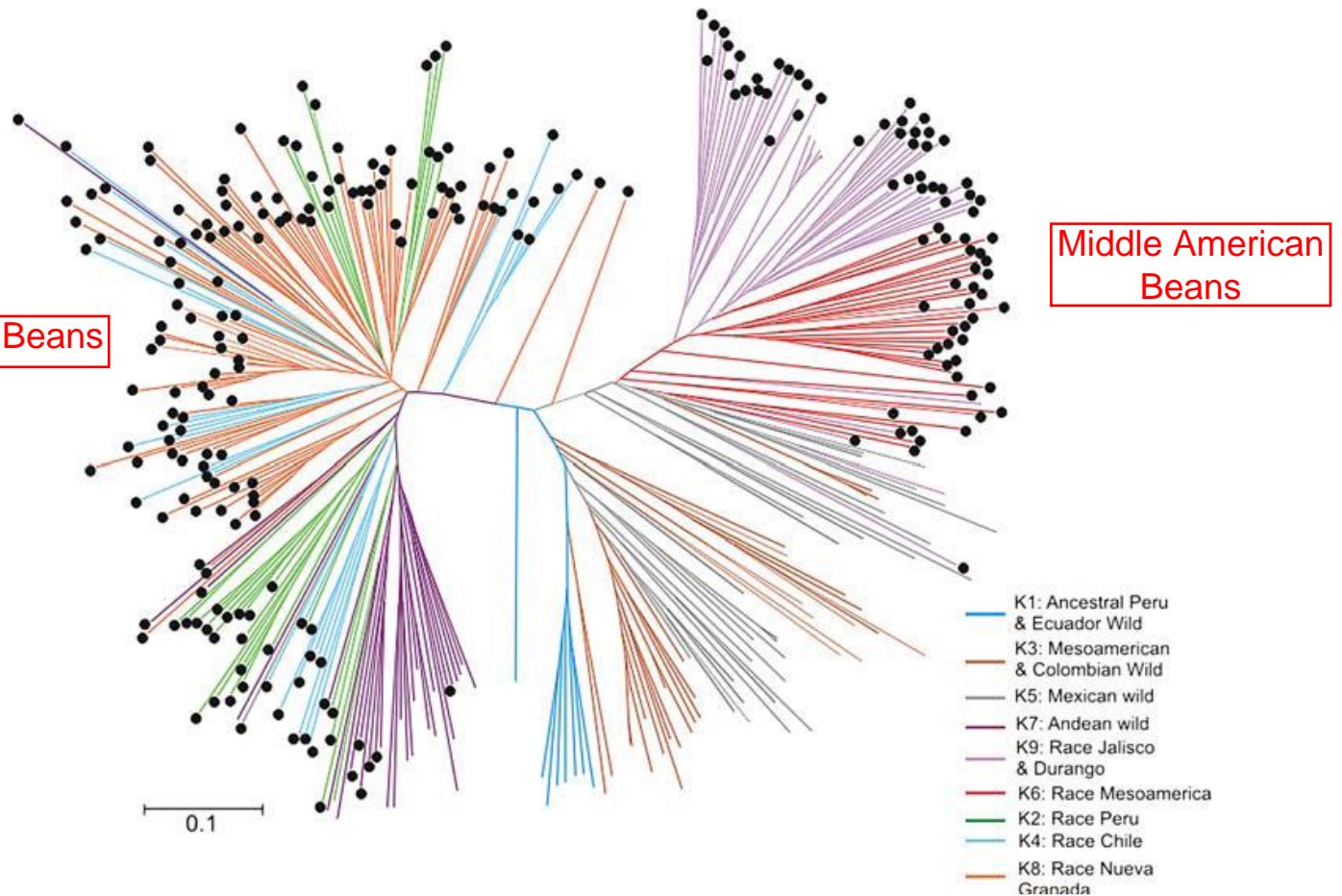
**Fig. 1** Hierarchical organization of genetic relatedness of 349 common bean accessions based on 26 microsatellite markers and analyzed by the STRUCTURE program as described in "Materials and methods" for K = 2 to 9. Bar graphs were developed with the program DISTRICT

A few markers can be informative!!!!



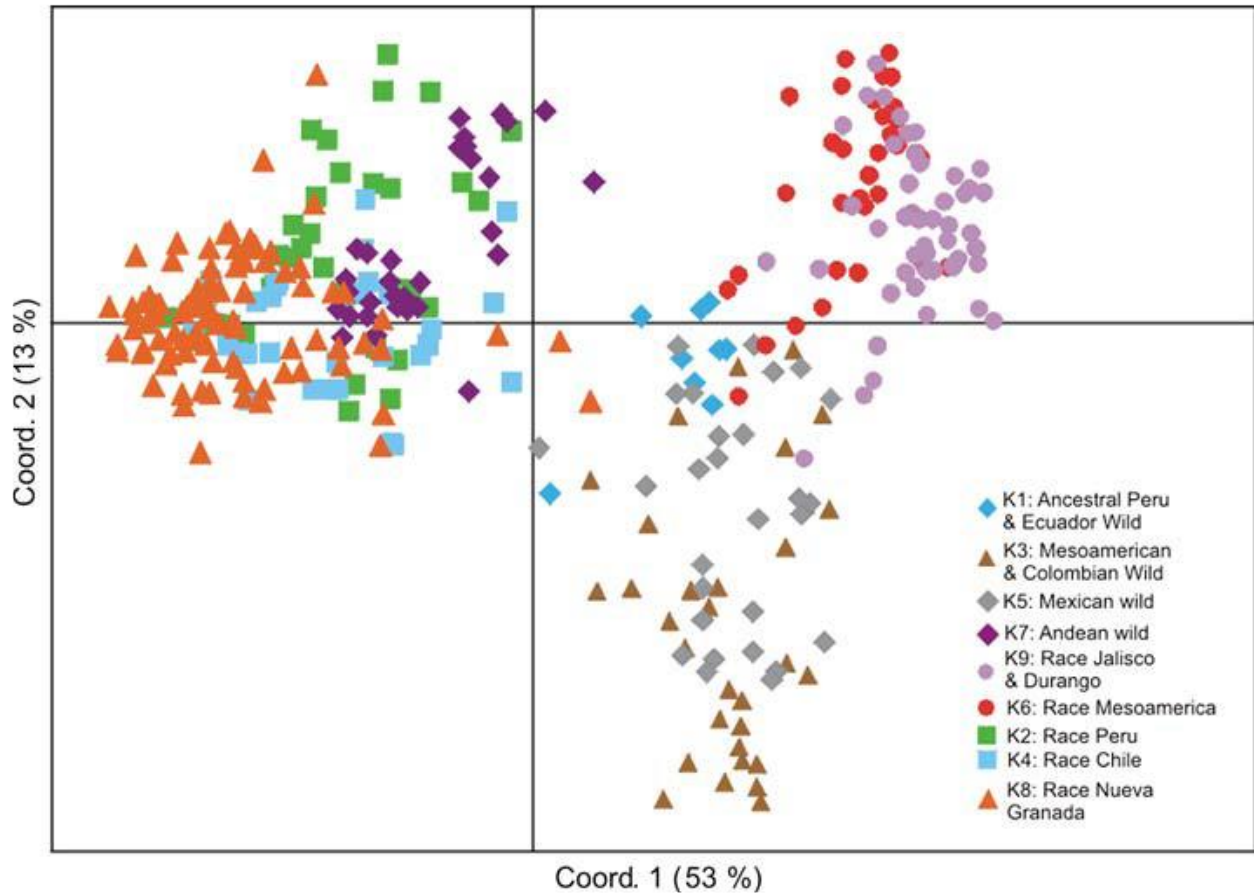
## NEIGHBOR-JOINING RESULT

**Fig. 2** Neighbor-joining tree of microsatellite diversity based on the C. S. Chord distance implemented in the Powermarker program. Each branch is color-coded according to membership into the K = 9 groups identified by STRUCTURE (same colors as in Fig. 1). Branches ending with black dots represent domesticated accessions, while those without dots are wild accessions.



# PRINCIPAL COMPONENT ANALYSIS RESULT

**Fig. 3** Principal coordinate analysis of microsatellite diversity based on the presence absence of alleles. Colors represent populations identified at K = 9 in Fig. 1



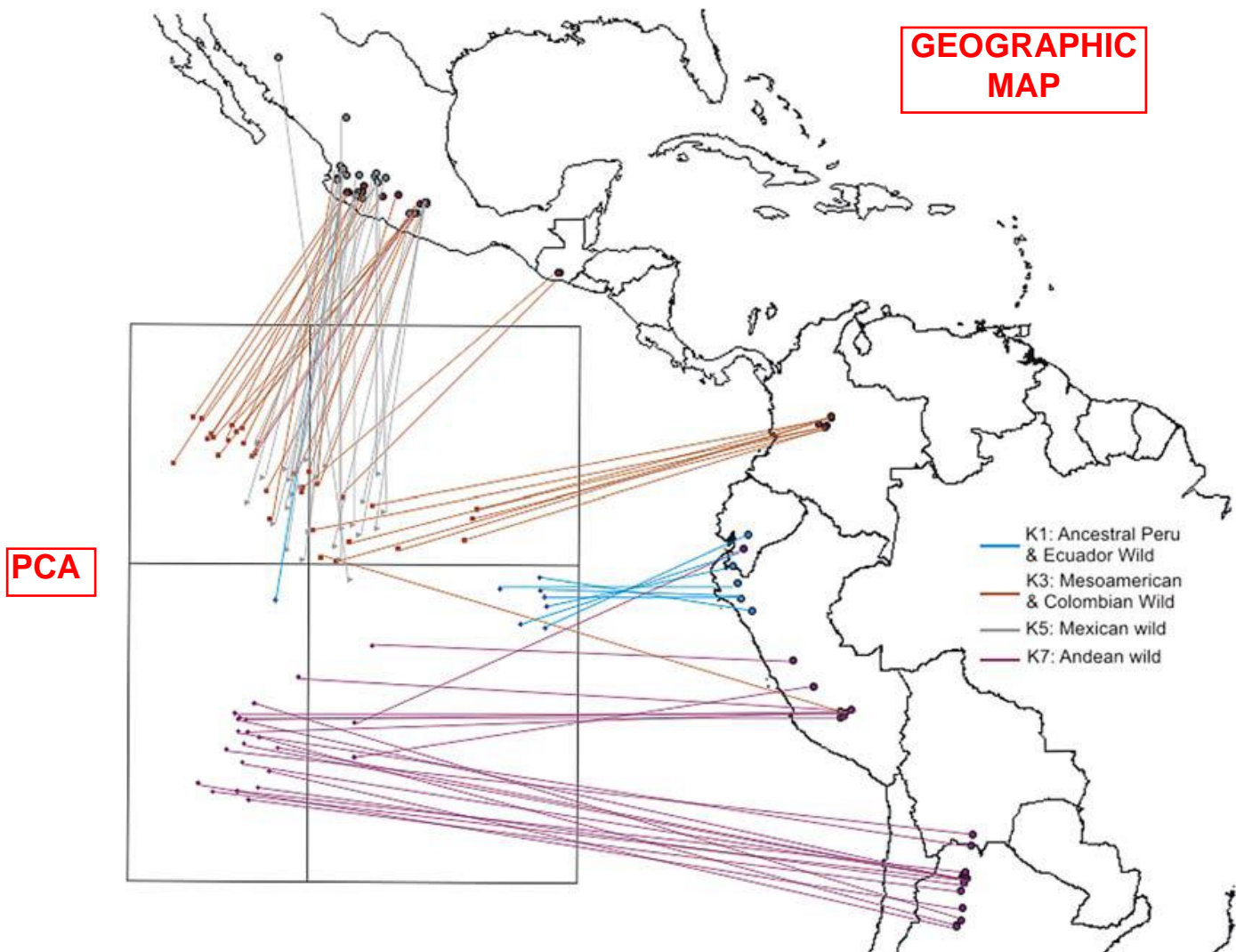
**PC1:**

\*\*Accounts for large amount of variation

\*\*Andean vs. Middle American

# PCA ANALYSIS MAPPED TO GEOGRAPHIC DISTRIBUTION

**Fig. 4** Geographical and genetic distributions of wild common bean accessions. The lower left plot is the result of a principal coordinate analysis involving wild accessions only (for which precise coordinates are available). The lines link positions of accessions in the PCA graph and their geographic origin on the map. The colors indicate population membership identified using STRUCTURE (same colors as in Fig. 1)



# $F_{ST}$ A Statistic That Measures Population Differentiation

## Fixation Index ( $F_{ST}$ )

### Subpopulation variation

- Important to know the degree to which specific subpopulations are different
- Subpopulation can evolve from other populations
  - Genetic drift
  - Selection
  - Mutation
  - Migration
  - Recombination

### When all subpopulations are considered together

- Effects working on each subpopulation are combined
- Sewell Wright developed a set of statistics that
  - Consider the variation within a subpopulation
  - Relative to the entire population
    - F-statistics

### Most widely used parameter

- The statistic  $F_{ST}$ 
  - A simple ratio of the following format

$$F_{ST} = \frac{X_T - X_S}{X_T}$$


$X_T$ : total population variation  
 $X_S$ : subpopulation variation

- Compares the ratio of a value for a subsection of population to the value for the whole population

Different values can be used



### Specific measures considered for this formula

- Classic definition of Wright based on 
  - Frequency of heterozygotes in total population relative to subpopulations
    - Greater the reduction of heterozygotes in a subpopulation
      - Larger the value of  $F_{ST}$
- Basing the values on heterozygotes, the above formula becomes:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

### How to interpret

- $H_T$ : this is proportion of the heterozygotes in full population
- $H_S$ : this is average proportion of heterozygotes in subpopulations
- If  $H_T$  is nearly equal to  $H_S$ , then subpopulations are similar
- If  $H_S$  is less in subpopulations, the subpopulations are different

### Sewell Wright example (Genetics (1943))

- Evaluated flower color of *Linanthus parryae* in S. California
- 30 zones, 100 flowers in each zone
- Collected frequency of heterozygotes over all zones and compared it to the entire region

$$H_S = 0.1424; H_T = 0.2371$$

$$F_{ST} = (0.2371 - 0.1424)/0.2371 = 0.3089$$

## Other variables

- Average number of pairwise differences

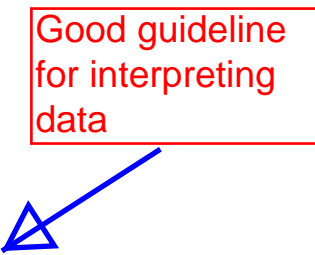
### $F_{ST}$ has a range

- 0 (no divergence) to
- 1 (complete divergence)

### $F_{ST}$ is often

- Well below 1
- How can  $F_{ST}$  be interpreted?
  - Wright suggestions

Good guideline  
for interpreting  
data



$F_{ST} = 0.00 - 0.05$	= little genetic divergence
$F_{ST} = 0.05 - 0.15$	= moderate degree of genetic divergence
$F_{ST} = 0.15 - 0.25$	= great degree of genetic divergence
$F_{ST} > 0.25$	= very great degree of genetic divergence

### These are suggestions

- The should be balanced against
  - What the researcher actually knows about a population

## $F_{ST}$ Example (Simple Case)

**Paper:** Wright S. 1943. An analysis of local variability of flower color in *Linanthus parryae*. Genetics 28:139.

**Species:** *Linanthus parryae*

**Location:** 80 mile long, 10.5 mile wide stretch of Piedmont north of San Gabriel/San Bernadino Mountains, California, USA

1	2	3	4	5	6
Zone	Subpopulation	Blue allele frequency (p)	Heterozygosity ( $2 \cdot p \cdot q$ = $2 \cdot p \cdot (1-p)$ )	Average zone blue allele frequency	Heterozygosity per zone
I	1	0.573	0.489	0.551	0.495
	2	0.717	0.406		
	3	0.657	0.451		
	4	0.504	0.500		
	5	0.302	0.422		
II	1	0.032	0.062	0.078	0.144
	2	0.007	0.014		
	3	0.005	0.010		
	4	0.339	0.448		
	5	0.008	0.016		
III	1	0.000	0.000	0.002	0.004
	2	0.000	0.000		
	3	0.009	0.018		
	4	0.000	0.000		
	5	0.000	0.000		
IV	1	0.000	0.000	0.017	0.033
	2	0.000	0.000		
	3	0.005	0.010		
	4	0.010	0.020		
	5	0.068	0.127		
V	1	0.002	0.004	0.026	0.051
	2	0.004	0.008		
	3	0.000	0.000		
	4	0.000	0.000		
	5	0.126	0.220		
VI	1	0.106	0.190	0.151	0.256
	2	0.224	0.348		
	3	0.014	0.028		
	4	0.000	0.000		
	5	0.573	0.489		

Avg  $p$  across all populations

Total heterozygosity

	Average Col 3	0.137			
	HT (2pg)	0.237			
	HS (ave col 4)		0.142		
	HR (ave col 6)				0.164

### $F_{ST}$

- Can be calculated at any level of population subdivision
  - Among populations
  - Among regions
- Each calculation uses different population measure
  - Wright's original data set used heterozygosity as a measure
  - Other parameters currently used today
- Formula has been revised to deal with different types of data sets

Calculations for *Linanthus parryae* data

General formula

$$F_{ST} = \frac{X_T - X_S}{X_T}$$

Heterozygosity formula for among subpopulations

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$H_T$  = heterozygosity of the whole population based on the average allele frequency across all subpopulations (average of all H values in column)

$$H_T = 2 * 0.137 * (1 - 0.137) = 0.237$$

$H_S$  = average heterozygosity of all subpopulations based on heterozygosity values calculated using the allele frequencies for each subpopulation

$$H_S = (\text{average of column 4}) = 0.142$$

**Subpopulation  $F_{ST} = (0.237 - 0.142) / 0.237 = 0.399$**

$H_R$  = average heterozygosity of all regions (zones) based on heterozygosity values calculated using the allele frequencies for each region (zone)

$$H_R = (\text{average of column 6}) = 0.164$$

**Region (Zone)  $F_{ST} = (0.237 - 0.164) / 0.237 = 0.309$**

# Classical Selection, Balancing Selection, and Neutral Mutations

## What Was the Classical Selection Perspective of the Fate of Mutations?

- All mutations are EITHER **beneficial or deleterious**
  - **Beneficial** mutations are **selected for** and **maintained** in the population
    - **Positive selection**
  - Mutation rapidly increases to a high frequency in the population
    - **Generate a new adaptive phenotype**
- **Deleterious** mutations are **selected against** and **eliminated** from the population
  - **Negative selection**

## Balancing Selection Perspective of the Fate of Mutations

- In general, agrees with Selection Perspective
  - But it was noted that some deleterious variation is maintained in the population
  - **How is the deleterious variation maintained?**
    - Selection for heterozygotes, one method
      - **Heterozygous** have a **fitness advantage** and undergo **natural selection**
        - Classic example: **human sickle cell anemia**
          - Homozygous normal  $\beta$ -globin allele
            - Proper oxygen carrying capacity, but susceptible to malaria
          - Homozygous mutant  $\beta$ -globin allele
            - Resistant to malaria but die young because of poor oxygen carrying capacity
          - Heterozygous  $\beta$ -globin allele individuals
            - Proper oxygen carrying capacity, and resistant to malaria

## **Surprise of 1960s and Onward**

- **Diversity** in populations ***much greater*** than predicted by either the classical or balancing selection theories
  - Based on protein electrophoresis and eventually sequence data

### **Neutral Mutations Recognized as a New Class of Mutations**

- **Allelic variation neither beneficial or deleterious**
  - **These alleles not provide any fitness difference among individuals in a population**
  - **Fate of an allele in a population is entirely random process**
    - **Allele can be maintained or eliminated**
      - **Controversial concept**

# Neutral Theory of Molecular Evolution

Kimura – Nature (1968) 217:624-626

King and Jukes – Science (1969) 164:788-798 (Non-Darwinian Evolution)

## Neutral Theory of Molecular Evolution

- Describes the source of variation in natural populations
- The majority of genetic differences between two populations are neutral
  - They have not effect on survival
- Predicts two factors are working
  - Mutation
    - Generates new variation
  - Random genetic drift
    - Fixes variation
    - Stochastic processes that lead to changes in gene frequency
    - Allele can be lost or maintained
- Also called the **Mutation-Drift Model**

## Definition of the Neutral or Mutation-Drift Model

- Population genetic variation results from the appearance of neutral mutations that are fixed (but usually lost) by genetic drift
- Why genetic drift?
  - A process in smaller populations
  - Random process will drive new alleles to fixation quicker

Assumption of random mating violated!!

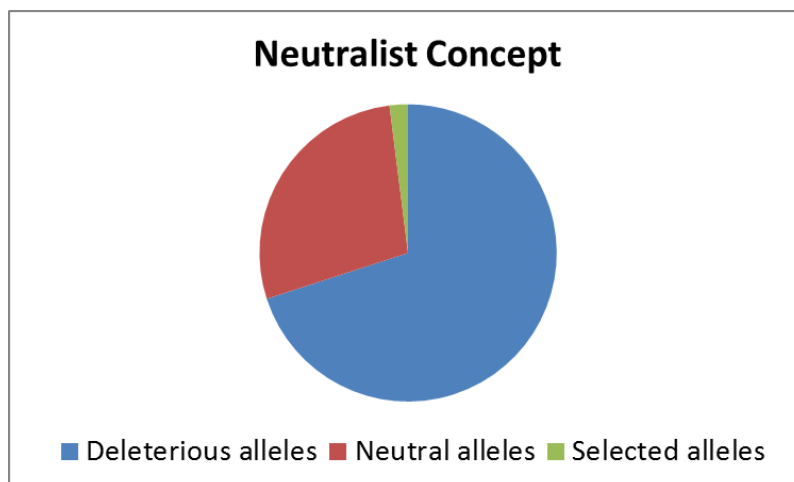
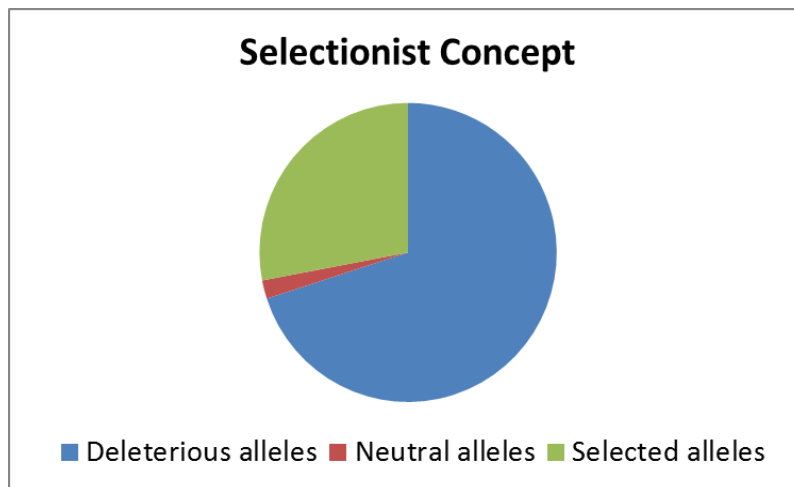
## Relationship to Selection Theory

- Most variation is the neutral, therefore **most population differences are not the result of adaptive selection**

**\*\*This was the controversy!!**  
**\*\*It contradicted long standing beliefs!!**



## Graphical Representation of the Alleles in a Population



### The Controversy and Debate

- Was there no role for selection?
  - What is the role (if any) of Darwinian selection for adaptation
  - Although argued, not really a debatable issue
- Some genes do under go selection
- What is the relative distribution of neutral and selected genes

## Why is the Neutral Theory so important?

- Ideal for mapping population structure and tracing ancestry
- Provide a null hypothesis for testing for selection

## Life Span of an Allele

- Drift will eventually lead to the fixation of one allele in a population
- How long before one allele is fixed (and the other lost)
  - $4N_e$  generations
    - $N_e$  is effective population size
      - The number of individuals, in an idealized ancient population that adheres to the assumptions of Hardy-Weinberg equilibrium, that would evolve to have the same level of diversity as the population that is being observed.
    - Usually less than the census size of population under study

## What can affect the life span of an allele?

### Balancing Selection

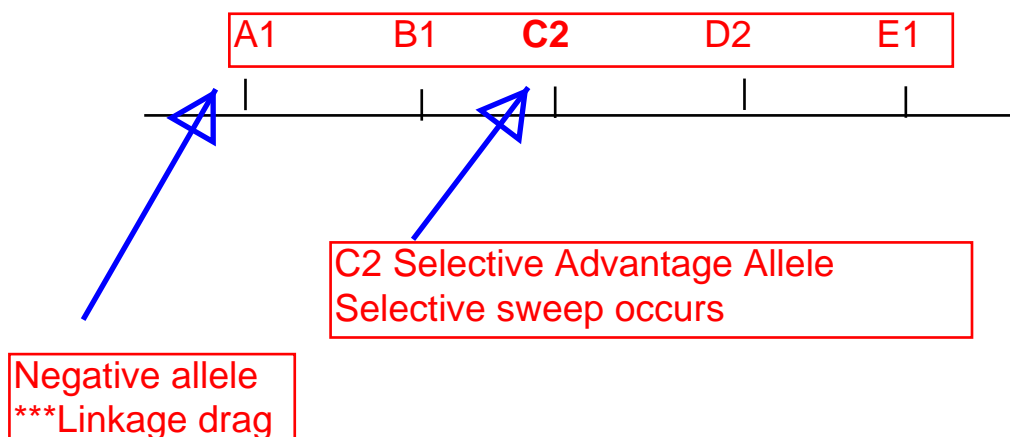
- The maintenance of multiple alleles within a population
- A mechanism
  - Heterozygote advantage
    - Heterozygote has a greater fitness
      - Sickle cell anemia
    - At least two alleles are maintained

### Selective Sweep

- A specific gene is the target of selection
- It becomes fixed in the population
- Linked genes become monomorphic (lose variation) during the selection process
  - A direct effect of effect of recombination
  - Hitchhiking
    - Neighboring genes are said to have “hitchhiked”

### Background Selection

- Result of eliminating a deleterious allele
  - Entire chromosome carrying deleterious allele is lost
  - Reduces diversity of all genes on that chromosome



# Pi and Theta

## Statistics That Measures Population Diversity

### Nucleotide Diversity Estimates for a Population of DNA Sequences

Problem 2.4. Principles of Population Genetics; Hartl and Clark; 1997; 3<sup>rd</sup> Edition  
Gene=*Rh3*; Species=*Drosophila simulans*; gene size=500 nucleotides; sample size=5 lines

	Polymorphic nucleotides															
Sample	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3
f	T	C	T	A	C	C	T	C	C	T	C	G	G	T	T	A
g	T	C	C	T	A	C	C	T	C	C	T	G	G	T	T	T
h	C	T	T	C	C	C	C	T	C	T	T	T	G	C	T	A
i	C	T	T	C	C	C	C	T	T	C	T	G	A	C	T	T
j	C	T	T	C	C	T	C	T	T	T	T	G	G	C	C	A

16 polymorphic sites of 500 sites

Pairwise differences

2X3	X	X								X	X				X	X
1X4			X		X	X	X	X			X	X	X		X	X
(1X4) + (1X3)				X												X

6 of 2x3

9 of 1x4

1 of 1x4 and 1x3

### Population Genetic Parameters

**S = frequency of polymorphic loci**

$$S = \# \text{ polymorphic nucleotides} / \text{total} \# \text{ of nucleotides}$$

**Example:**

$$S = 16 \text{ polymorphic nucleotides} / 500 \text{ nucleotide} \\ = 0.032$$

**$\pi$  = observed average pairwise nucleotide differences in a sample (nucleotide diversity)**

$$\pi = \frac{\text{sum pairwise differences}}{(\# \text{ pairs})(\text{sequence size})}$$

$$\# \text{ pairs} = \frac{n(n-1)}{2}$$

**Example:**  $n=5$  Number of samples in population

$$\# \text{ pairs} = \frac{5(5-1)}{2}$$

$$= 10$$

$$\pi = \frac{(6 \times 6) + (9 \times 4) + (1 \times 7)}{(10)(500)}$$

$$= 0.016$$

**$\theta$  = expectation of nucleotide polymorphism in a population when only mutation-drift are occurring (no selection)**

$$\theta = \frac{S}{a_1} \quad \text{where}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \quad \text{and}$$

n= sample size

**Example:**  $S = 0.032$

$$a_1 = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.083$$

$$\theta = \frac{0.032}{2.083}$$

$$= 0.015$$

## Tajima's D

- a test for neutrality
- a comparison of  $\pi$  and  $\theta$
- under neutrality, the mean of this value is 0
- significant deviation from 0 suggests the gene is undergoing selection

### Formula

$$D = \frac{\pi - \theta}{\sqrt{c_1 S + c_2 S(S - \frac{1}{k})}} \quad \text{where}$$

$$c_1 = \frac{b_1}{a_1} - \frac{1}{a_1^2} \quad \text{and}$$

$$c_2 = \left(\frac{1}{a_1^2 + a_2}\right) \left(b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}\right)$$

**Example:**  $c_1 = \frac{0.5}{2.083} - \frac{1}{(2.083)^2}$

$$= 0.0096$$

$$c_2 = \left( \frac{1}{(2.083)^2 + 1.42} \right) \left( 0.367 - \frac{5+2}{(2.083)(5)} + \frac{1.42}{(2.083)^2} \right)$$

$$= 0.0038$$

$$D = \frac{0.016 - 0.015}{\sqrt{(0.0096)(0.032) + (0.0038)(0.032)\left(0.032 - \frac{1}{500}\right)}}$$

$$= 0.0567$$

- ***D* not significant different than 0**
  - ***Physical region or gene evolving via the neutral theory***



What does a significantly **positive** or **negative** Tajima's  $D$  value indicate? The following quotes are directly from *A Primer of Population Genetics* (Hartl; 3<sup>rd</sup> edition):

### **Significantly positive $D$ value**

“The frequencies of polymorphic variants are too nearly equal. This pattern increases the proportion of pairwise differences over its neutral expectation, hence  $\pi - S/a_i (= \vartheta)$  is positive. The finding typically suggests either some type of balancing selection, in which heterozygous genotypes are favored, or some type of diversifying selection, in which genotypes carrying the less common allele are favored.”

- Multiple low frequencies alleles are maintained in population
  - Balancing selection at work in the population

### **Significantly negative $D$ value**

“The frequencies of the polymorphic variants are too unequal, with an excess of the most common type and a deficiency of the less common types. This pattern results in a decrease in the proportion of pairwise differences, so  $\pi - S/a_i (= \theta)$  is negative. Typical reasons for excessively unequal frequencies are:

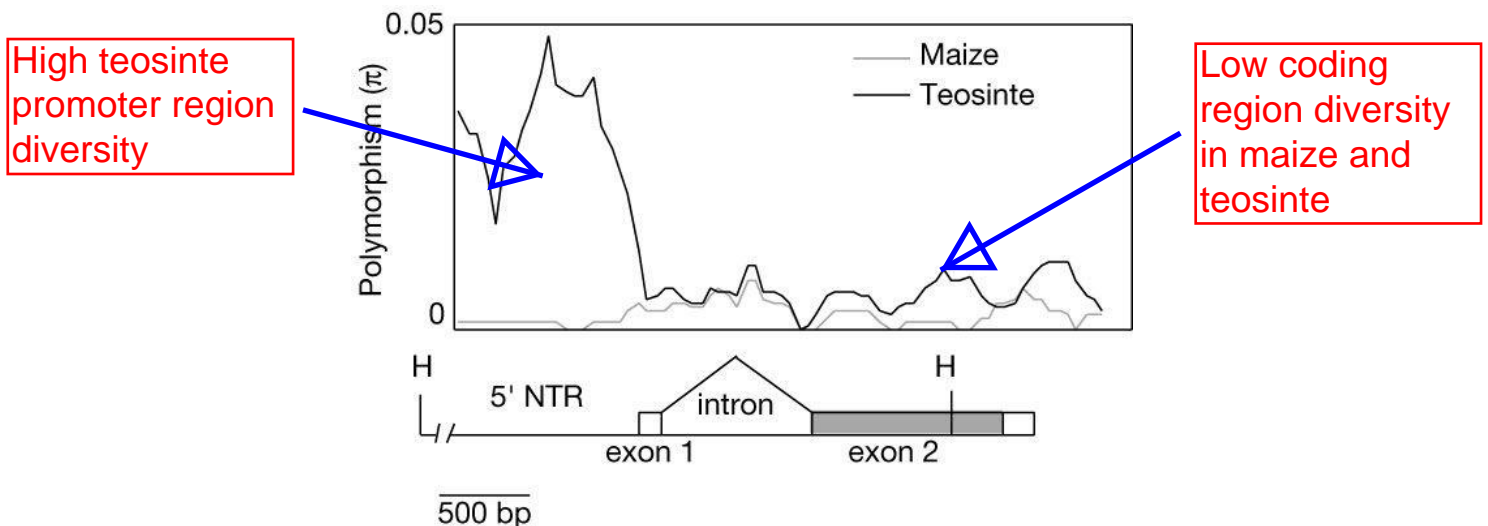
- Selection against genotypes carrying the less frequent alleles
  - Result of a recent bottleneck
  - A few distinct subpopulations appear
  - The bottleneck eliminates less frequent alleles, and insufficient time since the bottleneck to restore the equilibrium between mutation and random drift
  - Selective sweep regions can be observed

# The Effects of Domestication on Diversity

Application of population genetics statistics to identify important genetic factors

## Maize *tb1* gene studied

- Domestication gene of maize
  - Repressor element expressed higher in maize than teosinte
  - Suppresses branching in maize
- Polymorphism among 13 maize and 9 teosinte compared
  - Coding region
    - Nucleotide diversity ( $\pi$ ) was **low** for **both maize and teosinte**
  - 5' non-transcribed region
    - Nucleotide diversity ( $\pi$ ) was **much lower** for **maize than teosinte**
  - Selection acted on the maize 5' non-transcribed region



Wang et al. (1999) The limits of selection during maize domestication. Nature 398:236

Wang et al. (2001) The limits of selection during maize domestication. Nature 410:718.

## Follow up research on *tb1* 5' region

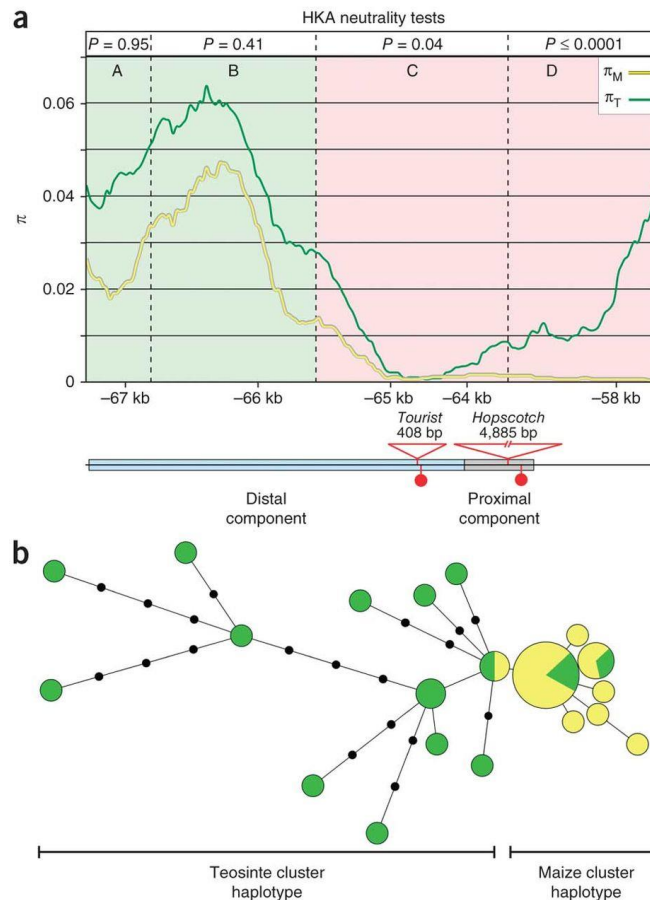
- Low diversity extends to 58.6 kb upstream of gene
- A selection sweep is observed in the upstream region of *tb1* of maize but not teosinte

Locus	Length, bp	Maize			Teosinte		
		n	$\theta \times 10^3$	$\pi \times 10^3$	n	$\theta \times 10^3$	$\pi \times 10^3$
162.9-kb	467	18	10.7	12.3	5	12.1	11.9
93.4-kb	485	14	27.1	20.8	8	38.6	37.5
58.6-kb	520	23	0.5	0.2	—	—	—
45.8-kb	1,003	24	1.1	0.3	9	31.1	32.9
35.6-kb	1,024	24	3.1	1.7	—	—	—
7.1-kb	842	24	6.7	4	8	17.6	12.7
2.5-kb	534	24	3.5	2.8	—	—	—
1.7-kb	935	24	0.6	0.3	8	34.1	34.9
0.4-kb	761	32	3.4	1.4	7	4.6	3.6
5' cDNA	839	32	1.8	1	7	6.8	5.2

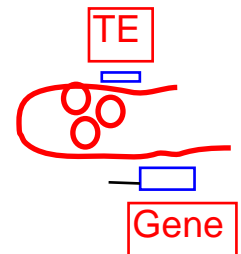
Clark et al. (2004) Proceedings of the National Academy of Science 101: 700

## Diversity extends to 65 Kb upstream of *tb1*

- Two transposable element are at the border of the low diversity
- *Tourist* element is older
- *Hopscotch* element is new and nearly completely fixed in all maize lines
  - *Is the transposable element the controlling element controlling the increased expression of the *tb1* gene product in maize?*



Functional element can be far away from gene!!!



(a) Nucleotide diversity across the *tb1* upstream control region. Base-pair positions are relative to AGPv2 position 265,745,977 of the maize reference genome sequence.  $P$  values correspond to HKA neutrality tests for regions A–D, as defined by the dotted lines. Green shading signifies evidence of neutrality, and pink shading signifies regions of non-neutral evolution. Nucleotide diversity ( $\pi$ ) for maize (yellow line) and teosinte (green line) were calculated using a 500-bp sliding window with a 25-bp step. The distal and proximal components of the control region with four fixed sequence differences between the most common maize haplotype and teosinte haplotype are shown below. (b) A minimum spanning tree for the control region with 16 diverse maize and 17 diverse teosinte sequences. Size of the circles for each haplotype group (yellow, maize; green, teosinte) is proportional to the number of individuals within that haplotype.

Studer et al. (2011) Nature Genetics 43:1160

## Whole Genome Application of Population Genetics Statistics

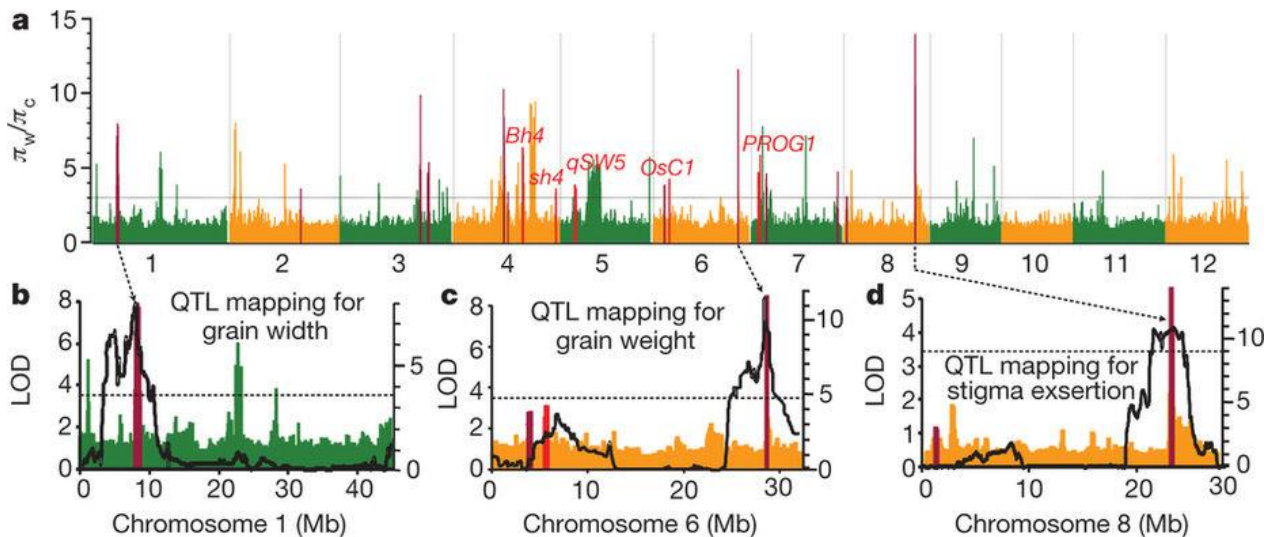
- The  $\pi$  statistics was estimated across the entire genome of rice
  - Statistic calculated for non-overlapping 100kb regions across all chromosomes
- Diversity ( $\pi$ ) between wild (*Oryza rufipogon*) and cultivated (*Oryza sativa*) were compared
  - The following  $\pi$  ratio was used

$$\pi_{\text{wild}}/\pi_{\text{cultivated}} = \pi_{O. \text{ rufipogon}} / \pi_{O. \text{ sativa}}$$

- Regions with high ratios are considered region that under went selection during domestication
  - Why???

Concept of importance!!!

- For domesticated lines, diversity was reduced in region with domestication genes whereas diversity was maintained in wild lines



**a**, Whole-genome screening of domestication sweeps in the full population of *O. rufipogon* and *O. sativa*. The values of  $\pi_w/\pi_c$  are plotted against the position on each chromosome. The horizontal dashed line indicates the genome-wide threshold of selection signals ( $\pi_w/\pi_c > 3$ ). **b–d**, A large-scale high-resolution mapping for fifteen domestication-related traits was performed in an *O. rufipogon*  $\times$  *O. sativa* population. The domestication sweeps overlapped with characterized domestication-related QTLs are shown in dark red, and the loci with known causal genes are shown in red. Among them, three strong selective sweeps were found to be associated with grain width (**b**), grain weight (**c**) and exerted stigma (**d**), respectively. In **b–d**, the likelihood of odds (LOD) values from the composite interval mapping method are plotted against position on the rice chromosomes. Grey horizontal dashed line indicates the threshold (LOD > 3.5).



100 kb window size; no slide

## How many windows were under selection?

- Cutoff set by permutation test
  - Indica and japonica rice combined
    - 55 windows
  - Indica rice alone
    - 60 windows
  - Japonica rice along
    - 62 windows

## Were domestication genes in the selection windows?

Yes!!!

- *Bh4*: hull color
- *sh4*: seed shattering
- *qSW5*: grain width
- *OsC1*: leaf sheath colour and apiculus colour
- *PROG1*: tiller angle
  - Located in windows with high  $\pi_{\text{wild}}/\pi_{\text{cultivated}}$  ratio