Molecular Markers and Phylogenetics

Markers can

- Indicate the haplotype state of an individual

Haplotype

- The specific combination of alleles across multiple adjacent loci in an individual
    - Whole genome level
        - Sequence the best indicator
        - Develops "hapmaps"
        - Species-wide effort to describe variation in the genome of a species
            - Human Hapmap
            - Medicago Hapmap
    - Collection of markers
        - Selected marker loci distributed "equally" across the genome
            - Why "equally" in parenthesis?
                - 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

Haplotype structure example

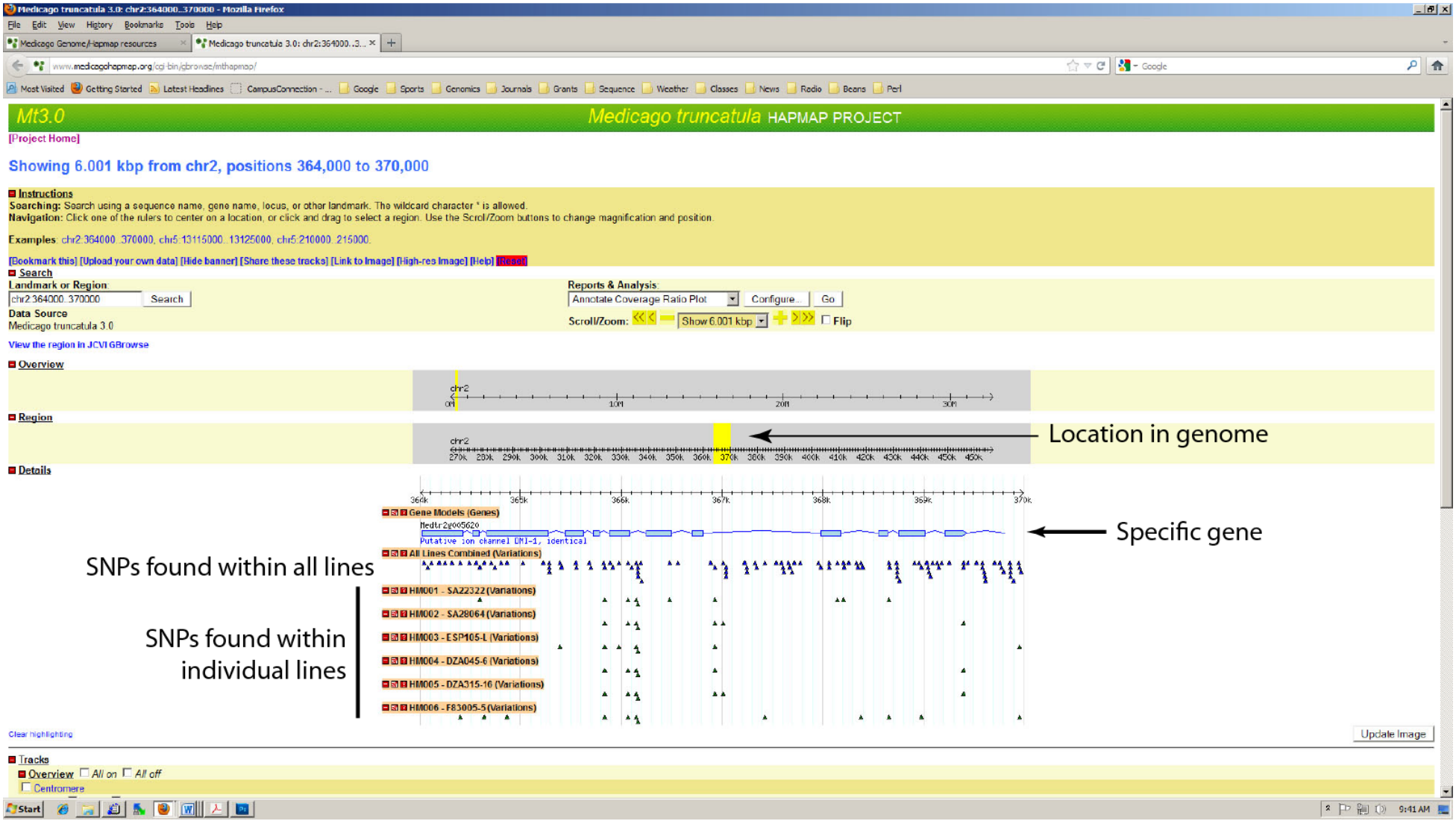
- Intron of chalcone isomerase intron 3 of common bean
- Sample of 67 individuals (landraces and cultivars)
    - 10 haplotypes observed

```
H
a      Position of variable SNPs in sample
p   1111111111111111111112222222233334445566
l 7833333344444599999999900034469126712907O3
o 6815678901234912345678901213675175661568l2
```

```
1 GCTTTTTTTTGTTGATACGAACACAGAAGTTCACTGTTCGAC
2 ..................................A..
3 T.....................................
4 .A..........T...........GGC.CTGTCA.-....
5 .A..........T...........GGCCCTGTCA.-....
6 AA..........T...........GGC.CTGTCA.-....
7 ..A----------.----------GGC........-T.G.
8 ..A----------.----------GGC.......A-T.G.
9 ..A----------.G---------GGC........-T...
```

Medicago HapMap GBrowser

www.medicagohapmap.org



Phylogenetics

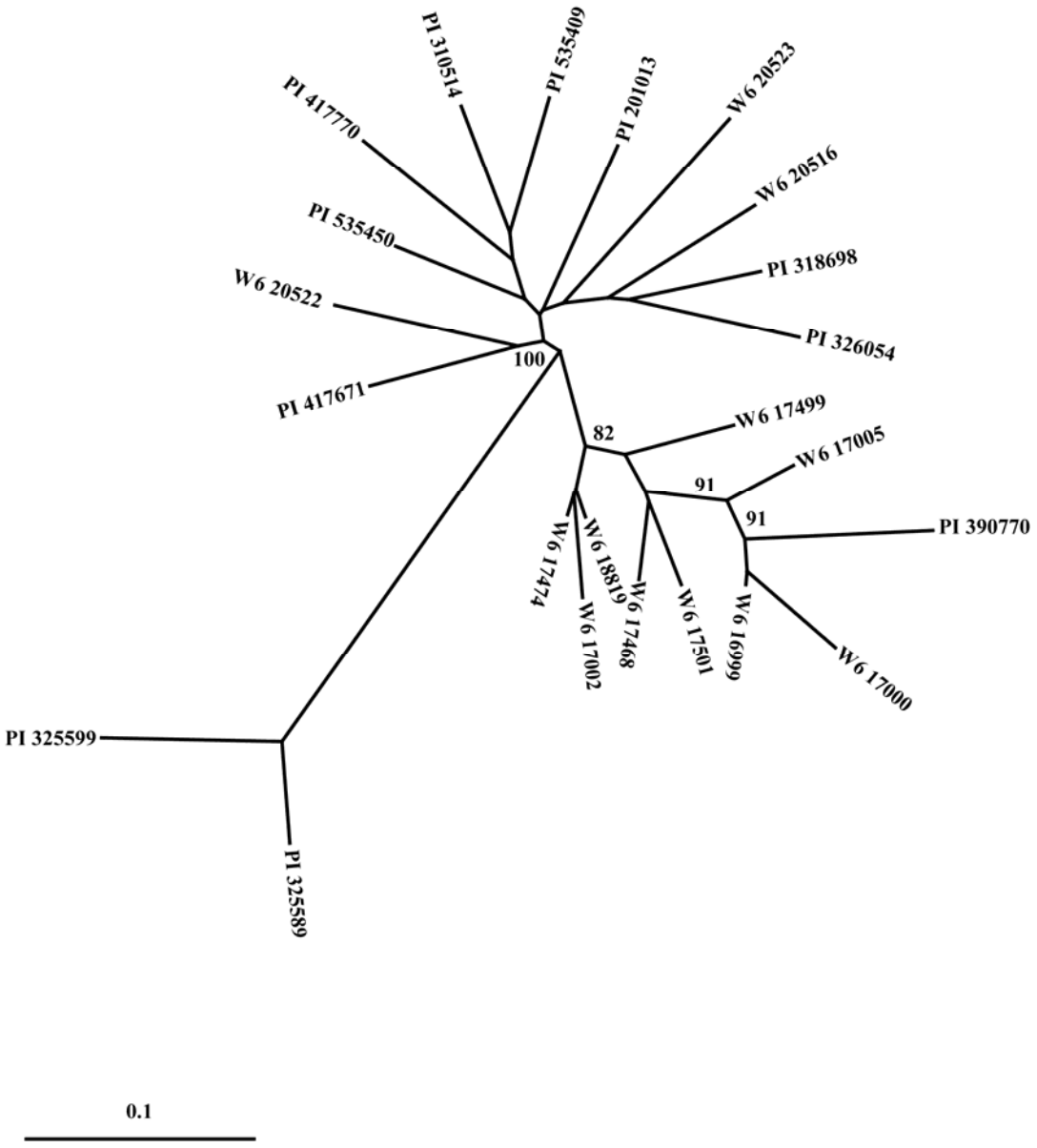
- Definition
    - The study of the evolutionary relationship between a collection of genotypes
- Can be based on
    - Phenotype
    - Molecular markers
    - Sequence data
- Creates a branching pattern that
    - Depicts the relationship of the members of the population

Two Major Approaches Used in Applied Crop Phylogenetics

Neighbor joining

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

NJ Phylogenetic Tree Example – Common Bean

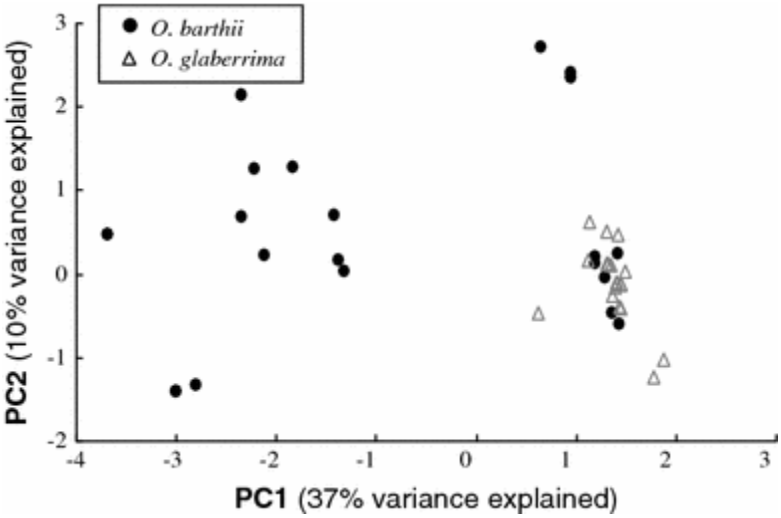


Principle component analysis

- "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 the geographical distribution of the population
        - Other PCs
            - Show the relationship among individuals within a location
            - Can show transition individuals between 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	75%	10%	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

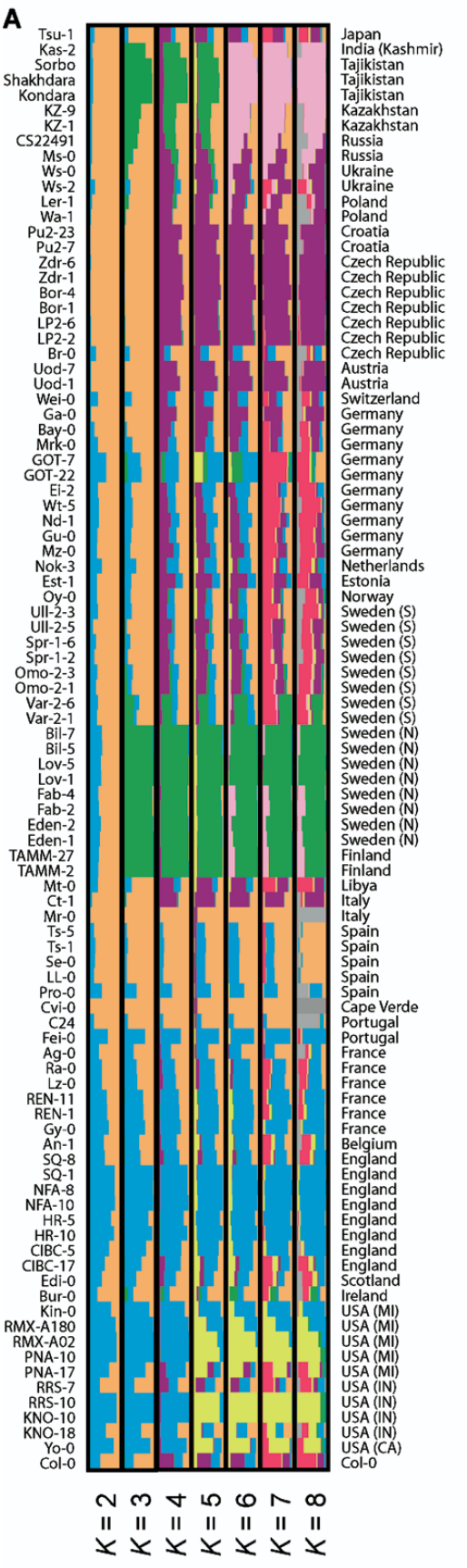
- All individuals in the full population are members of a specific k subpopulation (k_n)
- All loci within a subpopulation are in ***Hardy-Weinberg equilibrium***
- All loci within a subpopulation are in ***linkage equilibrium***
- A genotype can be defined relative to the percentage of each subpopulation 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

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

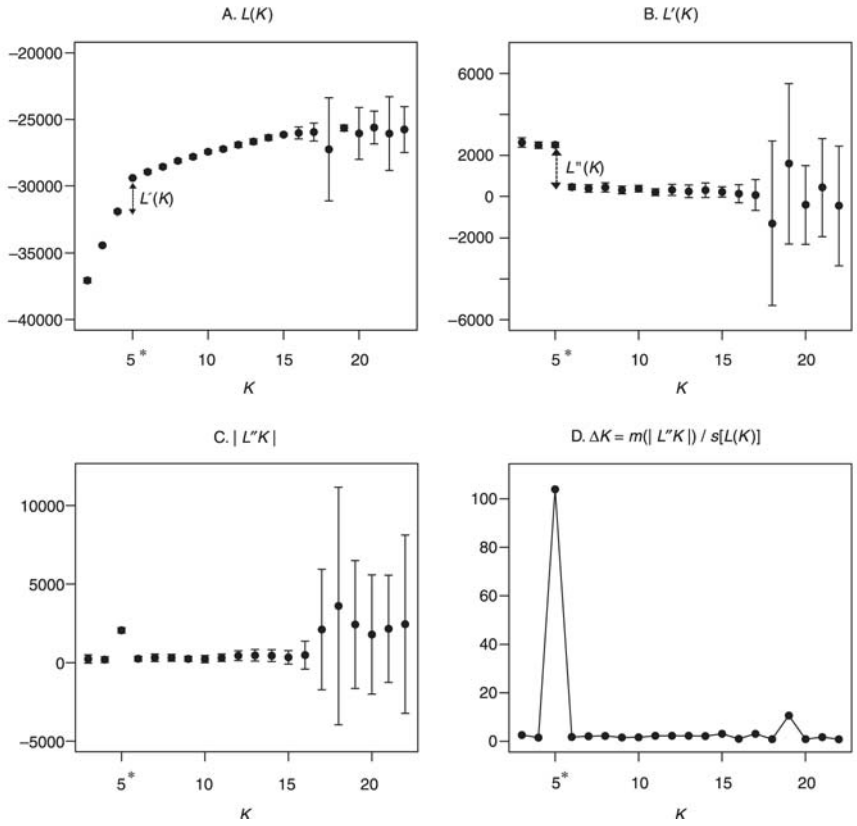
***k*=3-8**

- cluster split further along geographic borders

Determining the number of subpopulations in the sample

Evanno et al. 2005. Detecting the number of clusters of indivdiuals 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) (± 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 ± 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 ± SD) calculated according to the formula: $|L''(K)| = |L'(K+1) - L'(K)|$. (D) ΔK 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
- Follow the relationships of haplotyes
- 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 DISTRUCT

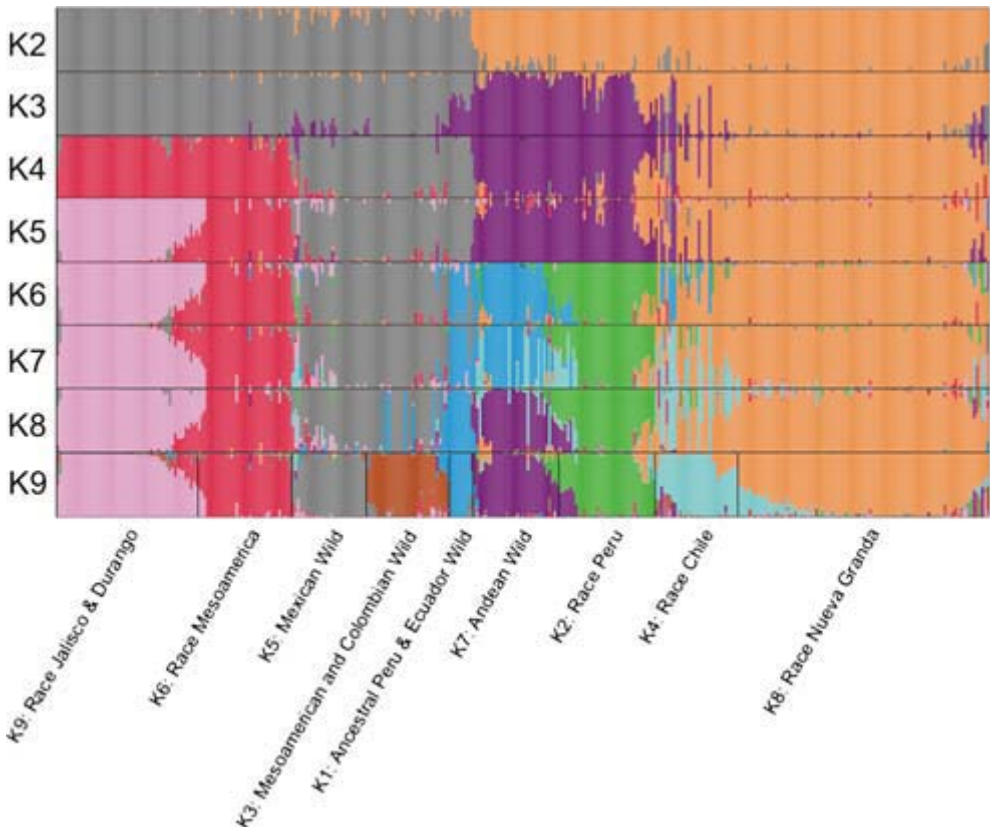


Fig. 2 Neighbor-joining tree of microsatellite diversity based on the C. S. Chord distance implemented in the Powermarker program. Each branch is color-coded according to membership into the K = 9 groups identified by STRUCTURE (same colors as in Fig. 1). Branches ending with black dots represent domesticated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

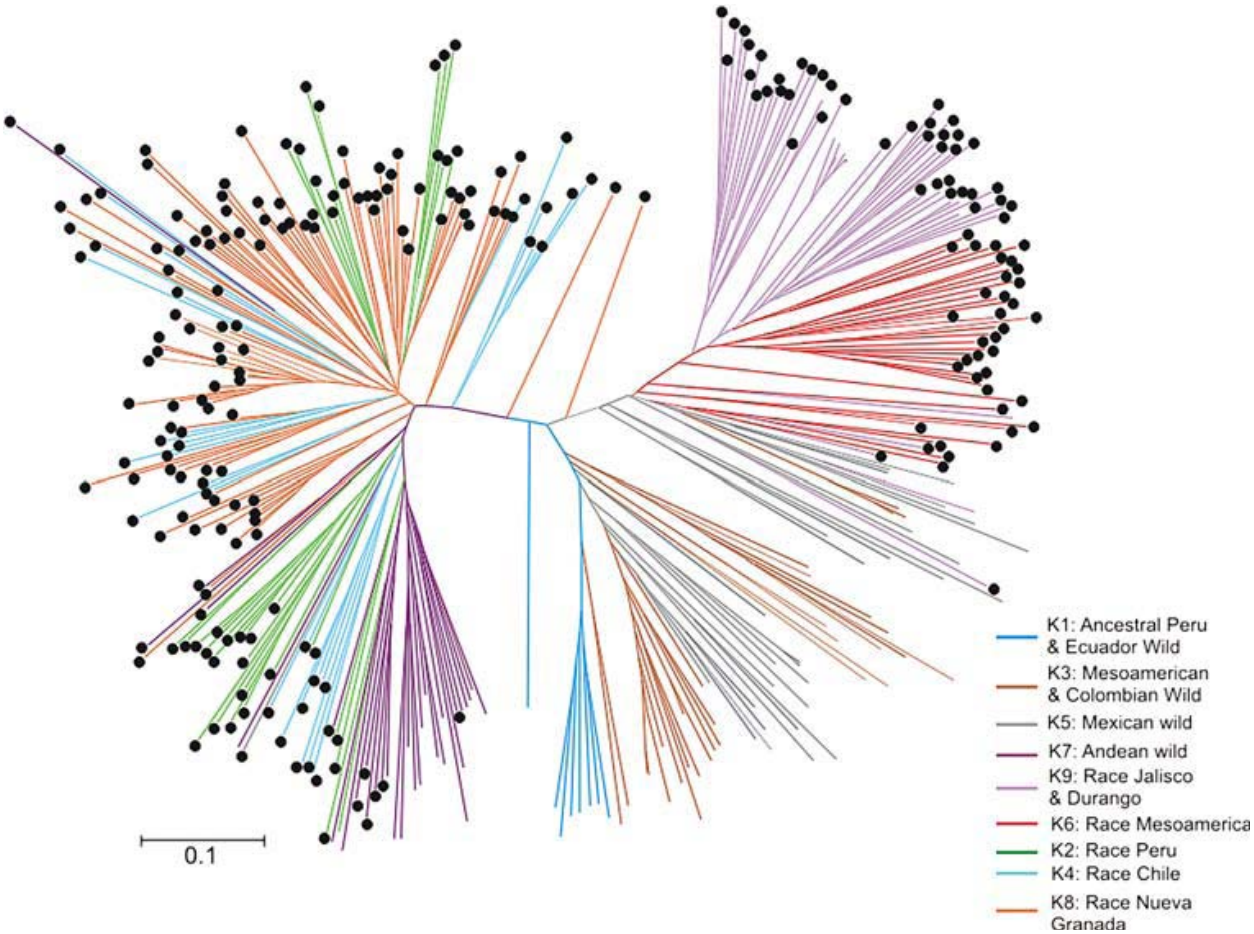


Fig. 3 Principal coordinate analysis of microsatellite diversity based on the presence absence of alleles. Colors represent populations identified at K = 9 in Fig. 1

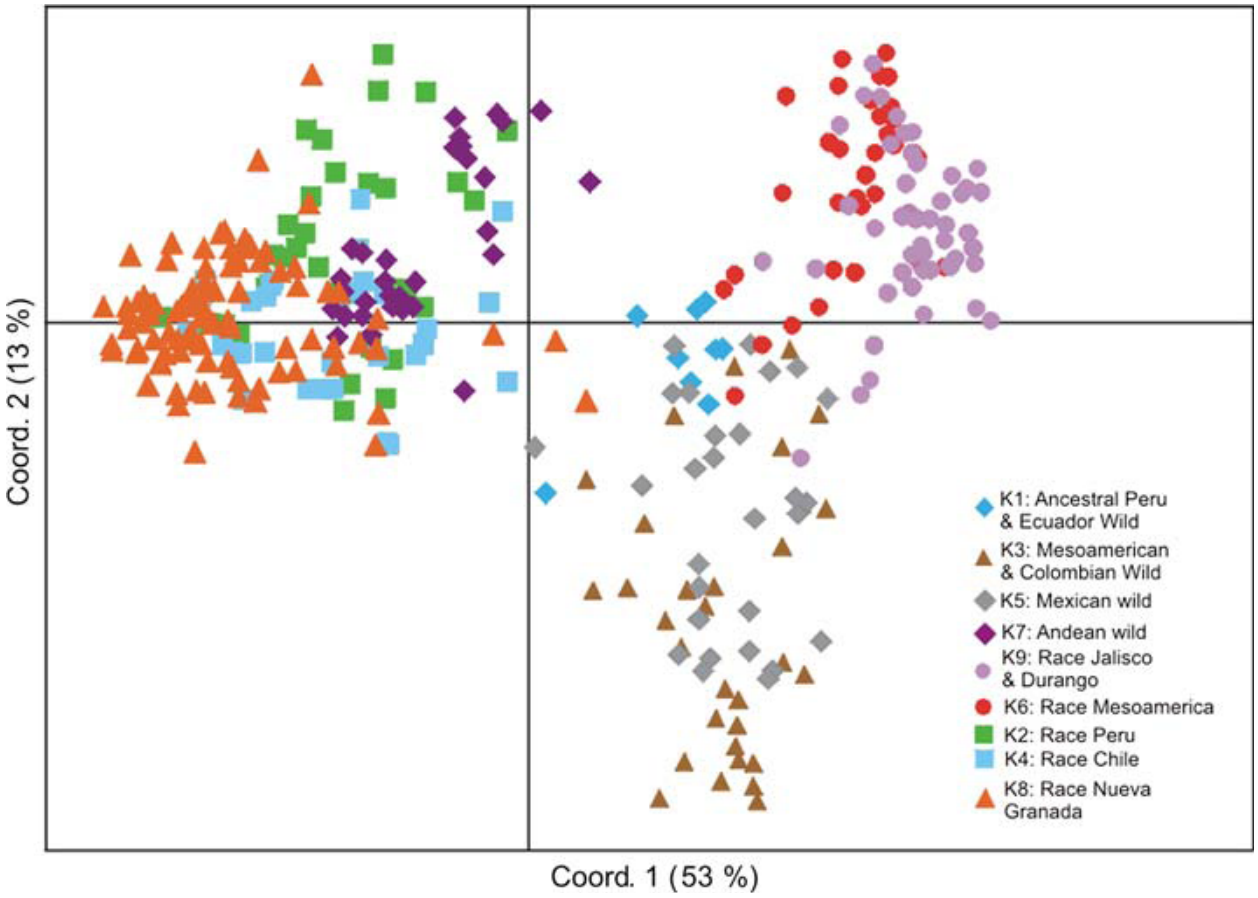
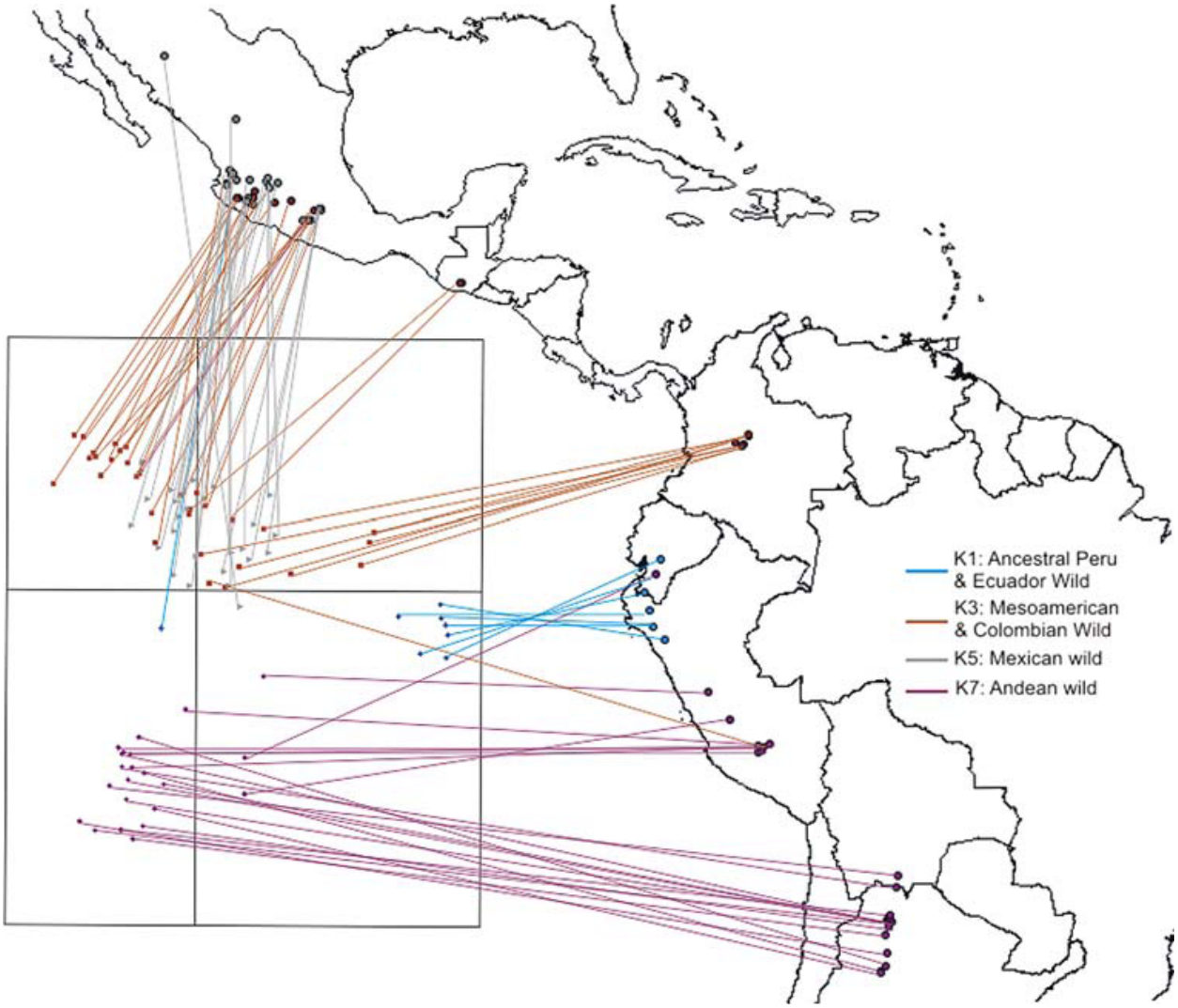


Fig. 4 Geographical and genetic distributions of wild common bean accessions. The lower left plot is the result of a principal coordinate analysis involving wild accessions only(for which precise coordinates are available). The lines link positions of accessions in the PCA graph and their geographic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



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

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

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

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. Califonia
- 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

$F_{ST} = 0.00 – 0.05$ = 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)	Heterozgosity (2*p*q = 2*p*(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		
	Average Col 3	0.137			
	HT (2pq)	0.237			
	HS (ave col 4)		0.142		
	HR (ave col 6)				0.164

$\boldsymbol{F_{ST}}$

- 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 subpopulatons based on heterozygosity values calculated using the allele frequencies for each subpopulation

H_S = (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 = (average of column 6) = 0.164

Region (Zone) F_{ST} = (0.237 – 0.164)/0.237 = 0.309

What is an informative marker?

Criteria

- Polymorphic between the parents of a mapping population
- Polymorphic among individuals within a population

Why is this important

- Maximize data generated per experiment
- While minimizing cost

How can we determine what is a useful marker

- Score parents of your mapping population, obviously
- Use a subset of a population and calculate a value that indicates the potential of a marker to detect

What markers should I use for a diversity study?

- Choose markers that provide the most information for discriminating among individuals in a population

Widely used Statistic

- **Polymorphism Information Content (PIC)**
    - Anderson et al. (1993) Genome 36:181 (as modified from)
        - Botstein et al. (1980) Am J of Human Genetics 32:314

Polymorphism Information Content

A widely used measure of the usefulness of a molecular marker

$$PIC_j = 1 - \sum_{i=1}^{n} p_i^2$$

i = the i^{th} allele of the j^{th} marker
n = the number of alleles at the j^{th} marker
p = allele frequency

Allele frequencies	Formula	PIC
Biallelic marker		
$p_1 = 0.5, p_2 = 0.5$	$1 - (0.5^2 + 0.5^2)$	0.50
$p_1 = 0.4, p_2 = 0.6$	$1 - (0.4^2 + 0.6^2)$	0.48
$p_1 = 0.3, p_2 = 0.7$	$1 - (0.3^2 + 0.7^2)$	0.42
$p_1 = 0.2, p_2 = 0.8$	$1 - (0.2^2 + 0.8^2)$	0.32
$p_1 = 0.1, p_2 = 0.9$	$1 - (0.1^2 + 0.9^2)$	0.18
Multiallelic marker		
$p_1 = 0.33, p_2 = 0.33, p_3 = 0.33$	$1 - (0.33^2 + 0.33^{2+} 0.33^2)$	0.67
$p_1 = 0.4, p_2 = 0.3, p_3 = 0.3$	$1 - (0.4^2 + 0.\ 3^{2+} 0.3^2)$	0.66
$p_1 = 0.7, p_2 = 0.2, p_3 = 0.1$	$1 - (0.7^2 + 0.\ 2^{2+} 0.1^2)$	0.46

Conclusions

- Markers with alleles of equal allelic frequencies within the population have higher PIC values
- Markers with multiple alleles generally have higher PIC values
    - But also allele frequencies dependent