Logistics

Readings for Week 5: Hahn Chapter 5 “Population Structure” (pp. 104-110)

Lawson et al. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE barplots. Nature Communications. 9:3258

Novembre et al. 2008. Genes mirror geography within Europe. Nature. 456: 98-101

Next Quiz: Wednesday 10/12/2022 12:30 - 1:45 pm (covers Week 4 and Week 5)

Assignment 1 due: Thursday October 6 at midnight.

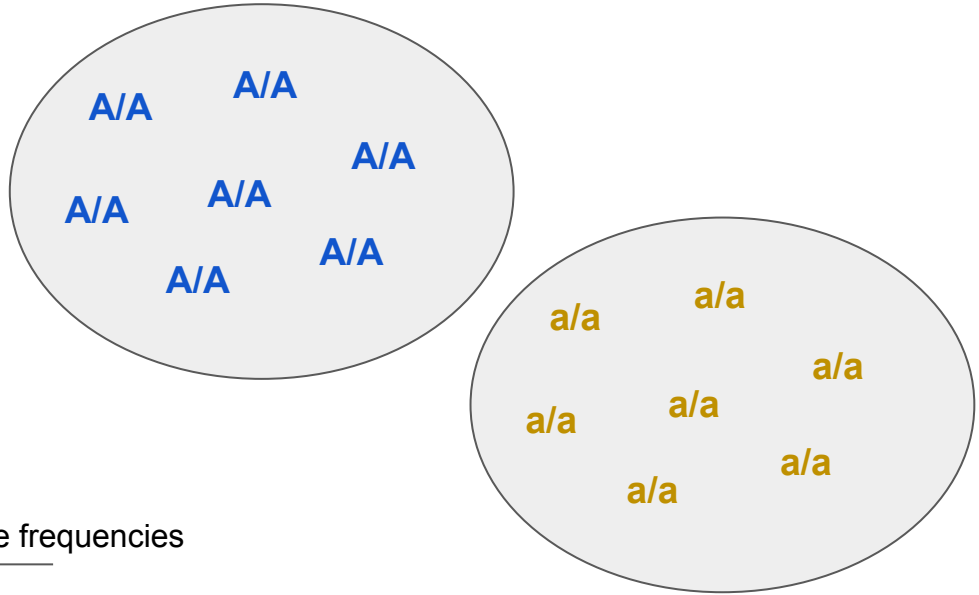
No Class next Monday, class next Tuesday

The Wahlund effect

Example: Consider two populations that are fixed for alternate alleles A with frequency p and a with frequency q

All individuals are homozygous for A in population 1 (p = 1, q = 0 in population 1) and a in population 2 (p = 0, q = 1 in population 2)

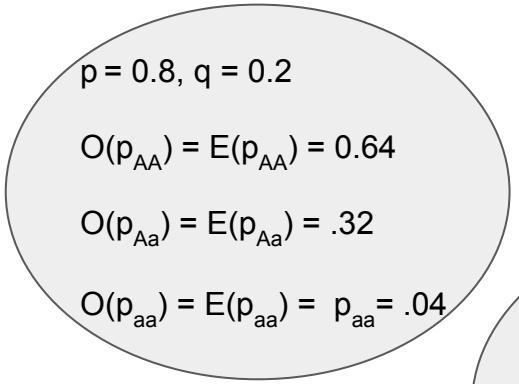
If we **combine** both populations, global allele frequencies are p = 0.5 and q = 0.5



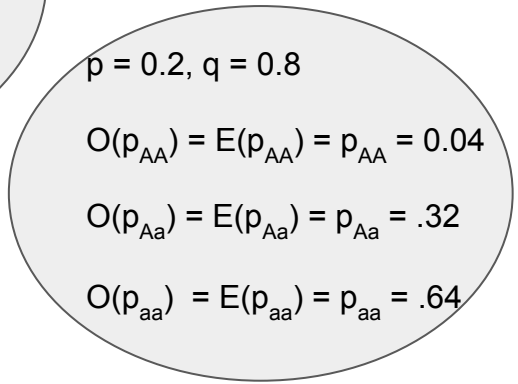
Expected genotype frequencies (under HWE)	Observed genotype frequencies
$E(p_{AA}) = p^2 = 0.25$	$Obs(p_{AA}) = 0.5$
$E(p_{Aa}) = 2pq = 0.5$	$Obs(p_{Aa}) = 0$
$E(p_{aa}) = 0.25$	$Obs(p_{aa}) = 0.5$

Wahlund Effect is observed when combining differentiated populations

The Wahlund effect



Two differentiated populations. Both are at HWE within populations because observed genotype frequencies match expectations under random mating within populations



p is frequency of A allele

q is frequency a allele

$O(p_{AA}, p_{Aa}, p_{aa})$ are observed genotype frequencies

$E(p_{AA}, p_{Aa}, p_{aa})$ are expected genotype frequencies

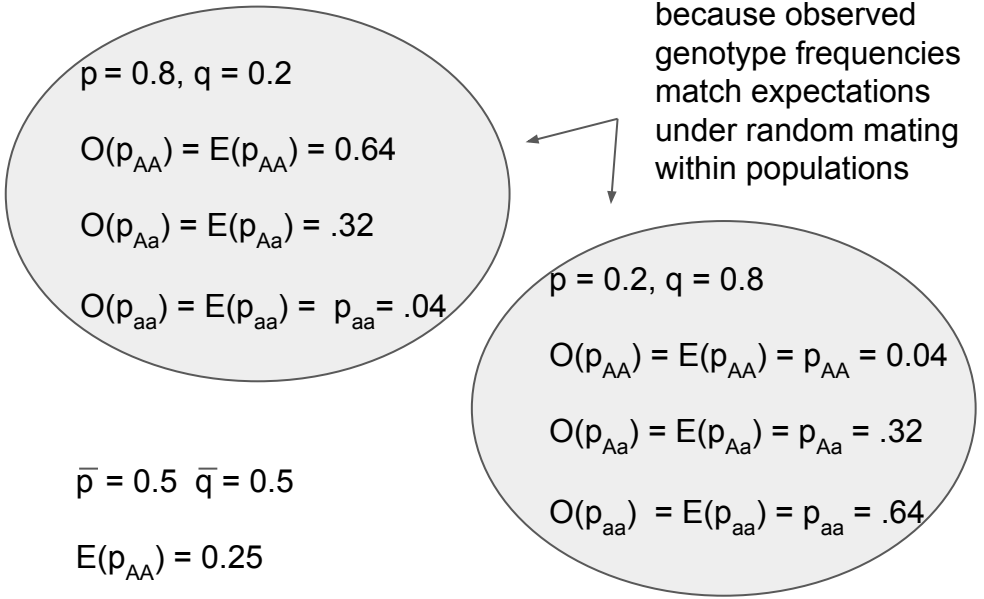
The Wahlund effect

Key point:

Note that each population is at HWE **within populations** because observed genotype frequencies match expected under random mating

However, the observed heterozygote frequency (0.32) is less than expected heterozygosity, $E(p_{Aa})$, of 0.5 under random mating among populations

The deficit of observed heterozygotes (considering the populations in aggregate as a single global population) is due to the Wahlund Effect (i.e., nonrandom mating among populations)



Two differentiated populations. Both are at HWE within populations because observed genotype frequencies match expectations under random mating within populations

Genotype expectations derived from global allele frequencies ($\bar{p}$ and $\bar{q}$), show that the expected frequency of heterozygotes is 0.5 **under HWE**

F_{st} is a measure of among population variance in allele frequencies

Recall that the among population variance in allele frequencies, σ^2, is a convenient way to quantify differences in allele frequencies among populations

This is because it measures the extent that populations considered in aggregrate deviate from HWE)

$$E(p_{AA}) = p^2 + \sigma^2$$

$$E(p_{Aa}) = 2\bar{p}\bar{q} - 2\sigma^2$$

The expected number of heterozygotes is less than HWE when $\sigma^2 > 0$

$$E(p_{aa}) = \bar{q}^2 + \sigma^2$$

*σ^2measures the extent populations differ from HWE

If σ^2 is 0, then genotype frequencies are at HWE.

The Wahlund effect

What is the among population variance in allele frequencies in this example?

$$\sigma^2 = \sigma_p^2 = \sigma_q^2 = \frac{\sum (p_i - \bar{p})^2}{n} = \frac{\sum p_i^2}{n} - \bar{p}^2$$

$$\sigma_p^2 = \frac{(0.8 - 0.5)^2 + (0.2 - 0.5)^2}{2} = 0.09$$

Therefore, we can calculate the expected number of heterozygotes given σ_p^2 as:

$$E(p_{Aa}) = 2\bar{p}\bar{q} - 2\sigma^2$$

$$= 2(0.5 * 0.5) - 2(0.09)$$

$$= 0.32$$

$p = 0.8, q = 0.2$

$O(p_{AA}) = E(p_{AA}) = 0.64$

$O(p_{Aa}) = E(p_{Aa}) = .32$

$O(p_{aa}) = E(p_{aa}) = p_{aa} = .04$

Two differentiated populations. Both are at HWE within populations because observed genotype frequencies match expectations under random mating within populations

$p = 0.2, q = 0.8$

$O(p_{AA}) = E(p_{AA}) = p_{AA} = 0.04$

$O(p_{Aa}) = E(p_{Aa}) = p_{Aa} = .32$

$O(p_{aa}) = E(p_{aa}) = p_{aa} = .64$

$\bar{p} = 0.5 \quad \bar{q} = 0.5$

$E(p_{AA}) = 0.25$

$E(p_{Aa}) = .5$

$E(p_{aa}) = .25$

Genotype expectations derived from global allele frequencies ($\bar{p}$ and $\bar{q}$), where $\bar{p}$ and $\bar{q}$ are average frequencies show that the expected frequency of heterozygotes is 0.5 under HWE

F_{st} is a measure of among population variance in allele frequencies

We can think of Fst as the percent of variation explained by population structure (i.e., the among-population component of variance)

A value of 0 indicates all diversity is explained by within population variation

A value of 1 indicates all genetic variation is explained by population structure

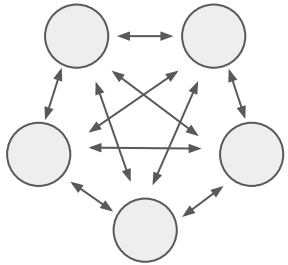
An F_{st} value of 0.15 typical of human populations means that 15% of diversity is partitioned between populations, while 85% of variation is within

Migration model of population differentiation

Under a set of assumptions about a population, we can draw inferences about migration rates from Fst

Wright's Infinite Island Model

1. Each population equally likely to exchange migrants
2. All populations are at migration-drift equilibrium
3. Alleles are shared because of migration not from common ancestry



Infinite island model

With these (and a few other) assumptions, the expected F_{ST} can be related to the effective number of migrants between populations as:

$$E(F_{ST}) = \frac{1}{1 + 4N_e m}$$

Under the migration model, low F_{ST} is due to high effective migration rates ($N_e m$), high F_{ST} is due to low $N_e m$

Migration model of population differentiation

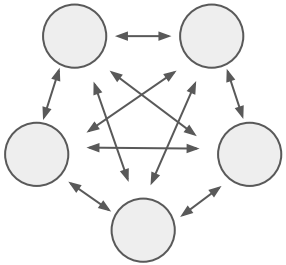
Do human populations conform to infinite island model?

Humans violate multiple assumptions of the Infinte Island Model, namely

1. human populations are not at migration-drift equilibrium

2. they are geographically structured such that geographically closer populations are less differentiated (implies migrants are not equally likely between all populations)

3. Allele sharing between populations likely due to common ancestry



Infinite island model

How is population differentiation impacted by natural selection?

Local adaption: positive selection restricted to a population(s)

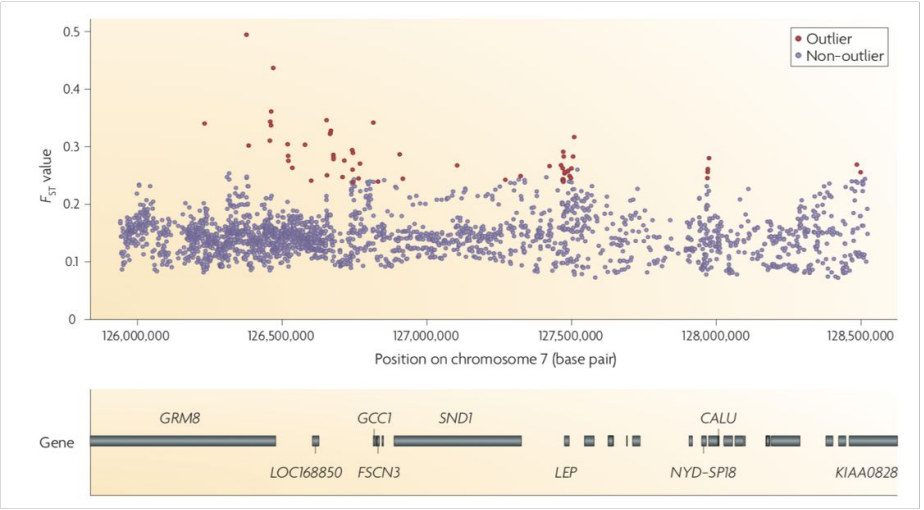
Positive selection increases frequency of locally adapted (beneficial) alleles

This increases the variance in allele frequencies between populations and increases F_{ST}

Genomic scans for selection using F_{ST}

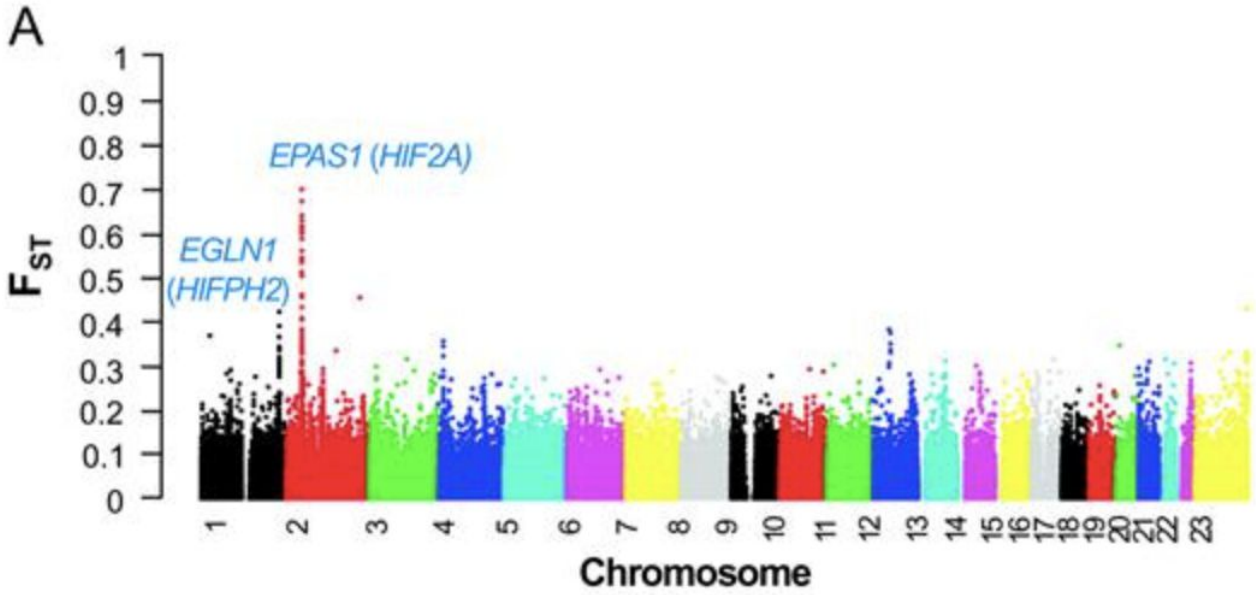
Genomic scans calculate F_{ST} for all SNP genomewide

SNPs subject to positive selection in a local population (but not all populations) are expected to have higher F_{ST}



Holsinger and Weir 2009

Example: local adaptation of hypoxia-related variants in Tibetans



Xu et al. (2011) Molecular Biology and Evolution

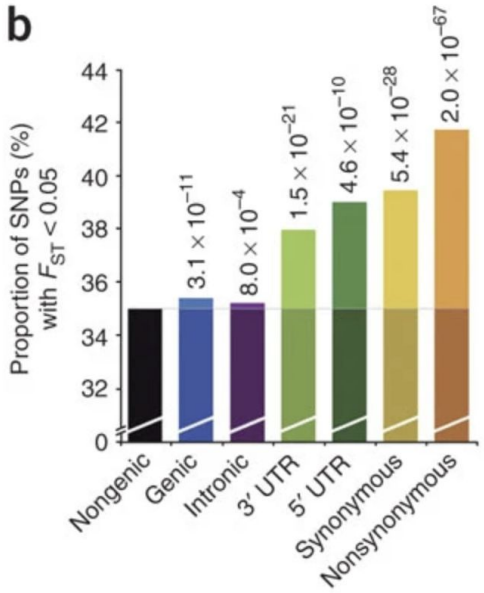
Negative selection on deleterious variants reduce F_{ST}

Negative selection against deleterious variants keeps allele frequencies low

Therefore allele frequencies are constrained from diverging between populations and FST is expected to be low

Example: Nonsynonymous (amino acid altering) SNPs have lower F_{ST} than other classes

Nonsynonymous SNPs are enriched in the low F_{ST} < 0.05 class



Barreiro et al. (2008) Nature Genetics

Example: Balancing selection at the Human Leukocyte Antigen (HLA) locus

A major constraint in whole genome scans of F_{ST} is that SNPs with low minor allele frequency have reduced maximum F_{ST}

One solution is to compare F_{ST} at SNPs with similar minor allele frequency

SNPs at HLA consistently have lower F_{ST} than the remainder of the genome

Consistent with balancing selection at HLA

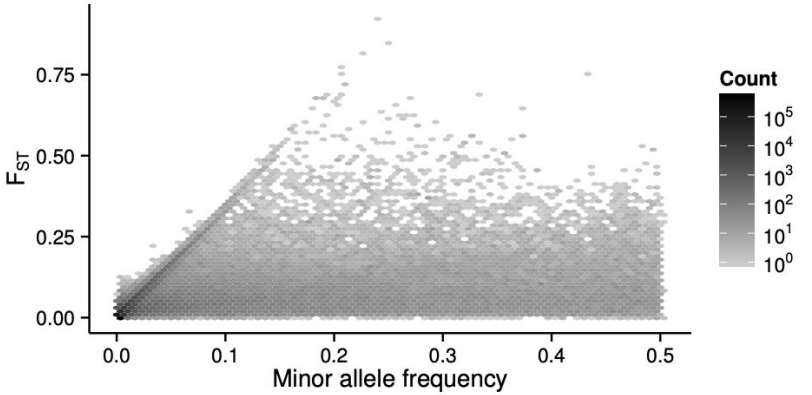


Figure 3 Population differentiation, measured by F_{ST}, as a function of minor allele frequency at biallelic exonic SNPs from the 1000 Genomes Project phase 3 data.

Brandt et al. (2018) G3

Example: Balancing selection at the Human Leukocyte Antigen (HLA) locus

A major constraint in whole genome scans of F_{ST} is that SNPs with low minor allele frequency have reduced maximum F_{ST}

One solution is to compare F_{ST} at SNPs with similar minor allele frequency

SNPs at HLA consistently have lower F_{ST} than the remainder of the genome

Consistent with balancing selection at HLA

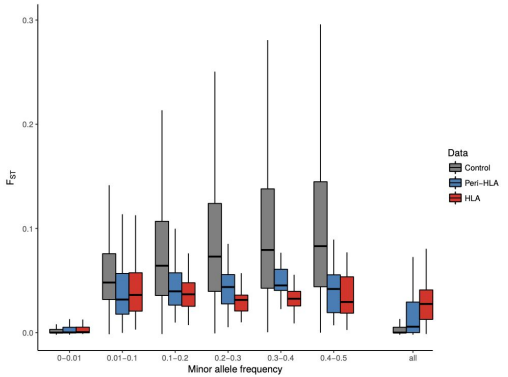


Figure 4 F_{ST} distributions per minor allele frequency (MAF) bin. HLA and Peri-HLA SNPs show lower F_{ST} than control SNPs in all bins with MAF > 0.01. Outliers (points above the 3rd quartile by 1.5 times the interquartile range, or below the 1st quartile by the same amount) were removed from figure, but not from statistical test, for better visualization. Figure S7 shows F_{ST} distributions including outliers.

Brandt et al. (2018) G3

How does natural selection impact F_{ST}?

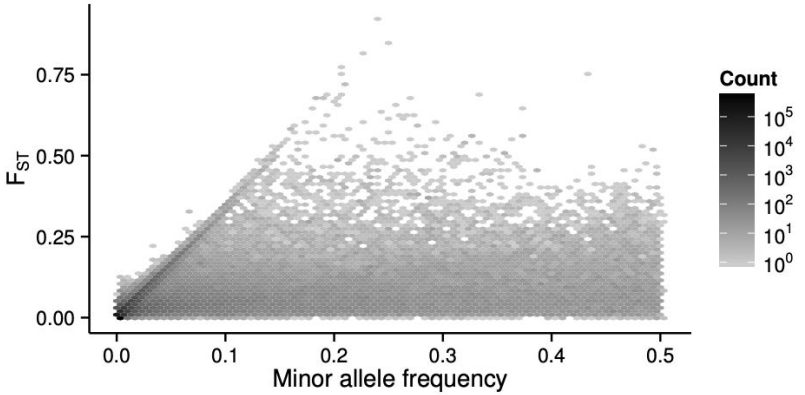
Type of selection	Impact on selected loci (relative to neutrality)	Cause
Positive selection (i.e., local adaptation)	Increases F_{ST}	Positive selection increases allele frequency, but only where allele is beneficial
Strong negative selection	Strong decrease in F_{ST}	Strongly deleterious Alleles are constrained to very low frequency classes in all populations
Weak negative selection	Slight decrease in F_{ST}	Weakly deleterious are not free to drift to high frequencies and therefore cannot diverge in frequencies compared to neutral
Balancing selection	Decrease in F_{ST} (but depends on scenario)	Assuming similar selection strength in different populations, balancing selection causes allele to reach similar frequencies in each population

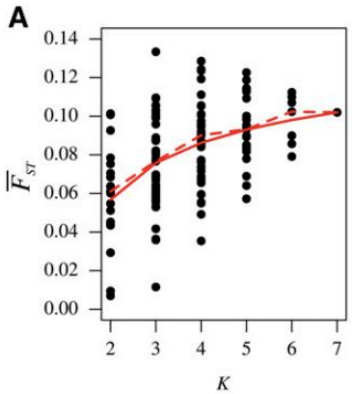
Problems with F_{ST}

F_{ST} is constrained by (1) number of populations sampled (2) minor allele frequency (3) depends on within population diversity (see Week 4)

F_{ST} calculated between pairs of populations is lower on average then F_{ST} calculated between >2 populations (Alcala and Rosenberg 2017, Genetics)

F_{ST} is constrained at low minor allele frequency (e.g., Brandt et al. 2018, G3)

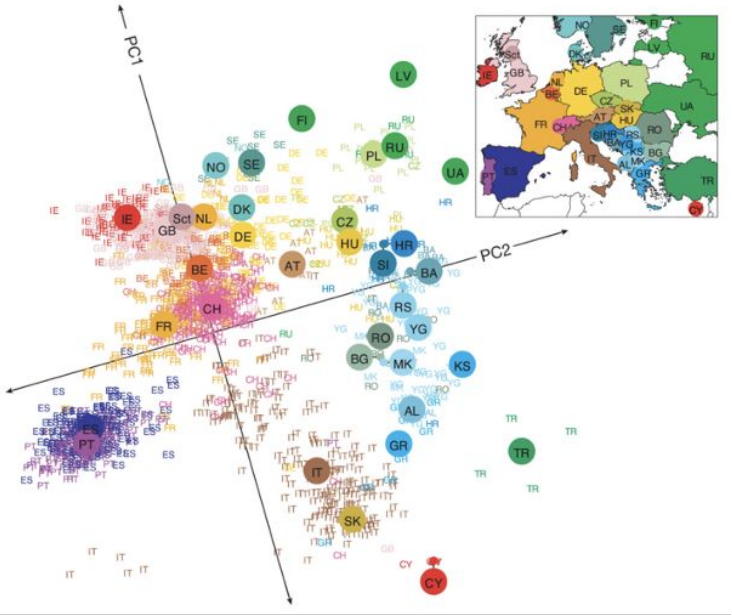


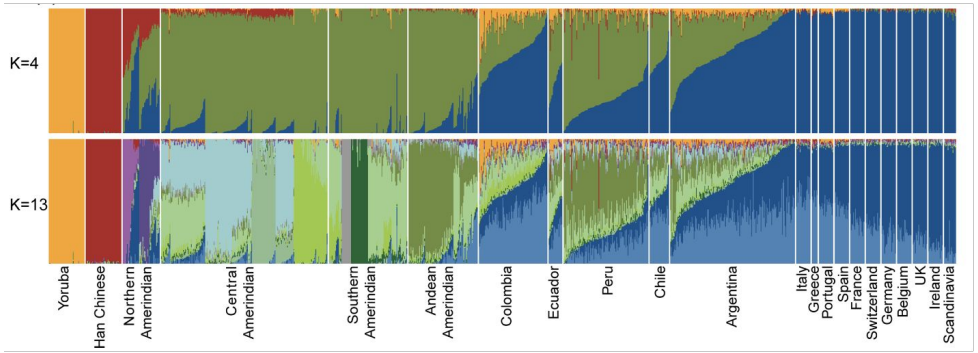


K is the number of populations used to calculate F_{ST}

Defining populations

Wednesday, October 6





Logistics

Readings for Week 5: Hahn Chapter 5 “Population Structure” (pp. 104-110)

Lawson et al. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE barplots. Nature Communications. 9:3258

Novembre et al. 2008. Genes mirror geography within Europe. Nature. 456: 98-101

Next Quiz: Wednesday 10/12/2022 12:30 - 1:45 pm (covers Week 4 and Week 5)

Assignment 1 due: Thursday October 6 at midnight.

No Class next Monday, class next Tuesday

Definitions

Admixture: the mixing of differentiated populations the product of which is individuals with mixed ancestry

Principal Component Analysis in Population Genomics

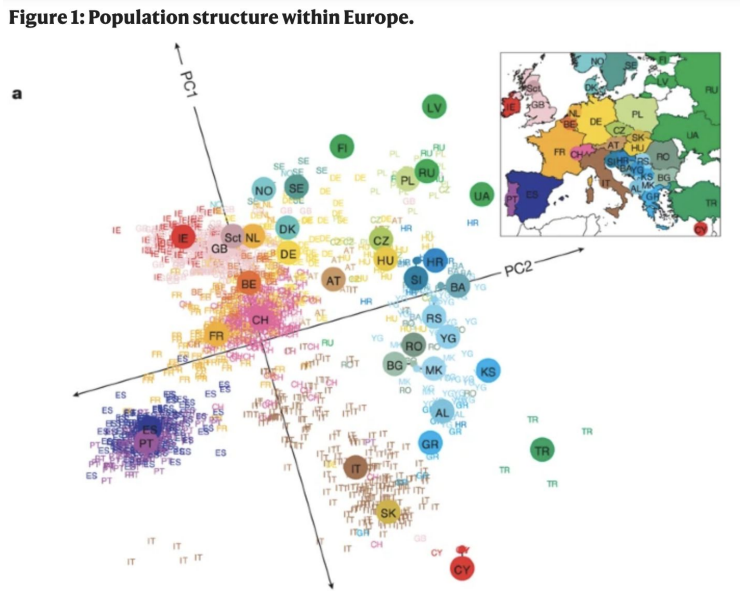
PCA reduces dimensions from millions of SNPs to a finite number of uncorrelated (orthogonal) variables termed principal components (PCs)

First PC captures the most variation, each PC describes a decreasing proportion of the genetic variation

Genotypes are then projected onto space spanned by the PC axes, which allows visualizing the samples and their distances from one another in a scatter plot

In this visualization, sample overlap is usually interpreted as identity, due to common origin or ancestry

PCA's most attractive property for population geneticists is that the distances between clusters allegedly reflect the genetic and geographic distances between them.



Novembre et al. (2008) Nature

Principal Component Analysis in Population Genomics

PCA has been a primary tool used by population geneticists originally 1963 but gained prominence in 2006

Often used for exploratory analysis but increasingly has been considered as part of primary results owing to the complexity of analyzing whole genome SNP data

Update: A recent paper (Elhaik 2022, Scientific Reports) has challenged the general the use of PCA arguing that results may not be robust or reproducible (i.e., resampling same populations may yield different results)

Elhaik (2022) Scientific Reports

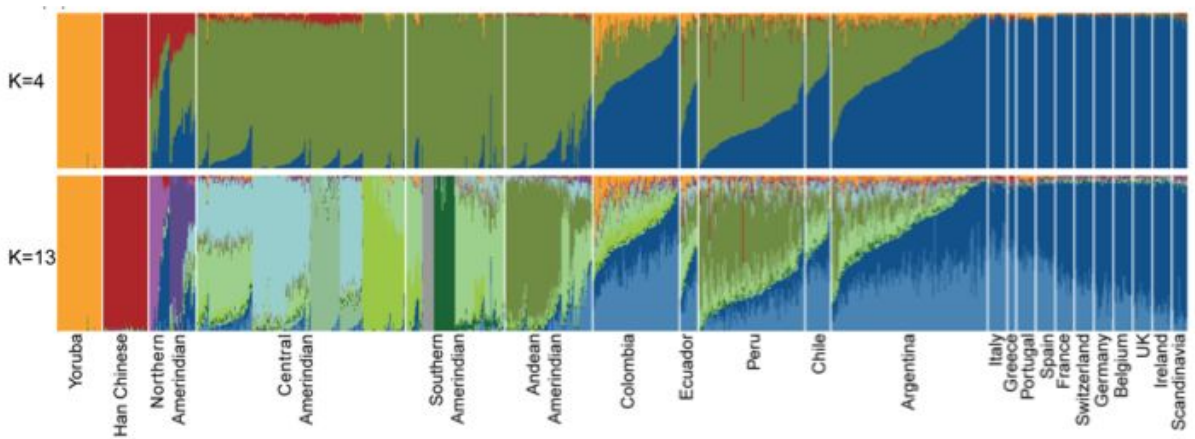
Population structure inference using model-based clustering

Clustering of individuals is based on a population genetic model

Ancestry is assigned to individuals in a way that minimizes deviations from Hardy-Weinberg and maximizes linkage equilibrium

The standard implementation of these algorithms is unsupervised, with no prior knowledge of ancestry

Samples are assigned to K clusters, where K is the desired number of groups specified by the user



Homburger et al. (2015) PLOS Genetics

Model-based clustering: the “ancestry bar plot”

Clustering of individuals produces as its primary output an “ancestry bar plot” (also referred to as a “STRUCTURE diagram”, “ADMIXTURE plot” or other related terms)

Each panel is a different K=4 and K=13 (a user-defined number of clusters)

Each vertical bar shows ancestry proportions (“admixture proportions”) for one individual

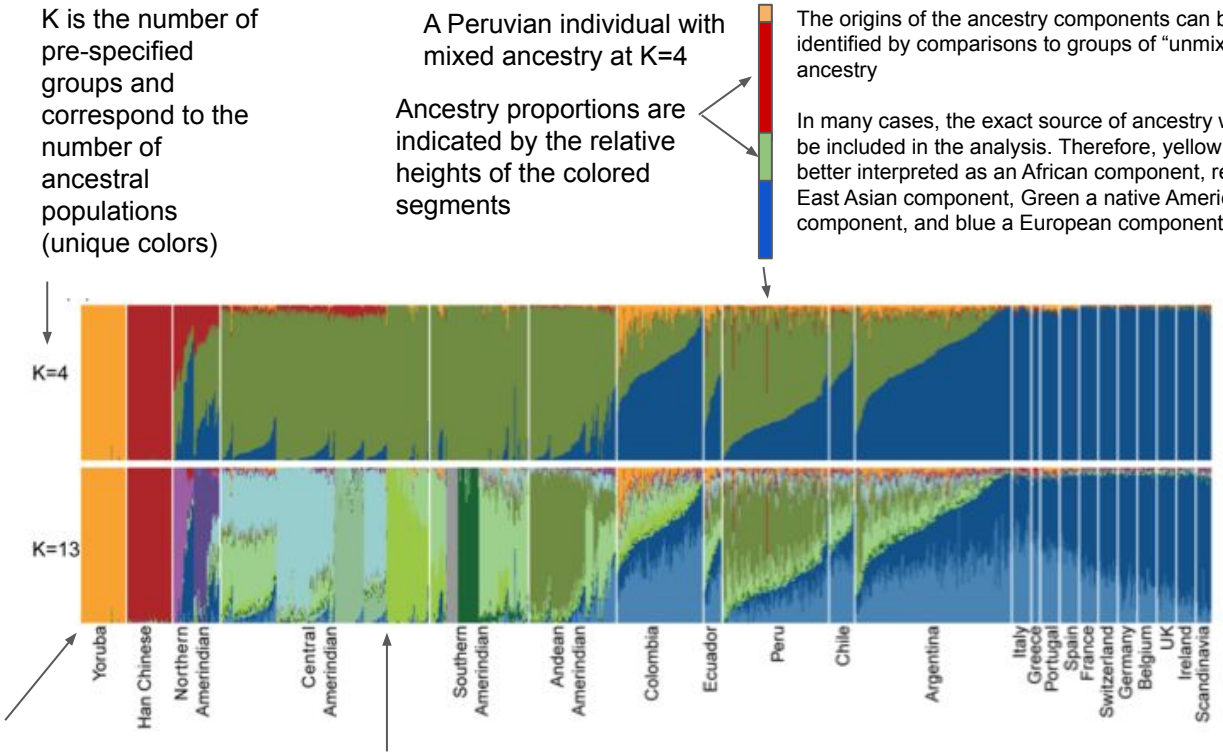
K is the number of pre-specified groups and correspond to the number of ancestral populations (unique colors)

A Peruvian individual with mixed ancestry at K=4

Ancestry proportions are indicated by the relative heights of the colored segments

The origins of the ancestry components can be identified by comparisons to groups of “unmixed” ancestry

In many cases, the exact source of ancestry won’t be included in the analysis. Therefore, yellow is better interpreted as an African component, red an East Asian component, Green a native American component, and blue a European component



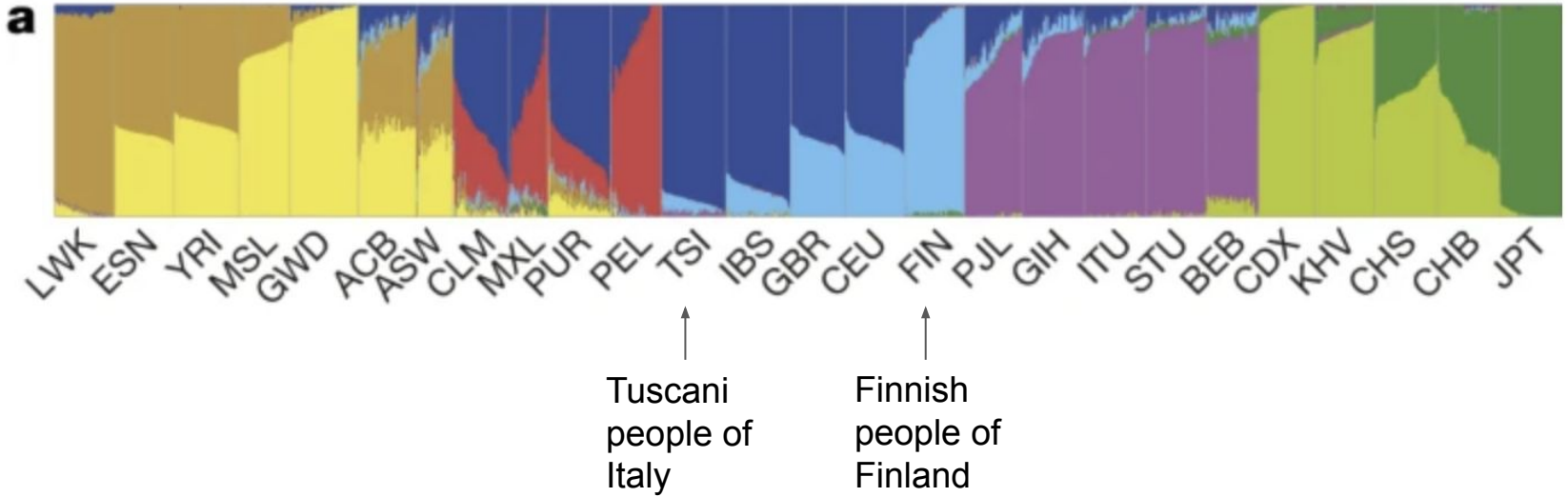
Once ancestries are estimated, like individuals are arranged together in the barplot for illustration purposes

Each individual is represented by a vertical bar

Homburger et al. (2015) PLOS Genetics

Model-based clustering of 1,000+ human genomes

Figure 2: Population structure and demography.



1000 Genomes Consortium (2015) Nature

Model-based clustering of population genomic data

Unsupervised model-based clustering approaches cluster individuals based on their genetic ancestry

Original method called STRUCTURE (Pritchard et al. 2000), FRAPPE (Tang et al. 2005) and ADMIXTURE (Alexander et al. 2009) are faster implementations

Use **population principles** to identify groups of individuals that meet expectations of random mating (e.g., maximize **Hardy-Weinberg equilibrium** and **linkage equilibrium** within populations)

Use a combination of Markov Chain Monte Carlo and Bayesian inference to estimate model parameters

What do the algorithms do?

(1) Estimate allele frequencies at each locus in *K* (a number defined by the user) ancestral populations

(2) Probabilitistically assign the membership of each sample to an ancestral population (or jointly to two or more populations if observed genotypes support mixed ancestry)

(3) Return the posterior probability Pr(Ancestry|K,Data) (i.e., the probability of the ancestry given K and Data) and parameter estimates (i.e., ancestry proportions)

What do the algorithms do?

These models are designed to capture two phases of population history

Phase I is a divergence phase where populations are isolated for some period of time and when allele frequencies divergence

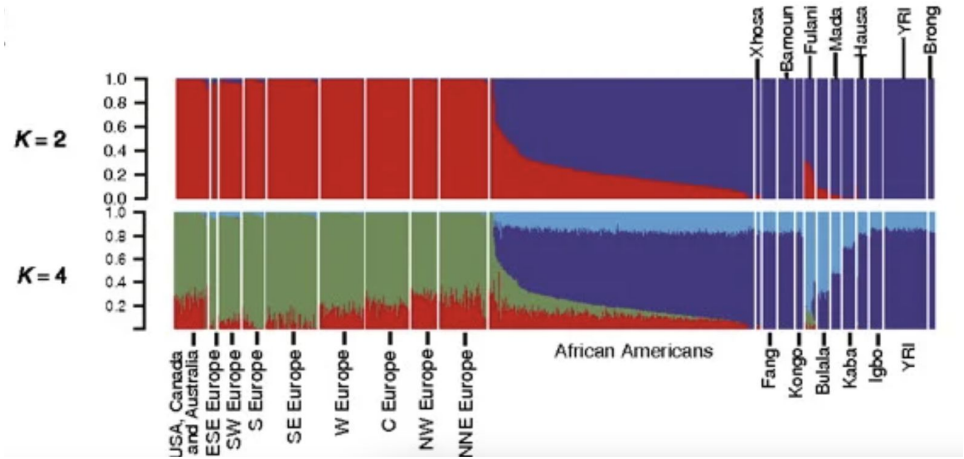
Phase II is an admixture phase where populations mixing (i.e., gene flow) occurs between the populations creating individuals with mixed ancestry

Example: African Americans

African Americans are an example well-suited to STRUCTURE or ADMIXTURE-type analysis

Many African Americans genetic ancestry traces to West Africa and Europe, two populations with unique histories (and different allele frequencies)

The genetic traces of this history are visible in the mixed ancestry inferred by STRUCTURE and ADMIXTURE-type methods



Bryc (2009) PNAS

Understanding the model(s)

Multiple models are detailed in the original STRUCTURE and subsequent papers

Admixture and Linkage models are most common

Most common admixture model assumes no linkage (hence the need for LD-pruned SNPs)

Models assumes a ‘divergence phase’ between discrete ancestral populations and an ‘admixture phase’ that produced observed samples

All sampled individuals are a result of *K* homogeneous ancestral populations with random mating within populations (inbreeding organisms are not accomodated)

Ancestral population should be well represented as unadmixed individuals in the data (no ghost populations)

Assume no drift after the admixture event(s)

LD Pruning

LD pruning is the process of removing highly correlated SNPs in tight linkage

This is done using sliding window based approach that removes all but one of the SNPs in a correlated set

PCA and STRUCTURE-type analysis are often best performed with LD pruned SNP data because over-representation of SNPs with high LD can impact the analysis (i.e., lead to inferring ancestry in the high LD regions and not the entire genome)

Example tools: SNPRelate (R package) or PLINK (Unix command line tool)

How to perform a STRUCTURE/ADMIXTURE-type analysis?

- Step 1: Prune genotype data to remove SNPs in linkage disequilbrium (e.g., using PLINK or SNPRelate)

- Step 2: Run clustering procedure for different K with sufficient MCMC chain length to allow convergence

Step 3: Repeat step 2 many times to ensure consistency (no multi-modality in parameter estimates)

Step 4: Evaluate to identify suitable K value(s) and evaluate the fit of the data to the underlying model

Implementing ADMIXTURE

Run ADMIXTURE software with 10 iterations of a cross-validation procedure to assist in identifying an appropriate K

Example:

At Unix command line:

```
for K in 2 3 4 5 6; do
admixture –cv=10 <genotype file> $K
done
```

Primary output is the “Q matrix”

K=2

Sample	cluster1	cluster2
1	0.99	0.01
2	0.49	0.51
3	1.0	0.0

The proportion of ancestry originating from population2

Admixed sample (at K=2)

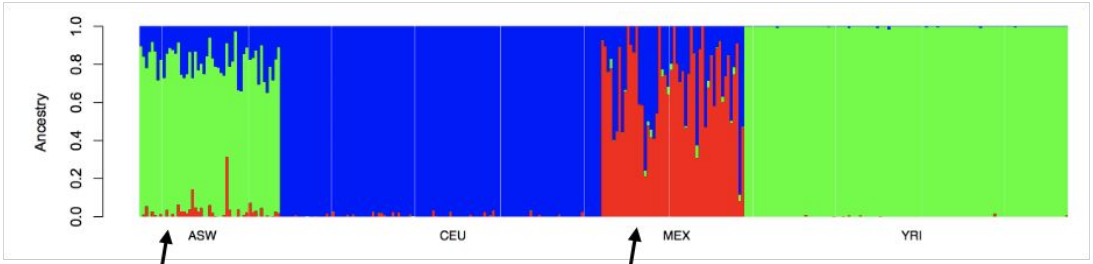
K=3

Sample	cluster1	cluster2	cluster3
1	0.01	0.98	0.01
2	0.45	0.55	0.0
3	0.01	0.99	0.0

The ancestry diagram is a visual representation of the Q matrix

Example: ADMIXTURE analysis of 4 human populations

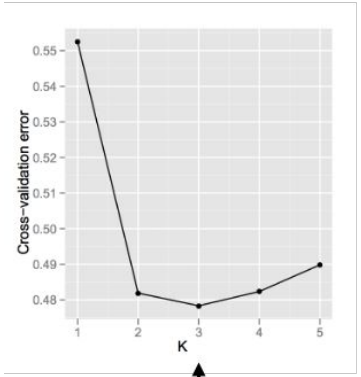
Ancestry diagram for K=3



Admixed population with YRI and CEU components

Admixed population with CEU and an unknown source

Cross-validation analysis



Cross-validation error minimized at K=3

Evaluating results: How many *K* are there?

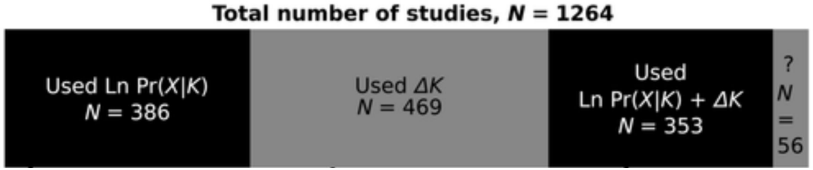
How to decide on how many *clusters?*

In theory, the Pr(X|K), (i.e., the likelihood of the data for a given K)

Typically average Ln Pr(X|K) across all iterations of the estimated for each K and to choose the maximum (if there is a clear maximum)

An alternative is the “Evanno method” that uses the rate of change in Pr(X|K) to determine an “optimal” K (Idea is to identify K where Pr(X|K) plateaus)

Other approaches involve statistical methods



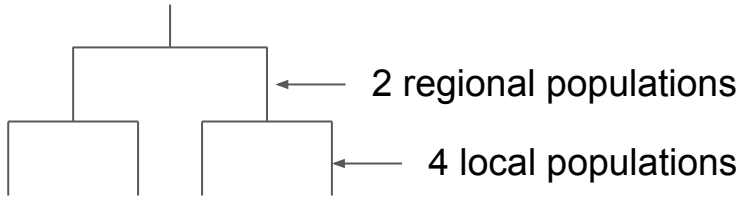
Janes et al. (2017) Molecular Ecology

Evaluating results: How many K are there?

In many cases, the biological interpretation of K may not be straightforward and there may not be a “true” K

For example, if structure is hierarchical, then knowledge of structure at lower K may provide different insight into the question then higher K

Interpretations frequently combine observations from multiple K values together with external knowledge



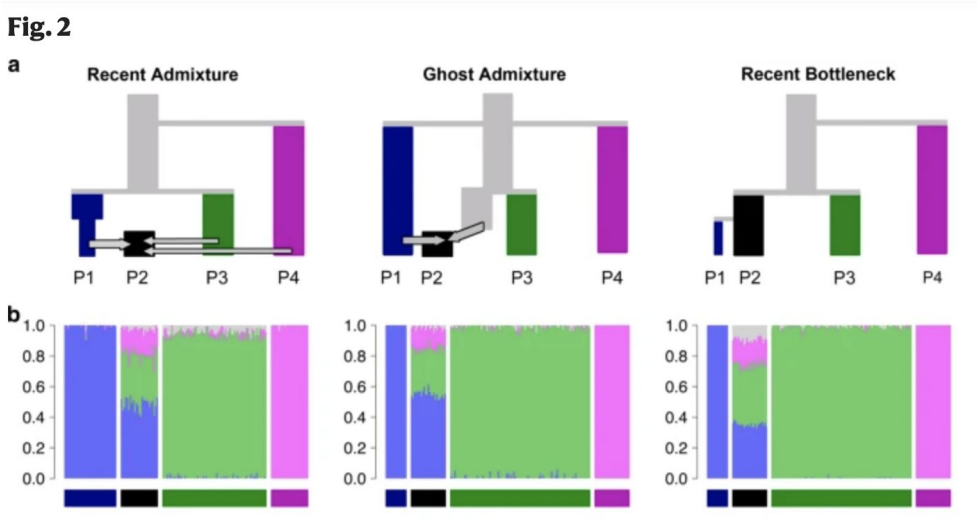
Example: hierarchically structured populations

*if we run STRUCTURE/ADMIXTURE at K=2 we would expect to see two groups representing two regional populations. With K=4 we would expect to see 4 groups. Both K values are interesting and telling us something about population structure in the data.

Different historical scenarios produce identical results

Example: Three very different simulated demographic scenarios, all yielding same STRUCTURE-type diagram

Population history often doesn't conform to a divergence phase-mixture phase scenario to which these methods are best suited



Lawson et al. 2018 Nature Communications

Different historical scenarios produce identical results

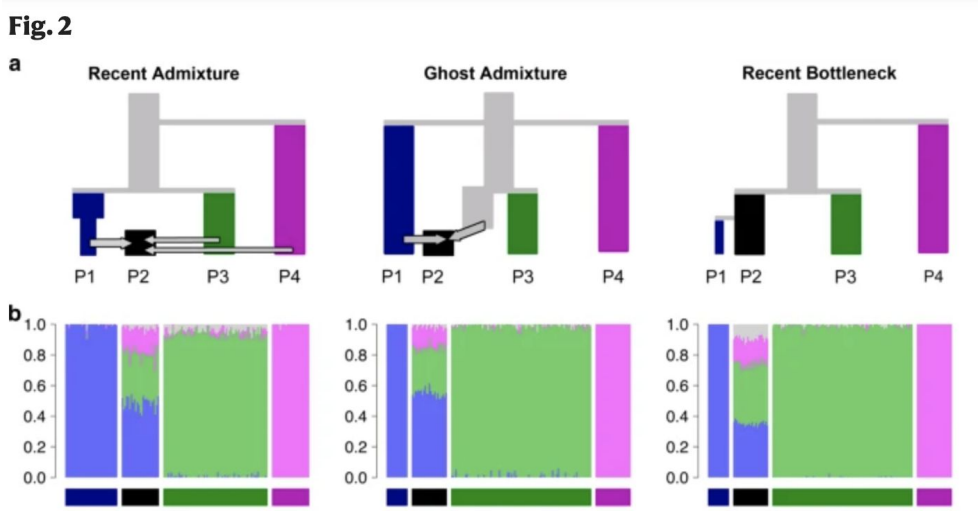
Example: Three very different simulated demographic scenarios, all yielding same STRUCTURE-type diagram

Population history often doesn't conform to a divergence phase-mixture phase scenario to which these methods are best suited

(1) Analysis is dependent on sampling relying on a majority of unadmixed individuals representing each of the ancestral populations

(2) Complex demographic scenarios can lead to ambiguous (which unfortunately are often treated as unambiguous by data analysts)

(3) STRUCTURE/ADMIXTURE are a tool, but often require additional statistical analysis to confirm admixture (see Week 6)



Lawson et al. 2018 Nature Communications

SMARTSNP Package (Herrando-Perez et al. 2021)

A new R-based implementation of the standalone SMARTPCA tool

https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13684