Population genetics 2: F-statistics and D-statistics

Fernando Racimo

Copenhagen, August 2018

Today

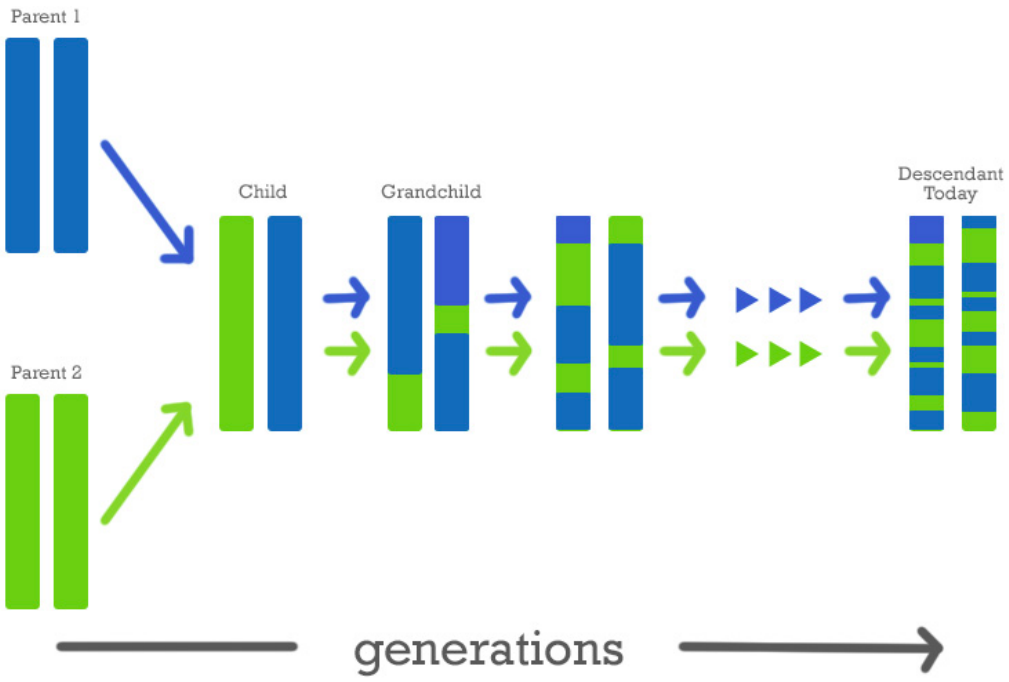
- Admixture
- D-statistics
- Genetic drift
- F-statistics
- Admixture graphs

Today

- **Admixture**
- D-statistics
- Genetic drift
- F-statistics
- Admixture graphs

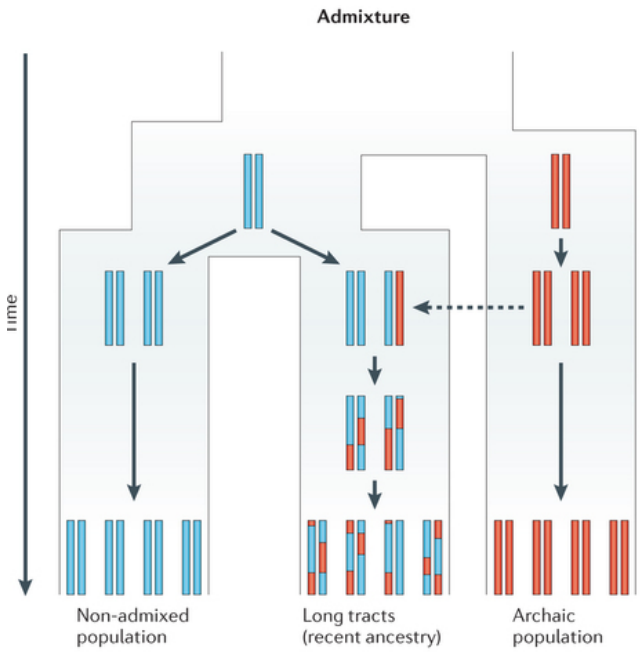
Admixture

- 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.

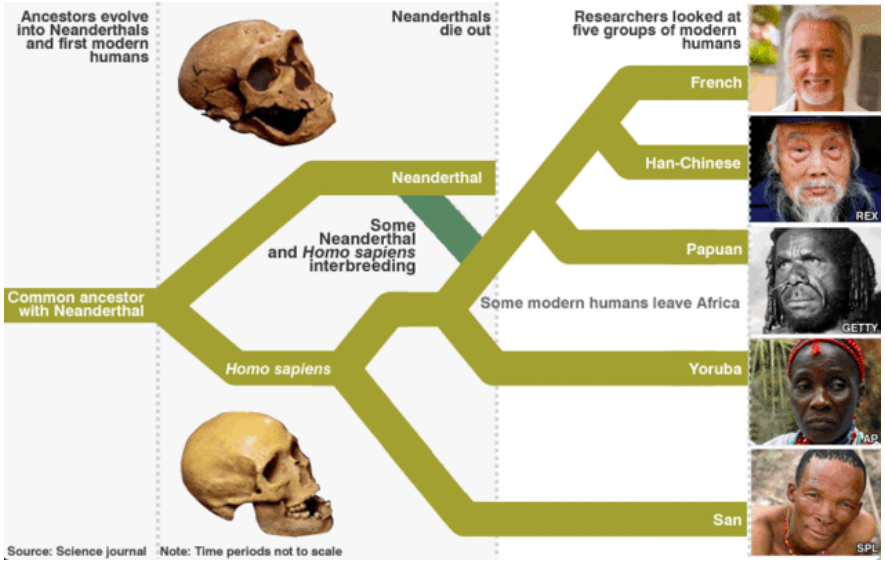


Today

- Admixture
- **D-statistics**
- Genetic drift
- F-statistics
- Admixture graphs

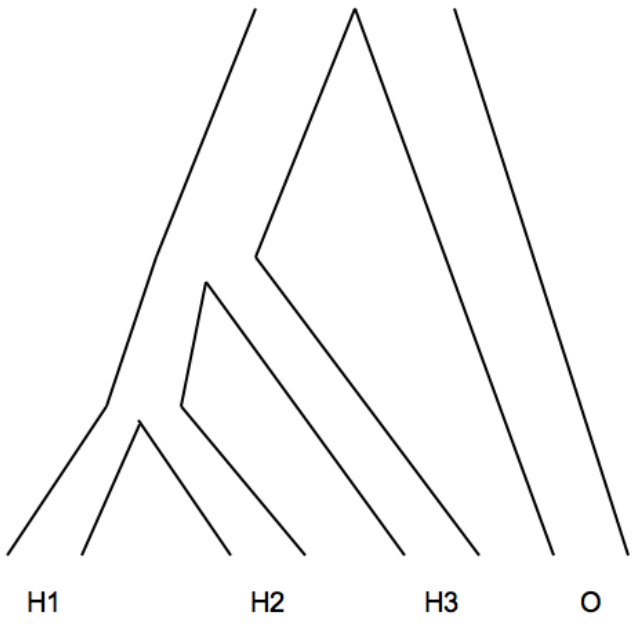
ABBA-BABA

- 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).



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

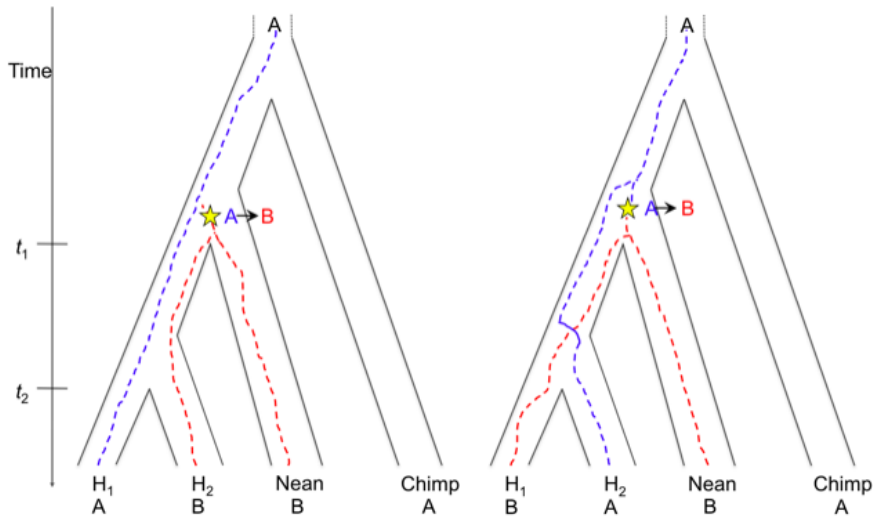
- Look at all diallelic loci where:
  - O and H3 have different alleles (called A and B)
  - H1 and H2 have different alleles
  - In other words, we look for sites where:
    - (H1,H2,H3,O) = (A,B,B,A)
    - (H1,H2,H3,O) = (B,A,B,A)
  - For example, (C,T,T,C) or (A,T,A,T).
- Count how many (A,B,B,A)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

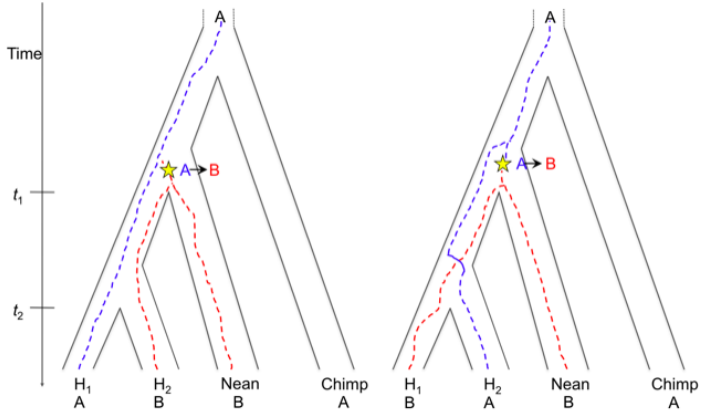
ABBA-BABA: rationale

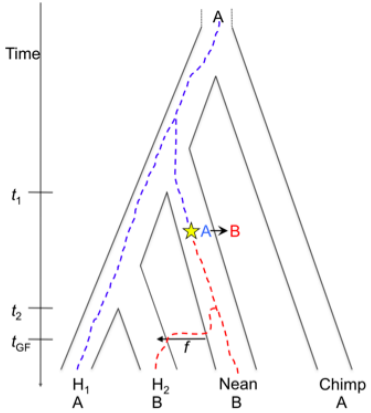
- If there was no admixture, the only way to generate coalescent trees consistent with ABBA or BABA is by incomplete lineage sorting (ILS).
- In that case, we expect the same number of ABBA trees as of BABA trees.



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

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?].

Abstract Info/History Metrics Supplementary material Preview PDF

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

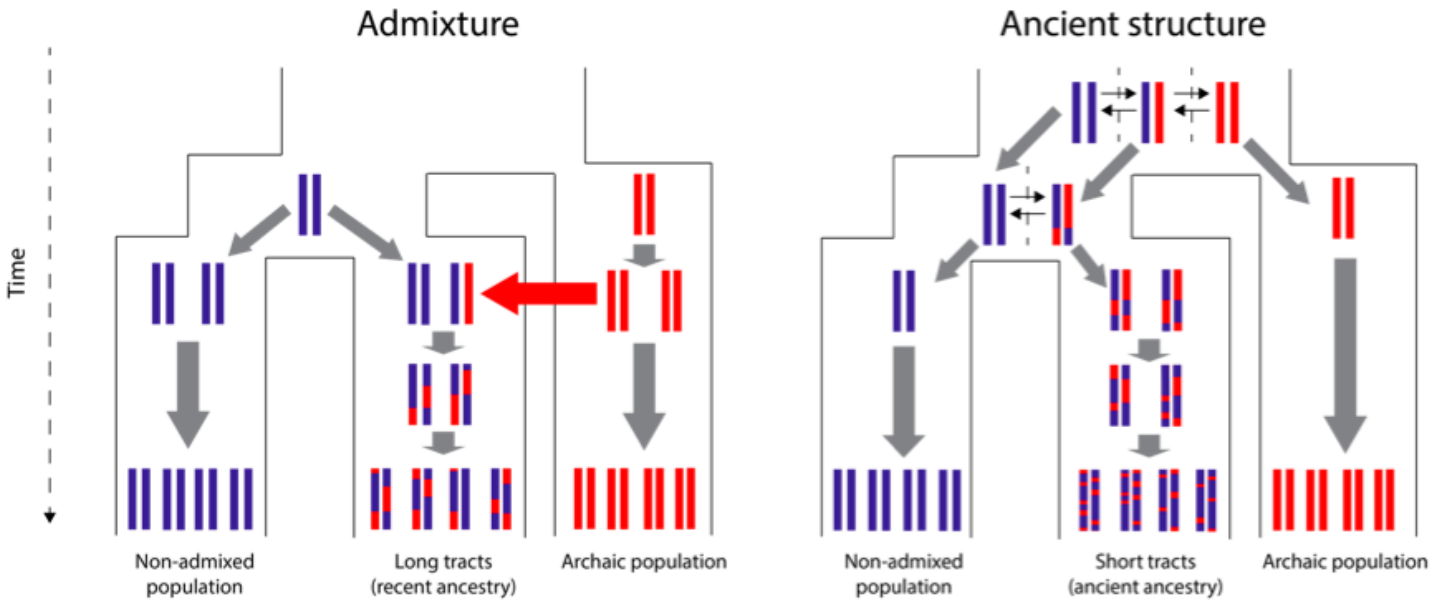
- Using sample allele frequencies (Durand et al. 2011)
- $D = \frac{\sum_{i=1}^{n}[(1-\hat{p_{i1}})\hat{p_{i2}}\hat{p_{i3}}(1-\hat{p_{i4}})-\hat{p_{i1}}(1-\hat{p_{i2}})\hat{p_{i3}}(1-\hat{p_{i4}})]}{\sum_{i=1}^{n}[(1-\hat{p_{i1}})\hat{p_{i2}}\hat{p_{i3}}(1-\hat{p_{i4}})+\hat{p_{i1}}(1-\hat{p_{i2}})\hat{p_{i3}}(1-\hat{p_{i4}})]}$
- $\hat{p_{i1}}$ 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.

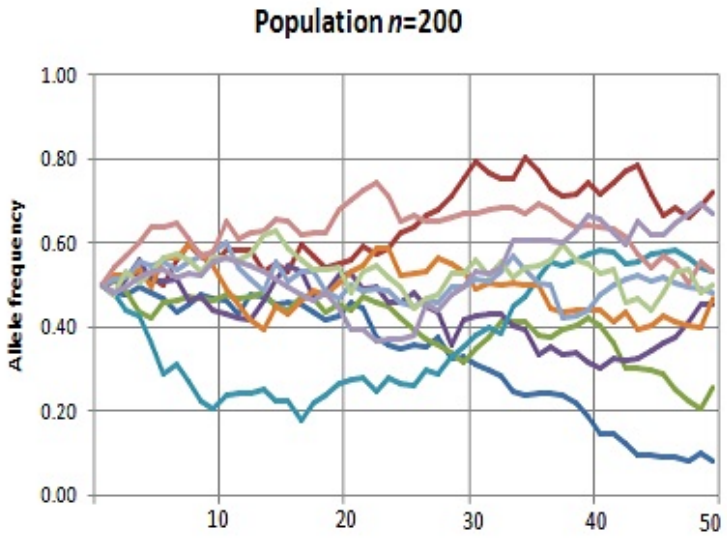


Today

- Admixture
- D-statistics
- **Genetic drift**
- F-statistics
- Admixture graphs

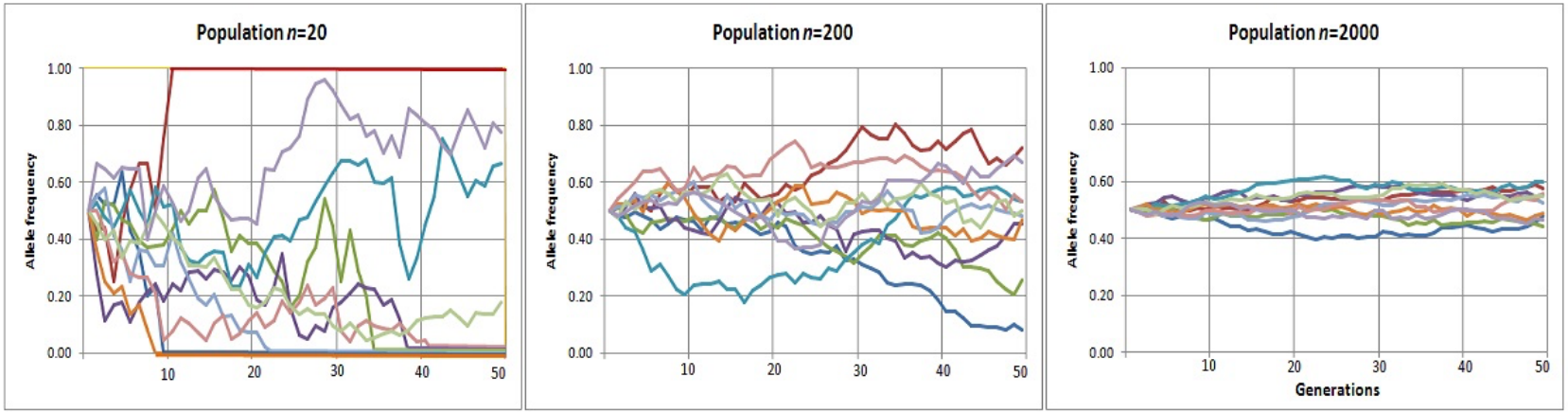
Genetic drift

- Genetic drift is the change in frequency of a genetic variant that occurs due to random sampling in finite populations
- Even under complete neutrality, at each generation, some individuals will die and others reproduce completely by chance
- Allele frequencies therefore fluctuate randomly over time



Genetic drift

- Drift increases with increasing time (more time for random flucturations to occur)
- Drift increases with decreasing population size (more stochasticity in smaller populations)

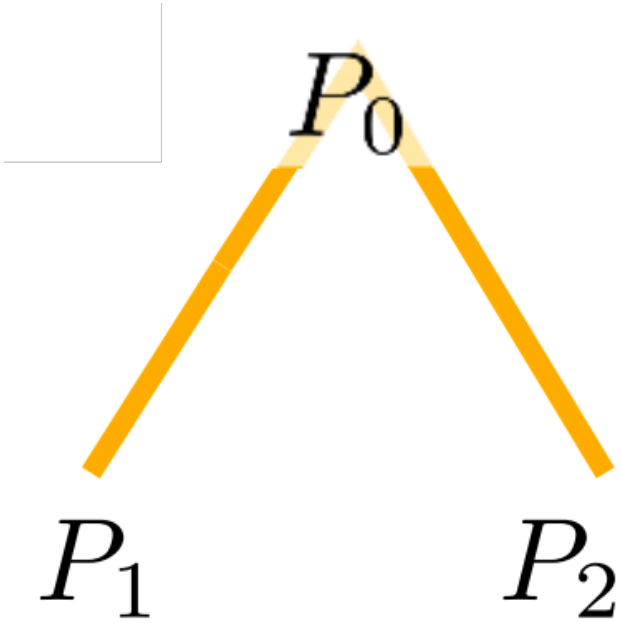


Today

- Admixture
- D-statistics
- Genetic drift
- **F-statistics**
- Admixture graphs

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
- $\mathbf{F_2(P_1, P_2) = E[(p_1 - p_2)^2]}$

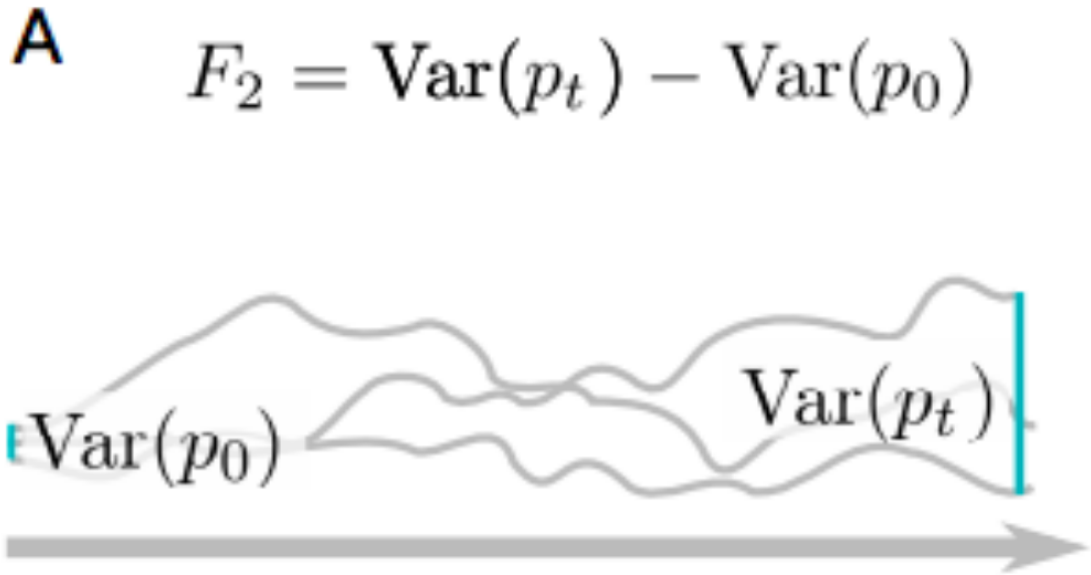


F_2 statistics

- $F_2(P_1, P_2) = E[(p_1 - p_2)^2]$
- 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

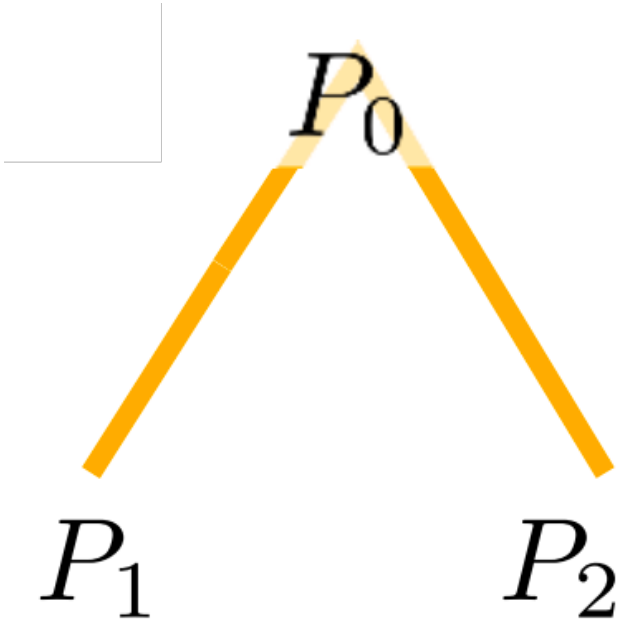
F_2 as a measure of genetic drift

- If we compute an F_2 statistic between an ancestral population P_t and a descendant population P_0, then $F_2(P_t, P_0) = Var[p_t] - Var[p_0]$
- We can therefore consider an F2 statistic to be a measure of the increase in allele frequency variance over time
- In essence, a measure of **genetic drift** (time scaled by population size)



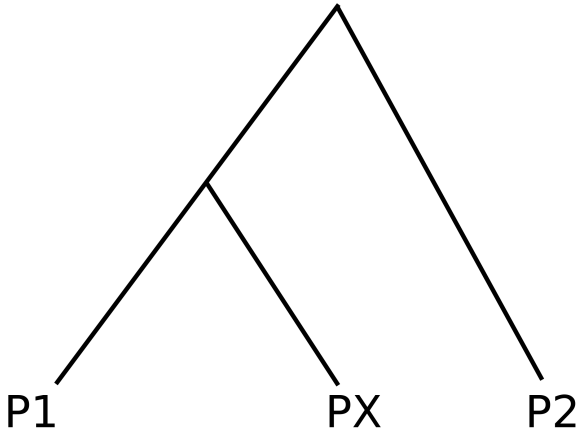
F_2 additivity

- 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_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 $F_2(p1, p2) \leq F_2(pX, p1) + F_2(pX, p2)$

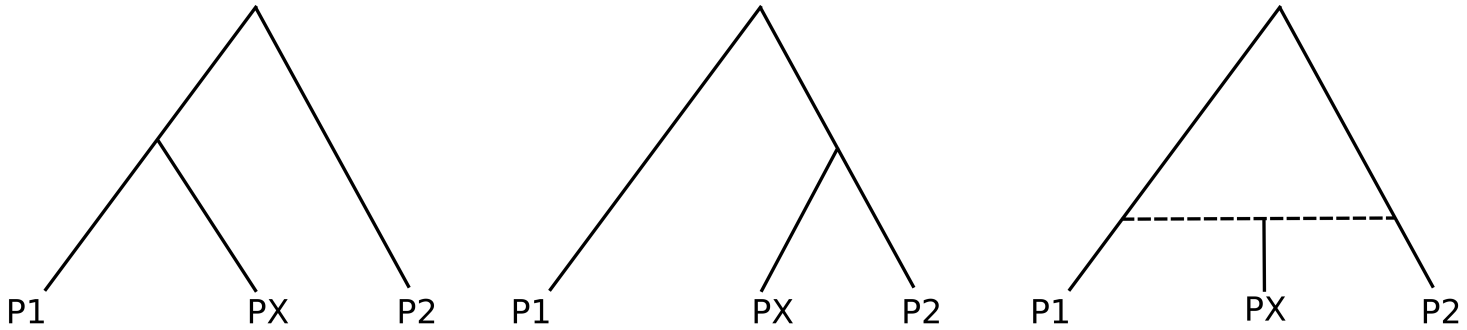


f3(PX; P1, P2) > 0

[1] Reich et al. (2009)

Admixture F_3 statistics

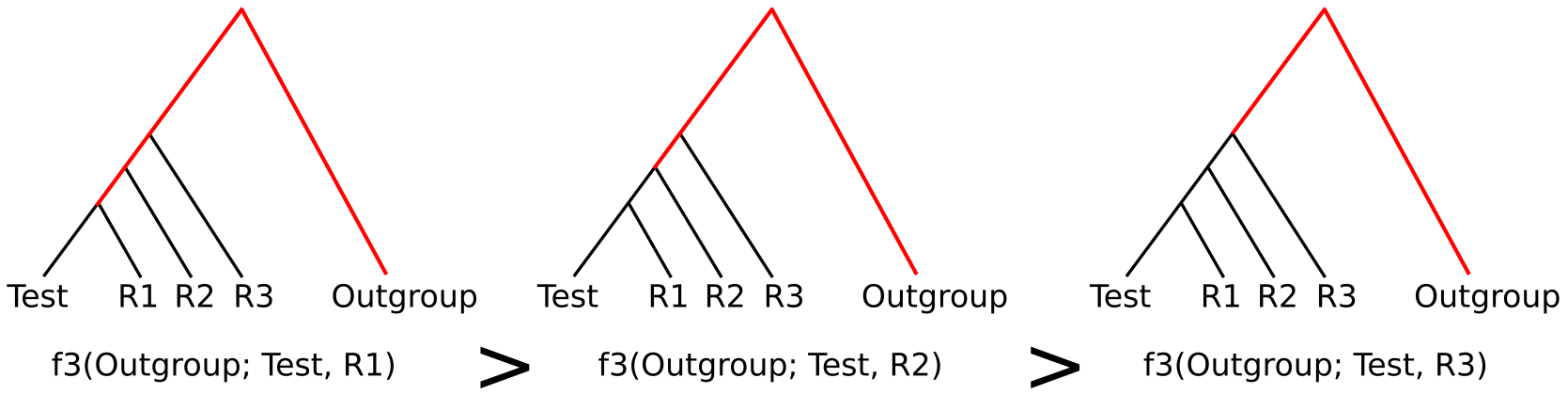
- One application of F3 is to detect violations in "treeness" (admixture or populations structure)
- If $F_2(p1, p2) > F_2(pX, p1) + F_2(pX, p2)$, then a tree is not a good descriptor of the populations, and $F_3(P_X; P_1, P_2) < 0$
- Run F3 statistics a Test population in the first position
- If the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



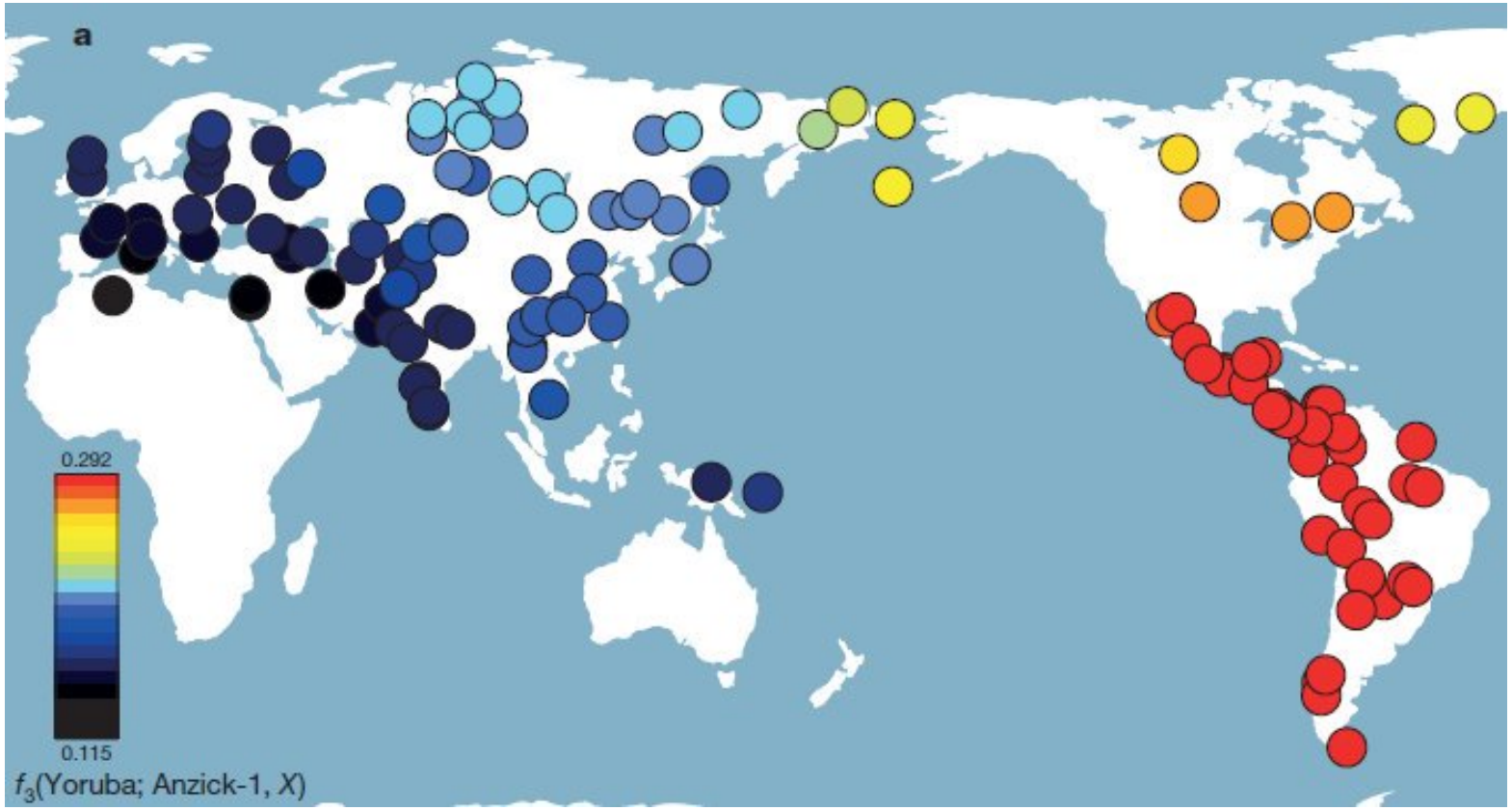
f3(PX; P1, P2) > 0 f3(PX; P1, P2) > 0 f3(PX; P1, P2) < 0

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



Outgroup F_3 statistics



An excellent resource for F-statistics

Admixture, Population Structure and F-statistics

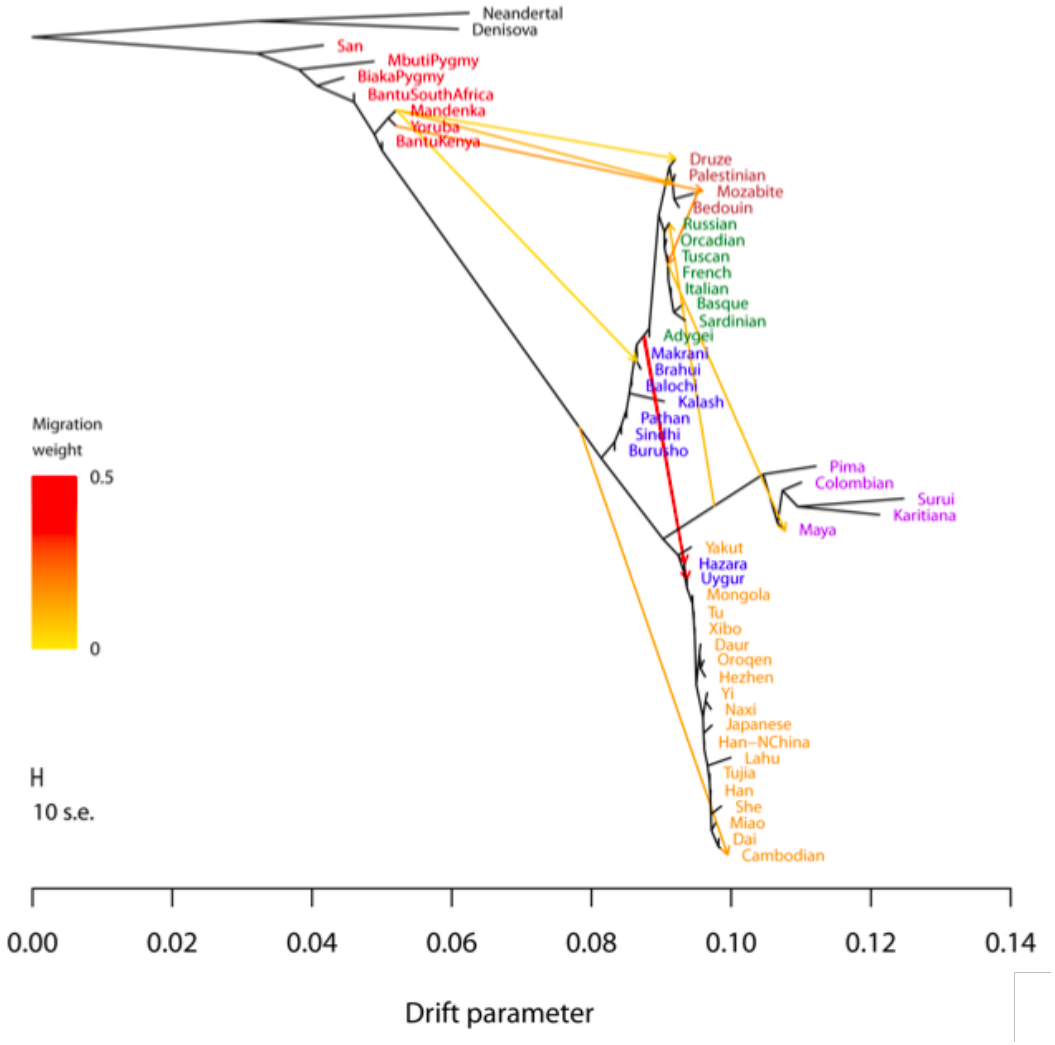
Benjamin M Peter[1]

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

Today

- Admixture
- D-statistics
- Genetic drift
- F-statistics
- **Admixture graphs**

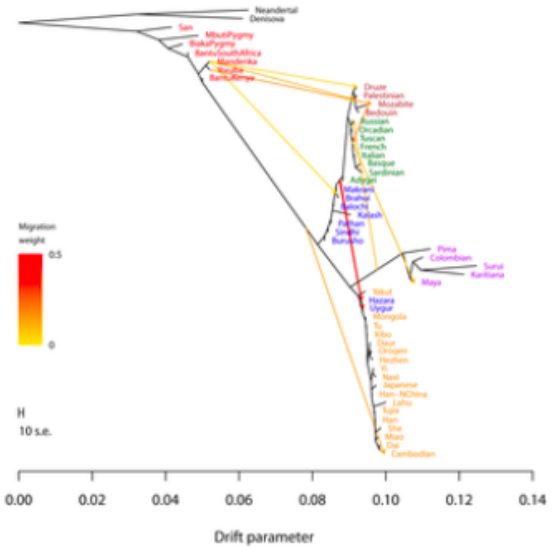
Dealing with many populations and admixture events



Admixture graph methods

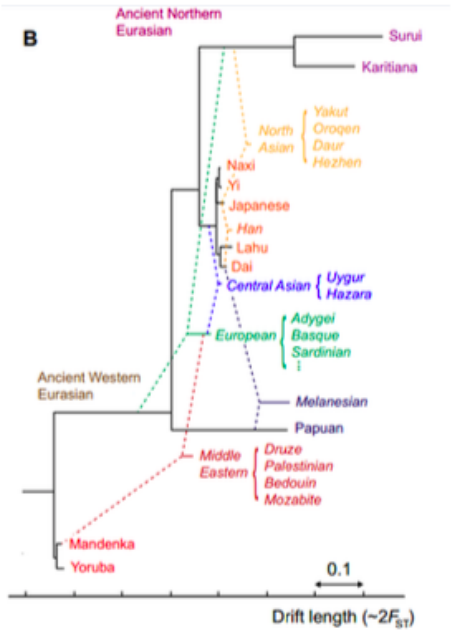
TreeMix

Pickrell and Pritchard 2012



MixMapper

Lipson et al. 2013



qpGraph

Patterson et al. 2012

Less supervised

More supervised

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!

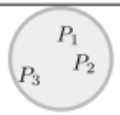
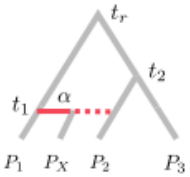
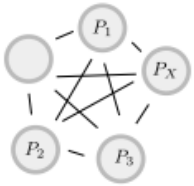
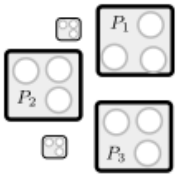
Model		$F_3(P_X; P_1, P_2)$	$F_4(P_1; P_X; P_2, P_3)$	Parameters
Panmictic		0	0	
Admixture Graph		$t_1 - 2\alpha(1-\alpha) \times (1-c_x)t_r$	$(1-\alpha)(t_2 - t_1)$	α: admixture ratio; t_1; admixture time; t_2 merging time of P_2 and P_3; t_r global ancestor
Island Model		$\frac{1}{M}$	0	M: Migration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

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