F-statistics and admixture

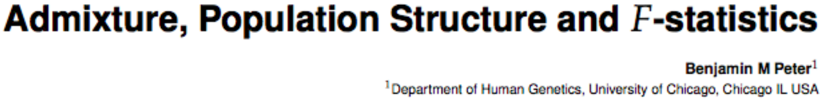
Fernando Racimo

Adelaide, January 2018

Today

- F2 statistics
- Outgroup F3 statistics
- Admixture F3 statistics
- F4 and D-statistics
- qpWave / qpAdm

An excellent resource!



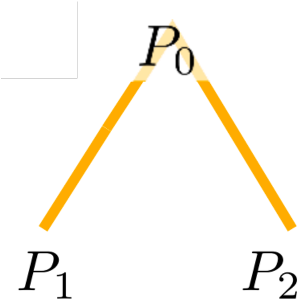
Admixture, Population Structure and F-statistics

Benjamin M Peter[1]

[1] Department of Human Genetics, University of Chicago, Chicago IL USA

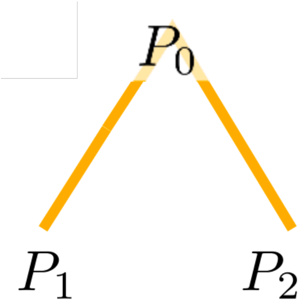
F_2 statistics

- Let's imagine we have two populations: P_1 and P_2
- At a particular site, the allele frequency of a (randomly chosen) allele is denoted as p
- $\mathbf{F_2(P_1, P_2) = E[(p_1 - p_2)^2]}$



F_2 statistics

- Let's imagine we have two populations: P_1 and P_2
- At a particular site, the allele frequency of a (randomly chosen) allele is denoted as p
- $\mathbf{F_2(P_1, P_2) = E[(p_1 - p_2)^2]}$



F_2 statistics

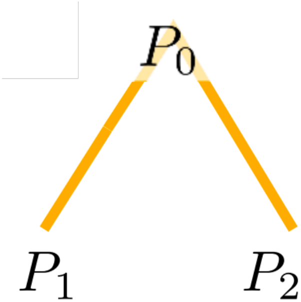
- $F_2(P_1, P_2) = E[(p_1 - p_2)^2]$
- E[ ] denotes an expectation.
- This expectation is over multiple **independent runs of the evolutionary process** of an allele. In practice, we don't have multiple runs.
- However, we can look at **multiple sites** across the genome
- Sites are not exactly independent (due to linkage), but we'll later see ways to account for this problem

F_2 statistics

- $F_2(P_1, P_2)$ can also be interpreted as a variance
- $Var[p_1 - p_2] = E[(p_1 - p_2)^2] - (E[p_1 - p_2])^2$
- But E[p1 - p2] = E[p1] - E[p2] = E[p0] - E[p0] = 0
- So Var[p1 - p2] = E[(p1 - p2)2], and therefore:
- $\mathbf{F2(P1, P2) = Var[p1 - p2]}$

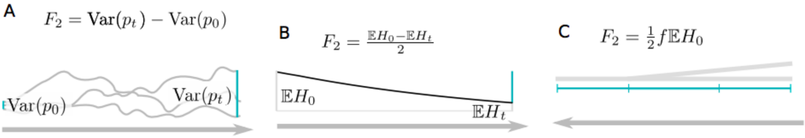
F_2 additivity

- If we consider a common ancestral population P_0, then:
- $F_2(P_1, P_2) = F_2(P_1, P_0) + F_2(P_2, P_0)$



F_2 as a measure of genetic drift

- If we compute an F_2 statistic between an ancestral and a descendant population, we can consider an F2 statistic to be:
  - A measure of the increase in allele frequency variance over time
  - A measure of the decrease in heterozygosity over time
  - A measure of the probability that two gene copies in the descendant population originate from a single copy in the ancestral population
  - In essence, a measure of **genetic drift** (time scaled by population size) or "population inbreeding"

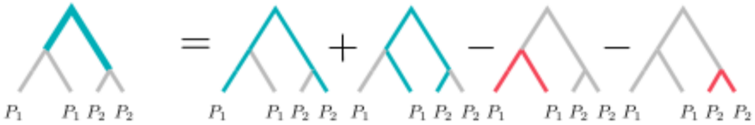

A
$F_2 = \mathrm{Var}(p_t) - \mathrm{Var}(p_0)$
$\mathrm{Var}(p_0)$
$\mathrm{Var}(p_t)$
B
$F_2 = \frac{\mathbb{E}H_0 - \mathbb{E}H_t}{2}$
$\mathbb{E}H_0$
$\mathbb{E}H_t$
C
$F_2 = \frac{1}{2} f \mathbb{E}H_0$

A coalescent interpretation

A. Equation

$$2\,F_2(P_1, P_2) = \theta\Big(\mathbb{E}T_{12} + \mathbb{E}T_{12} - \mathbb{E}T_{11} - \mathbb{E}T_{22}\Big)$$

B. Concordant genealogy

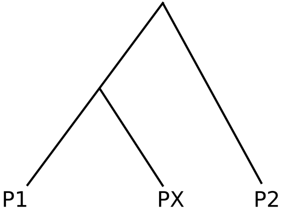


A coalescent interpretation

- The estimator for F2 can be written as a function of Tajima's estimator for θ (π)
- $F2 = \theta(E[T_{12}] - \frac{E[T_{11}]+E[T_{22}]}{2})$
- We know (from the first class) that θT is the expected number of differences between two sequences separated by time T, under the infinite sites model
- We also know (from the first class) that, an estimator for the expected number of differences for two sequences is π
- $\theta \hat{T_{12}} = \pi_{12}$
- $\theta \hat{T_{11}} = \pi_{11}$
- $\theta \hat{T_{22}} = \pi_{12}$
- $\hat{F}2 = \pi_{12} - \frac{\pi_{11}+\pi_{22}}{2}$

F_3 statistics

- F_3 statistics can be used to determine if a population X is admixed[1]
- $F_3(P_X; P_1, P_2) = E[(p_X - p_1)(p_X - p_2)]$
- They can also be expressed in terms of F_2 statistics
- $F_3(P_X; P_1, P_2) = \frac{1}{2}(F_2(pX, p1) + F_2(pX, p2) - F_2(p1, p2))$
- Note that if the populations can be described in terms of a tree, then $F_2(p1, p2) \leq F_2(pX, p1) + F_2(pX, p2)$

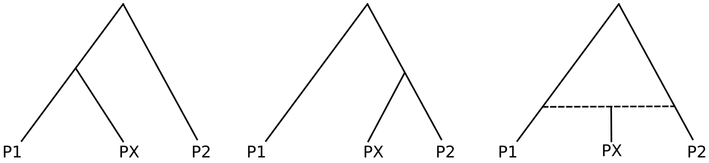


f3(PX; P1, P2) > 0

[1] Reich et al. (2009)

Admixture F_3 statistics

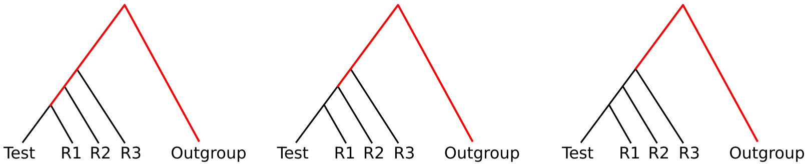
- One application of F3 is to detect violations in “treeness” (admixture or populations structure)
- If $F_2(p1, p2) > F_2(pX, p1) + F_2(pX, p2)$, then a tree is not a good descriptor of the populations, and $F_3(P_X; P_1, P_2) < 0$
- Run F3 statistics a Test population in the first position
- If the demographic history (with respect to 2 other populatiosn) can be described as a tree, then $F3 > 0$
- Violations in treeness result in $F3 < 0$



f3(PX; P1, P2) > 0 f3(PX; P1, P2) > 0 f3(PX; P1, P2) < 0

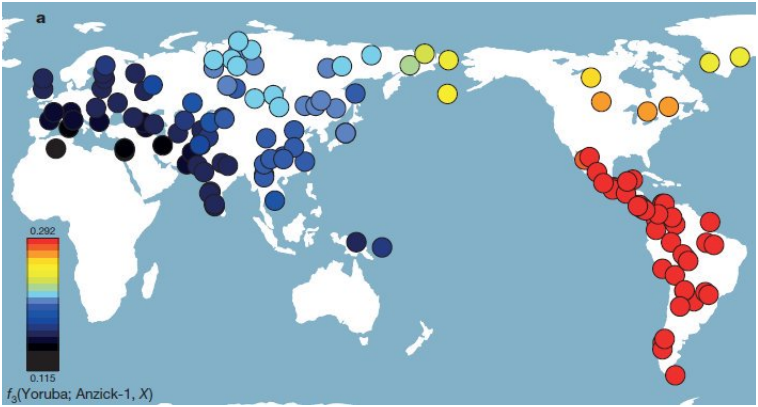
Outgroup F_3 statistics

- Another application of F3 is to determine which populations are closer (have more of a shared history) to a Test population
- Run F3 statistics with an Outgroup in the first position, followed by a Test population and several candidate Reference populatiosn
- F3 can be interpreted as the shared drift-path between a Test + Reference X and Test + Outgroup
- The more shared history between Test and Reference X, the larger the F3 statistic



f3(Outgroup; Test, R1) > f3(Outgroup; Test, R2) > f3(Outgroup; Test, R3)

Outgroup F_3 statistics



F_4 statistics

- F_4 statistics can be used to detect admixture and estimate admixture parameters.
- $F_4(P_1, P_2; P_3, P_4) = E[(p_1 - p_2)(p_3 - p_4)]$
- They can also be expressed in terms of F_2 statistics:
- $F_4(P_1, P_2; P_3, P_4) =$ $\frac{1}{2}(F_2(p1, p4) + F_2(p2, p3) - F_2(p1, p3) - F_2(p2, p4))$

F_4 statistics can be used to detect admixture

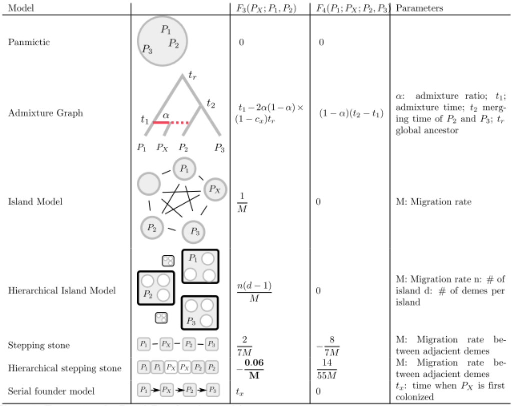
- A scaled version of the F_4 statistic (D) has been widely used to determine if admixture occurred in a population tree
- We'll describe D in detail in a few slides...

F-statistics vs. F_{ST}

- An F-statistic can be thought of as a covariance (or a linear combination of covariances) between population alllele frequencies
  - It ranges between $-\inf$ and inf
  - Easier to work with mathematically
  - It is additive: $F_2(P_1, P_2) = F_2(P_1, P_0) + F_2(P_2, P_0)$
  - Value **depends** on heterozygosity in the population
  - Highly used in models involving well-definied splits and admixture events
- F_{ST} can be thought of as an "absolute correlation" between population alllele frequencies
  - It ranges between 0 (panmixia) and 1 (complete divergence).
  - It is not additive: $F_{ST}(1, 2) \neq F_{ST}(1, 0) + F_{ST}(2, 0)$
  - Value does not depend on heterozygosity in the population
  - Highly used in stepping stone / migration models

Different models, different interpretations

- F-statistics will have different interpretations depending on underlying model
- Admixture graphs may not necessarily be the best descriptor of a biological system!



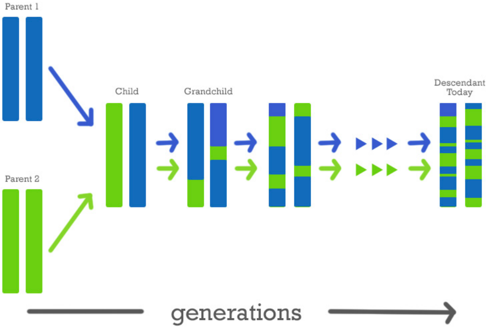
Model		$F_3(P_X;P_1,P_2)$	$F_4(P_1;P_X;P_2,P_3)$	Parameters
Panmictic		0	0	
Admixture Graph		$t_1-2\alpha(1-\alpha)\times(1-c_x)t_r$	$(1-\alpha)(t_2-t_1)$	α: admixture ratio; t_1; admixture time; t_2 merging time of P_2 and P_3; t_r global ancestor
Island Model		$\frac{1}{M}$	0	M: Migration rate
Hierarchical Island Model		$\frac{n(d-1)}{M}$	0	M: Migration rate n: # of island d: # of demes per island
Stepping stone		$\frac{2}{7M}$	$-\frac{8}{7M}$	M: Migration rate between adjacent demes
Hierarchical stepping stone		$-\frac{\mathbf{0.06}}{\mathbf{M}}$	$\frac{14}{55M}$	M: Migration rate between adjacent demes
Serial founder model		t_x	0	t_x: time when P_X is first colonized

P_1 P_2 P_3
t_r t_2 t_1 α
P_1 P_X P_2 P_3
P_1 P_X P_2 P_3
P_1 P_2 P_3
P_1 – P_X – P_2 – P_3
P_1 P_1 P_X P_X P_2 P_2
P_1 → P_X → P_2 → P_3

Figure 6. Expectations for F_3 and F_4 under select models.

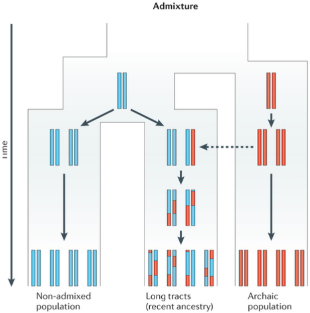
Admixture

- Admixture is the process by which two previously isolated populations interbreed.
- It results in the introduction of genetic material from a foreign source into a population.



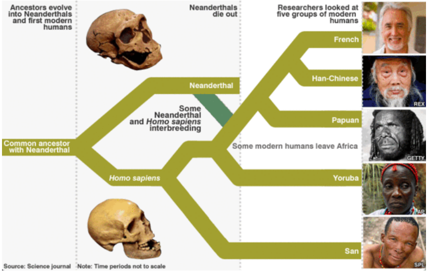
Admixture

- The signatures of admixture can be detected in the genomes of the descendants of the admixed individuals.



ABBA-BABA

- The ABBA-BABA test (or D-statistic) was developed to test for ancient gene flow between populations (Green et al. 2010, Durand et al. 2011, Patterson et al. 2012).
- Originally used as evidence for Neanderthal introgression into non-African modern humans (Green et al. 2010, Prufer et al. 2014).

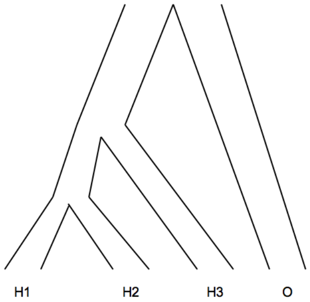

Ancestors evolve into Neanderthals and first modern humans
Neanderthals die out
Researchers looked at five groups of modern humans
French
Han-Chinese
REX
Neanderthal
Some Neanderthal and *Homo sapiens* interbreeding
Papuan
GETTY
Common ancestor with Neanderthal
Some modern humans leave Africa
Homo sapiens
Yoruba
AP
San
SPL

Source: Science journal

Note: Time periods not to scale

ABBA-BABA: assumptions

- We need to have sequence data from 3 populations (H1, H2 and H3) and an outgroup (O).
- The population tree should be known.
- There has been no recurrent mutations (short time-scales).
- Null hypothesis: no gene flow between H3 and H1 or between H3 and H2 after their respective splits.



ABBA-BABA: test using individual genomes

- Look at all diallelic loci where:
    - O and H3 have different alleles (called A and B)
    - H1 and H2 have different alleles
    - In other words, we look for sites where:
        - (H1,H2,H3,O) = (A,B,B,A)
        - (H1,H2,H3,O) = (B,A,B,A)
    - For example, (C,T,T,C) or (A,T,A,T).

ABBA-BABA: test using individual genomes

- Look at all diallelic loci where:
    - O and H3 have different alleles (called A and B)
    - H1 and H2 have different alleles
    - In other words, we look for sites where:
        - (H1,H2,H3,O) = (A,B,B,A)
        - (H1,H2,H3,O) = (B,A,B,A)
    - For example, (C,T,T,C) or (A,T,A,T).
- Count how many (A,B,B,A) sites and how many (B,A,B,A) sites there are

ABBA-BABA: test using individual genomes

- Look at all diallelic loci where:
    - O and H3 have different alleles (called A and B)
    - H1 and H2 have different alleles
    - In other words, we look for sites where:
        - (H1,H2,H3,O) = (A,B,B,A)
        - (H1,H2,H3,O) = (B,A,B,A)
    - For example, (C,T,T,C) or (A,T,A,T).
- Count how many (A,B,B,A) sites and how many (B,A,B,A) sites there are
- Calculate $D = \frac{\#ABBA - \#BABA}{\#ABBA + \#BABA}$

ABBA-BABA: test using individual genomes

- Look at all diallelic loci where:
    - O and H3 have different alleles (called A and B)
    - H1 and H2 have different alleles
    - In other words, we look for sites where:
        - (H1,H2,H3,O) = (A,B,B,A)
        - (H1,H2,H3,O) = (B,A,B,A)
    - For example, (C,T,T,C) or (A,T,A,T).
- Count how many (A,B,B,A) sites and how many (B,A,B,A) sites there are
- Calculate $D = \frac{\#ABBA - \#BABA}{\#ABBA + \#BABA}$
- (be careful with order of difference, some authors reverse it)

ABBA-BABA: test using individual genomes

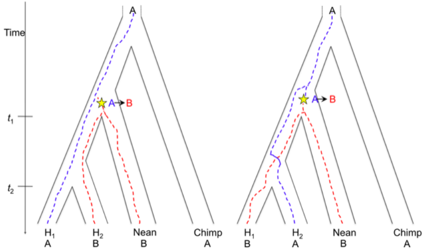
- Look at all diallelic loci where:
    - O and H3 have different alleles (called A and B)
    - H1 and H2 have different alleles
    - In other words, we look for sites where:
        - (H1,H2,H3,O) = (A,B,B,A)
        - (H1,H2,H3,O) = (B,A,B,A)
    - For example, (C,T,T,C) or (A,T,A,T).
- Count how many (A,B,B,A) sites and how many (B,A,B,A) sites there are
- Calculate $D = \frac{\#ABBA - \#BABA}{\#ABBA + \#BABA}$
- (be careful with order of difference, some authors reverse it)
- Test if D is significantly different from 0 (more on this in a second).

ABBA-BABA: test using individual genomes

- Look at all diallelic loci where:
  - O and H3 have different alleles (called A and B)
  - H1 and H2 have different alleles
  - In other words, we look for sites where:
    - (H1,H2,H3,O) = (A,B,B,A)
    - (H1,H2,H3,O) = (B,A,B,A)
  - For example, (C,T,T,C) or (A,T,A,T).
- Count how many (A,B,B,A) sites and how many (B,A,B,A) sites there are
- Calculate $D = \frac{\#ABBA - \#BABA}{\#ABBA + \#BABA}$
- (be careful with order of difference, some authors reverse it)
- Test if D is significantly different from 0 (more on this in a second).
- If so, reject the null hypothesis of no gene flow.

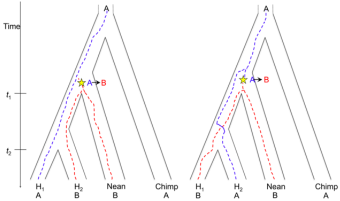
ABBA-BABA: rationale

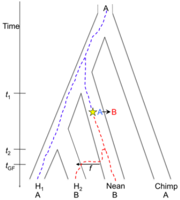
- If there was no admixture, the only way to generate coalescent trees consistent with ABBA or BABA is by incomplete lineage sorting (ILS).
- In that case, we expect the same number of ABBA trees as of BABA trees.



ABBA-BABA: rationale

- However, if there was gene flow from H3 to H2, we expect an excess of ABBA trees.
- Therefore, $\#ABBA > \#BABA$ and $D > 0$.





ABBA-BABA: testing for significance

- Perform block jacknife to get an estimate, $\hat{s}$, of the standard deviation of D.
- Assume that under the null hypothesis, $D \sim Normal(0, \hat{s}^2)$
- Use this distribution to calculate a Z-score
- Reject null hypothesis if $|Z| > 3$

ABBA-BABA: calculated from low-coverage data

- Look at 1 individual from each of the H1, H2, H3 and O populations

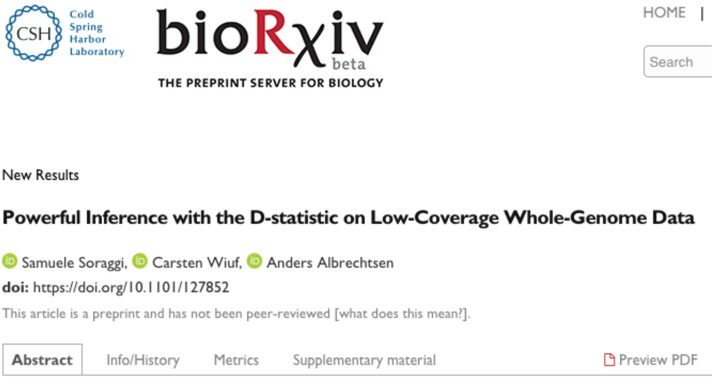
ABBA-BABA: calculated from low-coverage data

- Look at 1 individual from each of the H1, H2, H3 and O populations
- Randomly sample 1 read from each individual in each site

ABBA-BABA: calculated from low-coverage data

- Look at 1 individual from each of the H1, H2, H3 and O populations
- Randomly sample 1 read from each individual in each site
- Practical problems:
  - Not using all the information we could theoretically use
  - Bias can occur if H1 and H2 were sequenced using different platforms.
  - Bias can occur if H1 and H2 have different error rates.
  - SNP chip data is improperly used (without accounting for ascertainment bias).
  - With ancient genomes, increased error rates at specific positions (e.g. C-to-T) can also generate problems.

ABBA-BABA: calculated from low-coverage data


CSH
Cold Spring Harbor Laboratory
bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY
HOME |
Search

New Results

Powerful Inference with the D-statistic on Low-Coverage Whole-Genome Data

Samuele Soraggi, Carsten Wiuf, Anders Albrechtsen

doi: https://doi.org/10.1101/127852

This article is a preprint and has not been peer-reviewed [what does this mean?].

ABBA-BABA: example (from Green et al. 2010)

- We're testing for admixture from Neanderthals into French, using San Africans as the non-admixed sister population.

ABBA-BABA: example (from Green et al. 2010)

- We're testing for admixture from Neanderthals into French, using San Africans as the non-admixed sister population.
- $\#ABBA = 25,242$

ABBA-BABA: example (from Green et al. 2010)

- We're testing for admixture from Neanderthals into French, using San Africans as the non-admixed sister population.
- $\#ABBA = 25,242$
- $\#BABA = 22,982$

ABBA-BABA: example (from Green et al. 2010)

- We're testing for admixture from Neanderthals into French, using San Africans as the non-admixed sister population.
- $\#ABBA = 25,242$
- $\#BABA = 22,982$
- $D(San, French, Neanderthal, Chimpanzee) = 0.047$

ABBA-BABA: example (from Green et al. 2010)

- We're testing for admixture from Neanderthals into French, using San Africans as the non-admixed sister population.
- $\#ABBA = 25,242$
- $\#BABA = 22,982$
- $D(San, French, Neanderthal, Chimpanzee) = 0.047$
- After performing a block jackknife, $Z = 7.6$

ABBA-BABA: example (from Green et al. 2010)

- We're testing for admixture from Neanderthals into French, using San Africans as the non-admixed sister population.
- $\#ABBA = 25,242$
- $\#BABA = 22,982$
- $D(San, French, Neanderthal, Chimpanzee) = 0.047$
- After performing a block jackknife, $Z = 7.6$
- Conclusion: reject null hypothesis of no admixture.

ABBA-BABA: alternative formulation

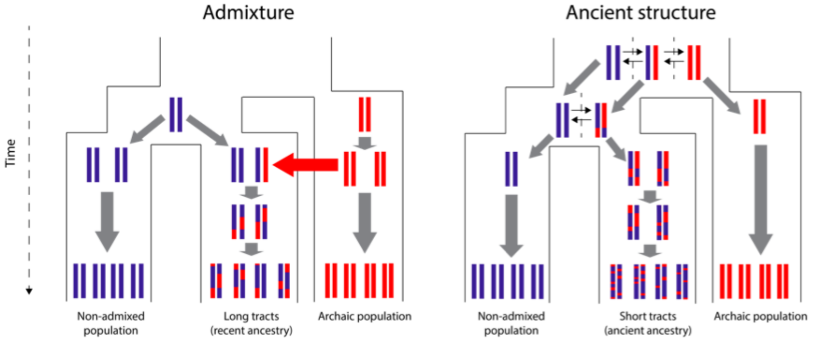
- Using sample allele frequencies (Durand et al. 2011)
- $D = \frac{\sum_{i=1}^{n}[(1-\hat{p_{i1}})\hat{p_{i2}}\hat{p_{i3}}(1-\hat{p_{i4}})-\hat{p_{i1}}(1-\hat{p_{i2}})\hat{p_{i3}}(1-\hat{p_{i4}})]}{\sum_{i=1}^{n}[(1-\hat{p_{i1}})\hat{p_{i2}}\hat{p_{i3}}(1-\hat{p_{i4}})+\hat{p_{i1}}(1-\hat{p_{i2}})\hat{p_{i3}}(1-\hat{p_{i4}})]}$
- $\hat{p_{i1}}$ is the sample allele frequency in H1 at SNP i.
- $\hat{p_{i2}}$ is the sample allele frequency in H2 at SNP i.
- $\hat{p_{i3}}$ is the sample allele frequency in H3 at SNP i.
- $\hat{p_{i4}}$ is the sample allele frequency in O at SNP i.

ABBA-BABA: caveats

- The value of D is not the same as the admixture rate!
- D depends on both the admixture rate AND the split times between the populations.
- Should not be deployed locally: ILS can generate local regions with $D \neq 0$.
- A genome-wide value of D significantly different from 0 could also be caused by ancestral population structure.

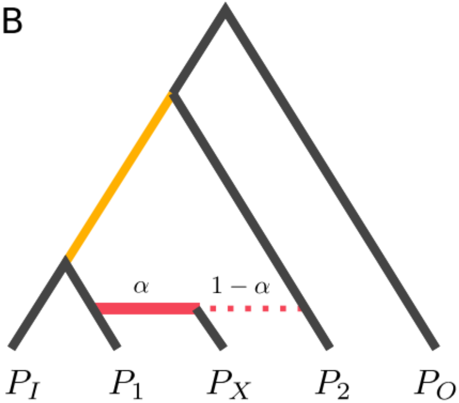
ABBA-BABA: caveats

- Important to find admixture tracts with lenghts consistent with introgression.
- Hard problem: requires probabilistic models like HMMs.



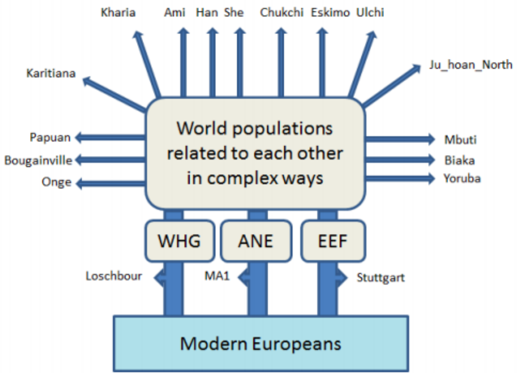
F_4 statistics can be used to estimate admixture proportions

- Assuming admixture occurred, F_4 statistics can be used to estimate the **amount** of admixture
- $\alpha = \frac{F_4(P_O, P_I; P_X, P_1)}{F_4(P_O, P_I; P_2, P_1)}$



qpWave / qpAdm

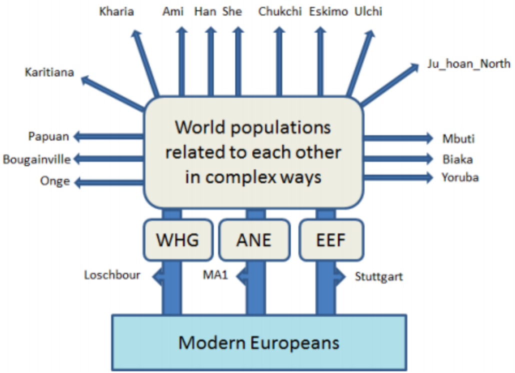
- The qpWave / qpAdm methodology[2]is a way to model admixture without detailed phylogenetic modeling
- This was originally used to argue for at least 3 highly-differentiated streams of ancestry contributing to present-day European genomes



[2] Lazaridis et al. (2014, 2016)

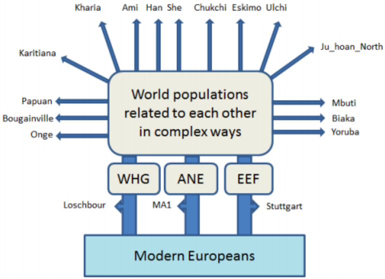
qpWave / qpAdm

- We need:
    - A) A Test population
    - B) A set of Outgroup populations
    - C) A set of Reference populations that are clades (with respect to the Outgroups) of populations potentially contributing ancestry to the Test



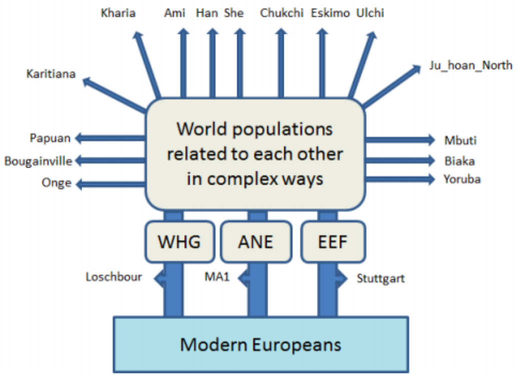
qpWave / qpAdm

- We can write F4 statistics for the Test as a weighted sum of N F4 statistics for the Reference populations
- $f_4(Test, O_1; O_2, O_3) = \sum_{i=1}^{N} \alpha_i f_4(Ref_i, O_1; O_2, O_3)$
- Given m Outgroups, there are $m\binom{m}{2}$ equations of the above form
- We can use regression to fit the mixture coefficients with the mixture coefficients (α_i) by regression



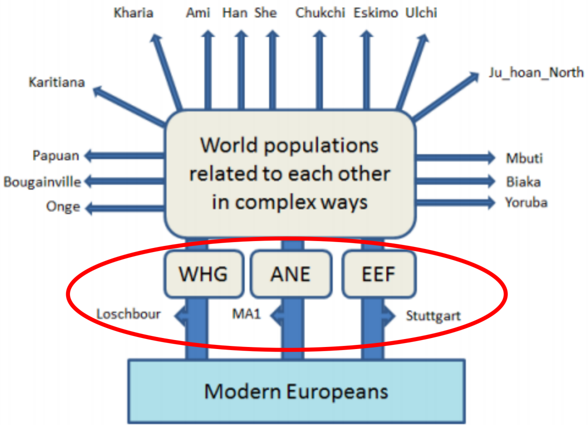
qpWave / qpAdm

- qpAdm is a program used to find the best-fitting admixutre coefficients under this framework



qpWave / qpAdm

- qpWave is a program used to find whether the Reference + Test populations can be modeled as being descended from as few as X source populations (that are differentially related to the Outgroups)
- Typicially one runs qpWave before qpAdm


Kharia
Ami
Han
She
Chukchi
Eskimo
Ulchi
Karitiana
Ju_hoan_North
Papuan
Bougainville
Onge
World populations related to each other in complex ways
Mbuti
Biaka
Yoruba
WHG
ANE
EEF
Loschbour
MA1
Stuttgart
Modern Europeans