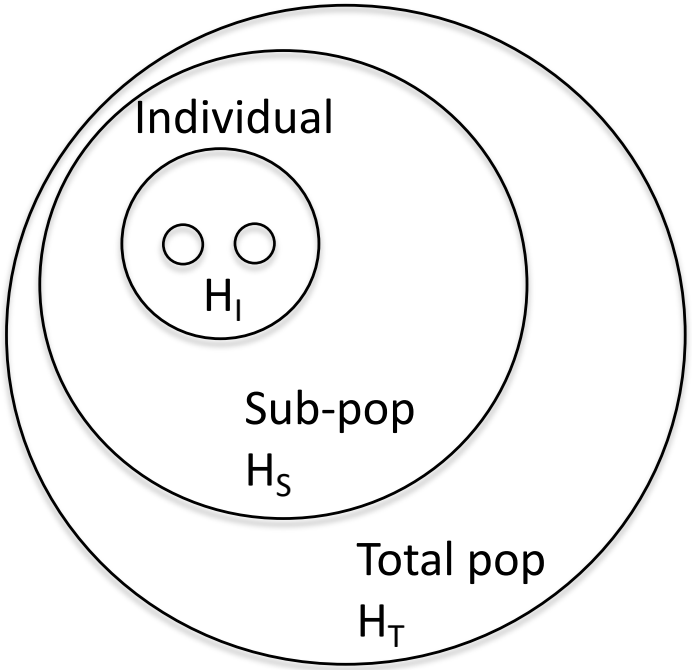
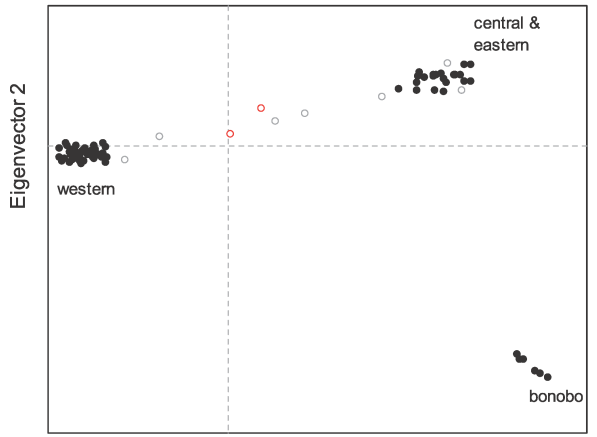
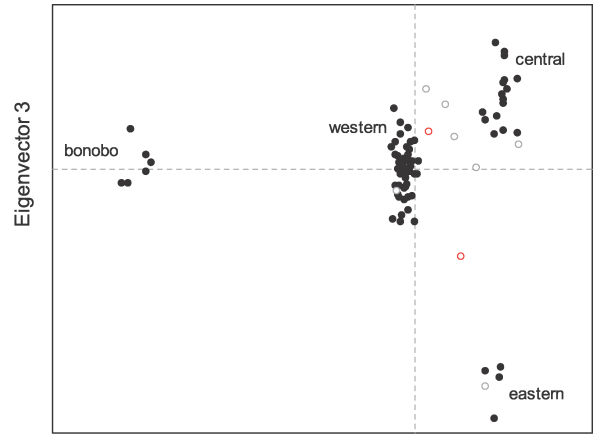
Coop, Chapter 3: Intro-3.0.1

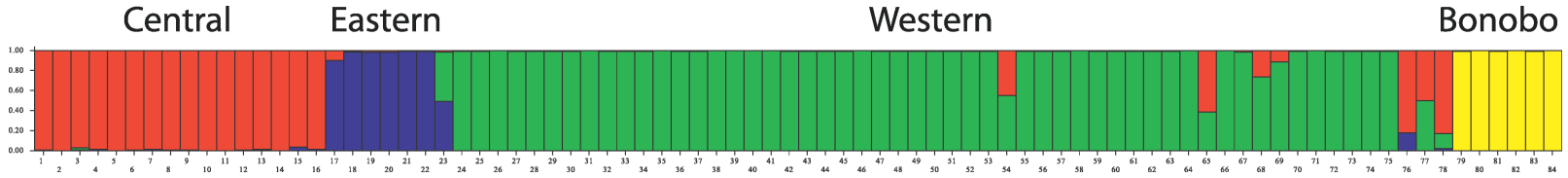
Population Structure and Correlations Among Loci

Inbreeding as a summary of population structure

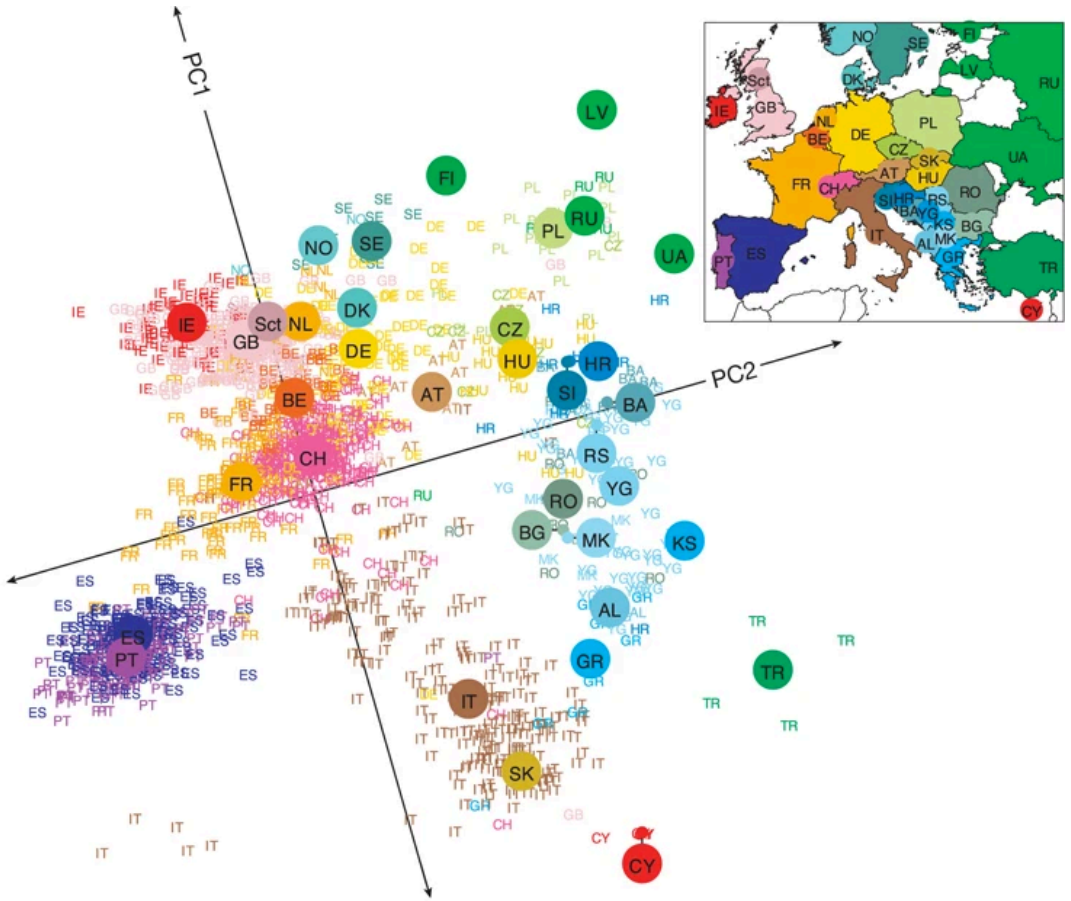








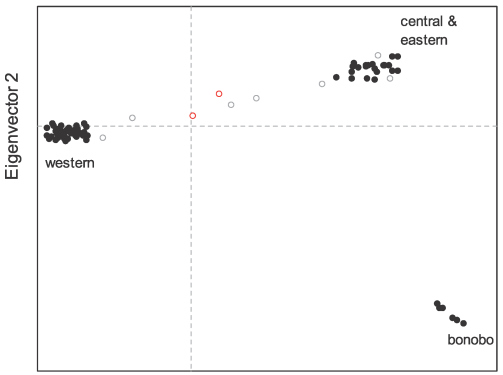
3: Introduction

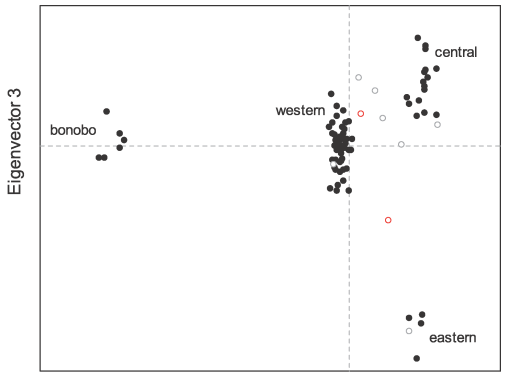


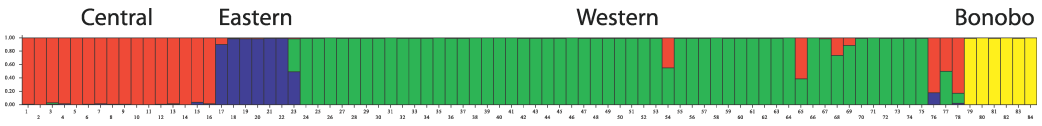
Novembre et al. 2008

- Individuals rarely mate randomly; for example individuals from Spain are more likely to mate with others from Spain than with those from the Ukraine
- This form of non-random mating is called population structure and has profound affects on the distribution of genetic variation in populations

3: Introduction

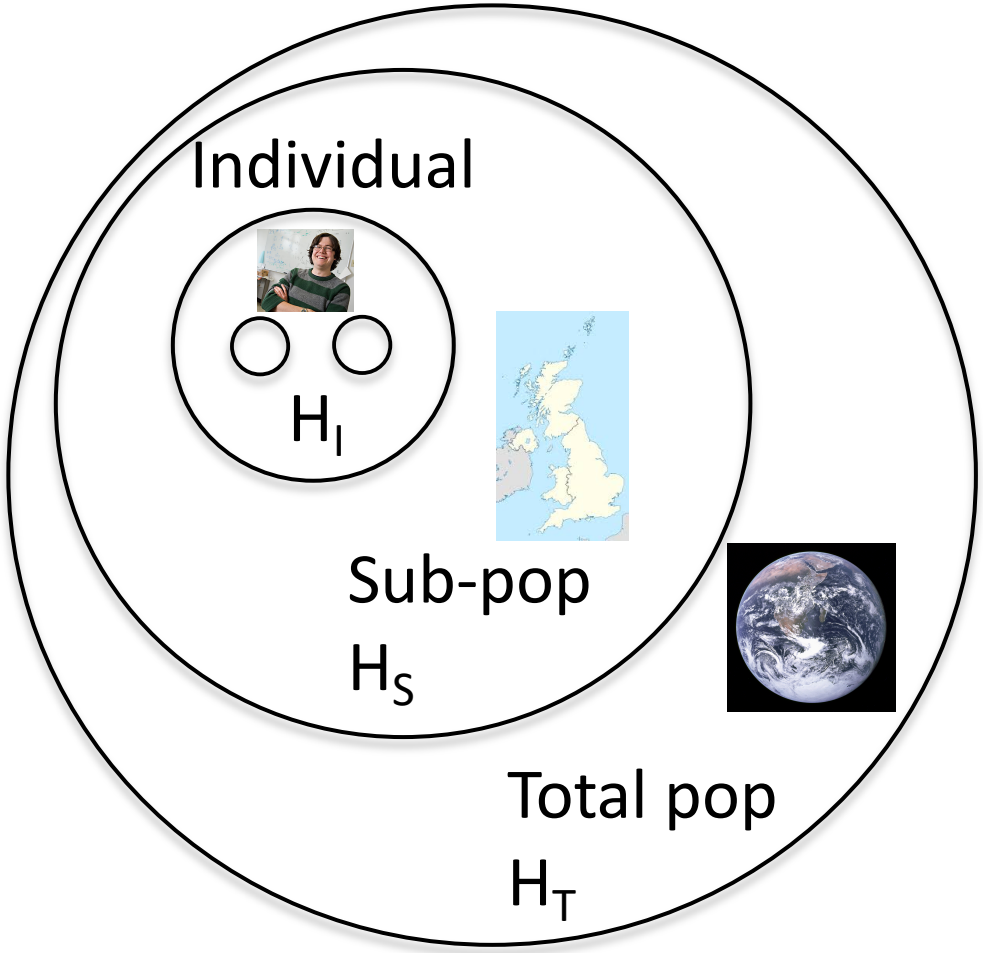






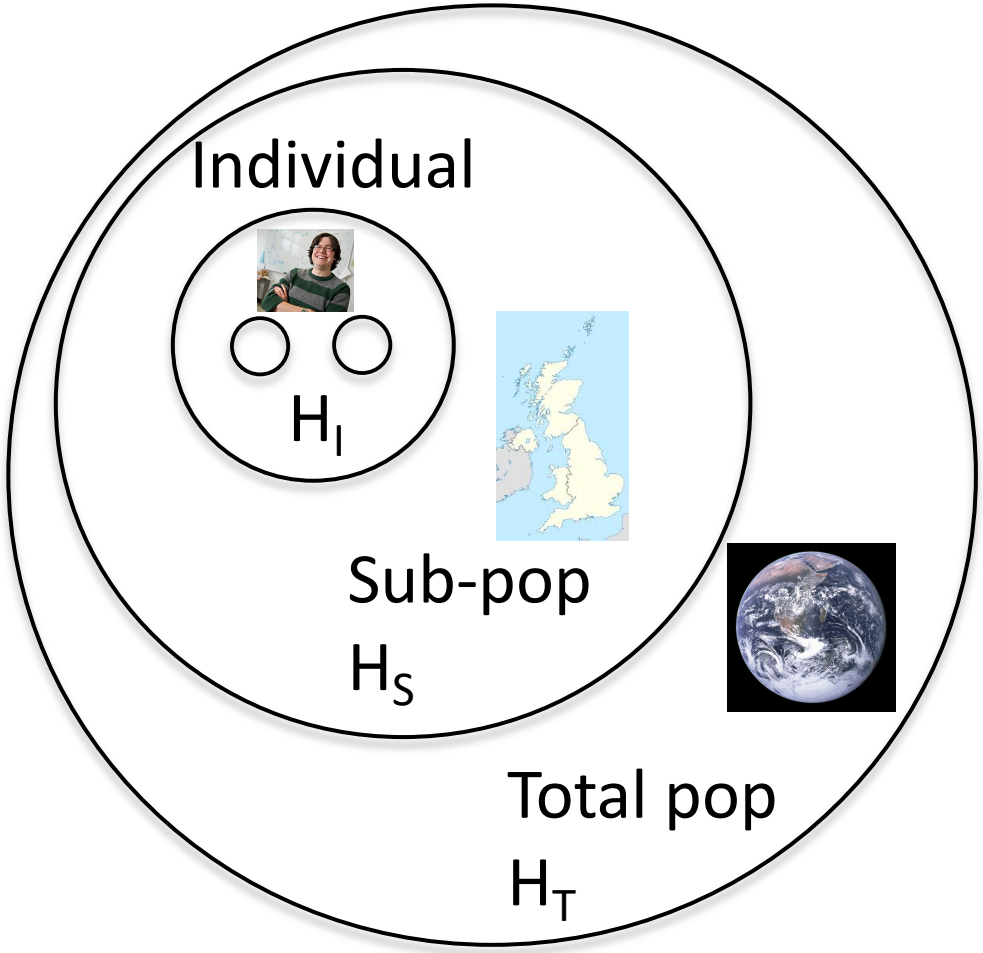
- This chapter walks through common ways to summarize and visualize population structure
- Toward the end of the chapter, we'll also consider how population structure drives correlations in allele state across "linked" loci
- Characterization of structure is often one of the first steps in a population genomic study

3.0.1: Inbreeding as a summary of population structure



- Inbreeding and inbreeding coefficients are one natural entry point for considering population structure
- Inbreeding: whether an individual's parents are more closely related than random draws from a reference population
- Choice of reference matters: Dr. Coop's parents may have been random draws from the UK, but likely not from the world

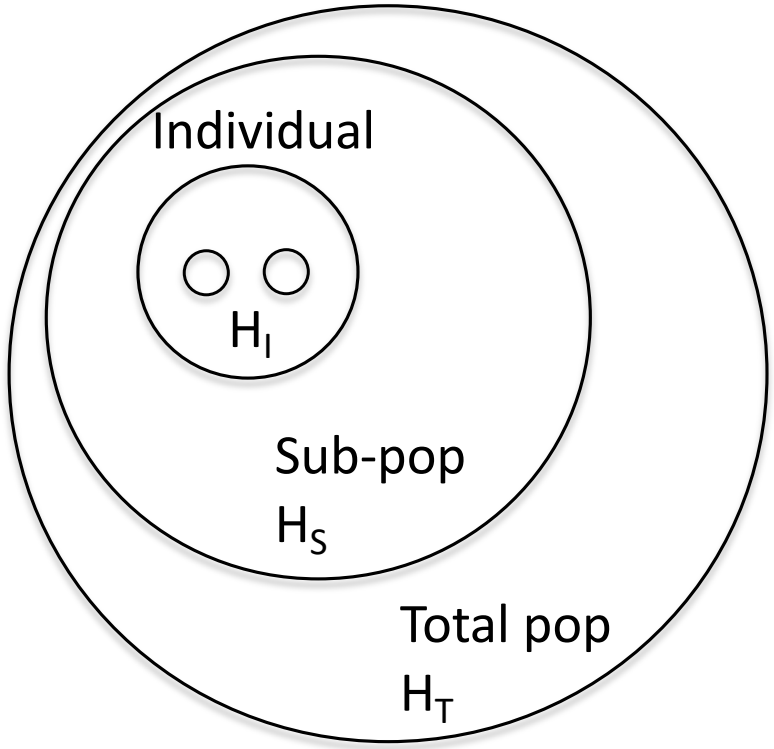
3.0.1: Inbreeding as a summary of population structure



- Dr. Coop's coefficient of inbreeding *(F_{ij})* would be nearly 0 if calculated with respect to individuals in the UK, but slightly higher with respect to the entire world's population

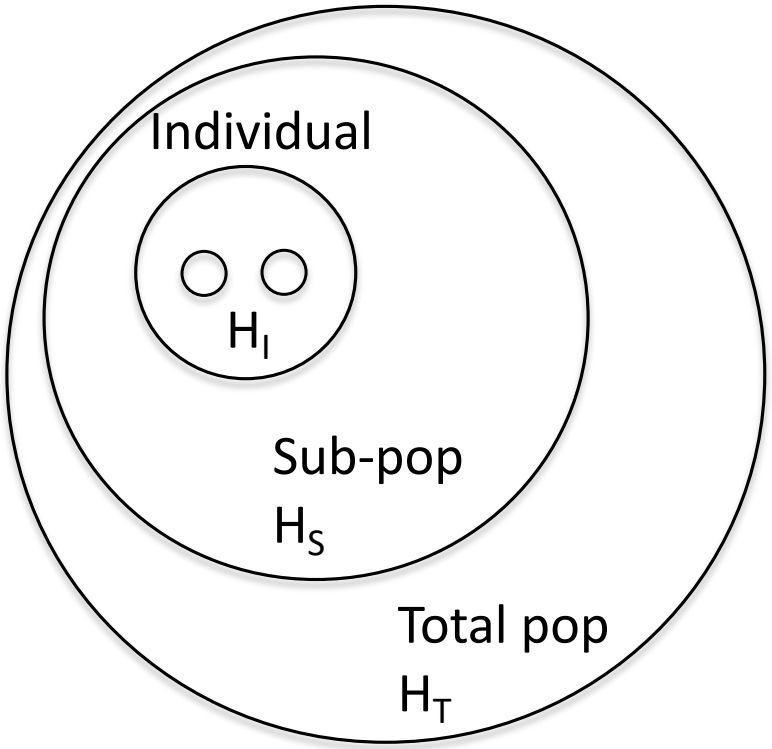
- There is lower heterozygosity within the UK's population than in the world's

3.0.1: Inbreeding as a summary of population structure



- Sewall Wright devised a set of *F*-statistics (*i.e.,* fixation indices) to consider inbreeding with respect to population structure
- F_{XY}: correlation between random gametes drawn from the same level X relative to level Y
- For example: F_{IS} = inbreeding coefficient of individual (*I*) relative to subpopulation (*S*)

3.0.1: Inbreeding as a summary of population structure

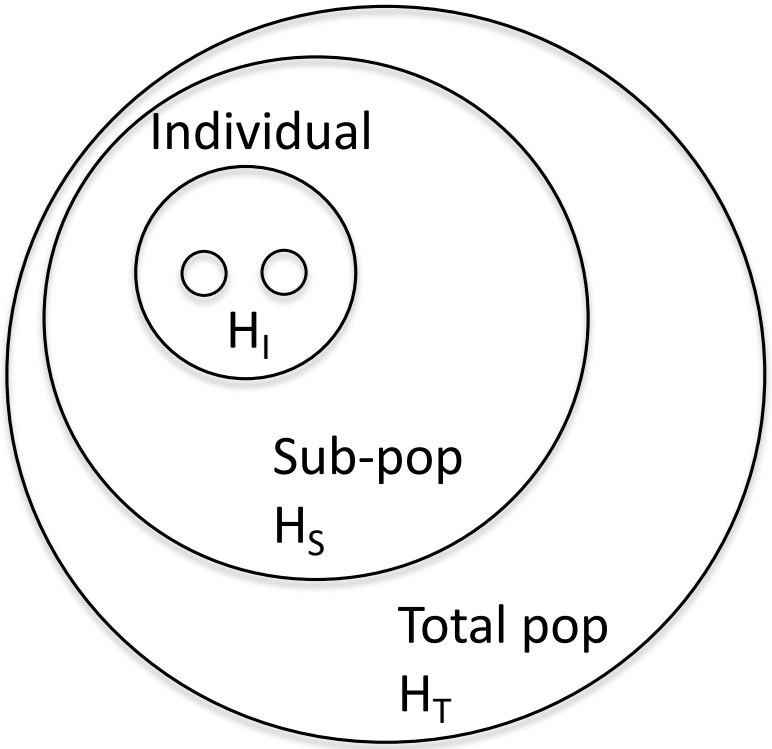


- Consider a locus where in subpopulation (*S*), a proportion $H_I = f_{12}$ of individuals are heterozygous:
  - The frequency of A_1 is p_S and expected heterozygosity is:
    $H_S = 2p_S(1 - p_S)$
  - Therefore:

$$F_{\mathrm{IS}} = 1 - \frac{H_I}{H_S} = 1 - \frac{f_{12}}{2p_S q_S}, \qquad (3.1)$$

- F_{IS} is the relative difference between observed and expected heterozygosity due to deviations from random mating in a subpopulation

3.0.1: Inbreeding as a summary of population structure



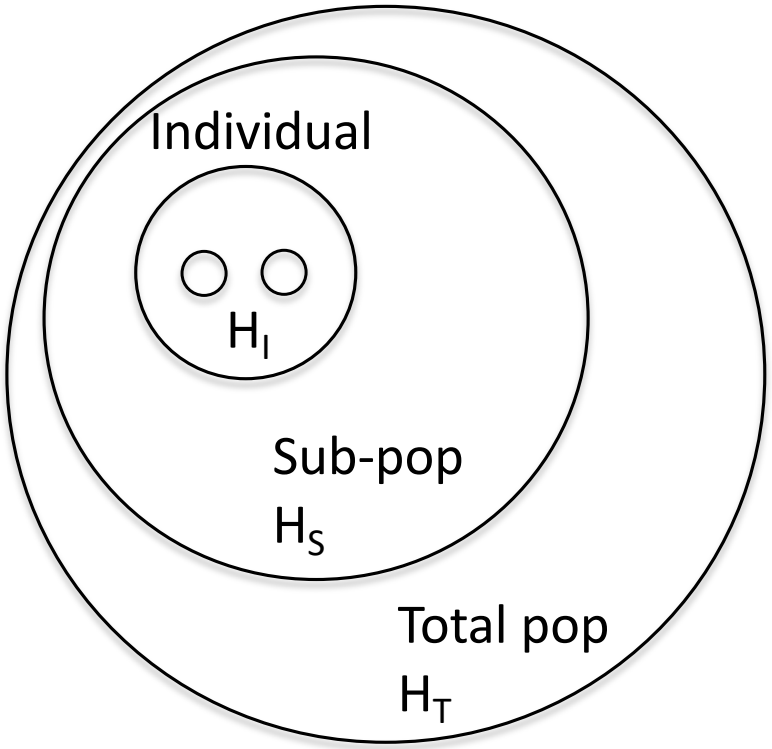
- We can also compare other levels within this hierarchy
- Deviations in observed and expected heterozygosity in individuals (H_I) vs. the total population (H_T)*:*

$$F_{\mathrm{IT}} = 1 - \frac{H_I}{H_T} = 1 - \frac{f_{12}}{2p_T q_T}, \qquad (3.2)$$

- Deviations in expected heterozygosity in subpopulations (H_S) vs. the total population (H_T)*:*

$$F_{\mathrm{ST}} = 1 - \frac{H_S}{H_T} = 1 - \frac{2p_S q_S}{2p_T q_T}. \qquad (3.3)$$

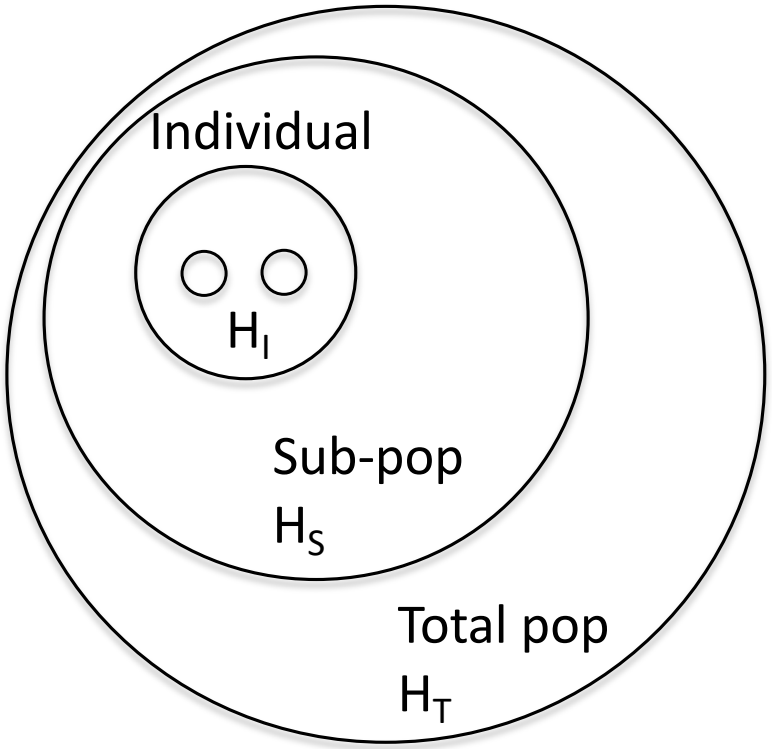
3.0.1: Inbreeding as a summary of population structure



- We can also show that the decrease in heterozygosity between the individual and total population levels can be decomposed into its component parts:

$$(1 - F_{\mathrm{IT}}) = \frac{H_I}{H_S}\frac{H_S}{H_T} = (1 - F_{\mathrm{IS}})(1 - F_{\mathrm{ST}}). \qquad (3.4)$$

3.0.1: Inbreeding as a summary of population structure



- F_{ST} can also be calculated as an average across multiple subpopulations:

$$F_{\mathrm{ST}} = 1 - \frac{\bar{H}_S}{H_T}, \qquad (3.5)$$

  with each H_S calculated using allele frequencies from each subpopulation

- measurements of F_{ST} can also be based on many loci across the genome by using average values of H_I, H_S, and H_T across loci

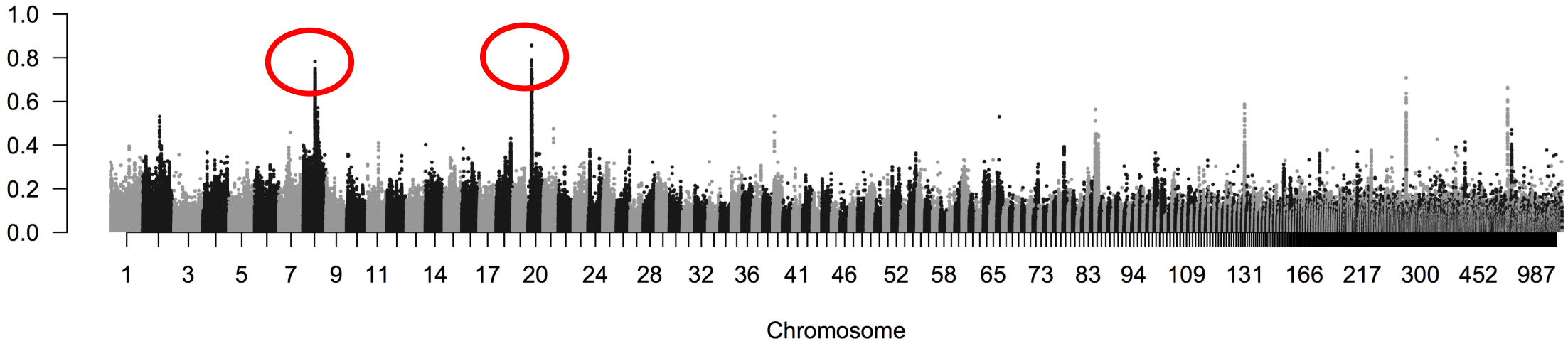
3.0.1: Inbreeding as a summary of population structure

- Let's consider an empirical example of how F_{ST} can vary across a genome by looking at blue- and golden-winged warblers
- Native to eastern North America, with the golden-winged warbler having a more northerly distribution
- Where their ranges overlap, these species readily hybridize

3.0.1: Inbreeding as a summary of population structure

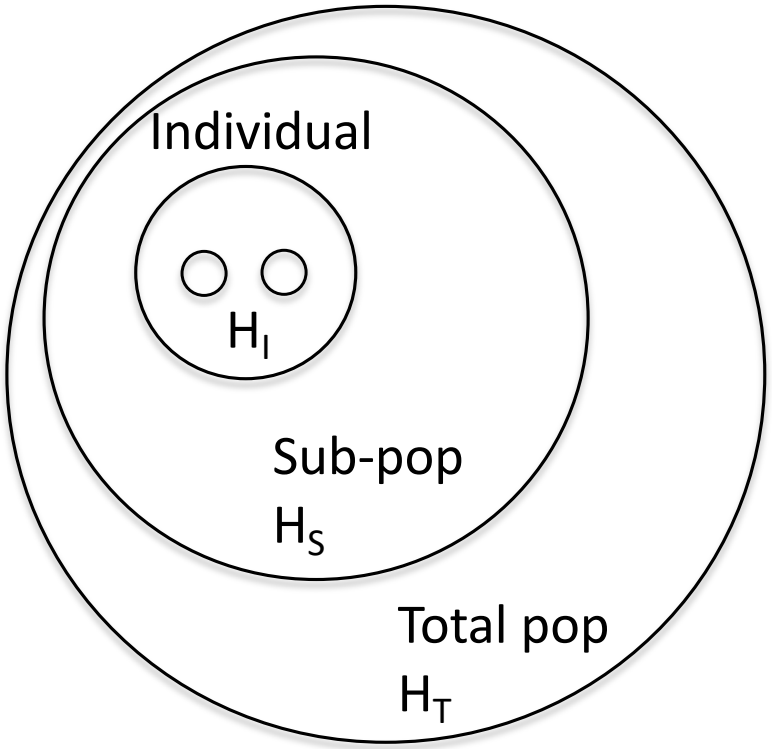
- The golden-winged warbler is listed as a threatened species in Canada and has received population genetic study motivated by conservation
- After sequencing thousands of loci across the genome, Toews and co-authors (2016) found that the genome-wide average of F_{ST} was very low (0.0045) between the two species
- Even at the species level, these warblers appeared to be randomly mating
- However…

3.0.1: Inbreeding as a summary of population structure



- When the landscape of F_{ST} was plotted across the genome, certain high F_{ST} regions were identified
- These corresponded to genes underlying plumage coloration

3.0.1: Inbreeding as a summary of population structure

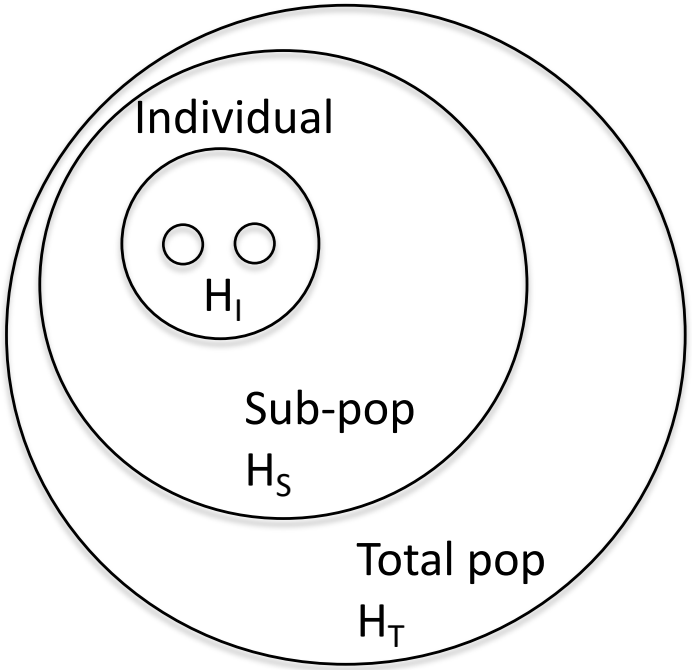


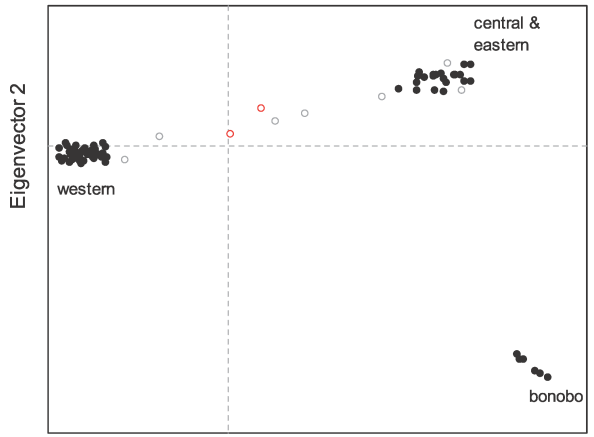
- Based on further derivations of *F*-statistics, Coop demonstrates that they can be viewed as:
  1. The correlation between alleles drawn from a population or individual beyond that expected by chance.
  2. The proportion of variance explained by, for example, subpopulation labels in the case of F_{ST}

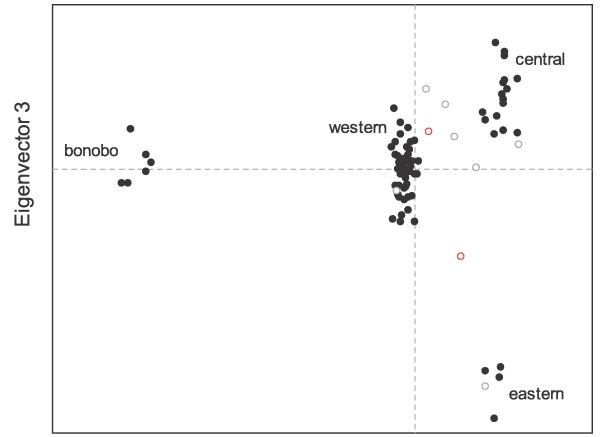
Coop, Chapter 3: 3.0.2-3.0.3

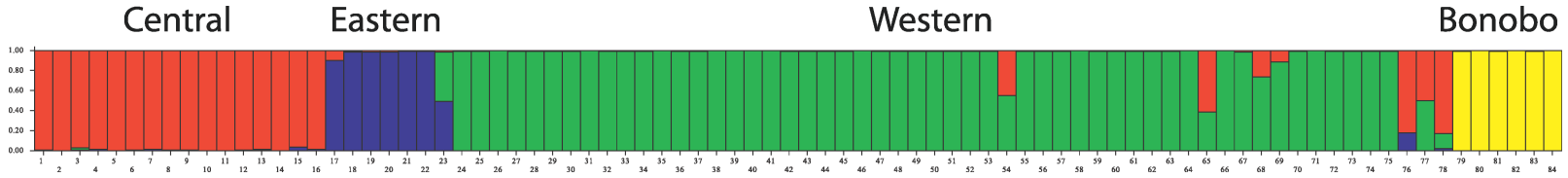
Population Structure and Correlations Among Loci

Assignment Methods









3.0.3 Assignment Methods

In conservation genetics, often the goal is to determine where an individual came from (e.g., poaching)

- P(pop K | ind.) = Probability of specific population K being the source of origin for the genotyped individual, given this individual's genotype

Pan troglodytes; © Ian Nichols

3.0.3 Assignment Methods

Say you have genotyped an individual from an unknown population at a single locus. The probability of this individual's genotype at locus l, given that the individual comes from population k, is:

$$P(g_l|\text{pop k}) = \begin{cases} (1-p_{k,l})^2 & g_l = 0 \\ 2p_{k,l}(1-p_{k,l}) & g_l = 1 \\ p_{k,l}^2 & g_l = 2 \end{cases} \quad (3.9)$$

Where:

$g_l =$ unknown individual's genotype **g** at locus **l**

$p_{k,l} =$ allele frequency of allele **p** in population **k** at locus **l**

$1 - p_{k,l} =$ allele frequency of allele **q** in population **k** at locus **l**

3.0.3 Assignment Methods

Say you have genotyped an individual from an unknown population at a single locus. The probability of this individual's genotype at locus l, given that the individual comes from population k, is:

$$P(g_l|\text{pop k}) = \begin{cases} (1-p_{k,l})^2 & g_l = 0 \\ 2p_{k,l}(1-p_{k,l}) & g_l = 1 \\ p_{k,l}^2 & g_l = 2 \end{cases} \tag{3.9}$$

This is just Hardy-Weinberg genotype frequencies for a specific population

3.0.3 Assignment Methods

As we often have multiple loci, the probability of the individual's genotype across all loci genotyped, given the individual comes from population k, is:

$$P(\text{ind.}|\text{pop k}) = \prod_{l=1}^{S} P(g_l|\text{pop k}) \qquad (3.10)$$

Which is the product of the probabilities of the genotypes at each locus given population k

3.0.3 Assignment Methods

As we often have multiple loci, the probability of the individual's genotype across all loci genotyped, given the individual comes from population k, is:

$$P(\text{ind.}|\text{pop k}) = \prod_{l=1}^{S} P(g_l|\text{pop k}) \qquad (3.10)$$

In other words,

$$\mathrm{P}(g_1|pop\ k) \times \mathrm{P}(g_2|pop\ k) \times \mathrm{P}(g_3|pop\ k) \ldots \times \mathrm{P}(g_l|pop\ k)$$

3.0.3 Assignment Methods

Now we can calculate the probability that the individual is from population k, given the individual's genotype at multiple loci, using Bayes' rule:

$$P(\text{pop k}|\text{ind.}) = \frac{P(\text{ind.}|\text{pop k})P(\text{pop k})}{P(\text{ind.})} \qquad (3.11)$$

Where:

$$P(ind.\,|\,pop\ k) = \prod_{l=1}^{S} P(g_l|\text{pop k})$$

$P(pop\ k) = {}^{1}/_{K}$, and K is the number of populations

$$P(\text{ind.}) = \sum_{k=1}^{K} P(\text{ind.}|\text{pop k})P(\text{pop k}) \qquad (3.12)$$

3.0.3 Assignment Methods

$P(pop\ k \mid ind.)$ = posterior probability that our new individual comes from each of our K populations

$$P(\text{pop k}|\text{ind.}) = \frac{P(\text{ind.}|\text{pop k})P(\text{pop k})}{P(\text{ind.})} \qquad (3.11)$$

3.0.3 Assignment Methods

Question 2:

Returning to our chimp example, imagine that we have genotyped a set of individuals from the Western and Eastern populations at two SNPs (we'll ignore the central population to keep things simpler). The frequency of the capital allele at two SNPs (A/a and B/b) is given by

Population	locus A	locus B
Western	0.1	0.85
Eastern	0.95	0.2

Sequenced individual's genotype is AA/bb. Is it from the Eastern or Western population?

3.0.3 Assignment Methods

Question 2:

Population	locus A	locus B
Western	0.1	0.85
Eastern	0.95	0.2

Genotype is AA/bb

Calculate:

1. $P(ind. \mid pop\ k)$ for k = Western and k = Eastern populations

Where:

$$P(ind. \mid pop\ k) = \prod_{l=1}^{S} P(g_l|\text{pop k})$$

In words: we're finding the product of the probabilities of the genotypes of this individual at independent loci coming from each population

From the Western pop.: $(0.1)^2$ X $(0.15)^2$ = 0.000225

From the Eastern pop.: $(0.95)^2$ X $(0.8)^2$ = 0.5776

3.0.3 Assignment Methods

Question 2:

Population	locus A	locus B
Western	0.1	0.85
Eastern	0.95	0.2

Genotype is AA/bb

Calculate:

2. $P(ind.)$

Where:

$$P(ind.) = \sum_{k=1}^{K} P(\text{ind.}|\text{pop k})P(\text{pop k})$$

$P(pop\ k) = {}^{1}/_{K}$, and K is the number of populations

In words: we're finding the sum of the probabilities of the chimp coming from each population under the prior that both populations are equally likely

0.000225(0.5) + 0.5776(0.5) = 0.2889

3.0.3 Assignment Methods

Question 2:

Population	locus A	locus B
Western	0.1	0.85
Eastern	0.95	0.2

Genotype is AA/bb

Calculate:

3. Use Bayes' Rule to find probability of individual belonging to each population

$$P(\text{pop k}|\text{ind.}) = \frac{P(\text{ind.}|\text{pop k})P(\text{pop k})}{P(\text{ind.})} \qquad (3.11)$$

From the Eastern pop.: (0.5776)0.5/0.2889 = 0.9997

From the Western pop.: (0.000225)0.5/0.2889 = 0.0003

Our individual was from the Eastern population

3.0.3 Assignment Methods

When we want to assign individuals based on their genotypes into K populations:

1. Randomly assign individuals to K populations
2. Estimate allele frequencies at all loci for each of K populations
3. Reassign each individual to a population k with a probability equal to:

$$P(ind. \mid pop\ k) = \prod_{l=1}^{S} P(g_l|\text{pop k})$$

4. Repeat steps 2-3.

3.0.3 Assignment Methods

- Bayesian clustering algorithm
- STRUCTURE Webpage
- Cited > 28,000 times!

[HTML] Inference of population **structure** using multilocus genotype data

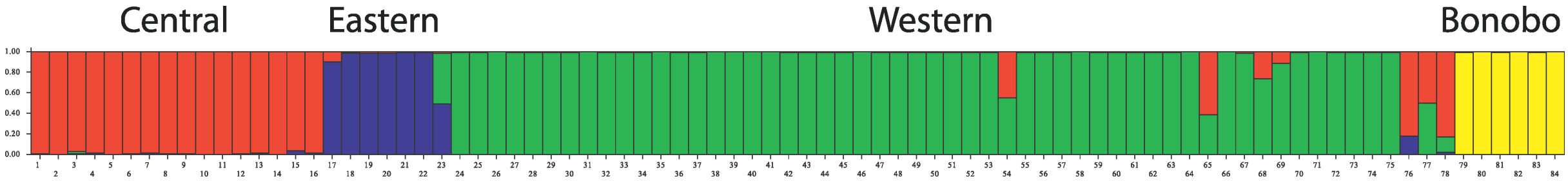
JK **Pritchard**, M Stephens, P Donnelly - Genetics, 2000 - Genetics Soc America

We describe a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations. We assume a model in which there are K populations (where K may be unknown), each of which is characterized by a set …

☆ Cited by 28197 Related articles All 51 versions

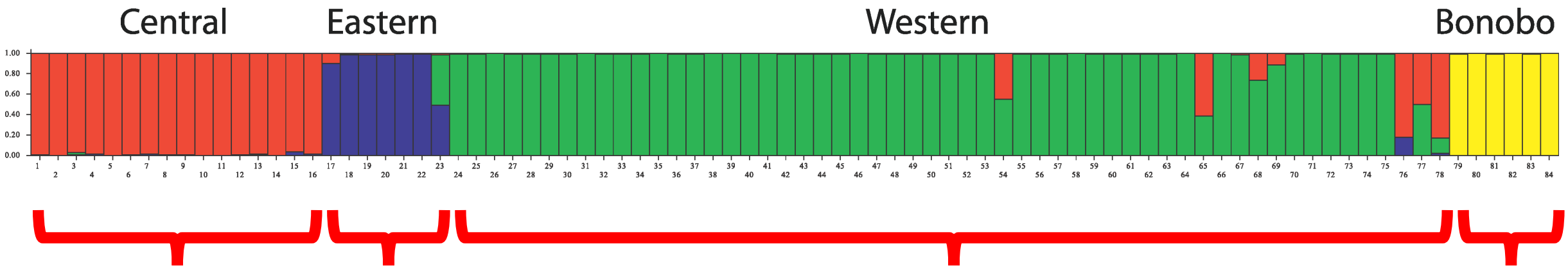
3.0.3 Assignment Methods

Becquet et al. 2007 STRUCTURE Analysis (K=4):



3.0.3 Assignment Methods

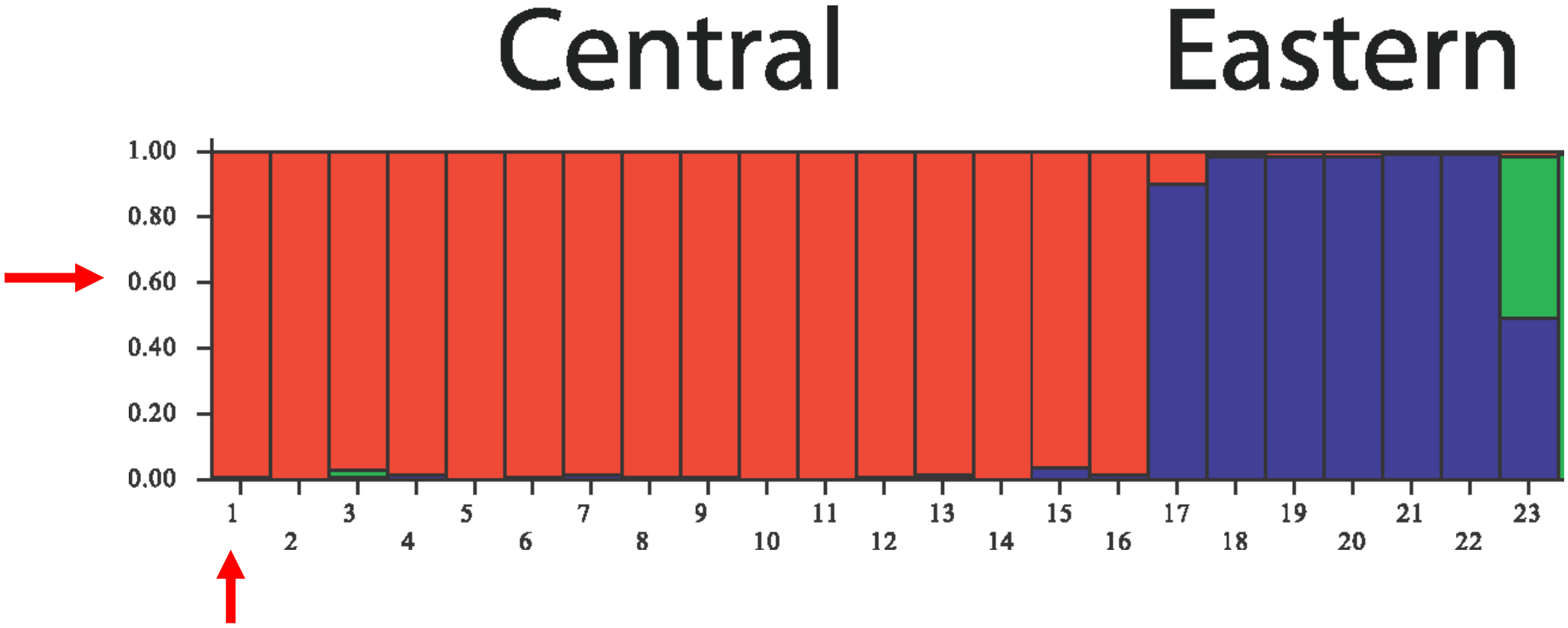
Becquet et al. 2007 STRUCTURE Analysis (K=4):



3.0.3 Assignment Methods

Proportion of ancestry from certain k, based on genotype

Becquet et al. 2007 STRUCTURE Analysis (K=4):



Individual 1: Columns are individuals

3.0.3 Assignment Methods

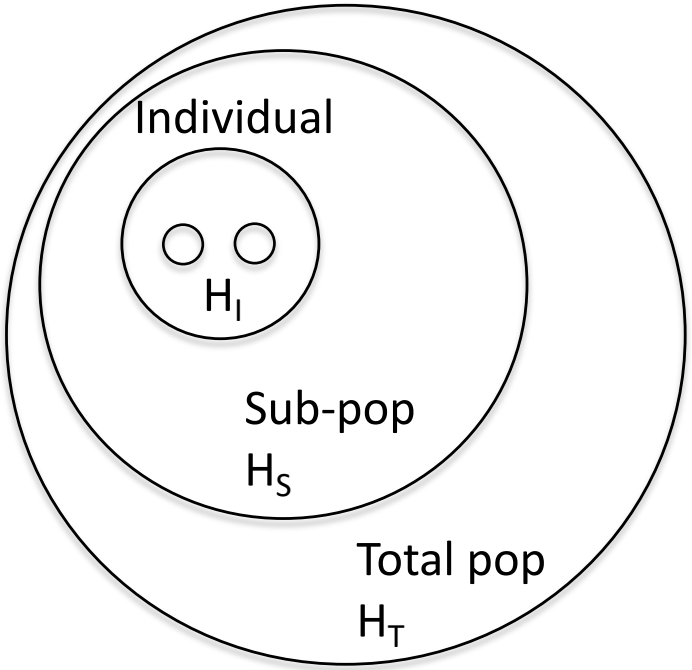
Important notes:

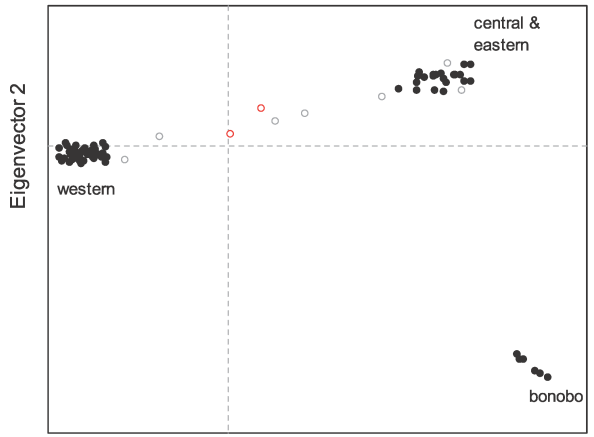
- STRUCTURE alone does not tell you the best K
- The colors denoted on the plot do not tell you that these clusters are ‘pure’ ancestral populations

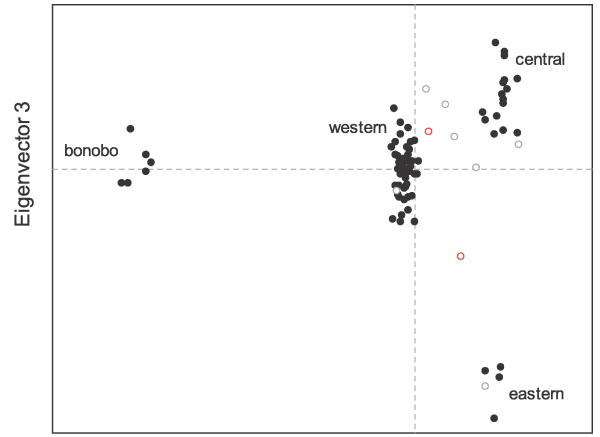
Coop, Chapter 3: 3.0.4

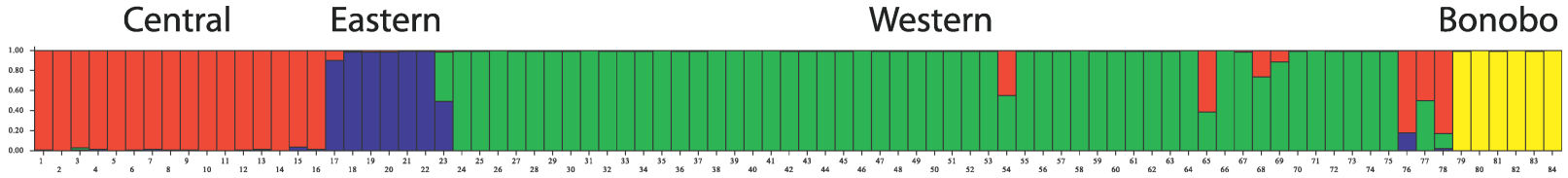
Population Structure and Correlations Among Loci

Principal Components Analysis

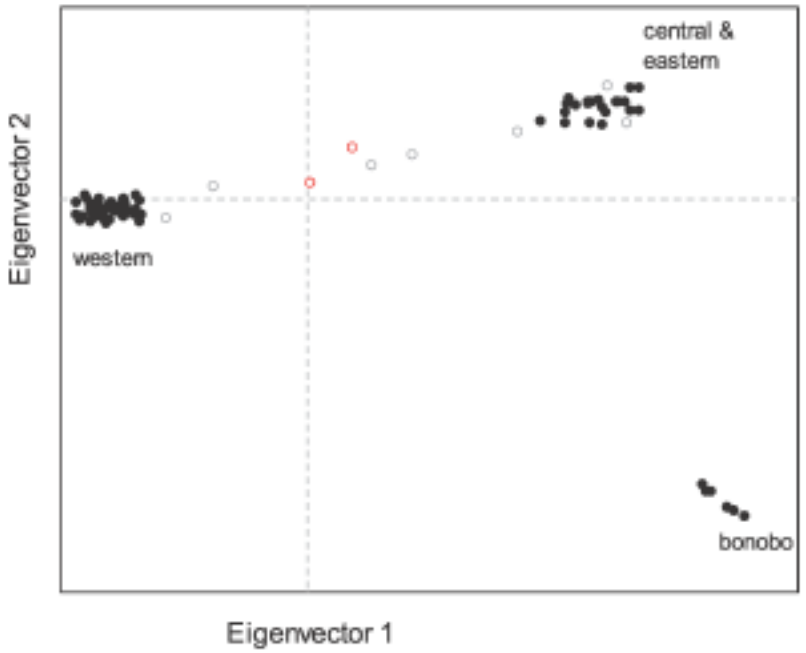






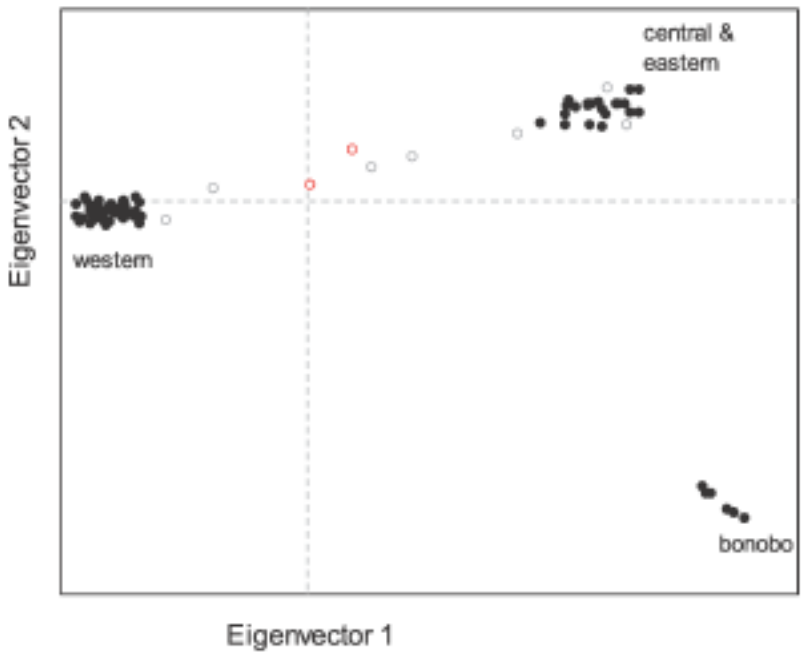


3.0.4 Principal Components Analysis



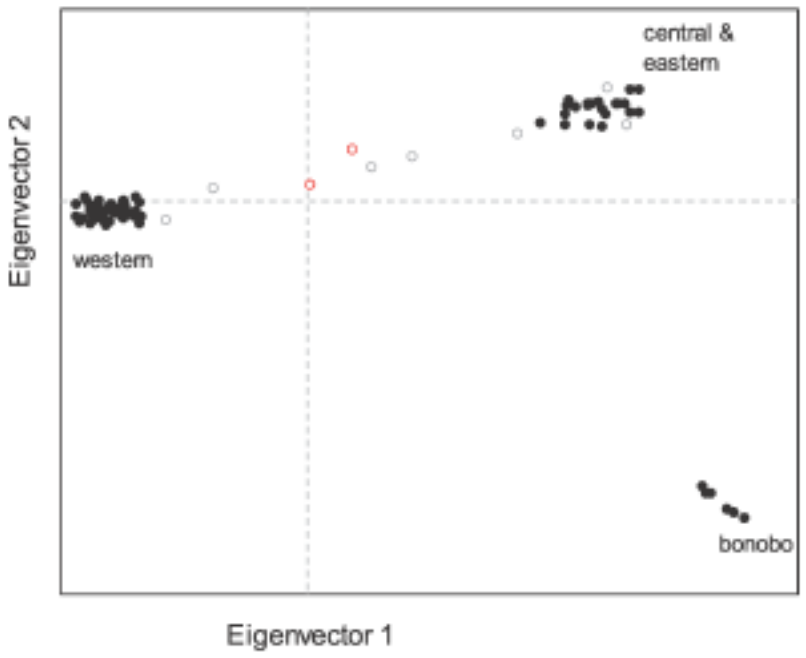
- PCA is a common statistical approach to “reduce the dimensionality” of complex data sets
- For example, we can take genetic relatedness from 1000’s of SNPs and reduce this down to two axes
- PC1 describes the most amount of variation, PC2 second-most, etc.

3.0.4 Principal Components Analysis



- Let's think about a data set with *N* individuals genotyped at *S* loci
- The i^{th} individual's genotype at locus *l* is $g_{i,l} = 0$, 1, or 2 based on the number of copies of the A_1 allele
- These values are placed in an *N* x *S* matrix

3.0.4 Principal Components Analysis

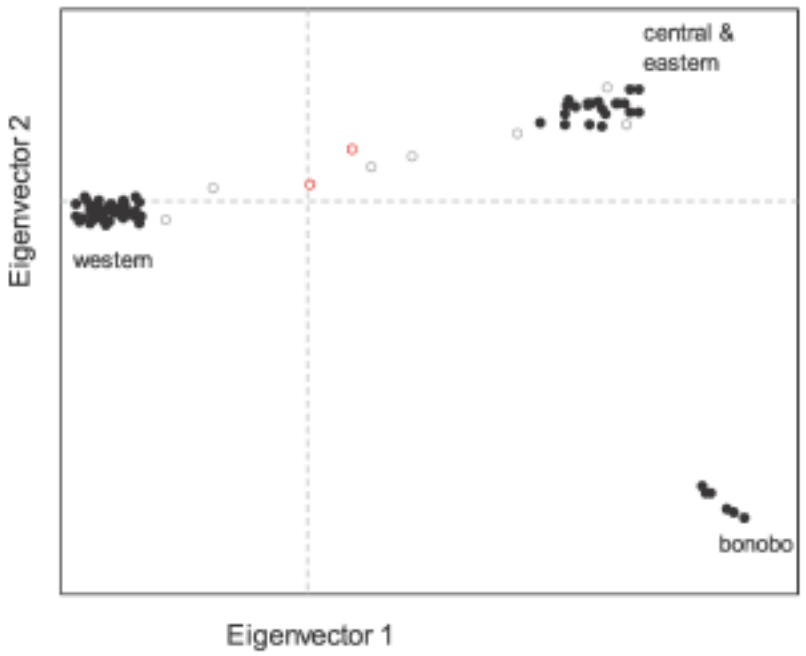


- Genotypes are then standardized:

$$\frac{g_{i,\ell} - 2p_\ell}{\sqrt{2p_\ell(1-p_\ell)}} \qquad (3.14)$$

- where p_l is the mean frequency of SNP *l* and the denominator is the square root of the expected variance under binomial sampling
- These normalized values are then used for the PCA

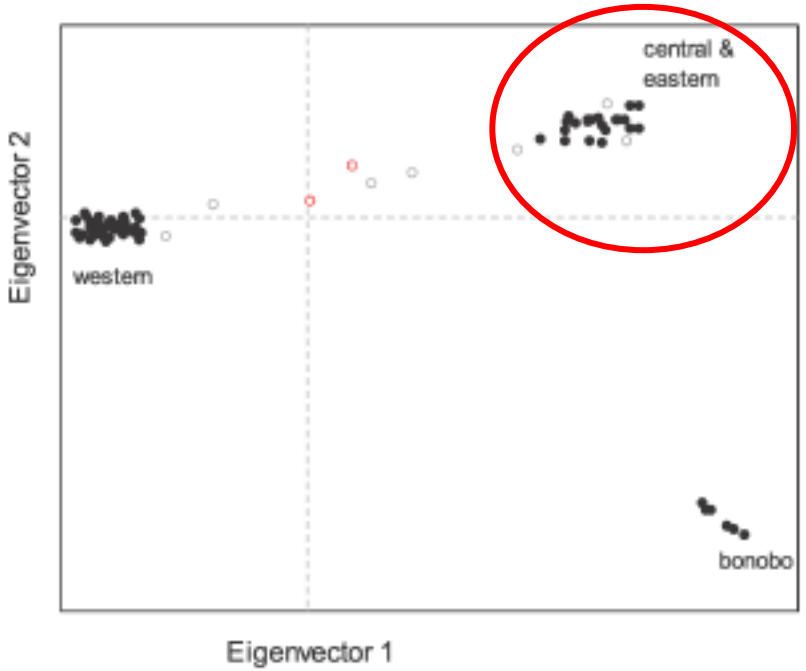
3.0.4 Principal Components Analysis



- Equation 3.15 tells us that we are looking for covariance between samples above what we might expect if they were drawn randomly from the total population:

$$\frac{1}{S-1}\sum_{\ell=1}^{S}\frac{(g_{i,\ell}-2p_\ell)(g_{j,\ell}-2p_\ell)}{2p_\ell(1-p_\ell)} \tag{3.15}$$

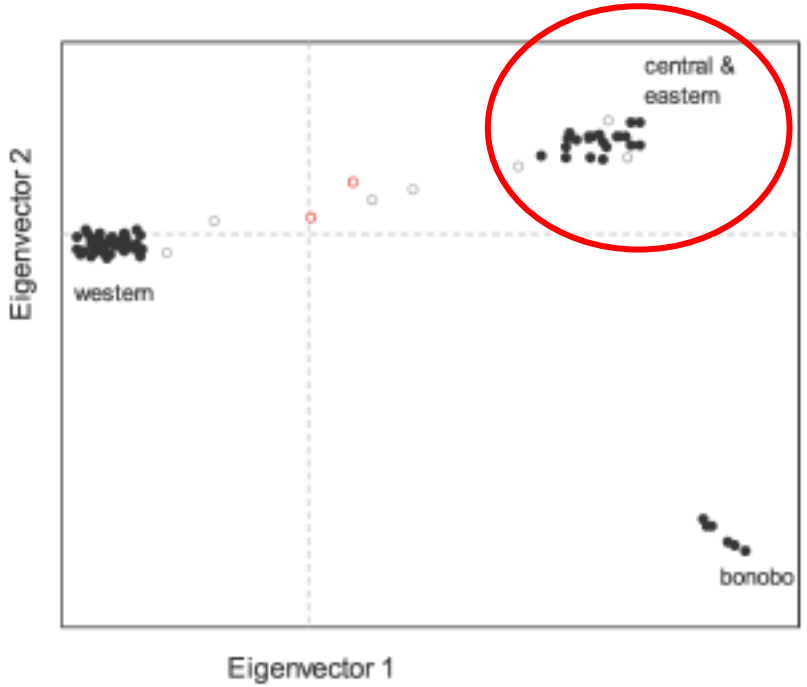
3.0.4 Principal Components Analysis

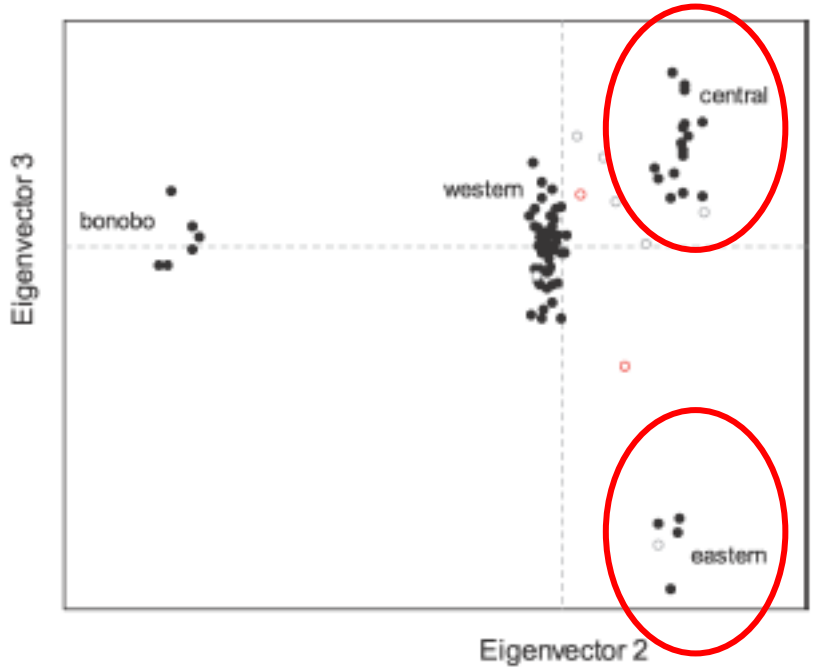


- Let's focus on interpretation, going back to the chimpanzee data
- Based on Eigenvectors (PCs) 1 & 2, would you say that the central and eastern populations are distinct?
- Now let's look at additional axes…

3.0.4 Principal Components Analysis

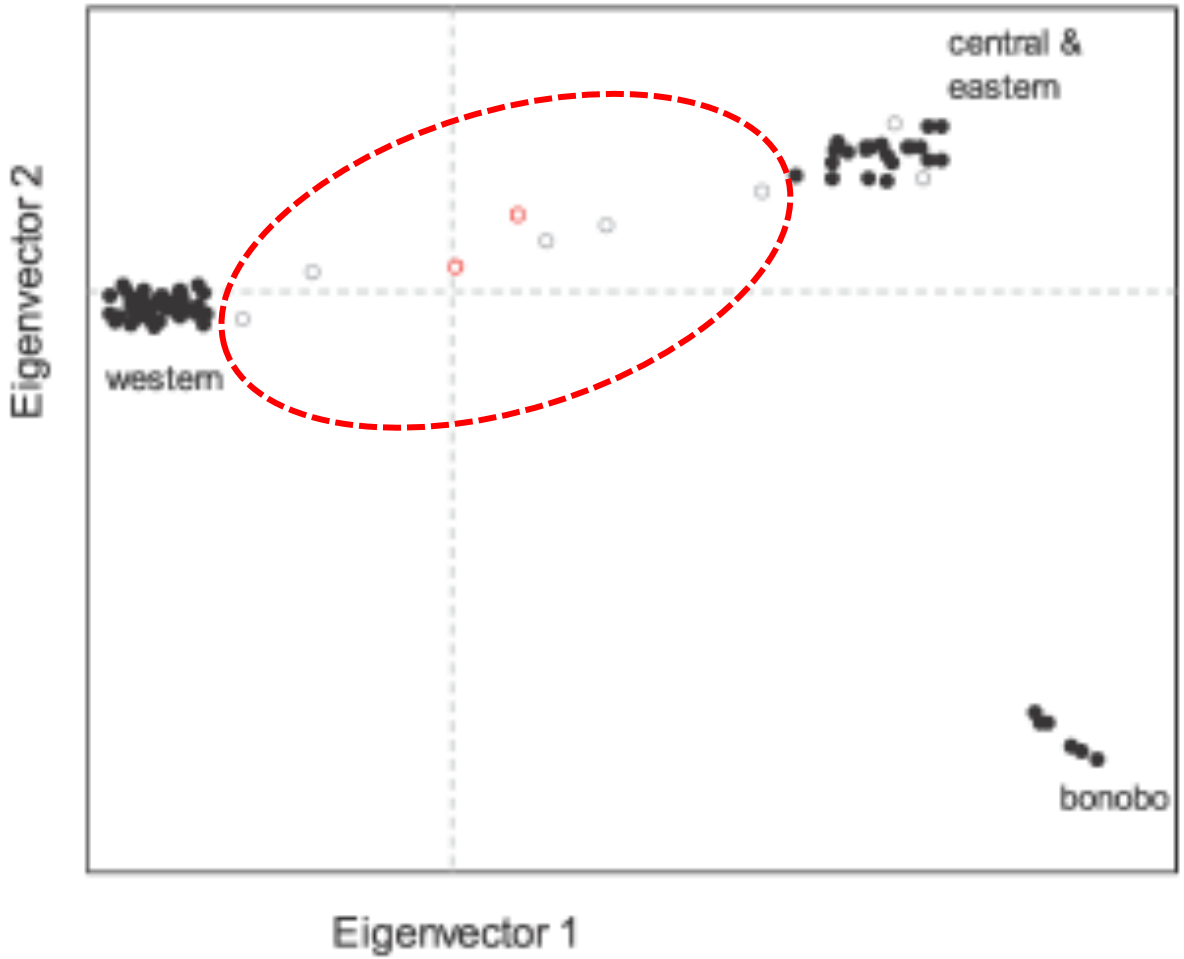
What about now?





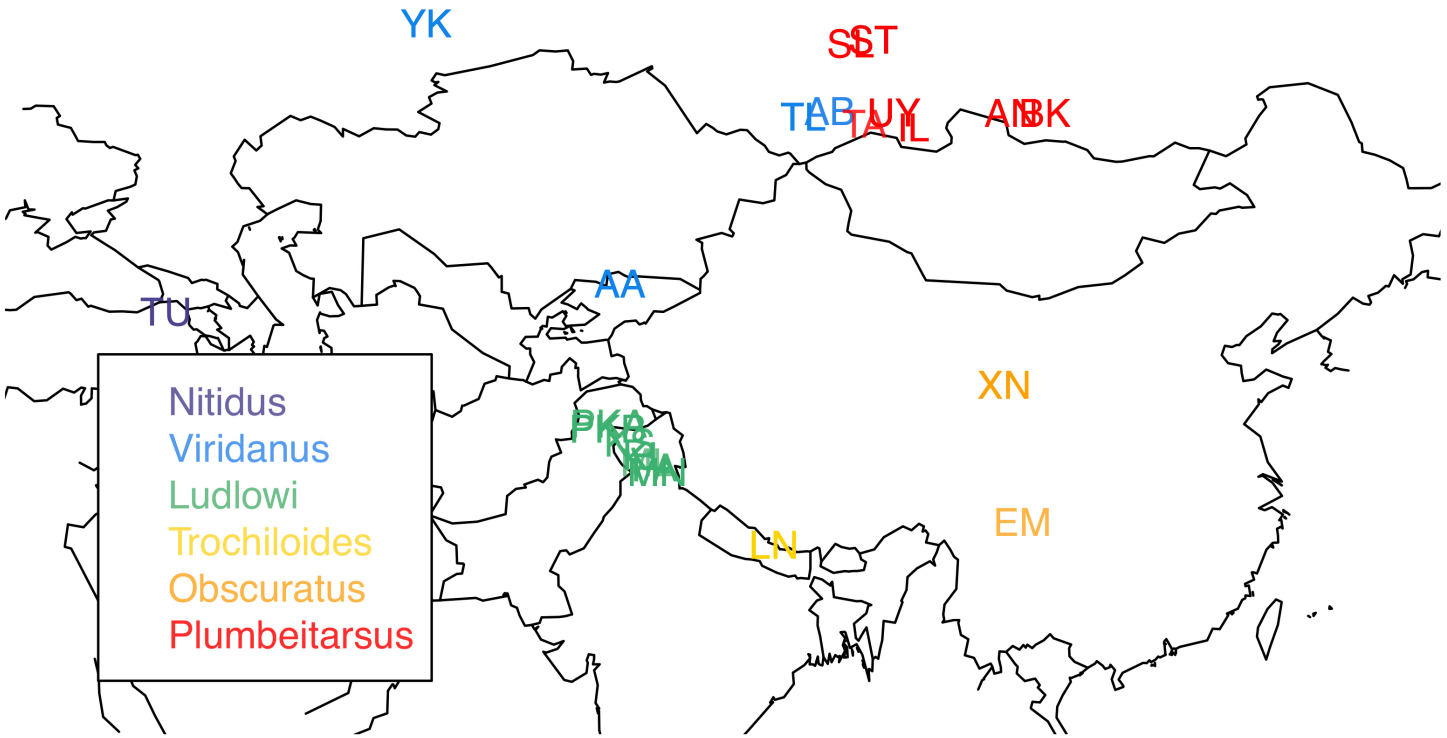
3.0.4 Principal Components Analysis

Thoughts about what might be going on with these individuals?

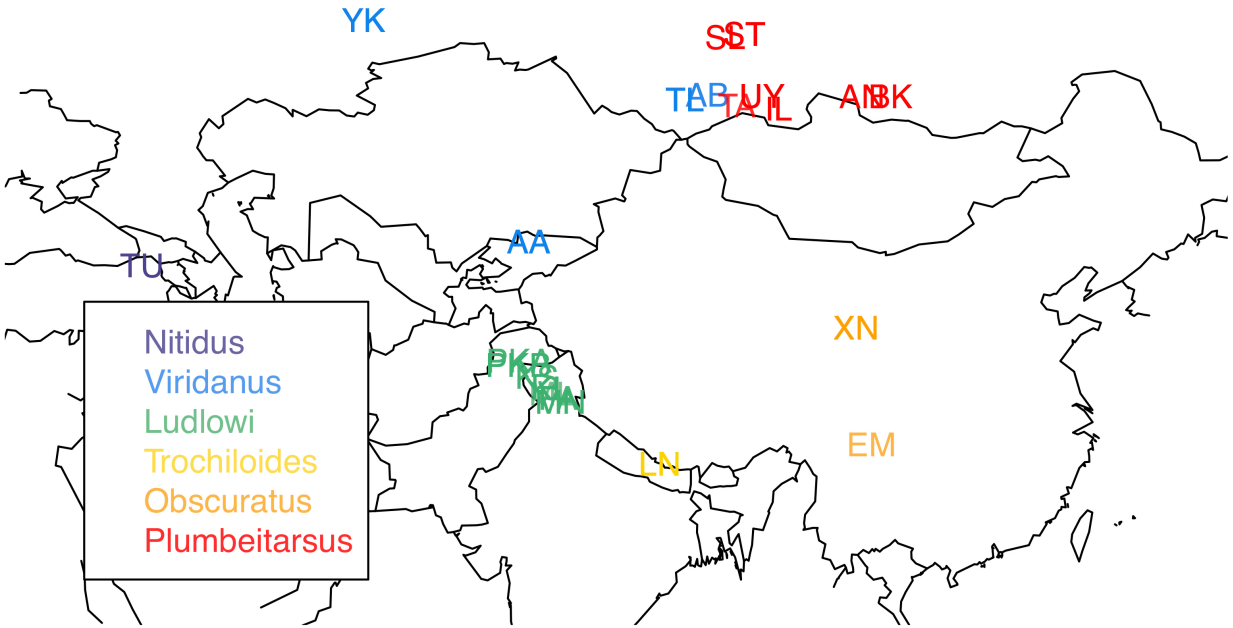


3.0.4 Principal Components Analysis

Let’s consider another empirical example, the greenish warbler, a putative ring species:

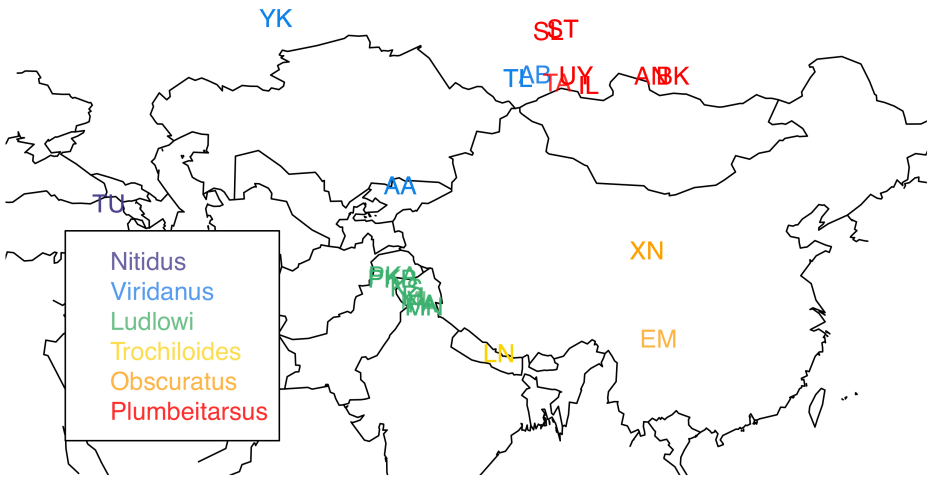


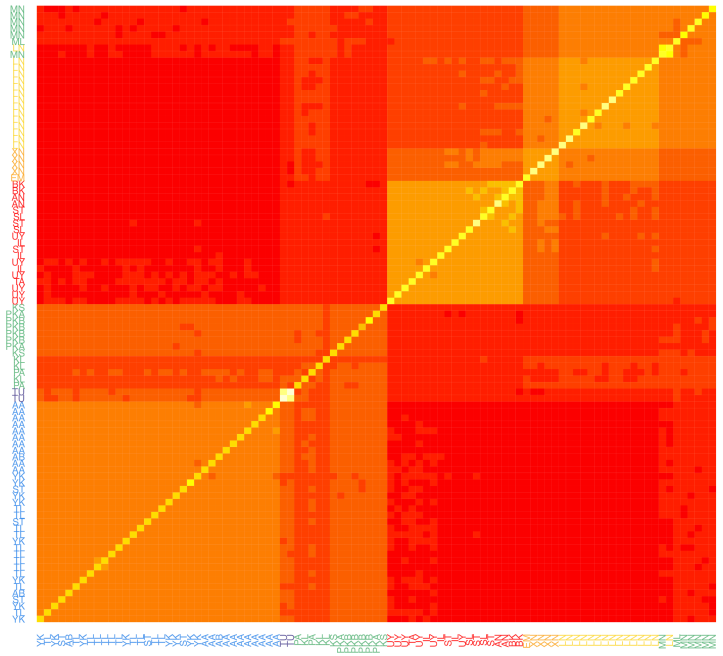
3.0.4 Principal Components Analysis



- Alcaide *et al.* (2014) collected 95 birds from 22 sites and genotyped them for 2,334 SNPs
- This species is thought to have originated in south Asia and then dispersed in a ring around the inhospitable Himalayas
- Given this history, what would be your expectations regarding relatedness and population structure?

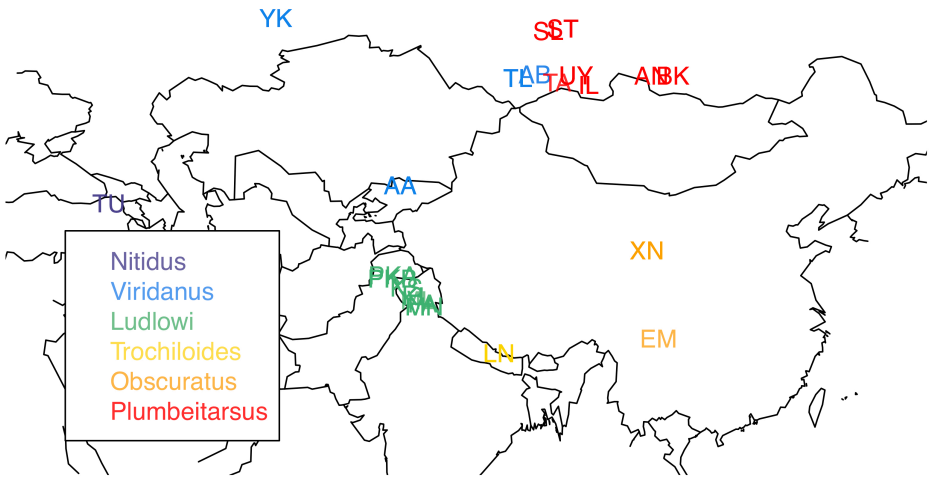
3.0.4 Principal Components Analysis

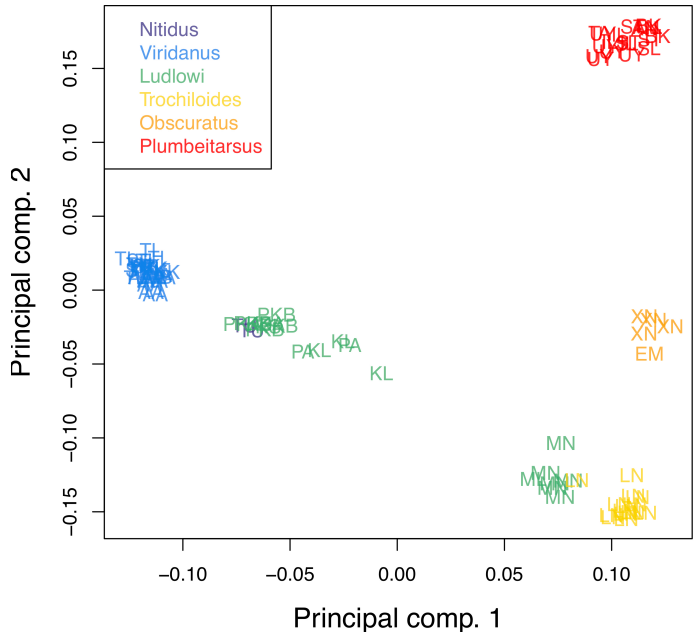




- The covariance matrix here shows the level of relatedness between individuals across populations
- Lighter colors indicate greater kinship
- Green and blue individuals show relatedness (western route around Himalayas) as do yellow and red individuals (eastern route)

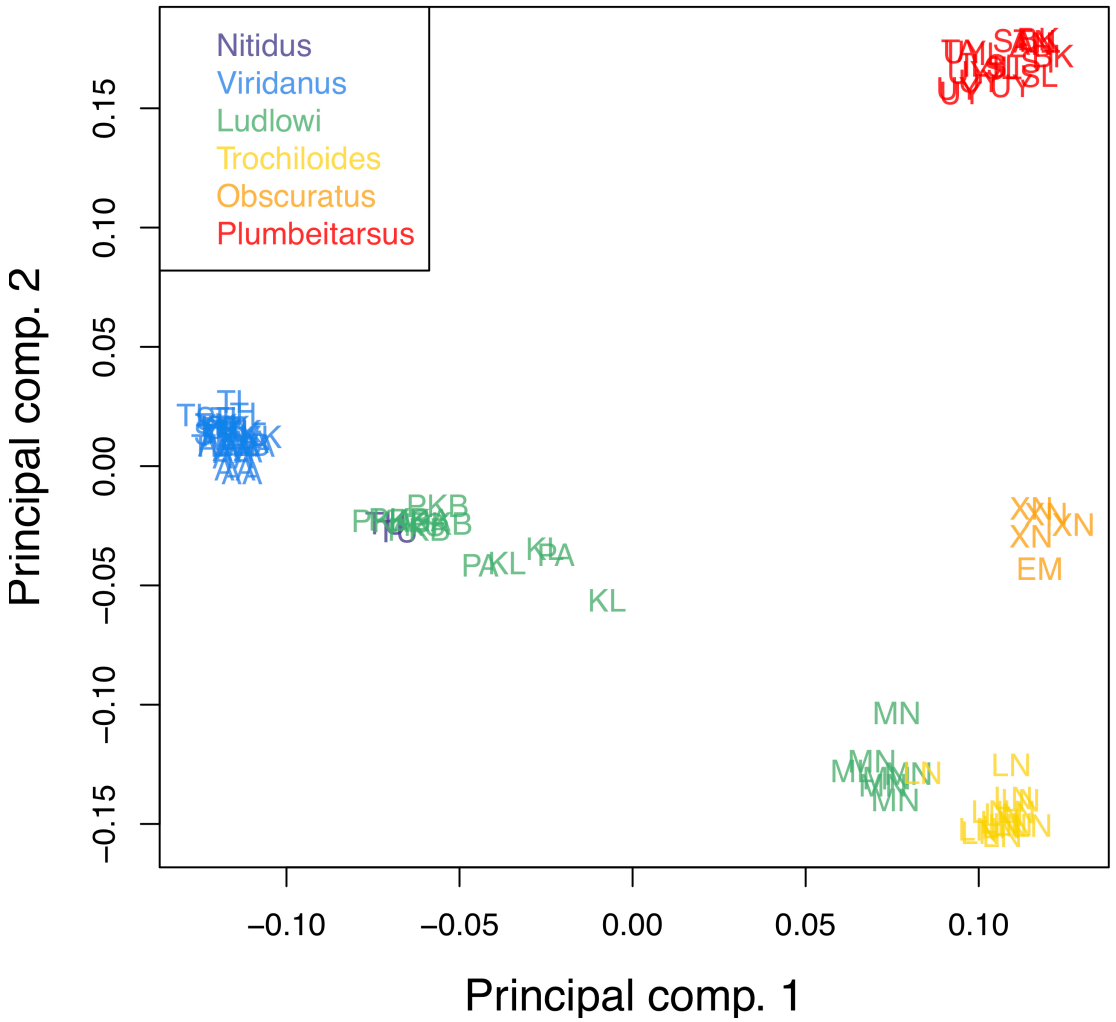
3.0.4 Principal Components Analysis





- The PCA shows two diverging sets of populations: the western route of migration around the Himalayas and the eastern route
- Even though the blue and red populations are close geographically, they are the most distant populations genetically due to their history
- Interesting break in the green Ludlowi species suggests local barriers to migration

3.0.4 Principal Components Analysis

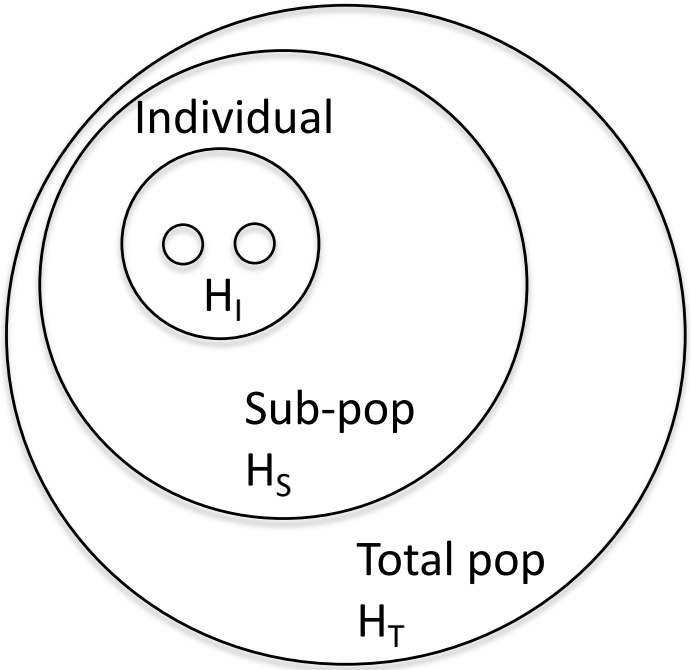


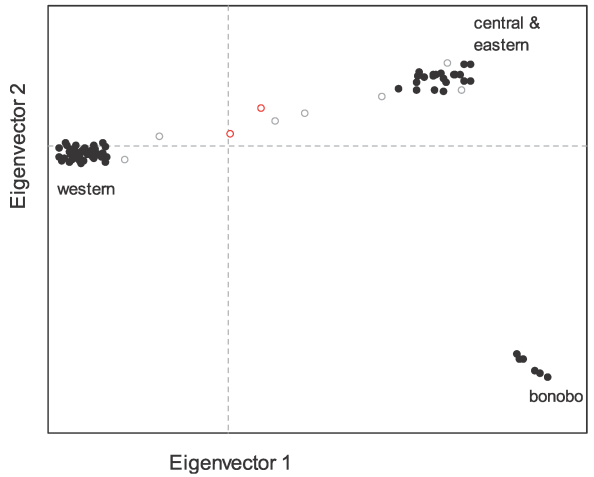
- While PCA can be an excellent tool for interpreting structure in genetic data, care needs to be taken in interpretation
- Complex geometric structures can arise in PC space even under simple geographic models, for example, when populations are arrayed along a line
- Multiple tools should be used to clarify the geographic and population-genetic history of a species

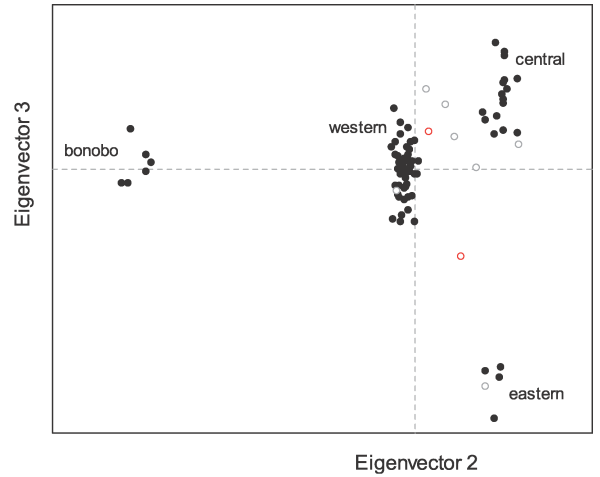
Coop, Chapter 3: 3.0.5

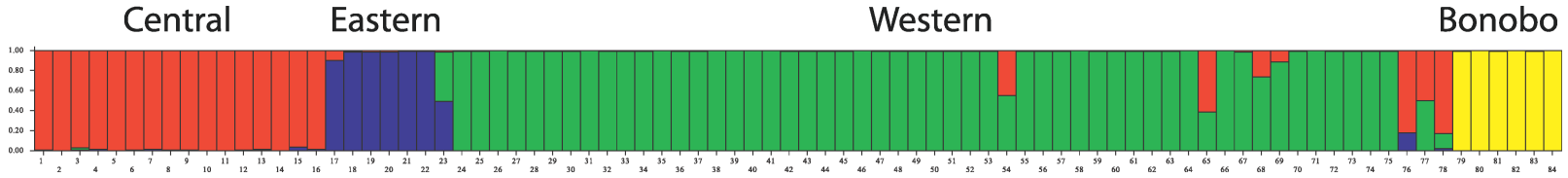
Population Structure and Correlations Among Loci

Correlations between loci, linkage disequilibrium, and recombination



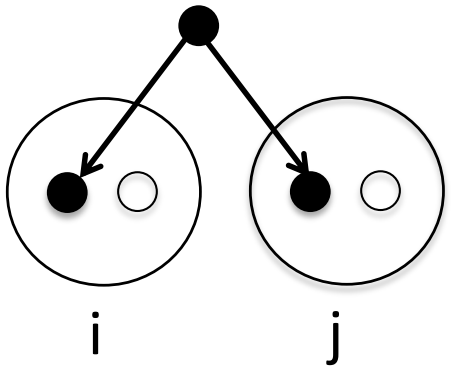




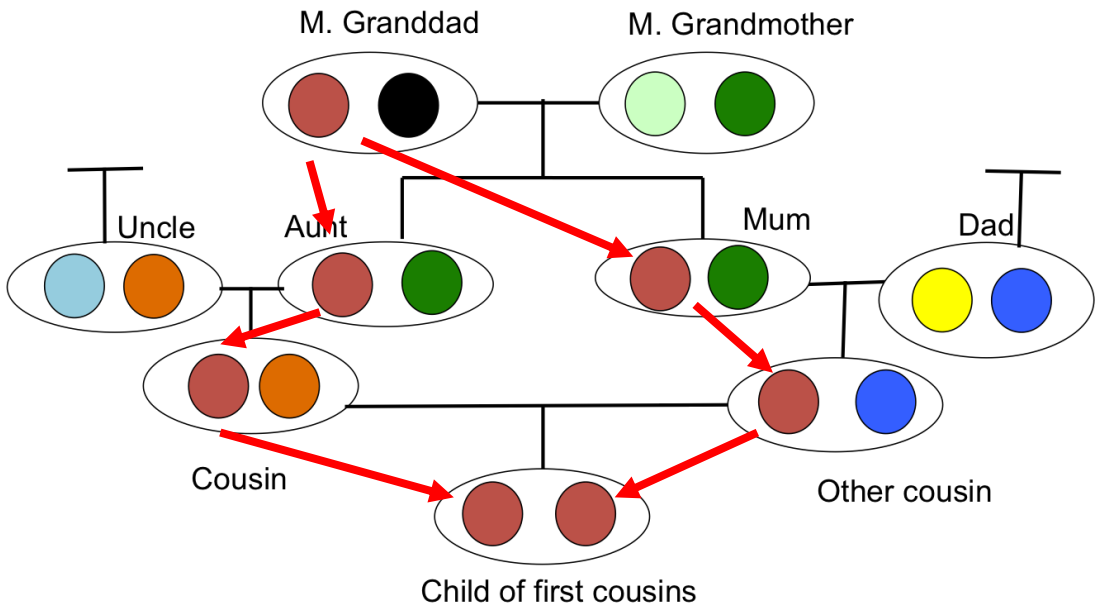


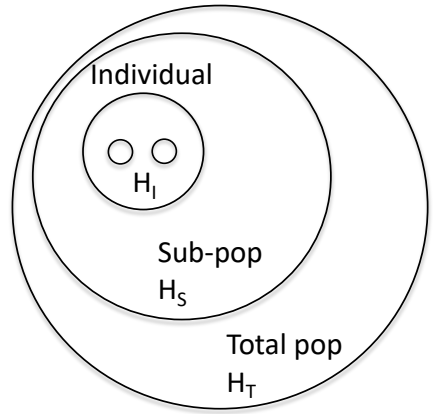
3.0.5 Correlations between loci

Up until now we have been looking at correlations at a single locus, within an individual (inbreeding) or between individuals (relatedness)



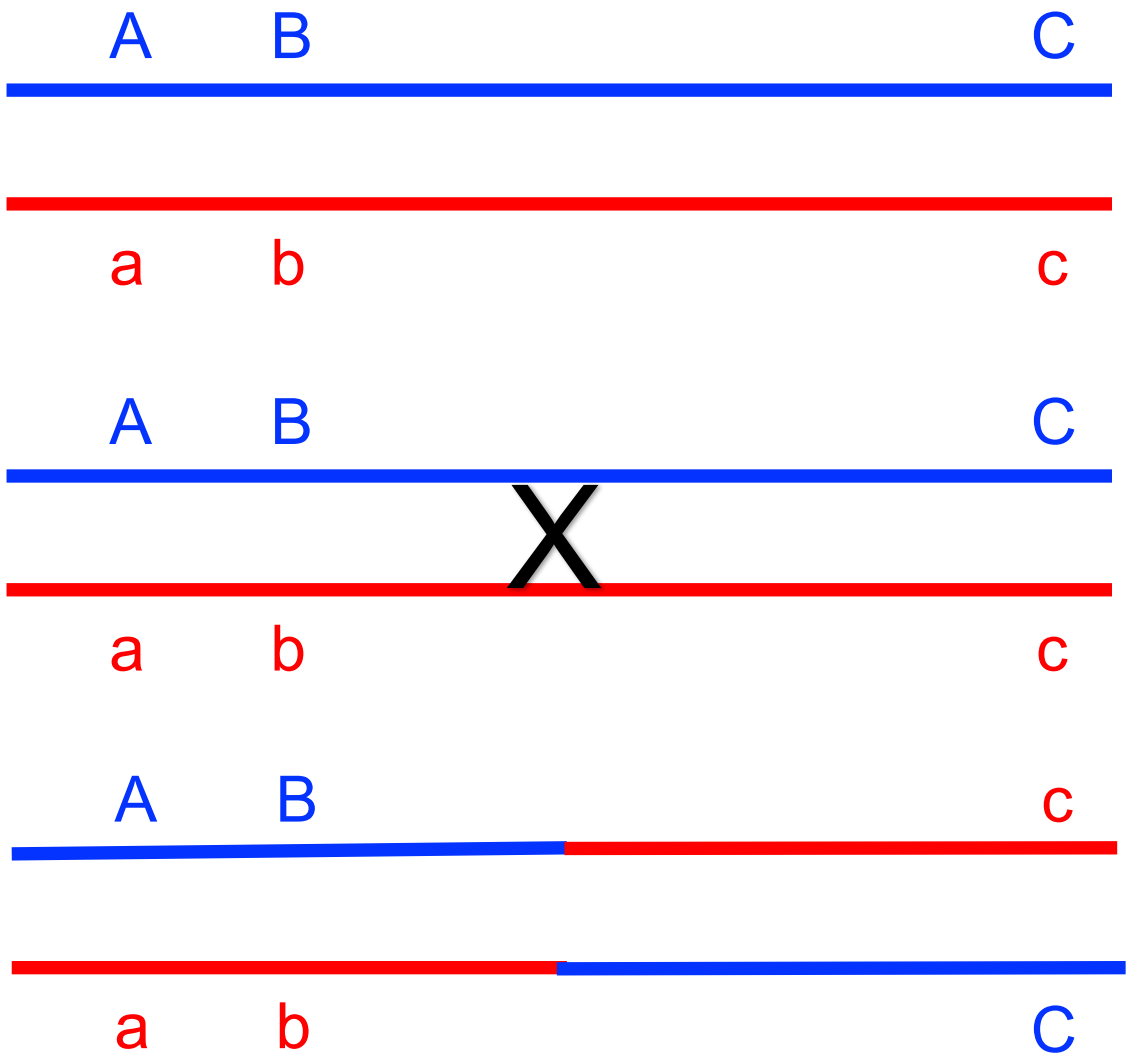
r_1





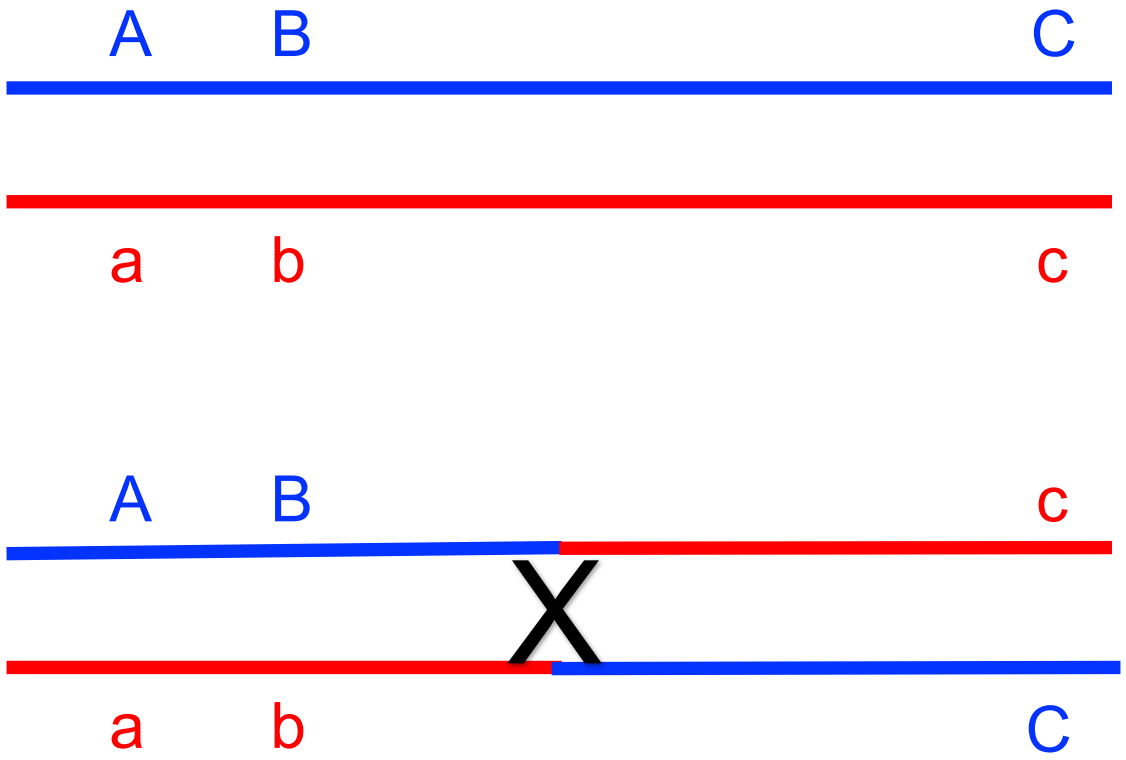
Now we'll consider correlations across loci…

3.0.5 Correlations between loci



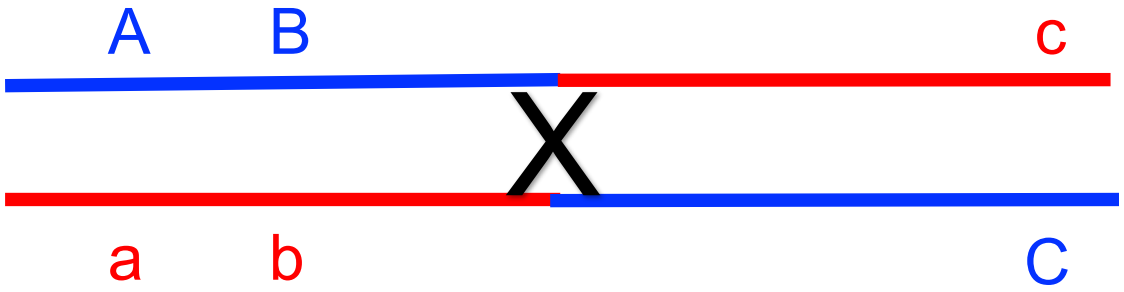
- Imagine a diploid organism with two chromosomes distinct at loci A, B, and C
- If recombination occurs at random spots along the chromosome, just by chance, it is more likely in the interval between B and C
- Alleles at A and B are therefore more correlated than those at B and C

3.0.5 Correlations between loci



- When recombination does not occur, offspring receive alleles intact from their parents
- When recombination does occur, alleles are shuffled into unique combinations not found in parental chromosomes, creating novel variation

3.0.5 Correlations between loci



r_{BP} = recombination rate (in Morgans) per basepair

L = Number of basepairs between loci

- Recombination, therefore, breaks up correlation/association between loci.
- The recombination fraction (r) is the probability of an odd number of crossing over events between loci in a single meiosis
- Typically, we'll be talking about very short chromosomal regions and recombination is rare:

$$r = r_{BP} L \ll \frac{1}{2}$$

3.0.5 Correlations between loci

- The term linkage disequilibrium (as Dr. Coop mentions, this is an awful, non-intuitive term) refers to the statistical non-independence of alleles at different loci within a population
- Let's consider two, biallelic loci that segregate for the alleles *A/a* and *B/b*
- We can represent the frequency of the two-locus haplotype AB, for example, as p_{AB} (and can do the same for the other 3 possible haplotypes)
- If these loci (A and B) are independent, then $p_{AB} = p_A p_B$, otherwise $p_{AB} \neq p_A p_B$
- The covariance between the *A* and *B* alleles is:

$$D_{AB} = p_{AB} - p_A p_B \qquad (3.16)$$

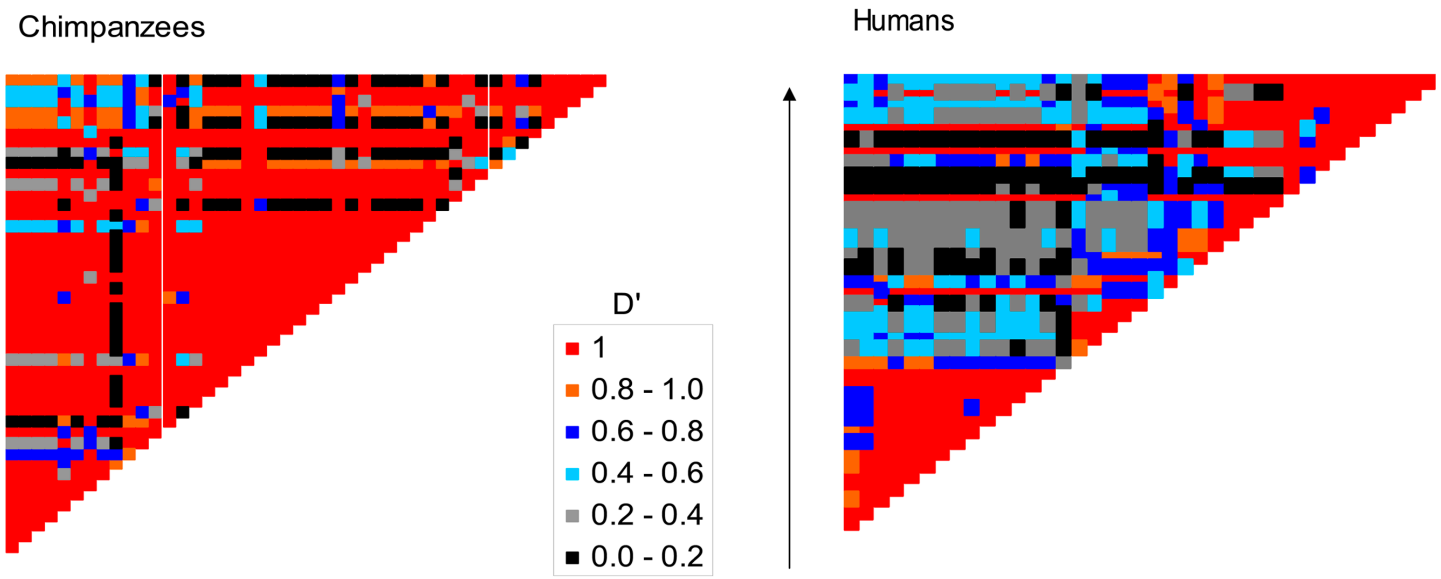
3.0.5 Correlations between loci

- The gametes that form with unrecombined alleles *(AB* and *ab*) are coupling gametes; those with recombined alleles (*Ab* and *aB*) are repulsion gametes
- Therefore, our *D* statistic measures the excess of coupling relative to repulsion gametes
- Only one *D* (D_{AB}, D_{ab}, etc…) needs to be known to know them all, so we'll simplify this to *D*

$$p_{AB} = p_A p_B + D. \qquad (3.17)$$

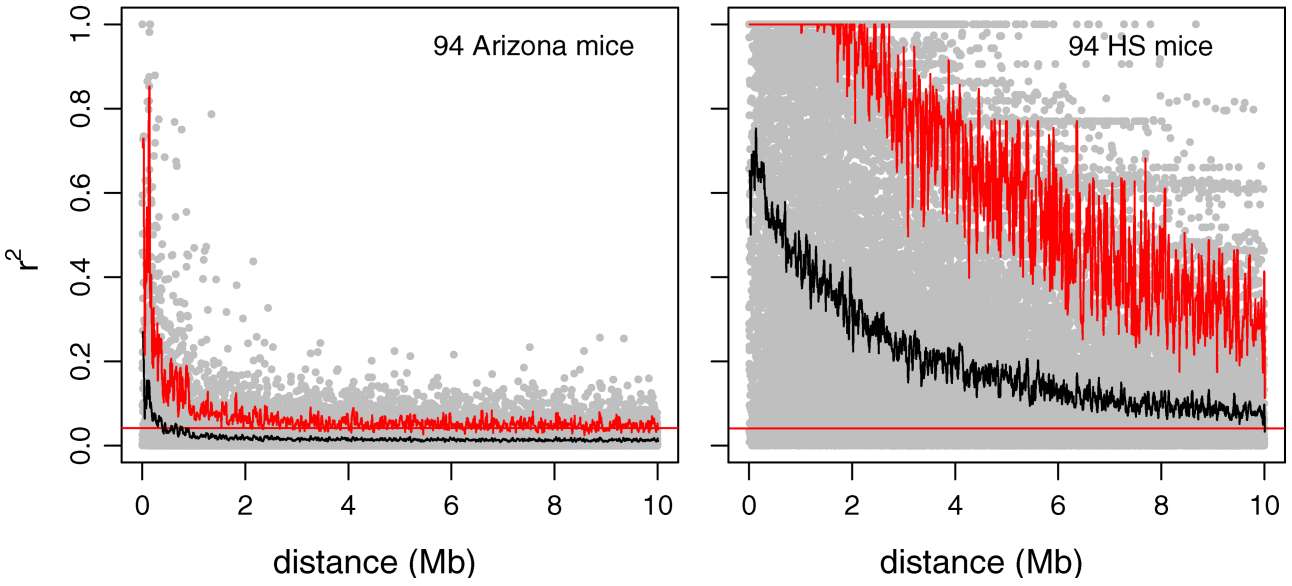
- If *D* is anything other than 0, there is linkage disequilibrium (LD) in our population
- Be aware that LD can also occur between loci on different chromosomes (more later…)

3.0.5 Correlations between loci



- *D* can vary dramatically across loci due to relative frequencies of alleles
- To address this, *D* is normalized to *D'* by dividing by it's maximum value given frequency; *D'* varies between -1 to 1
- Example from the *TAP2* locus

3.0.5 Correlations between loci



- Another very commonly used statistic to measure LD is the squared correlation coefficient, r^2 (note that this is NOT the recombination fraction r)

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \qquad (3.18)$$

- This is an example from mouse showing the decay of LD based on r^2 over distance along a chromosome

3.0.5 Correlations between loci

- Linkage disequilibrium can be created by natural selection favoring two alleles co-occurring, genetic drift randomly causing two alleles to co-occur, or through gene flow between populations with distinct combinations of alleles
- Let's consider, however, how LD can break down in the absence of these when only recombination is occurring
- Let's consider the frequency of the *AB* haplotype in the next generation to be p'_{AB}
- We lose a fraction *r* of our *AB* haplotypes per generation due to recombination, but gain a fraction rp_Ap_B due to the alleles recombining together. Thus:

$$p'_{AB} = (1 - r)p_{AB} + rp_Ap_B \qquad (3.19)$$

3.0.5 Correlations between loci

- A bit of rearranging and substituting shows that the Δp_{AB} is:

$$\Delta p_{AB} = p'_{AB} - p_{AB} = -rp_{AB} + rp_A p_B = -rD \qquad (3.20)$$

- Recombination will decrease the frequency of p_{AB} when the *AB* haplotype is common (*D > 0*) and increase the frequency when *AB* haplotypes are rare (*D < 0*)

- *D* in the next generation can be calculated as:

$$\begin{aligned} D' &= p'_{AB} - p'_A p'_B \\ &= (p_{AB} + \Delta p_{AB}) - (p_A + \Delta p_A)(p_B + \Delta p_B) \\ &= p_{AB} + \Delta p_{AB} - p_A p_B \\ &= (1 - r)D \end{aligned} \qquad (3.21)$$

3.0.5 Correlations between loci

- The change in *D* across generations can then be written as:

$$D_t = (1 - r)^t D_0 \qquad (3.22)$$

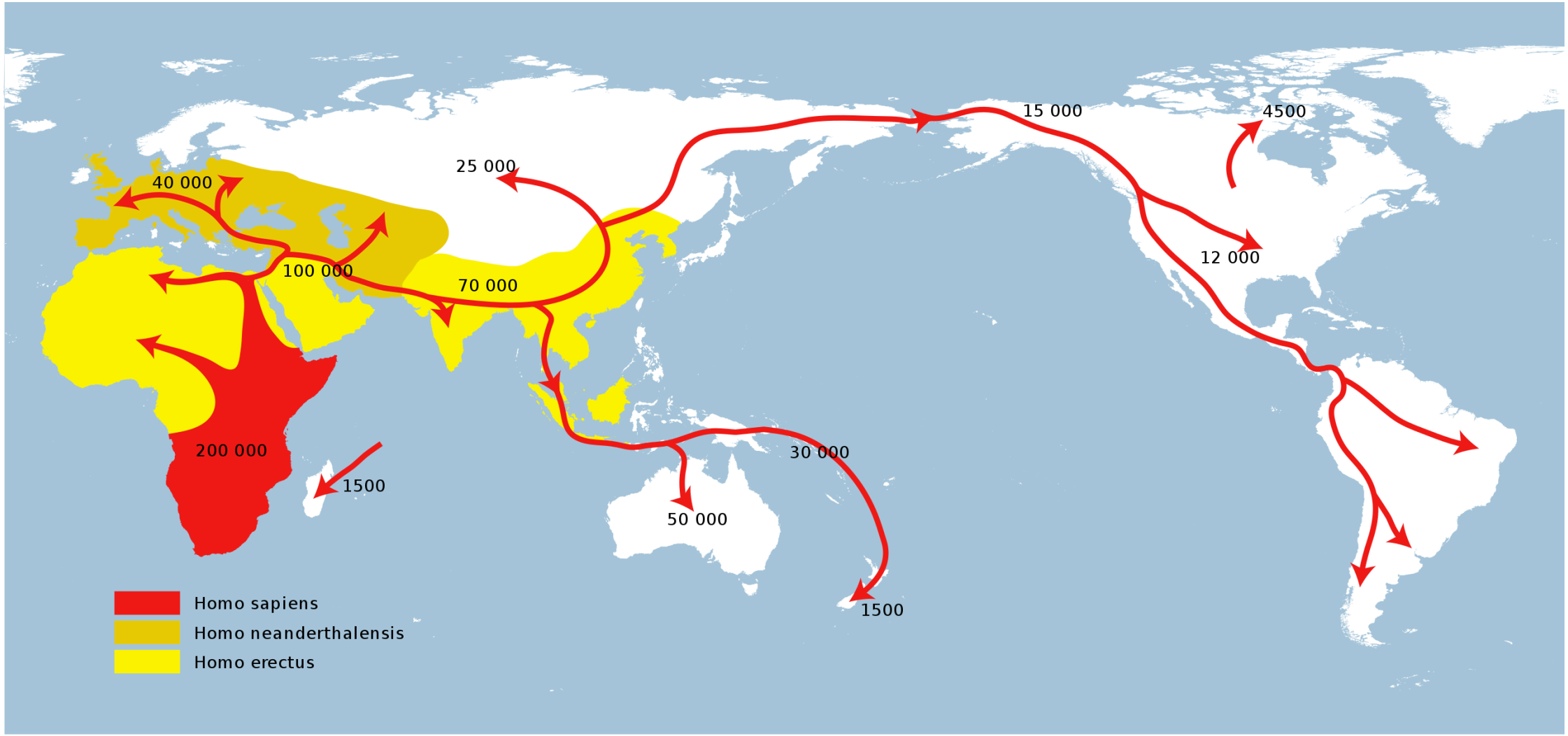
- This shows that recombination is acting to decrease LD, and this occurs geometrically at a rate of (1 - *r*).

- Since *r* is typically very small ($r \ll 1$), this can be approximated with an exponential:

$$D_t \approx D_0 e^{-rt} \qquad (3.23)$$

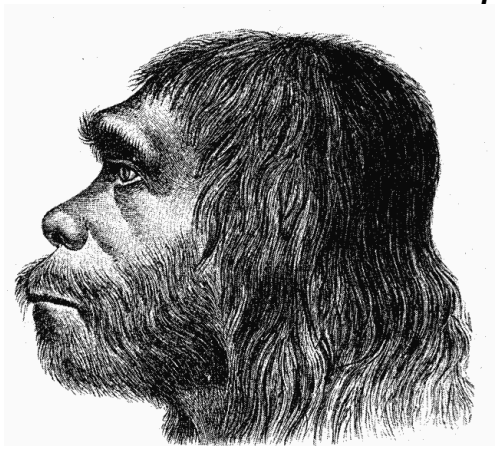
3.0.5 Correlations between loci

Empirical example: using LD to infer human history



3.0.5 Correlations between loci

Empirical example: using LD to infer human history



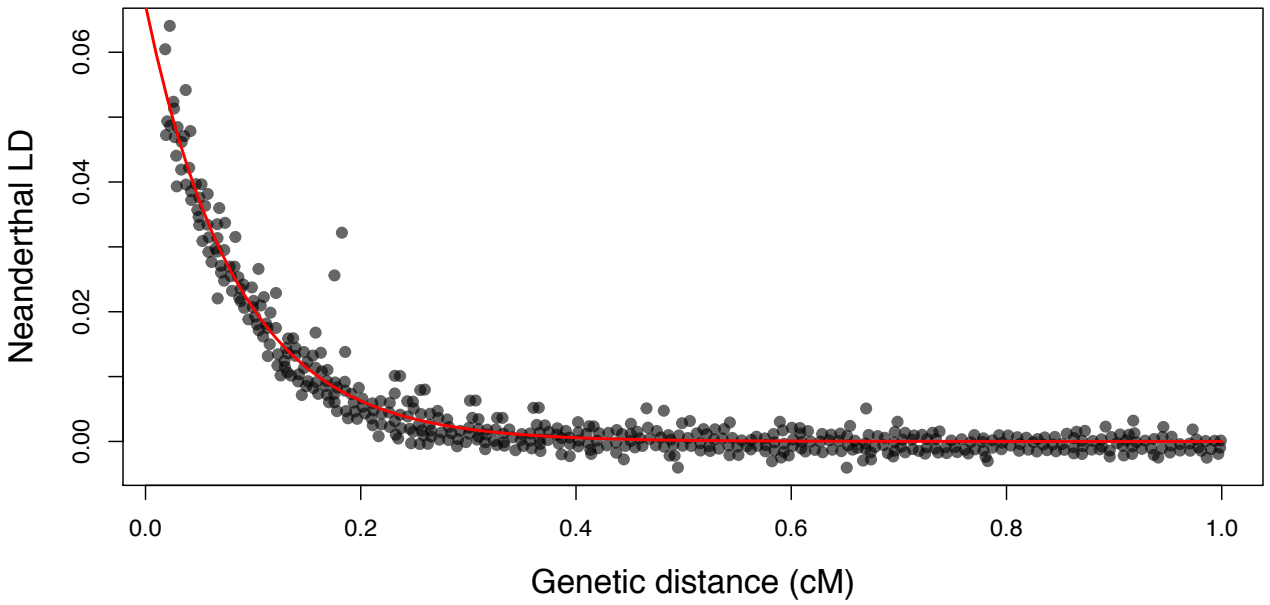
Neanderthal 1 by Hermann Schaaffhausen

Skullcap of Neanderthal 1
Musée de l'Homme, Paris

- When modern humans migrated out of Africa, they encountered and intermated with Neanderthals that were broadly distributed in Europe and Asia at that time
- Over time, these modern human-Neanderthal hybrids continued to mate with modern humans and Neanderthal tracts of DNA were broken up and reduced within the genome
- Humans from outside Africa typically have a small percentage of their genome that can be traced to Neanderthals

3.0.5 Correlations between loci

Empirical example: using LD to infer human history



- Sankararaman and colleagues (2012) were able to show the decay of LD in Neanderthal DNA in human genomes
- Based on this information, the human recombination rate *r,* and the use of equation 3.23, Coop estimates the timing of interbreeding was approximately 1200 generations in the past or about 35,000 BP
- Recombination has slowly broken up associations

3.0.5 Correlations between loci

Empirical example: using LD to infer human history

A video overview of Neanderthal population genetic research:

https://www.sciencelearn.org.nz/videos/1603-palaeogenomics-and-interbreeding