Models for learning about history from genetics

Joe Pickrell
Harvard Medical School

How Can We Infer Geography and History from Gene Frequencies?

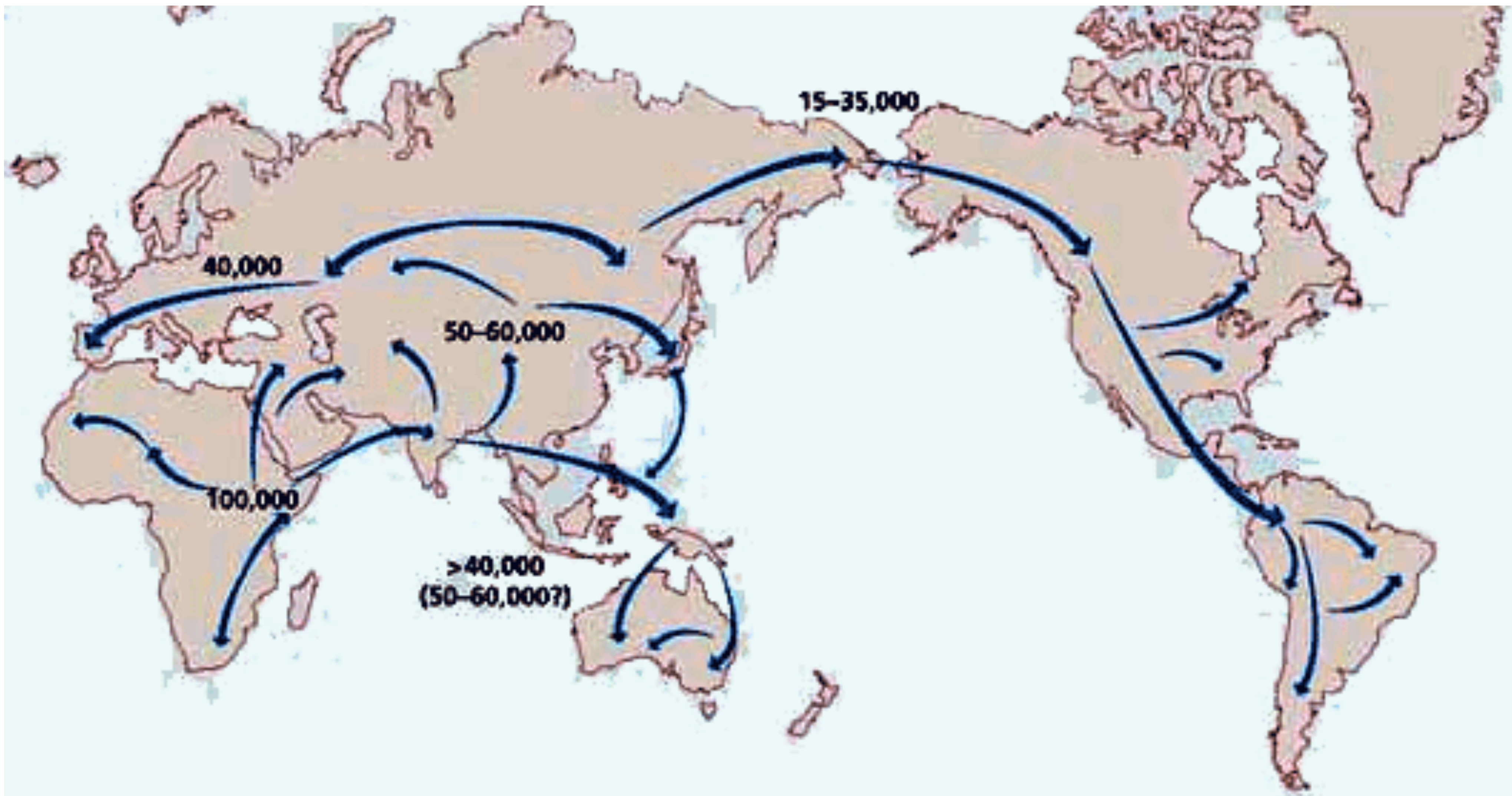
JOSEPH FELSENSTEIN

Department of Genetics, University of Washington, Seattle, Washington 98195, U.S.A.

(*Received* 21 *September* 1981)

In principle, patterns of migration and historical branching events can be reconstructed from gene frequency data, but we still lack most of the techniques necessary to do this. This is a fairly clearly defined problem with a variety of interesting subcases. It appears likely to raise interesting mathematical and statistical questions. The amount of data potentially available is very large. The problem is reviewed in this paper, but no new solutions are proposed.

Humans expanded out of Africa to occupy nearly the entire globe over the last 60-100k years

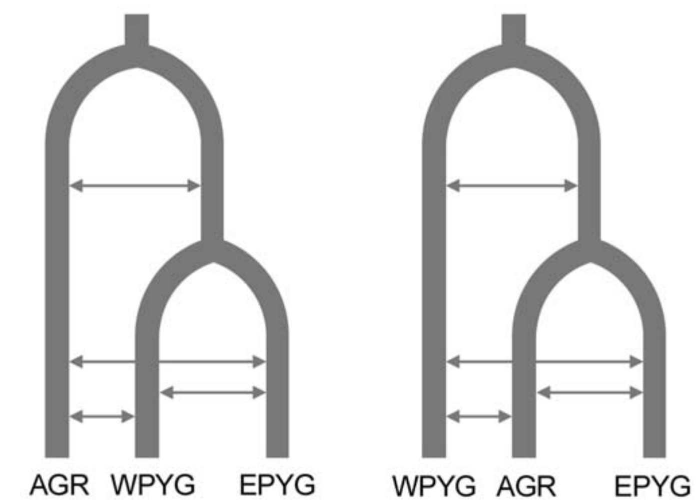


Details in this figure are are vague. How can we figure out population sizes, timings of population movements, etc?

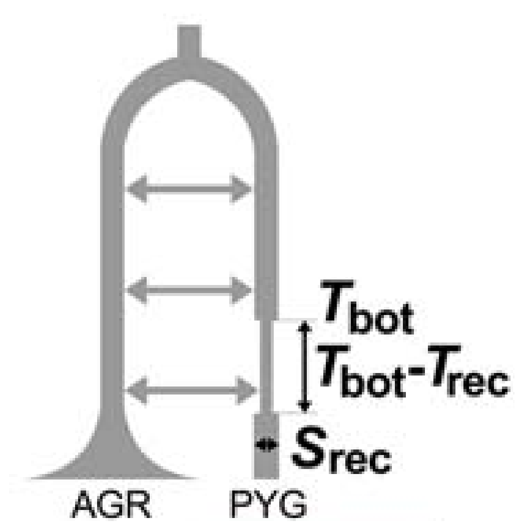
Humans are intensively studied, yet we still know only vague details. What about other species?

What can we hope to learn?

- Topology
  - what is the branching structure of populations?



- Demography
  - when did demographic events occur?
  - what were population sizes?



Patin et al. (2009)

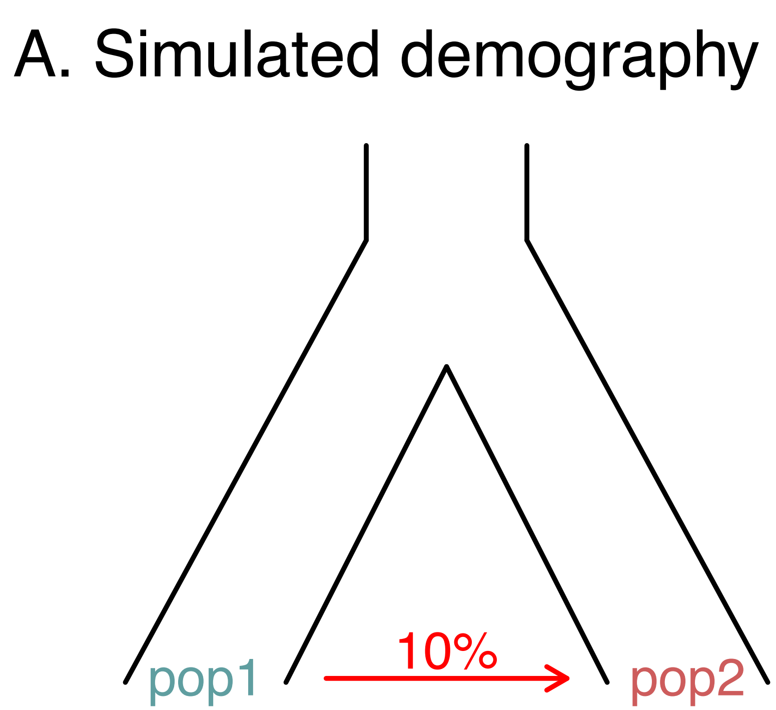
What sources of information do we have?

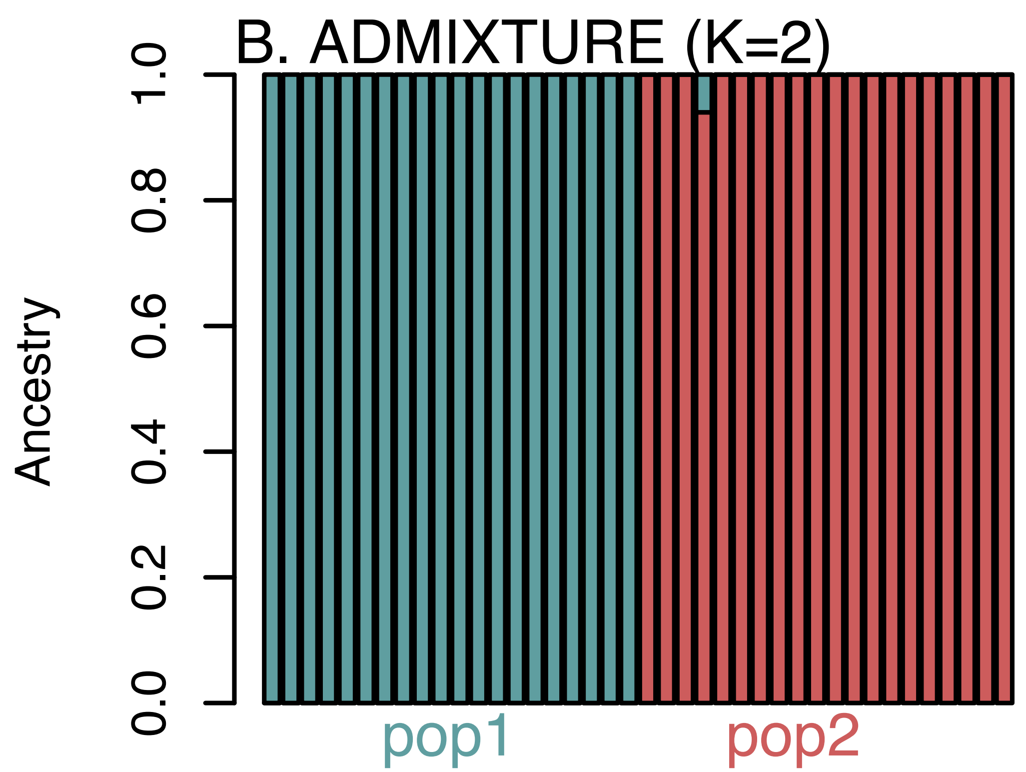
- Allele frequencies--more closely related population have more similar allele frequencies. E.g. clustering algorithms (STRUCTURE/PCA), tree-building algorithms

- Linkage disequilibrium--influenced by mixture between populations. E.g. local ancestry inference, ROLLOFF

- Mutations--shared rare mutations between populations indicate shared history. E.g. mtDNA trees (not going to cover this)

Caution: all methods can be misinterpreted

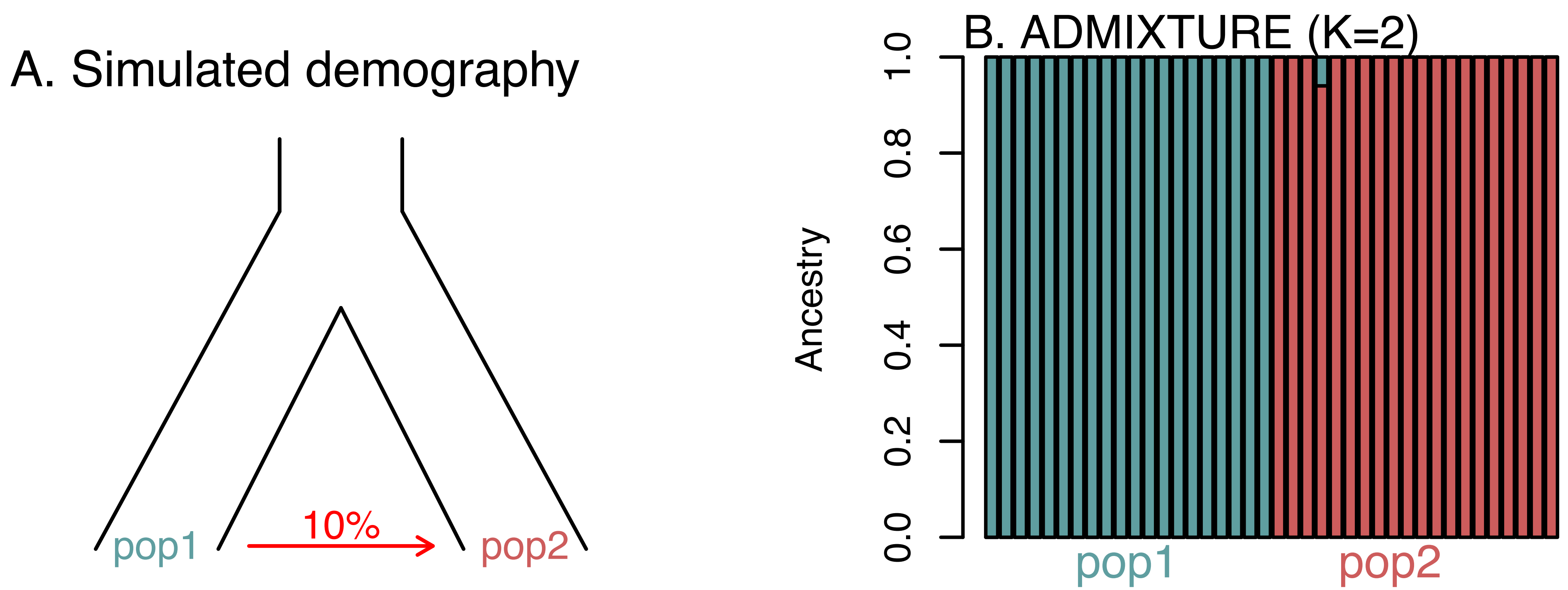




- Why does ADMIXTURE not identify admixture?

Pickrell et al. (2012)

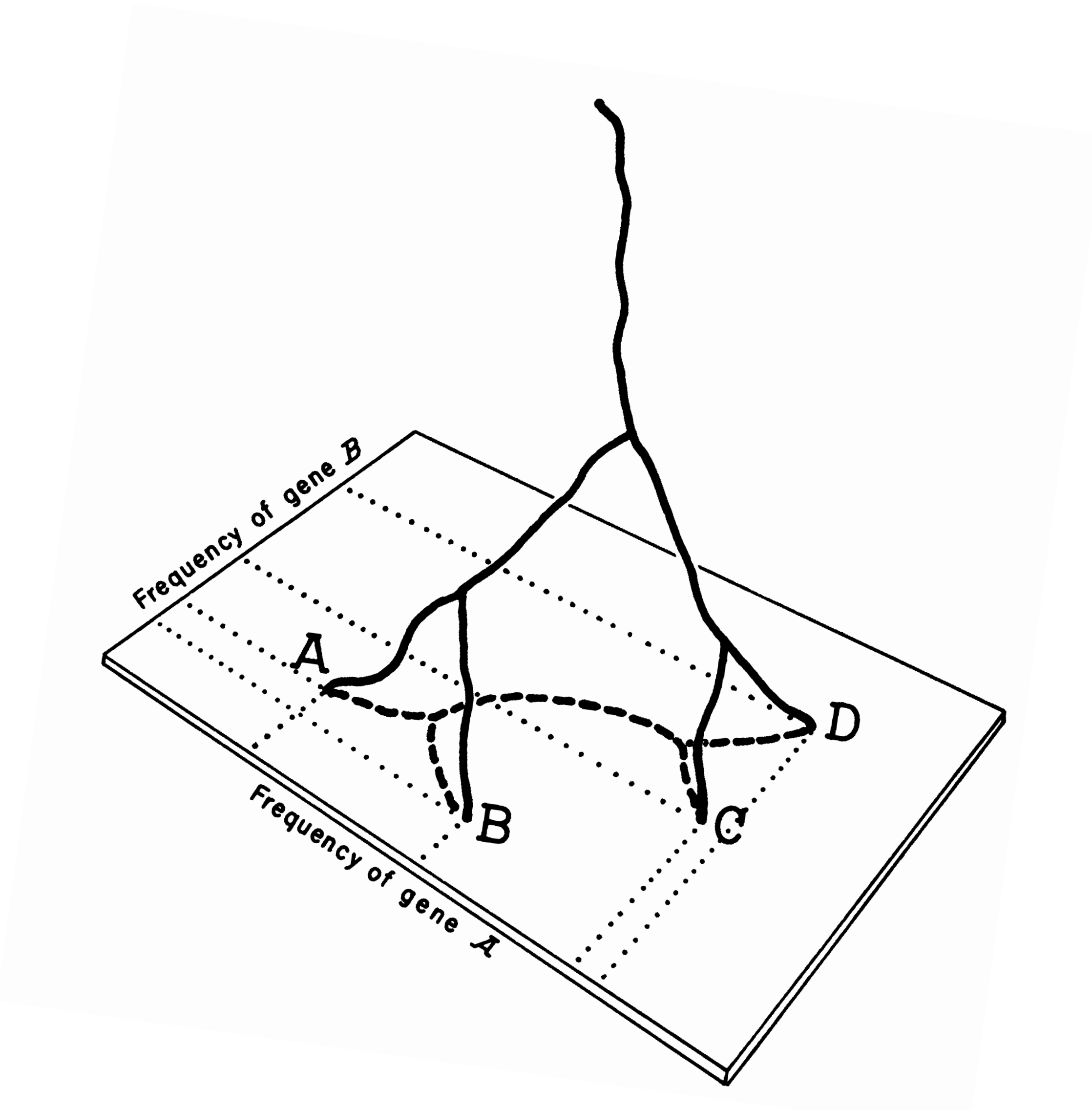
Caution: all methods can be misinterpreted



- Need somewhat-more model based methods for inferring/testing tree topologies

Pickrell et al. (2012)

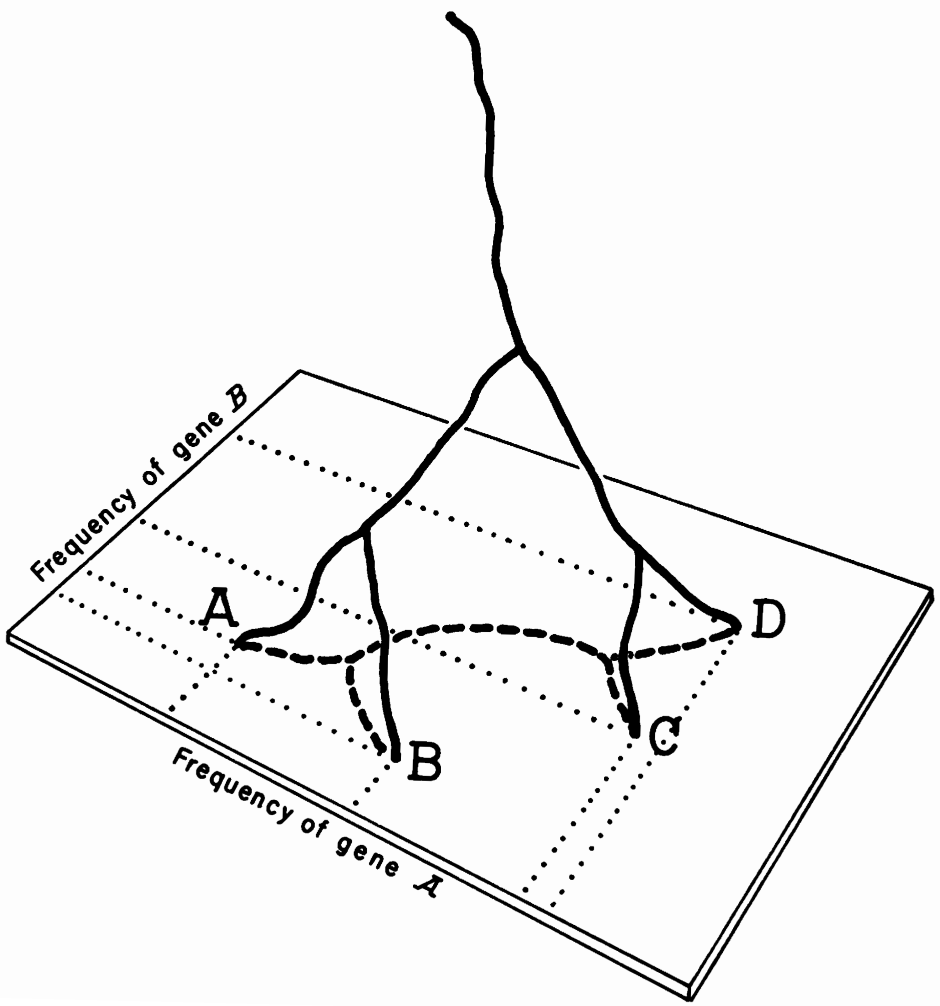
Inferring and testing topologies



- Cavalli-Sforza and Edwards (1967) point out that allele frequencies drift randomly, might be used to reconstruct population trees

- What type of model?

Cavalli-Sforza and Edwards (1967)



- The first proposal for treating inference of population phylogenies as a statistical problem
- Consider an allele with frequency x
- What is its frequency t generations later?
- Can be written down analytically (Kimura 1955), but an approximation is useful

Normal approximation to drift

time

x_A

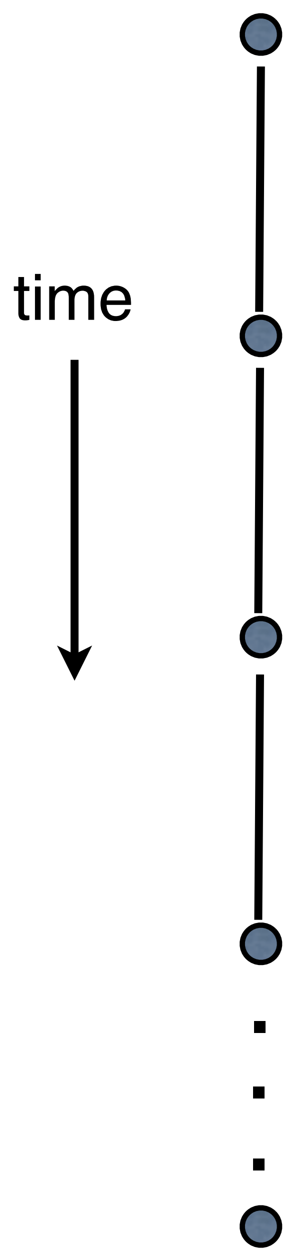
$$X_1 \sim \frac{Bin(2N_e, x_A)}{2N_e}$$

$$X_2 \sim \frac{Bin(2N_e, x_1)}{2N_e}$$

$$\vdots$$

$$X_t \sim \frac{Bin(2N_e, x_{t-1})}{2N_e}$$

Normal approximation to drift



$$x_A$$

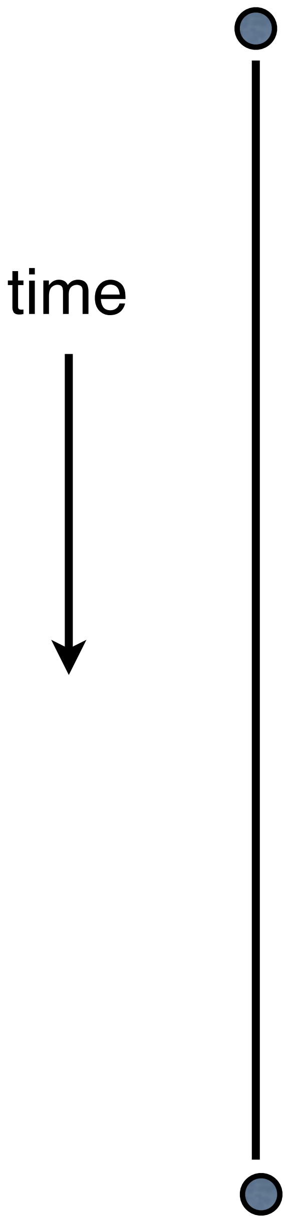
$$X_1 \sim N\left(x_A, \frac{x_A[1 - x_A]}{2N_e}\right)$$

$$X_2 \sim N\left(x_1, \frac{x_1[1 - x_1]}{2N_e}\right)$$

$$X_t \sim N\left(x_{t-1}, \frac{x_{t-1}[1 - x_{t-1}]}{2N_e}\right)$$

Normal approximation to drift

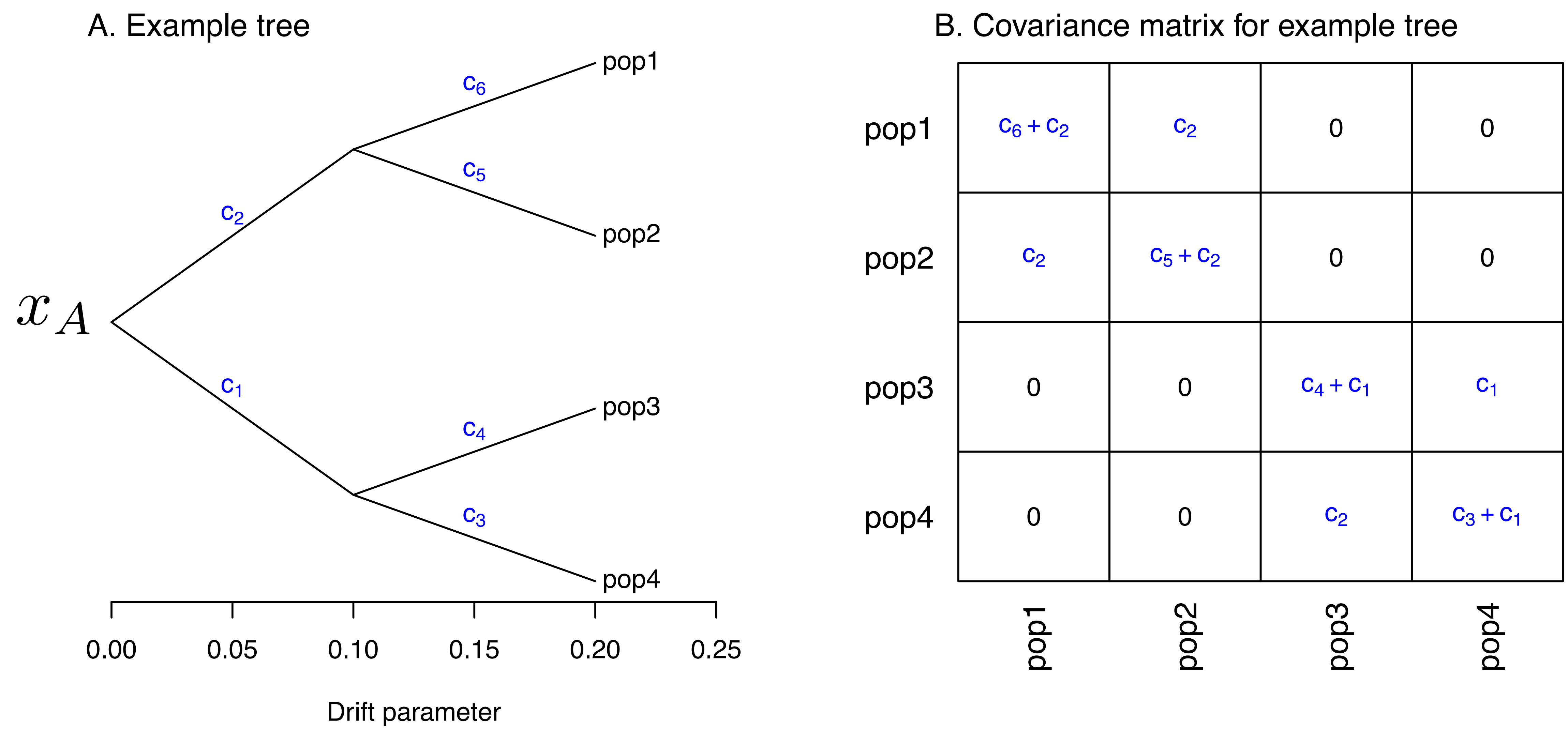
x_A



$X_t \sim N(x_A, c x_A[1 - x_A])$

$$c \approx \frac{t}{2N_e}$$

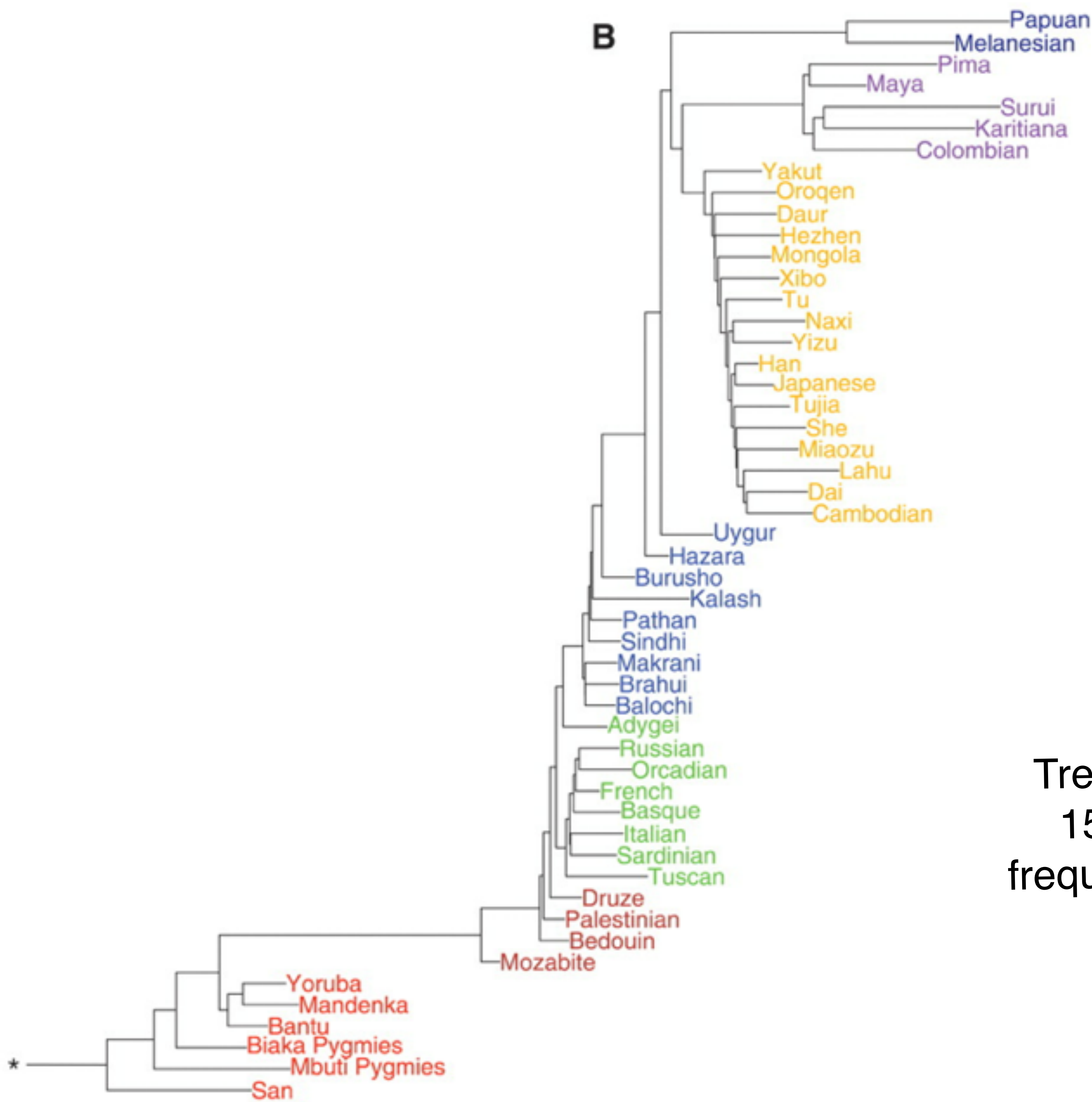
Natural way of learning about tree structure from allele frequencies



$$[X_1, X_2, \ldots] \sim MVN(x_A, V)$$

Felsenstein (1973), figure from
Pickrell and Pritchard (2012)

Application to humans



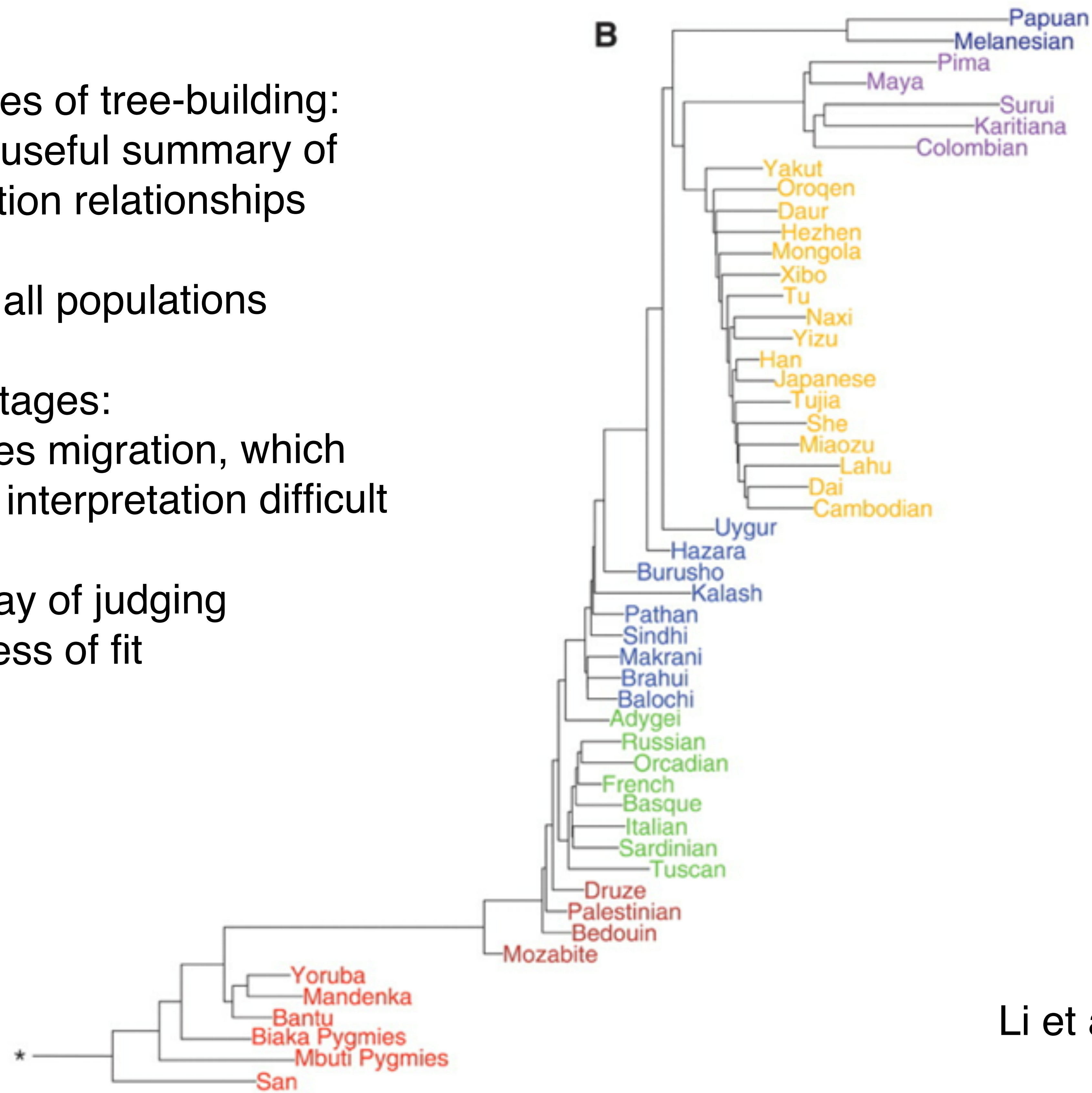
Tree constructed from 150,000 SNP allele frequencies using contml

Li et al. (2008)

Application to humans

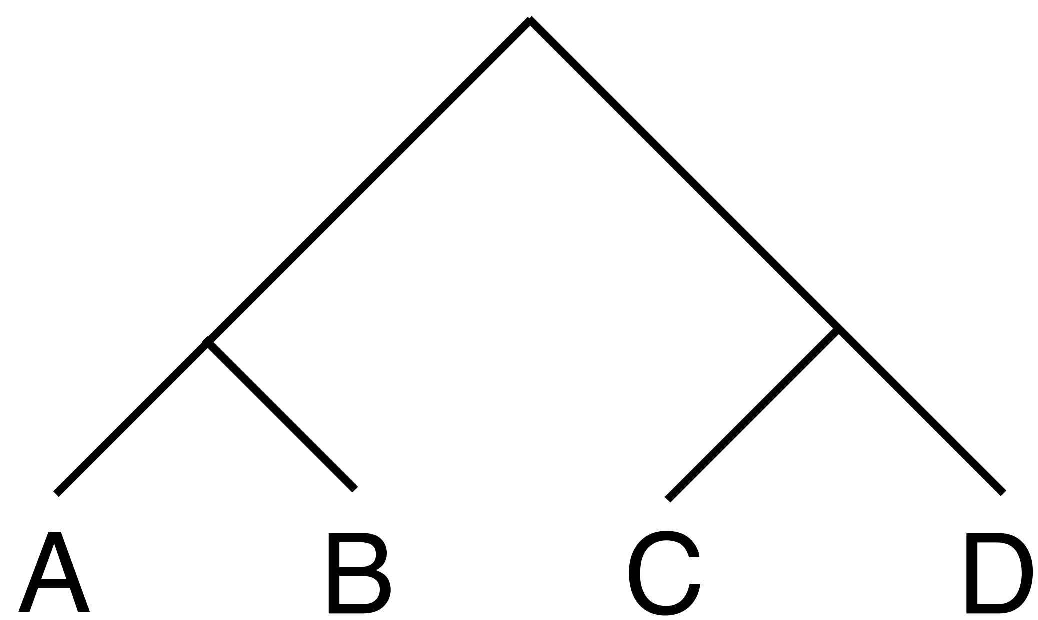
Advantages of tree-building:
- Fast, useful summary of population relationships
- Uses all populations

Disadvantages:
- Ignores migration, which makes interpretation difficult
- No way of judging goodness of fit

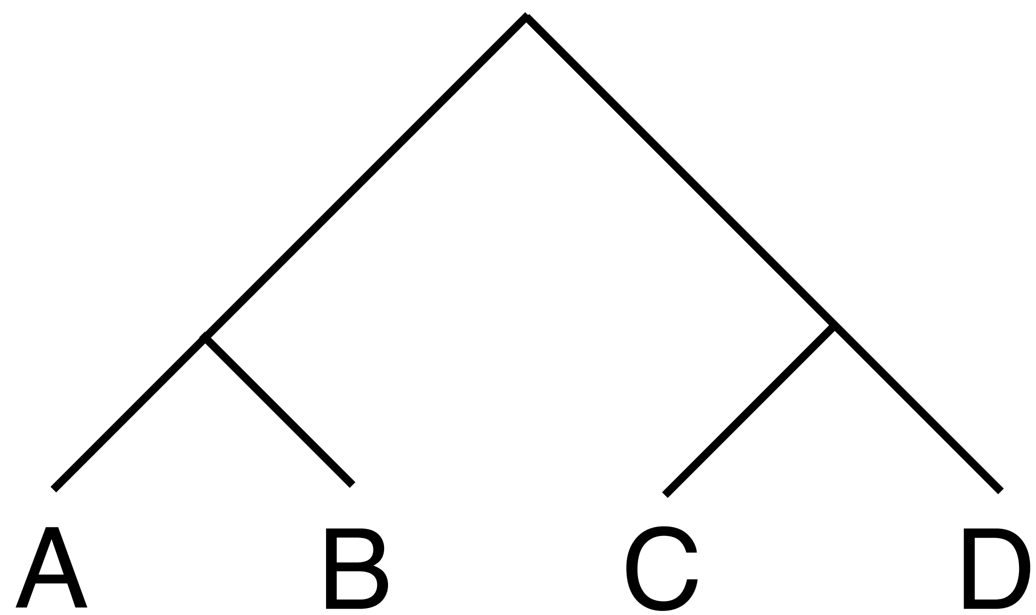


Li et al. (2008)

Is a tree a good fit?



Is a tree a good fit?

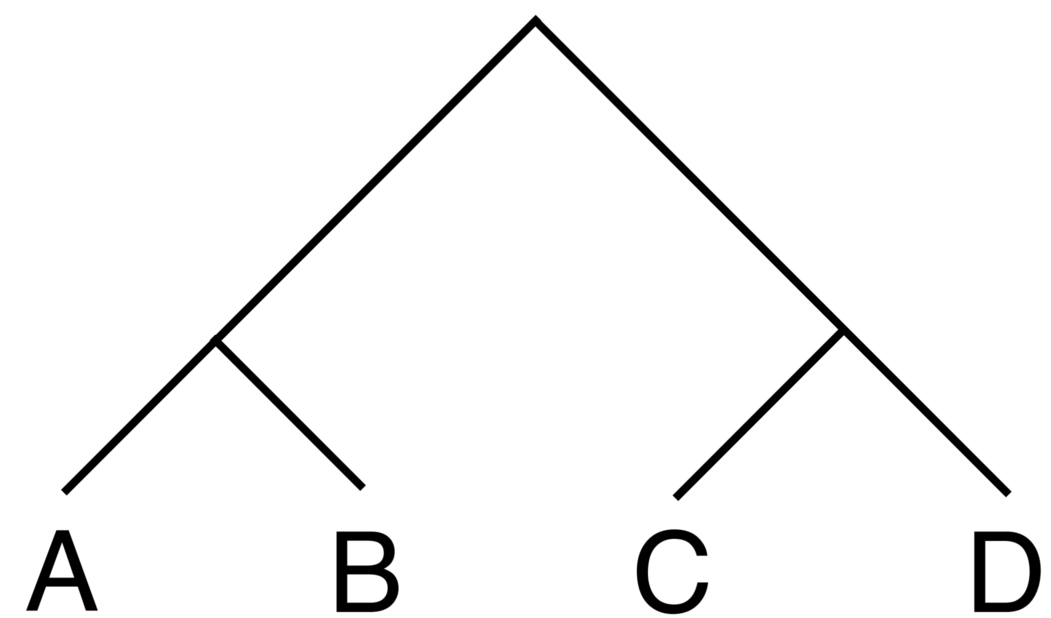


Four-population test: consider the following statistic, averaged over all SNPs in a genome

$$f_4 = (f_A - f_B)(f_C - f_D)$$

Reich et al. (2009)

Is a tree a good fit?



Four-population test: consider the statistic

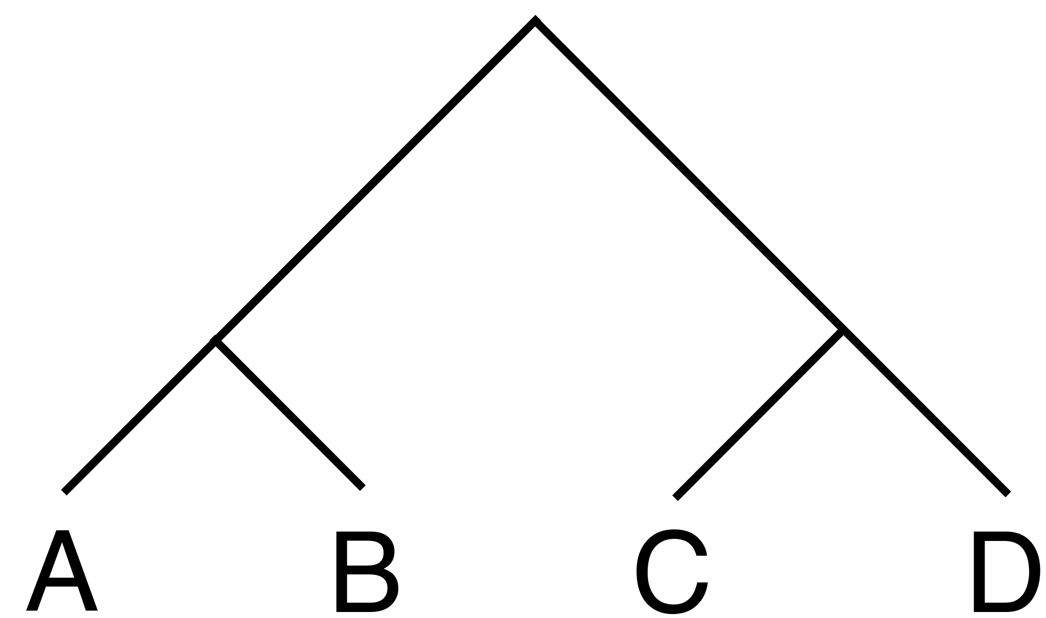
$$f_4 = (f_A - f_B)(f_C - f_D)$$

This is equivalent to the following expression in terms of covariances:

$$f_4 = V_{AC} - V_{BC} - V_{AD} + V_{BD}$$

Reich et al. (2009)

Is a tree a good fit?



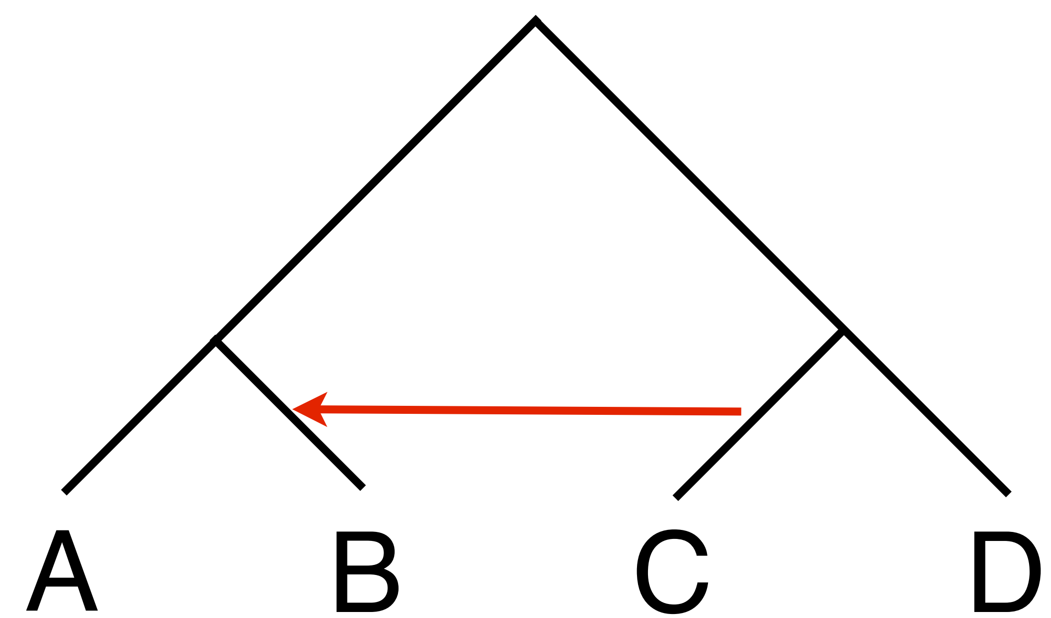
Four-population test: consider the statistic

$$f_4 = V_{AC} - V_{BC} - V_{AD} + V_{BD}$$

In the absence of migration (i.e. if the tree is correct), the expected value of this statistic is 0

Reich et al. (2009)

Is a tree a good fit?



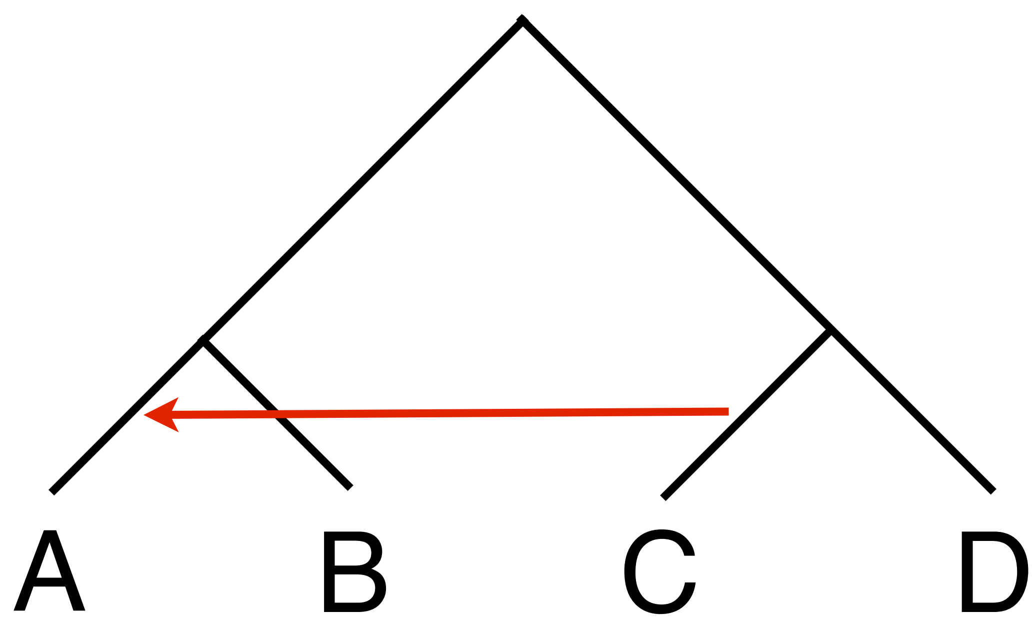
Four-population test: consider the statistic

$$f_4 = V_{AC} - \boxed{V_{BC}} - V_{AD} + V_{BD}$$

Negative value: gene flow between
(populations related to) B and C or A and D

Reich et al. (2009)

Is a tree a good fit?



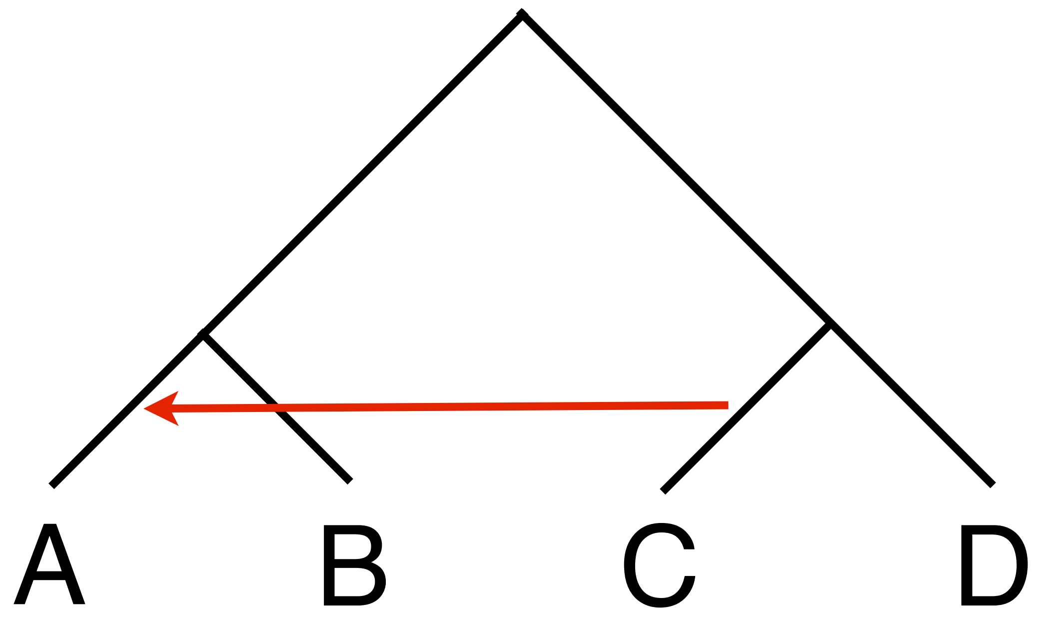
Four-population test: consider the statistic

$$f_4 = \boxed{V_{AC}} - V_{BC} - V_{AD} + V_{BD}$$

Positive value: gene flow between (populations related to) A and C or B and D

Reich et al. (2009)

Is a tree a good fit?



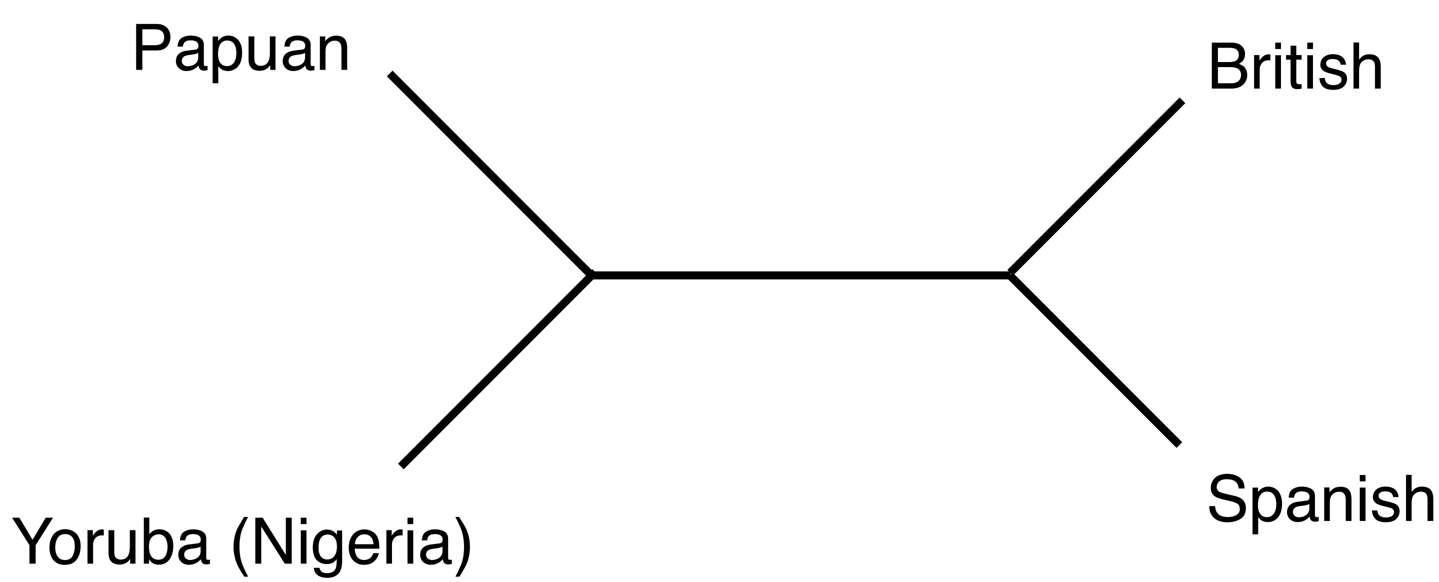
How to test significance? Compute statistic many times, dropping out large genomic regions, compute standard error in estimate, compute Z-score

(This is a standard statistical technique known as the jackknife)

Reich et al. (2009)

Application to humans

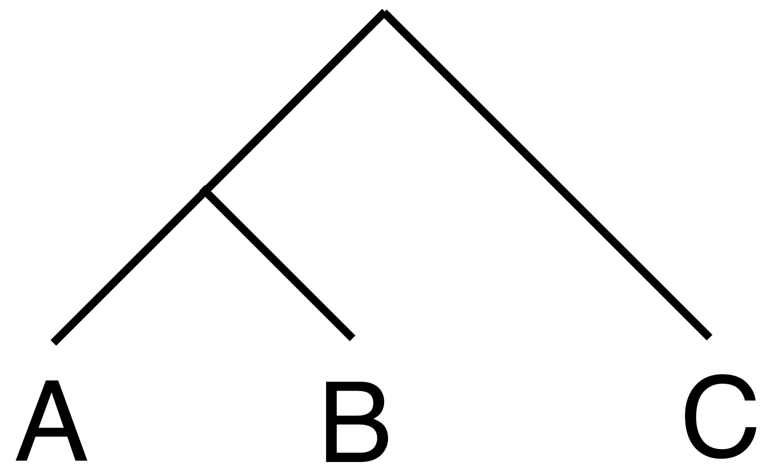
- Consider the following unrooted tree



- Does this tree work?
- Test (P-Y)(B-S), get value greater than 0, with a Z-score around 12 (p-value negligible)
- What is the interpretation?

Moorjani et al. (2011)

Is a tree a good fit?



Four-population tests are often hard to interpret. Consider a three-population test for admixture in population A:

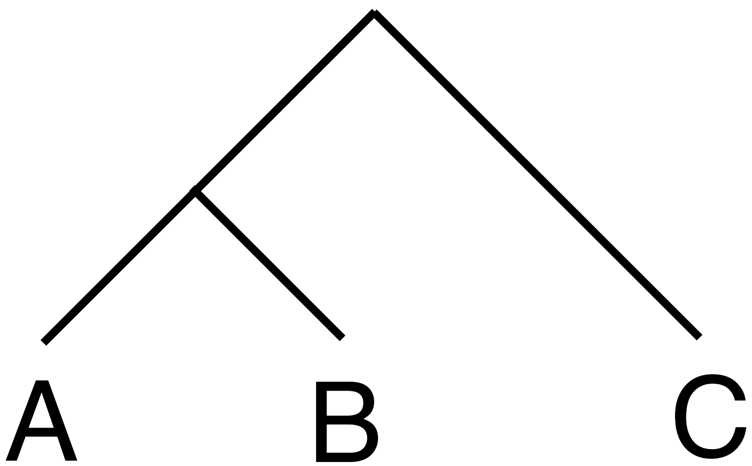
$$f_3 = (f_A - f_B)(f_A - f_C)$$

Which is equivalent to:

$$f_3 = V_{AA} - V_{AB} - V_{AC} + V_{BC}$$

Reich et al. (2009)

Is a tree a good fit?

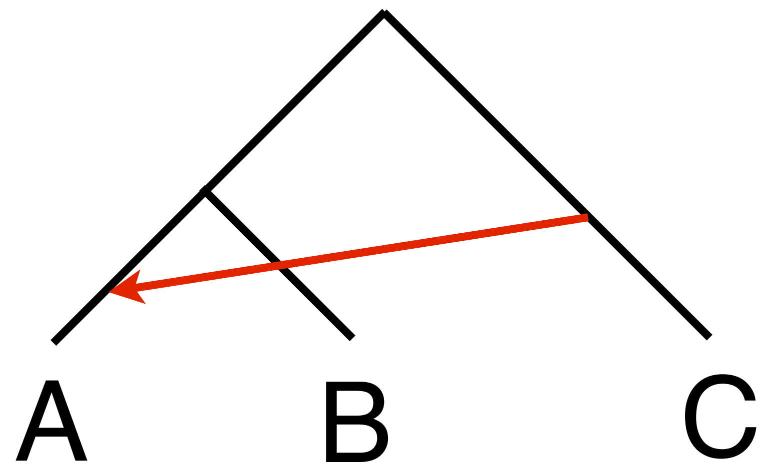


Three population test: in the absence of admixture in population A, this statistic is necessarily greater than zero

$$f_3 = V_{AA} - V_{AB} - V_{AC} + V_{BC}$$

Reich et al. (2009)

Is a tree a good fit?



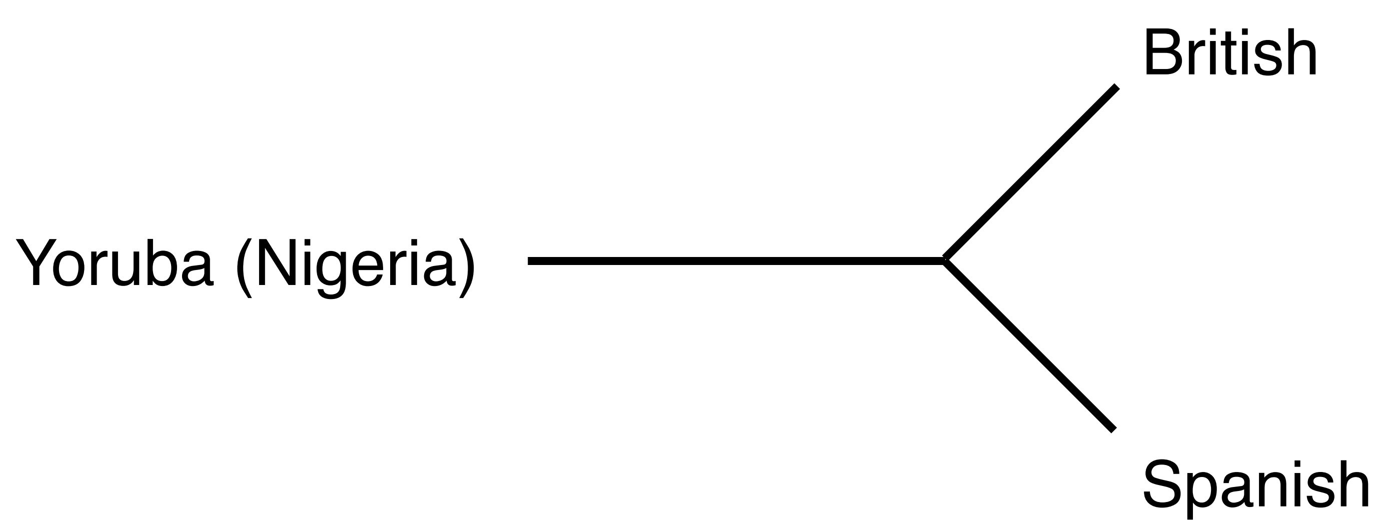
Three population test: in the presence of admixture in population A, this statistic can be less than zero

$$f_3 = V_{AA} - V_{AB} - \boxed{V_{AC}} + V_{BC}$$

Reich et al. (2009)

Application to humans

- Consider the following unrooted tree



- Does this tree work?

- Test Spanish for admixture; ie. calculate (S-Y)(S-B). Negative three-population test; Z-score around -20 (p-value negligible)

- What is the interpretation?

Moorjani et al. (2011)

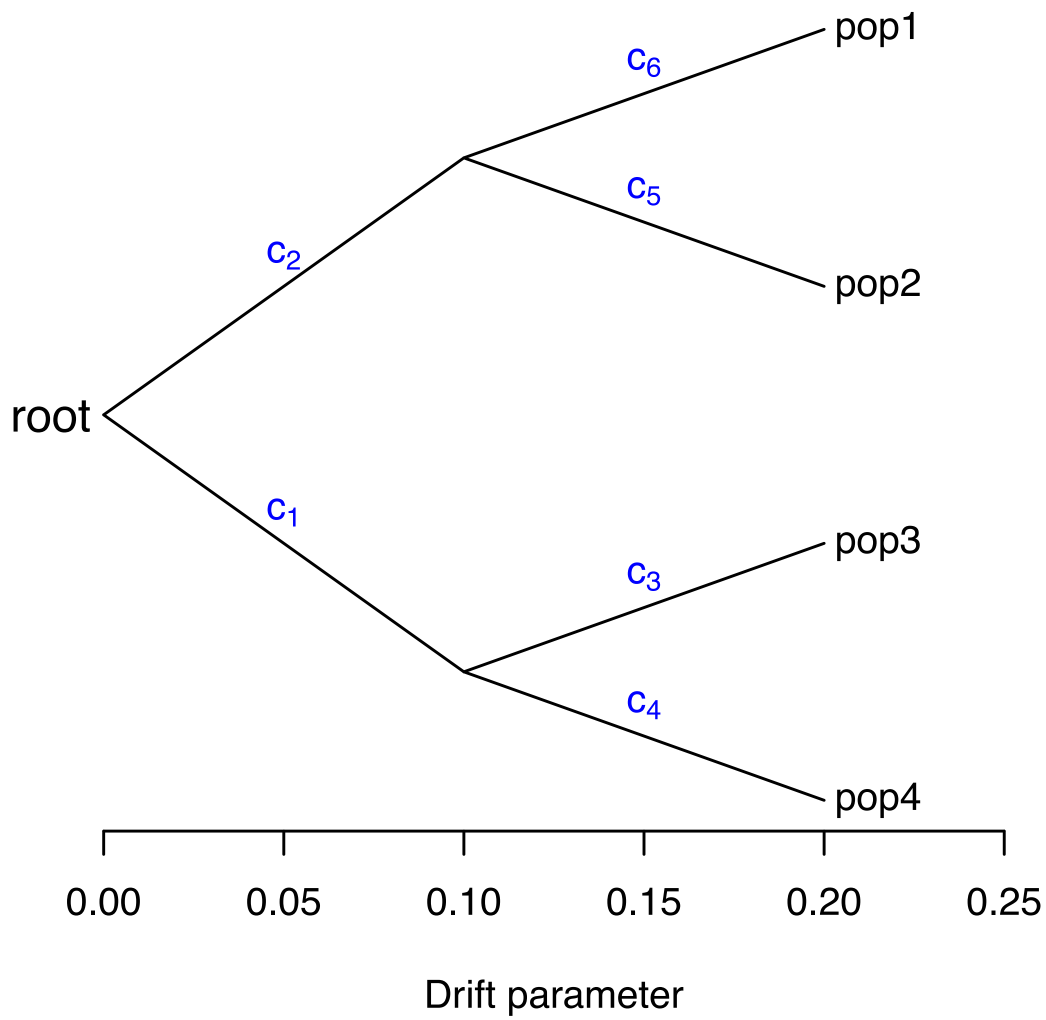
Can we incorporate migration into trees?

- Trees of populations can be constructed efficiently for many populations, three- and four-population tests indicate places where a tree fails

- Is there a model that can handle many populations with migration?

How are the allele frequencies in different populations related?

A. Example tree



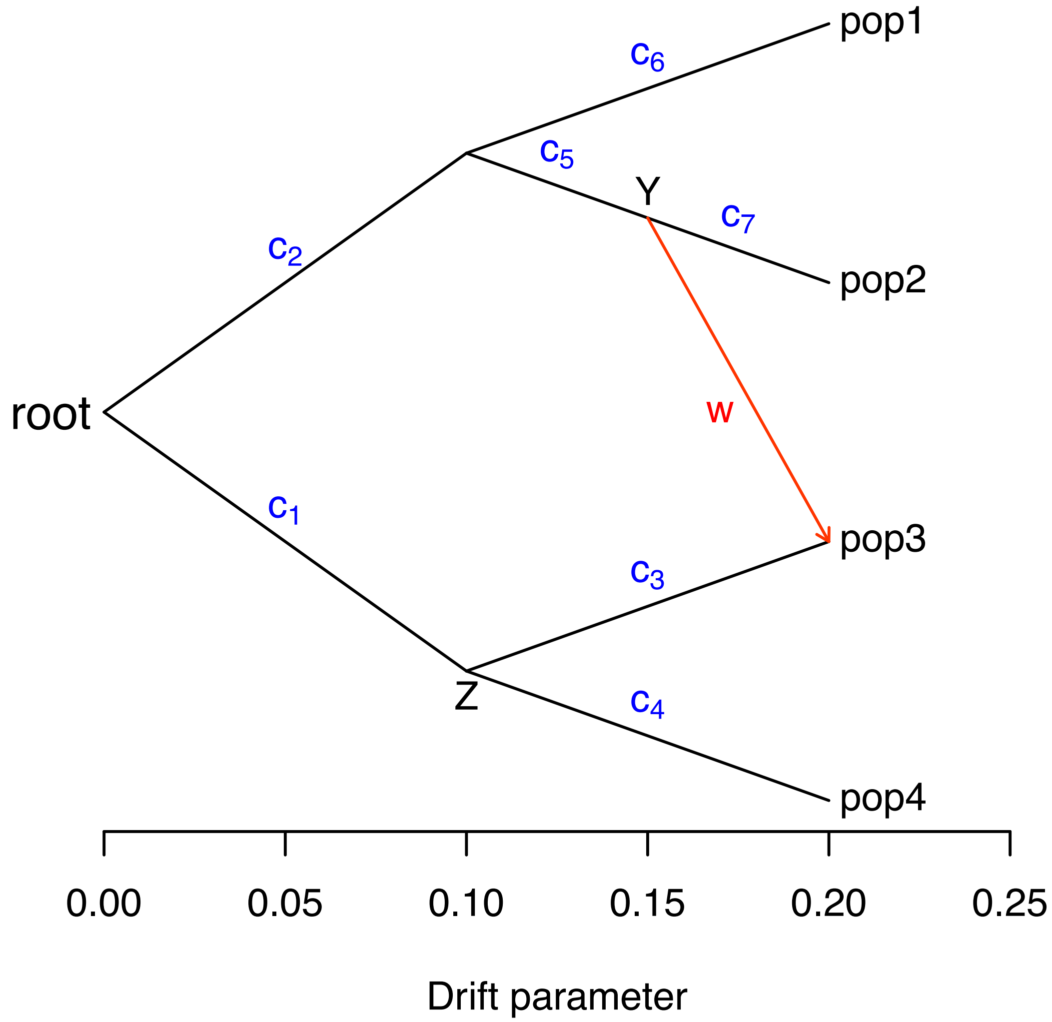
B. Covariance matrix for tree in A.

	pop1	pop2	pop3	pop4
pop1	$c_2 + c_6$	c_2	0	0
pop2	c_2	$c_2 + c_5$	0	0
pop3	0	0	$c_1 + c_3$	c_1
pop4	0	0	c_1	$c_1 + c_4$

Pickrell and Pritchard (2012)

How are the allele frequencies in different populations related?

C. Example graph



D. Covariance matrix for graph in C.

	pop1	pop2	pop3	pop4
pop1	$c_2 + c_6$	c_2	wc_2	0
pop2	c_2	$c_2 + c_5 + c_7$	$w(c_2 + c_5)$	0
pop3	wc_2	$w(c_2 + c_5)$	$w^2(c_2 + c_5) + (1-w)^2(c_1 + c_3)$	$(1-w)c_1$
pop4	0	0	$(1-w)c_1$	$c_1 + c_4$

Pickrell and Pritchard (2012)

See also Patterson et al. (2012)

Fit the observed matrix to the one predicted by the tree/graph

Graph-based prediction

	pop1	pop2	pop3	pop4
pop1	$c_2 + c_6$	c_2	wc_2	0
pop2	c_2	$c_2 + c_5 + c_7$	$w(c_2 + c_5)$	0
pop3	wc_2	$w(c_2 + c_5)$	$w^2(c_2 + c_5) + (1-w)^2(c_1 + c_3)$	$(1-w)c_1$
pop4	0	0	$(1-w)c_1$	$c_1 + c_4$

Observed data

	pop1	pop2	pop3	pop4
pop1	$\hat{\mathbf{W}}_{11}$	$\hat{\mathbf{W}}_{12}$	$\cdots$	
pop2	$\hat{\mathbf{W}}_{21}$	$\hat{\mathbf{W}}_{22}$	$\cdots$	
pop3	$\cdots$	$\cdots$		
pop4				

Algorithm: find a graph that best fits the data

How to quantify fit?

Graph-based prediction (W)

	pop1	pop2	pop3	pop4
pop1	$c_2 + c_6$	c_2	wc_2	0
pop2	c_2	$c_2 + c_5 + c_7$	$w(c_2 + c_5)$	0
pop3	wc_2	$w(c_2 + c_5)$	$w^2(c_2 + c_5)$ $+(1-w)^2(c_1 + c_3)$	$(1-w)c_1$
pop4	0	0	$(1-w)c_1$	$c_1 + c_4$

Observed data

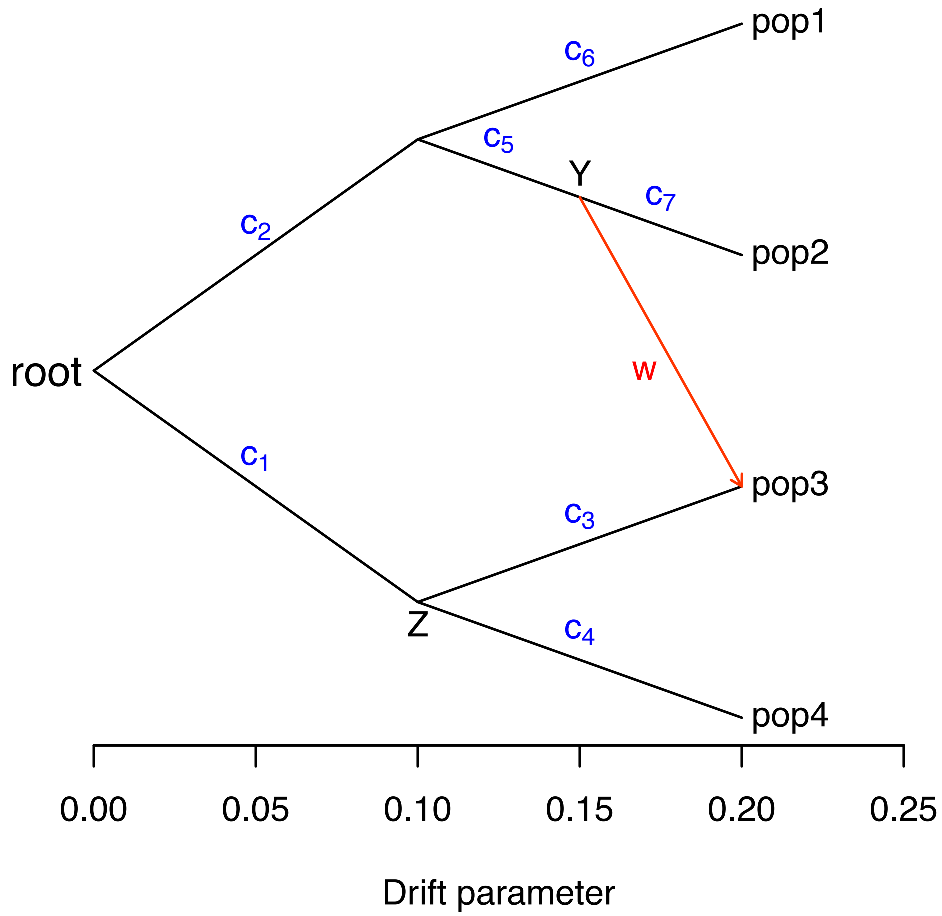
	pop1	pop2	pop3	pop4
pop1	$\hat{\mathbf{W}}_{11}$	$\hat{\mathbf{W}}_{12}$	$\cdots$	
pop2	$\hat{\mathbf{W}}_{21}$	$\hat{\mathbf{W}}_{22}$	$\cdots$	
pop3	$\cdots$	$\cdots$		
pop4				

Composite likelihood:

$$l(\hat{W}|W) = \sum_{i=0}^{m} \sum_{j=i}^{m} N(\hat{W}_{ij}|W_{ij}, \hat{\sigma}_{ij}^2)$$

Estimation

C. Example graph



D. Covariance matrix for graph in C.

	pop1	pop2	pop3	pop4
pop1	$c_2 + c_6$	c_2	wc_2	0
pop2	c_2	$c_2 + c_5 + c_7$	$w(c_2 + c_5)$	0
pop3	wc_2	$w(c_2 + c_5)$	$w^2(c_2 + c_5)$ $+(1-w)^2(c_1 + c_3)$	$(1-w)c_1$
pop4	0	0	$(1-w)c_1$	$c_1 + c_4$

$$C_{11} = c_6 + c_2$$

$$C_{12} = c_2$$

$$C_{13} = wc_2$$

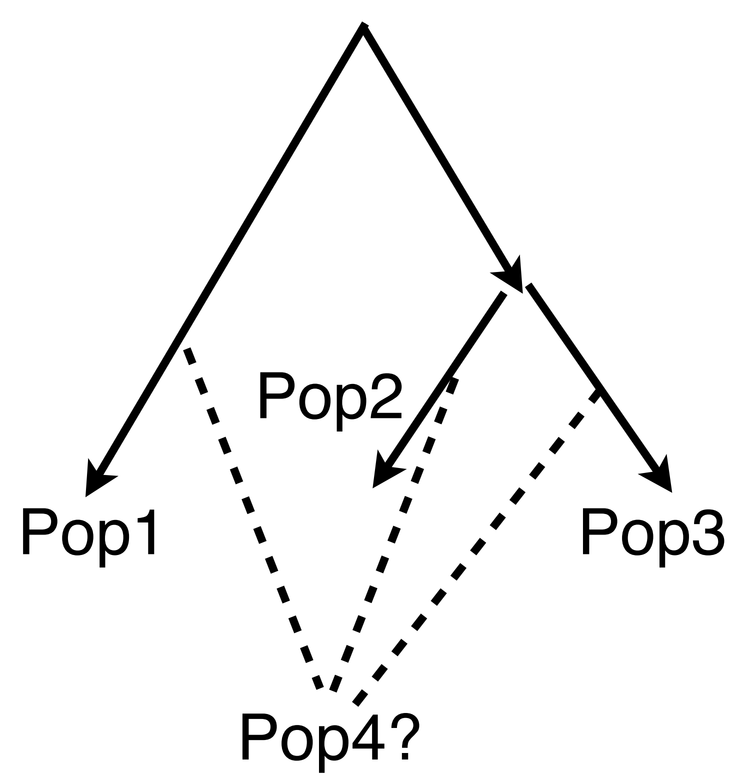
$$\dots$$

Fix w, solve c's by (non-negative) least squares

Search over w to get MLEs for a given topology

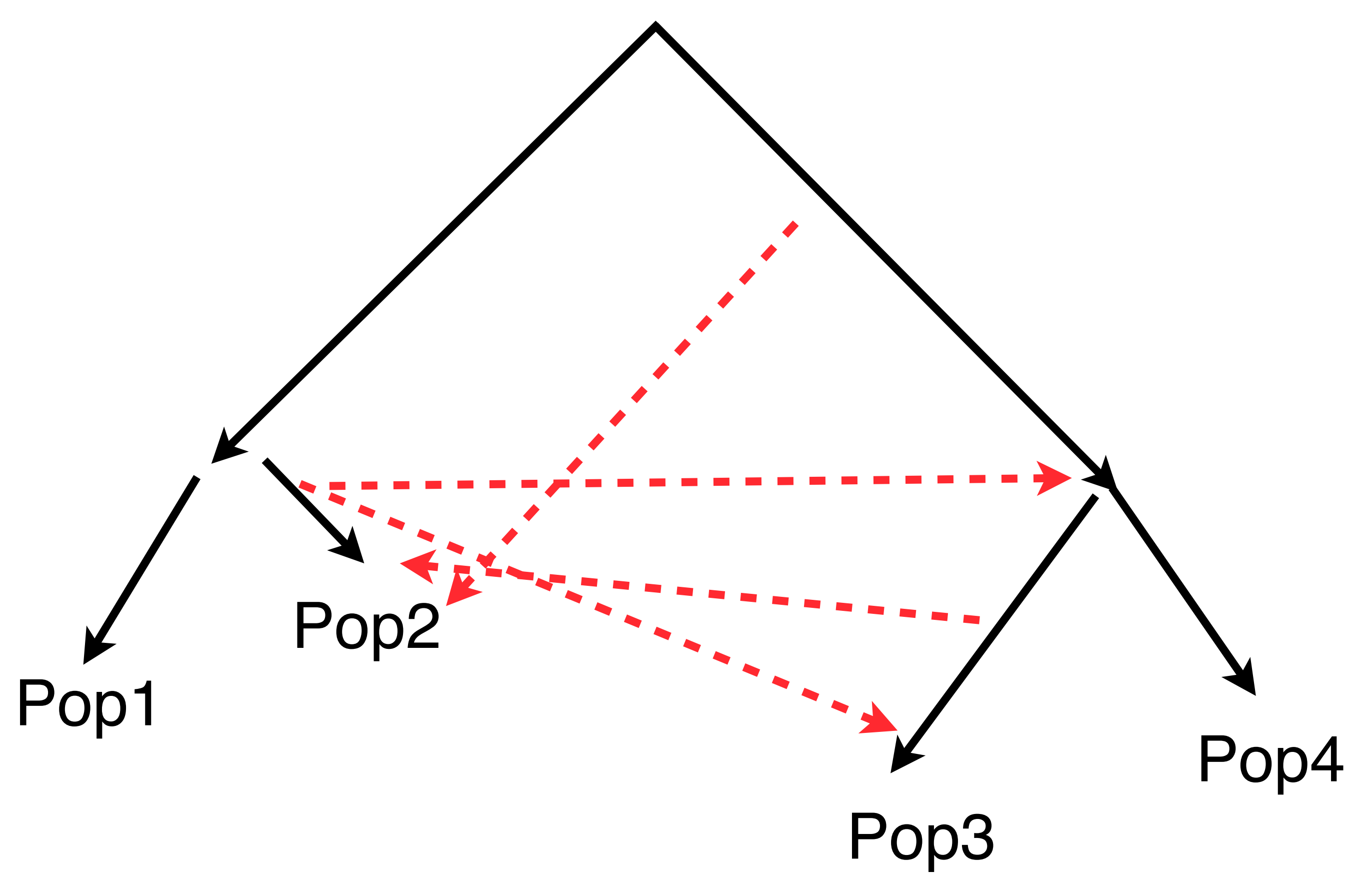
Estimation

1. Estimate tree without migration

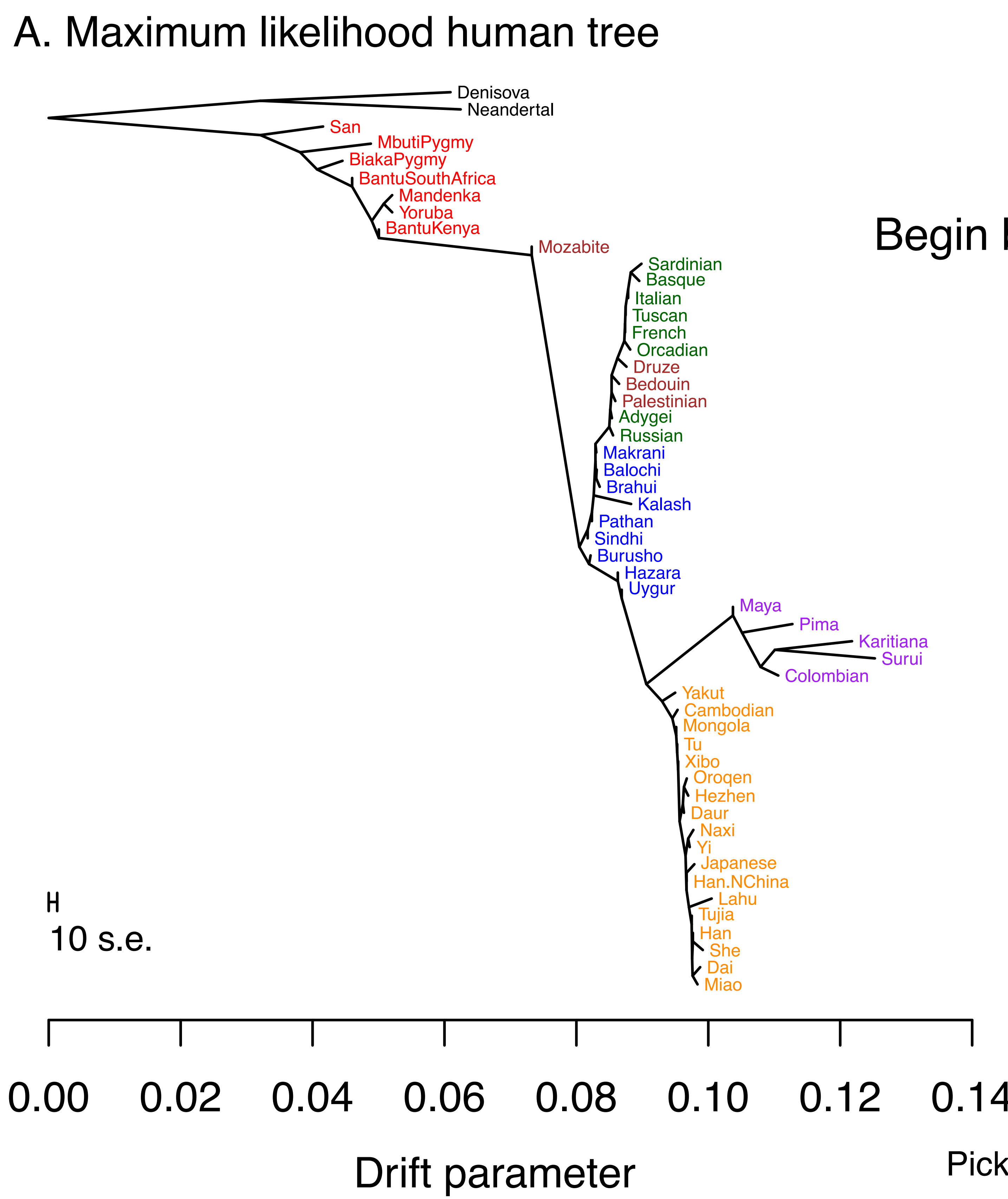


Estimation

2. Find poorly fitted populations, try migration events



Application 1: worldwide sample of humans

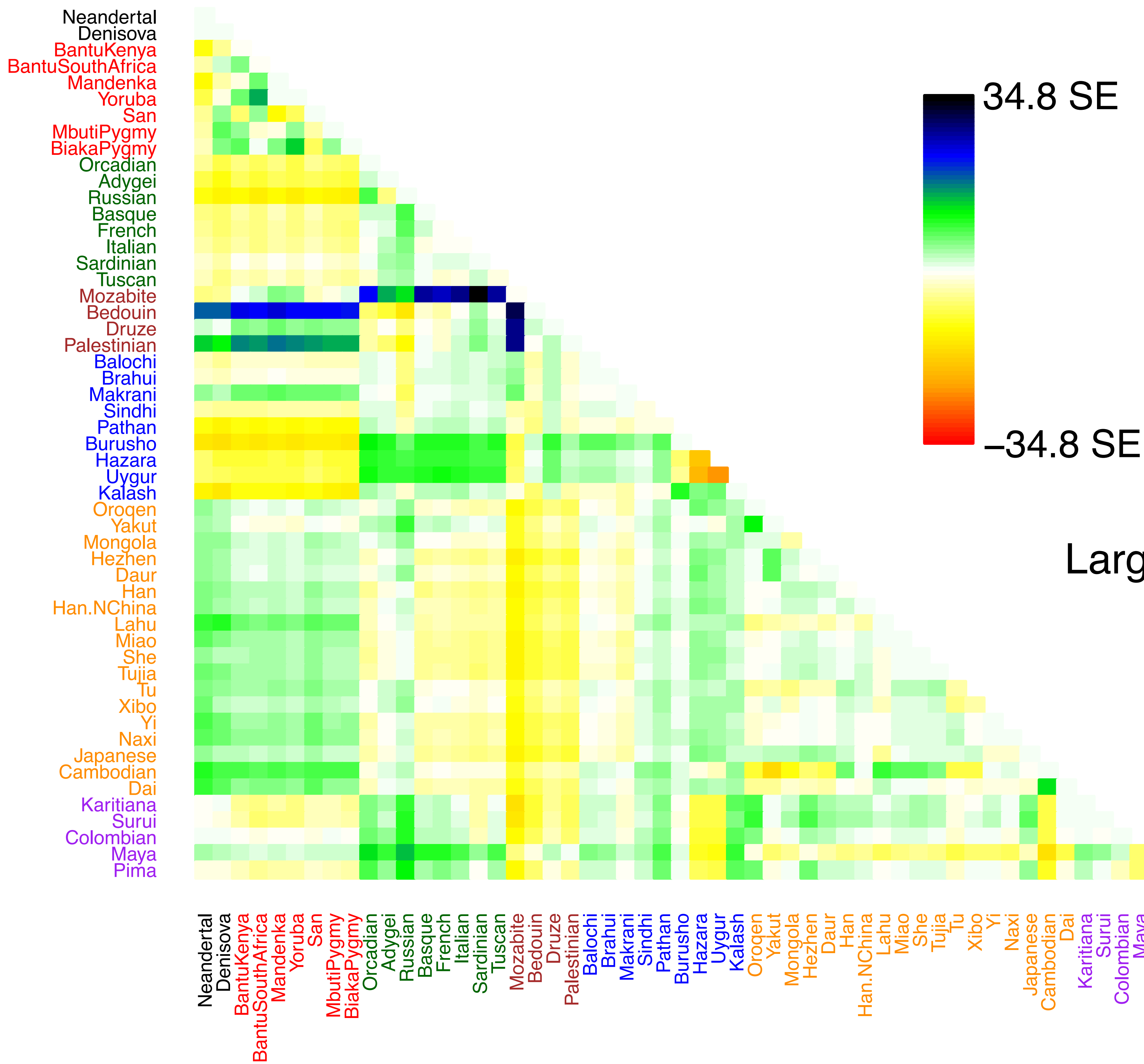


Begin by building tree

Pickrell and Pritchard (2012)

Application 1: worldwide sample of humans

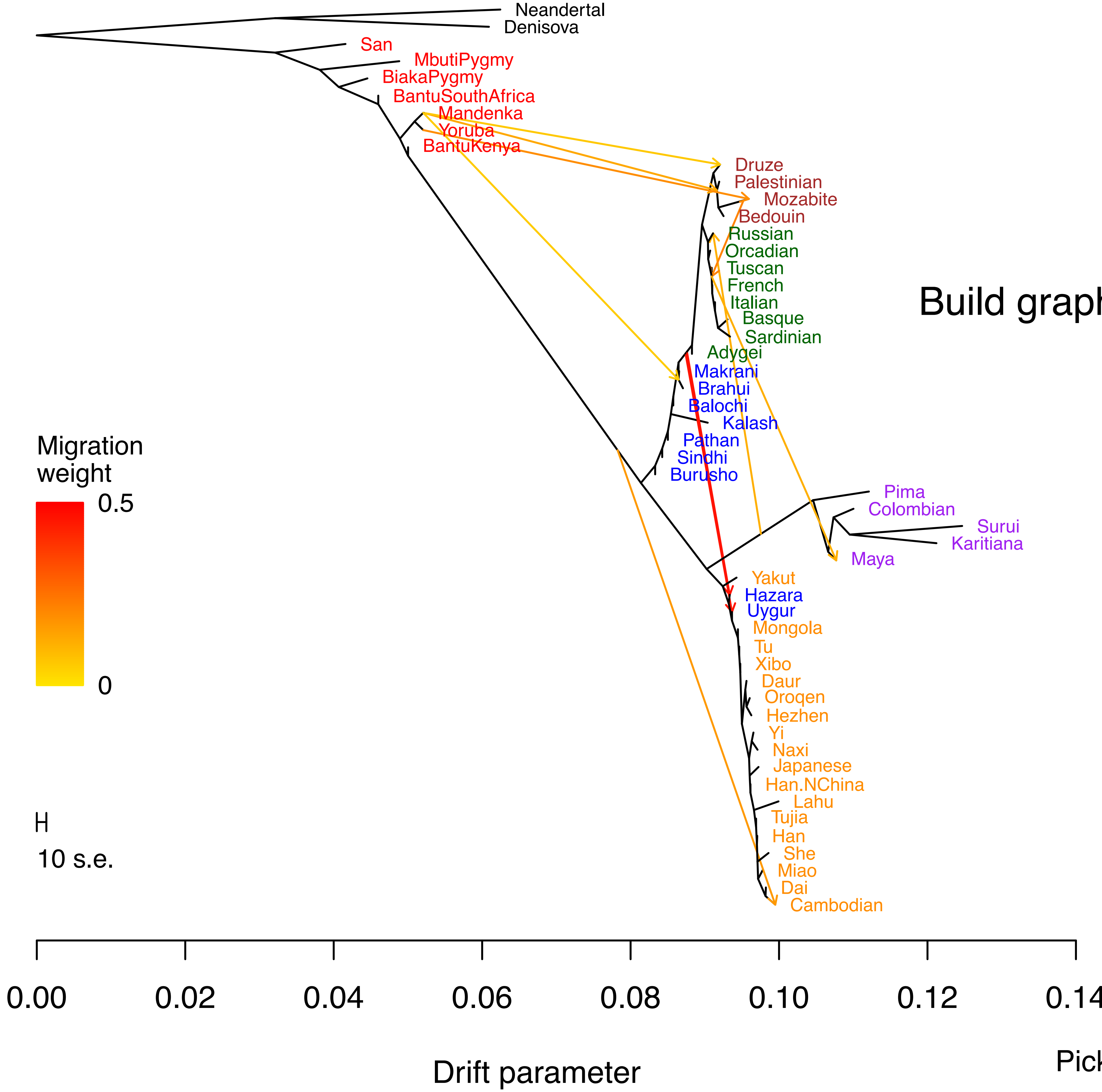
B. Residual fit from tree



Large residuals indicate poorly fitting populations

Pickrell and Pritchard (2012)

Application 1: worldwide sample of humans



Build graph of populations

Pickrell and Pritchard (2012)