

# Quantifying relationships among populations: Simple tests



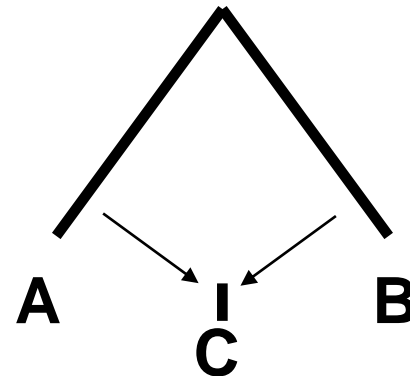
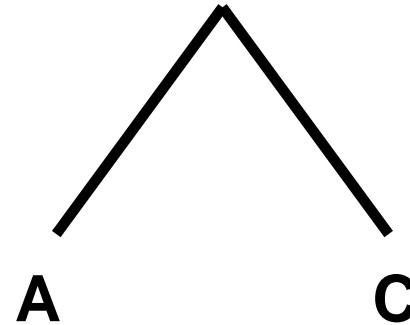
Andrea Manica



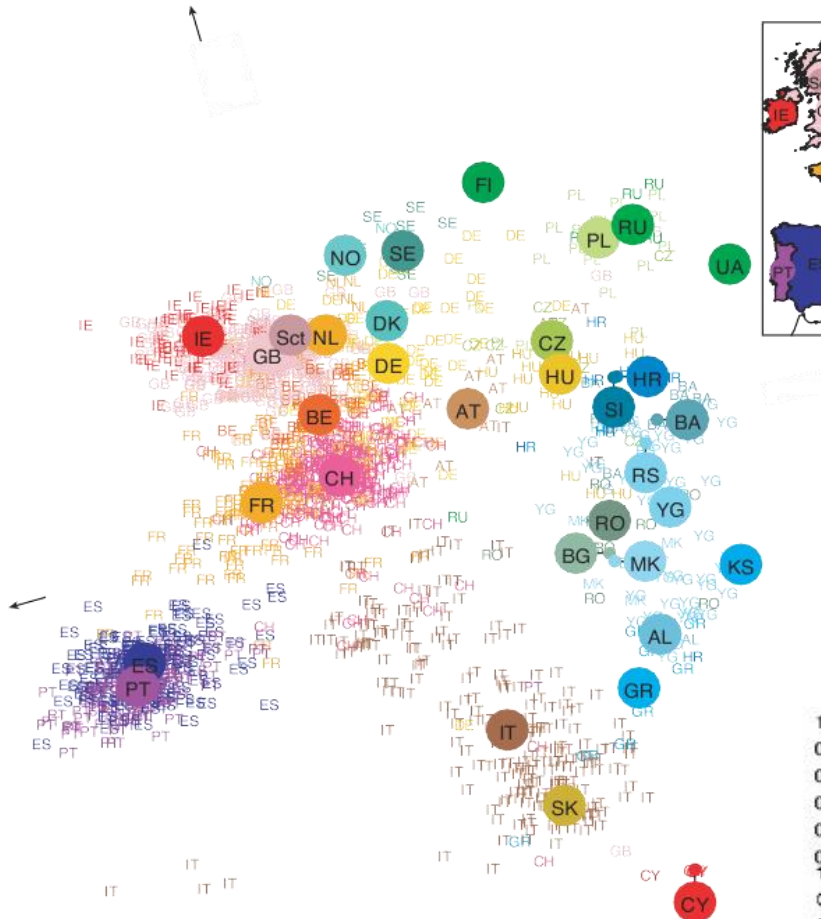
UNIVERSITY OF  
CAMBRIDGE

# Outline

- Using  $f$  statistics (measure of drift) to compare populations
- Relationships among populations on a simple tree
- Admixture

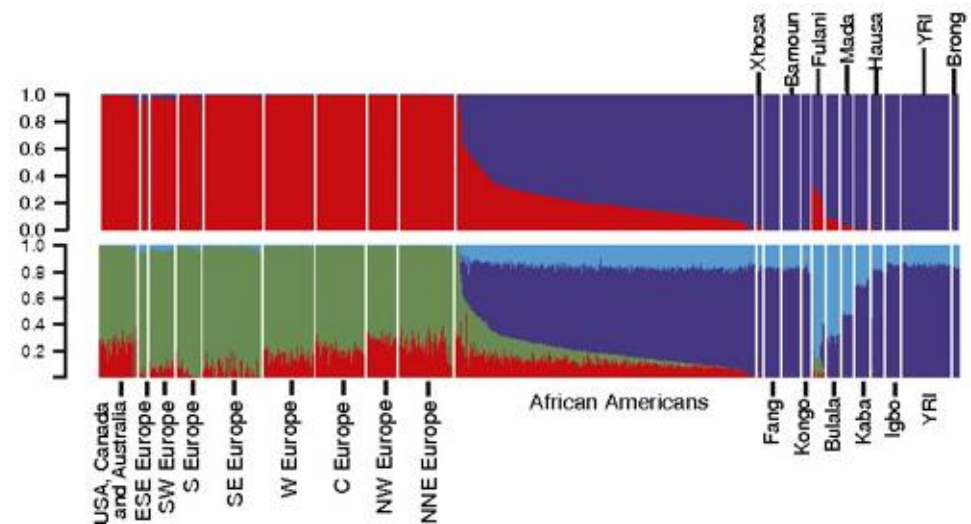


# Relationships among populations



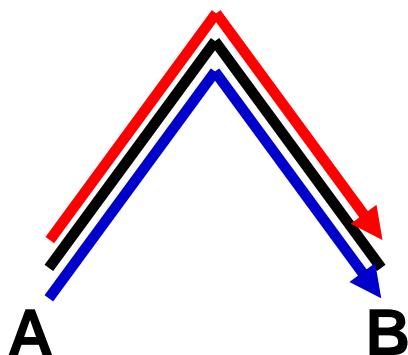
Proximity in  
PCA

Clustering algorithms  
(Leo's lectures)



## $f_2$ as a measure of distance

$$\begin{aligned} f_2(A,B) &= E[(p_A - p_B)^2] \\ &= E[(p_A - p_B) (p_A - p_B)] \end{aligned}$$



## $f_2$ as a measure of distance

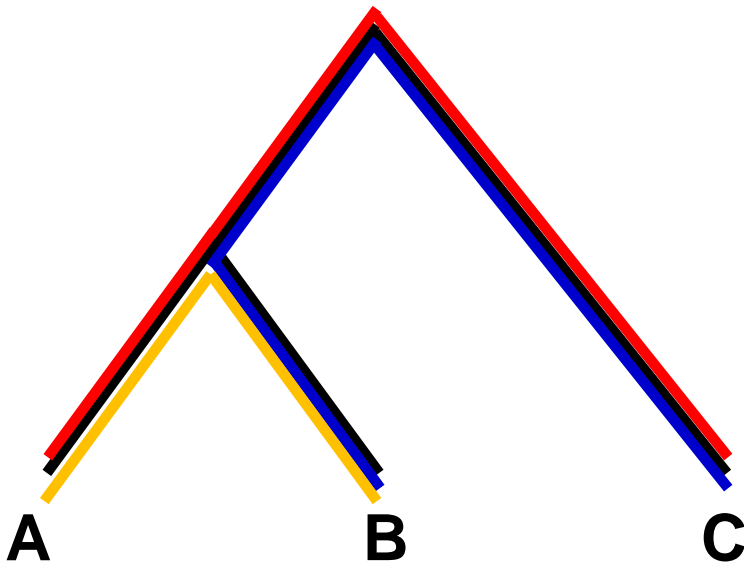
$$\begin{aligned} f_2(A,B) &= E[(p_A - p_B)^2] \\ &= E[(p_A - p_B) (p_A - p_B)] \end{aligned}$$



## Outgroup $f_3$ to define similarity

$$f_2(A,B) = E[(p_A - p_B)^2]$$

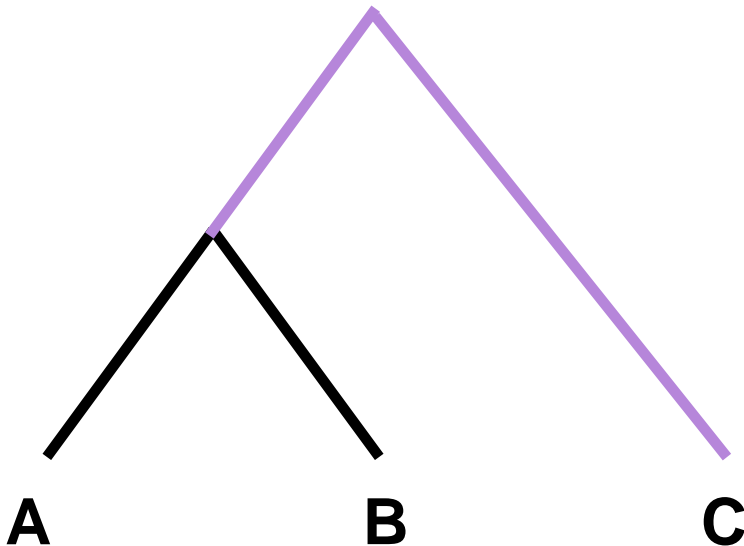
$$\begin{aligned} f_3(C; A,B) &= E[(p_C - p_A) [(p_C - p_B)]] = \\ &= \frac{1}{2} [f_2(C,A) + f_2(C,B) - f_2(A,B)] \end{aligned}$$



## Outgroup $f_3$ to define similarity

$$f_2(A,B) = E[(p_A - p_B)^2]$$

$$\begin{aligned} f_3(C; A,B) &= E[(p_C - p_A) [(p_C - p_B)]] = \\ &= \frac{1}{2} [f_2(C,A) + f_2(C,B) - f_2(A,B)] \end{aligned}$$

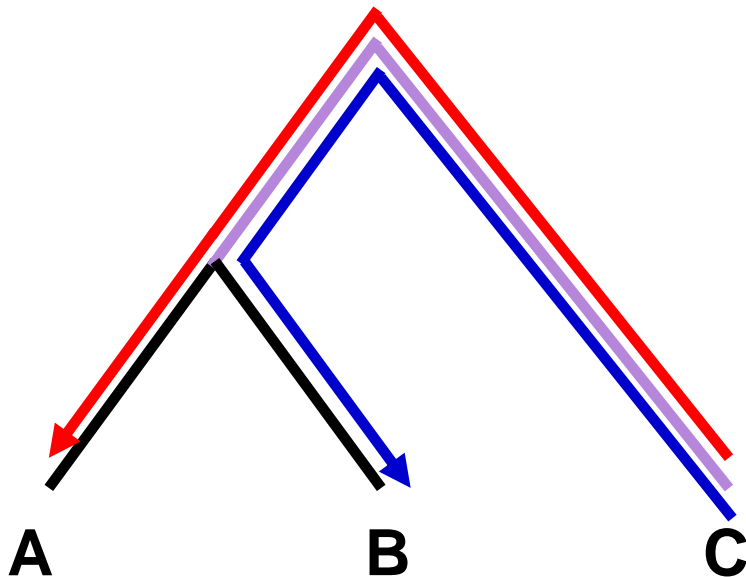


Shared drift for A and  
B from C

## Outgroup $f_3$ to define similarity

$$f_2(A,B) = E[(p_A - p_B)^2]$$

$$\begin{aligned} f_3(C; A,B) &= E[(p_C - p_A) [(p_C - p_B)] = \\ &= \frac{1}{2} [f_2(C,A) + f_2(C,B) - f_2(A,B)] \end{aligned}$$



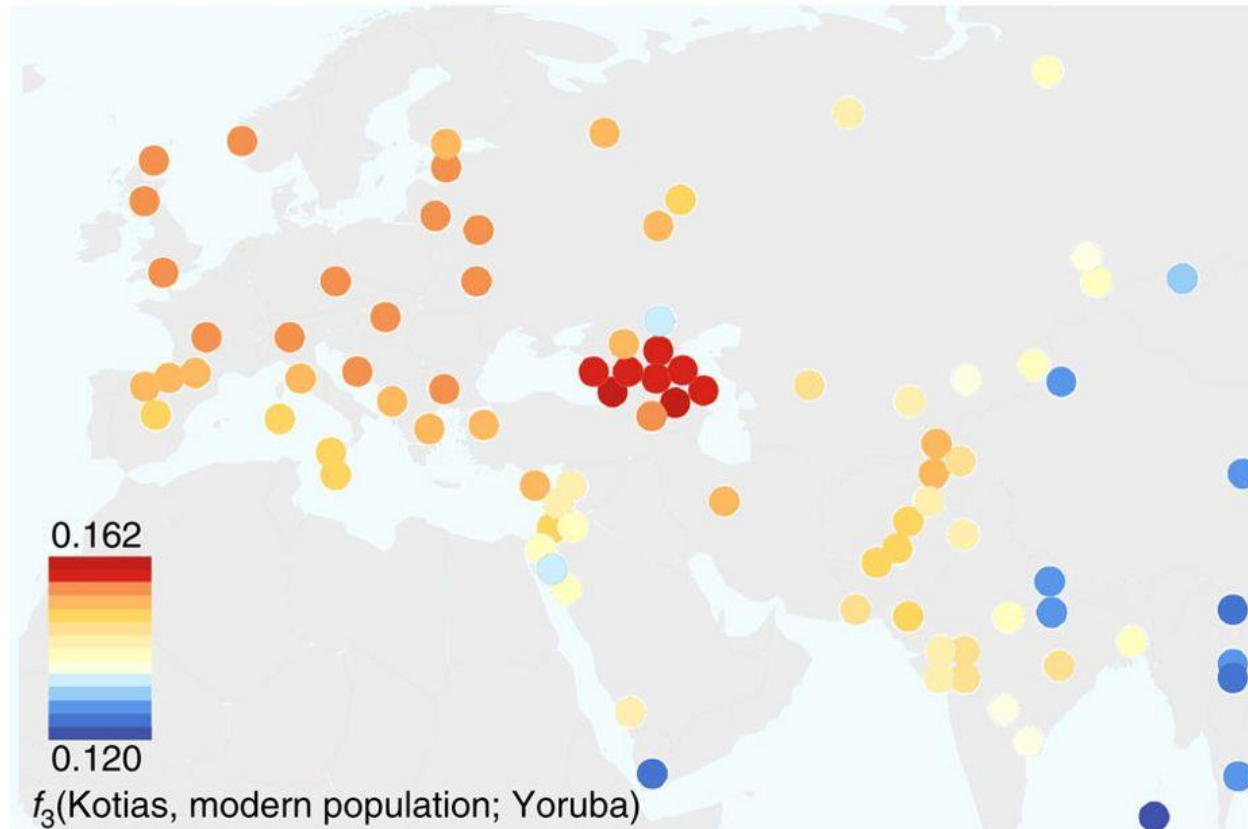
Shared drift for A and  
B from C

Overlap of paths C- $\rightarrow$ A  
and C- $\rightarrow$ B



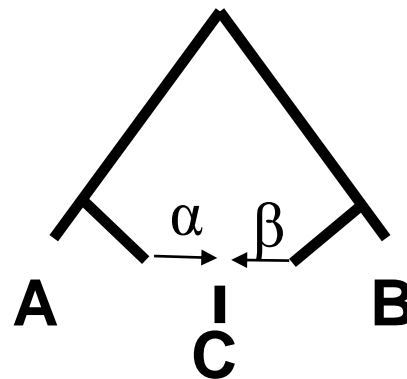
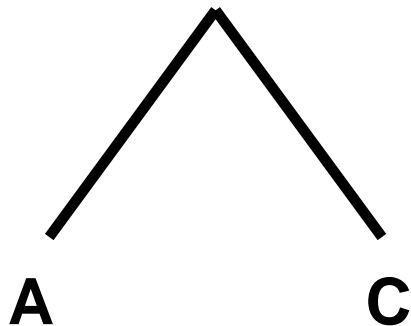
# Outgroup $f_3$ to define similarity

a



Admixture  $f_3$  if topology is not a simple tree

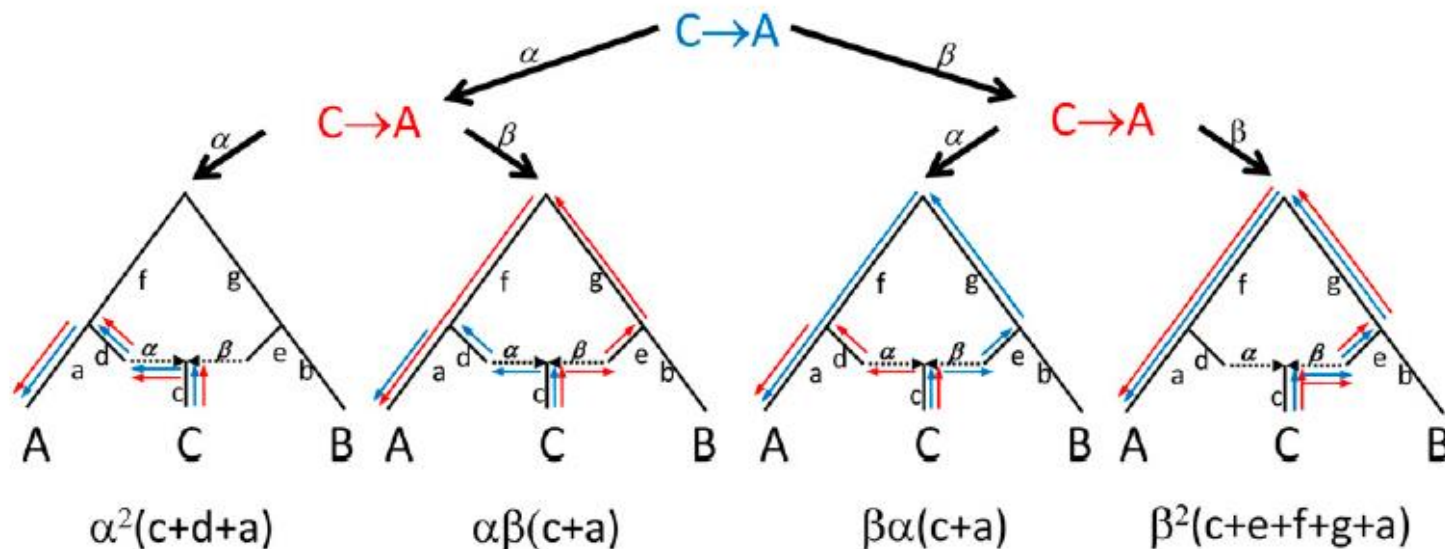
$$f_2(C,A) = E[(p_C - p_A)^2] = E[(p_C - p_A)(p_C - p_A)]$$



## Admixture $f_3$ if topology is not a simple tree

$$f_2(C,A) = E[(p_C - p_A)^2] = E[(p_C - p_A)(p_C - p_A)]$$

$$F_2(C,A) = a + c + \alpha^2 d + \beta^2(e + g + f)$$

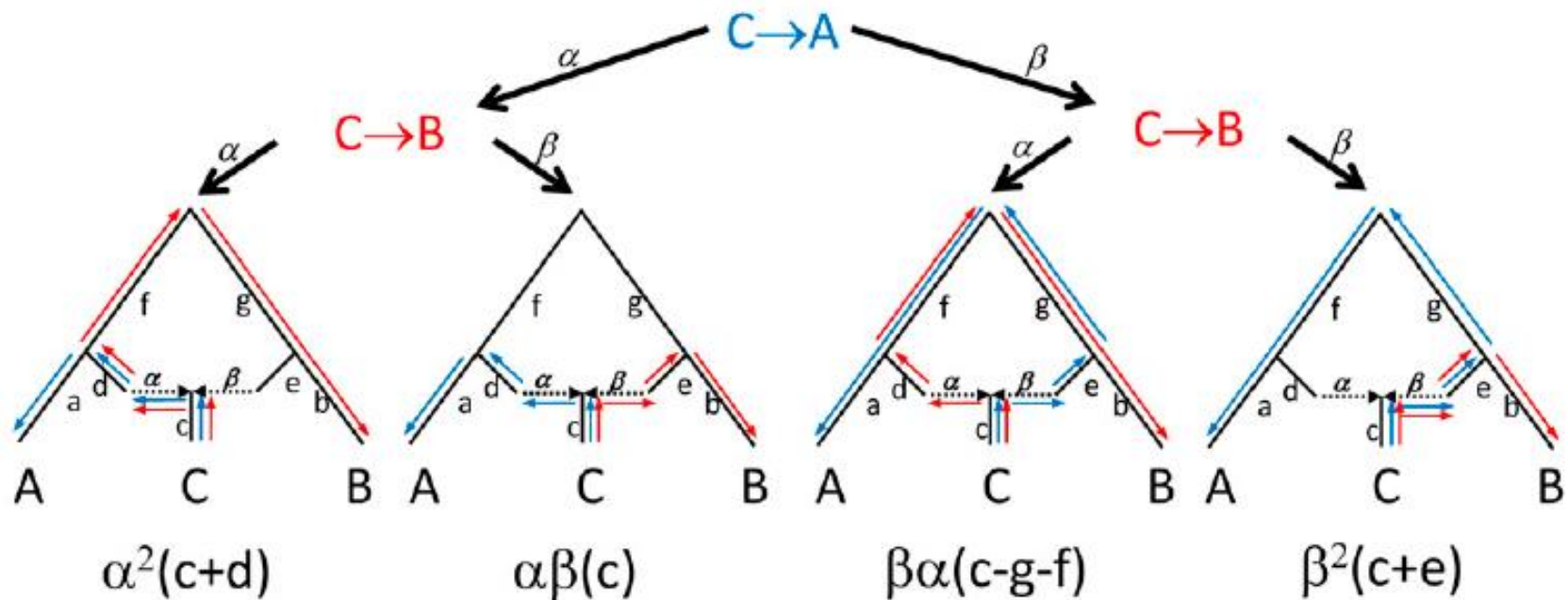


Tip:  $f$  statistics can be seen as overlap of paths from one sample to the other in the quadratic term

## Admixture $f_3$ if topology is not a simple tree

$$f_3(C; A, B) = E[(p_C - p_A) [(p_C - p_B)]] =$$

$$F_3(C; A, B) = c + \alpha^2 d + \beta^2 e - \alpha\beta(g+f)$$

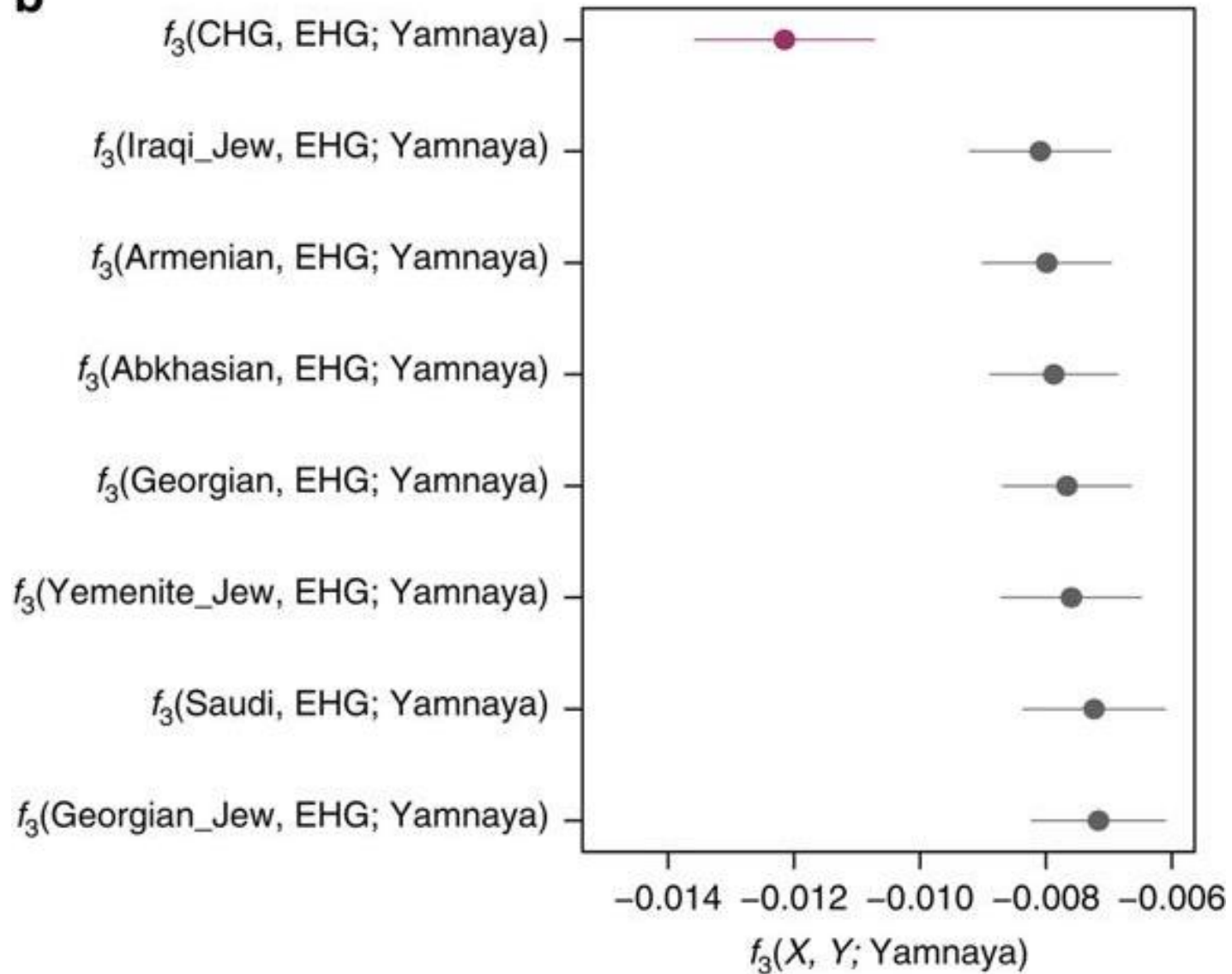


Negative  $f_3$  is a sign of admixture

But note that a positive  $f_3$  does not prove absence of admixture (e.g. large  $c$ , drift to C)

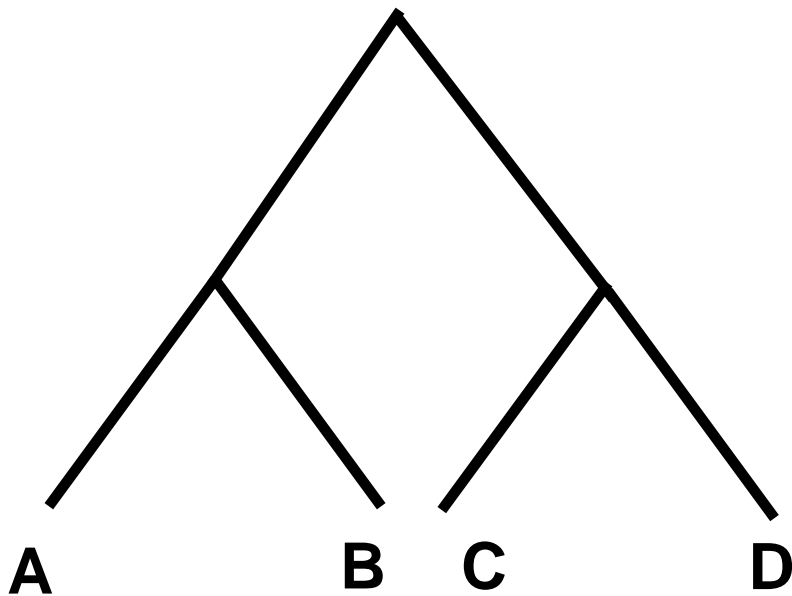
## Admixture $f_3$

**b**



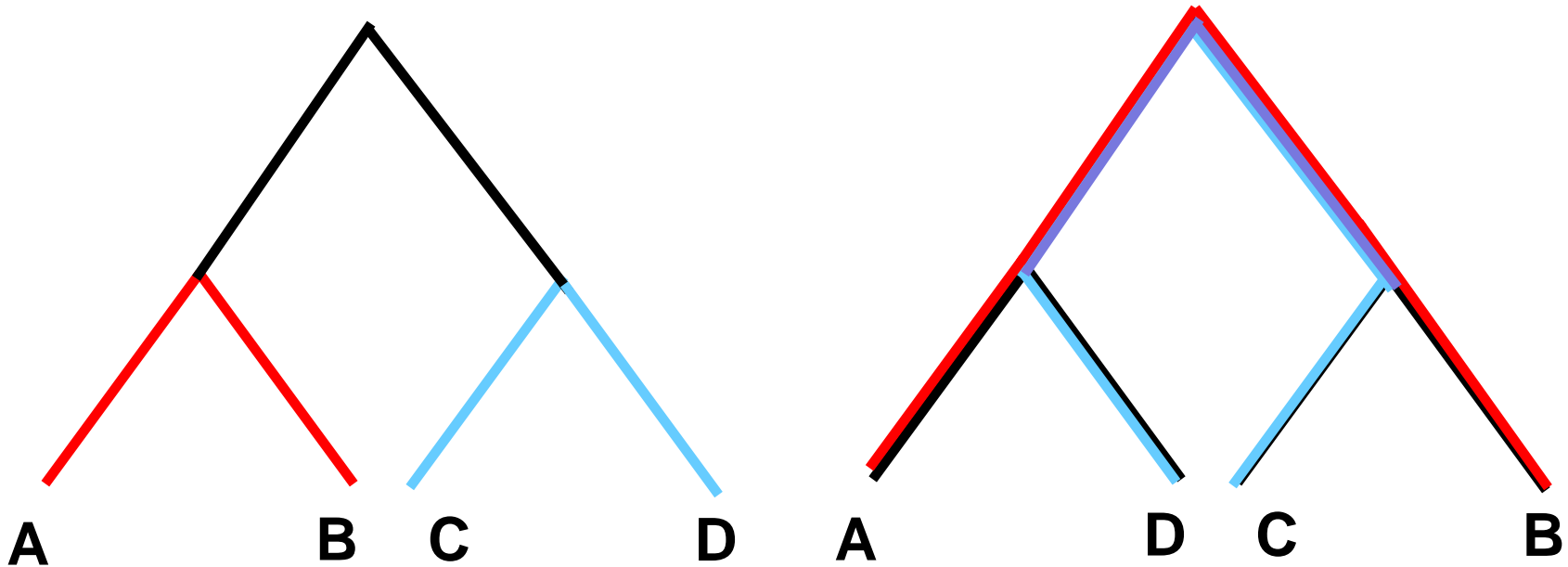
## $f_4$ for a simple tree

$$f_4(A,B; C,D) = E[(p_A - p_B) [(p_C - p_D)] =$$



## $f_4$ for a simple tree

$$f_4(A,B; C,D) = E[(p_A - p_B)(p_C - p_D)]$$



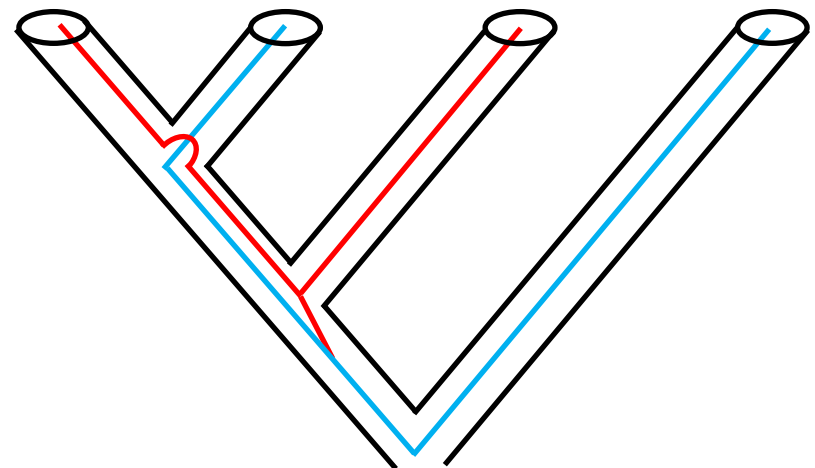
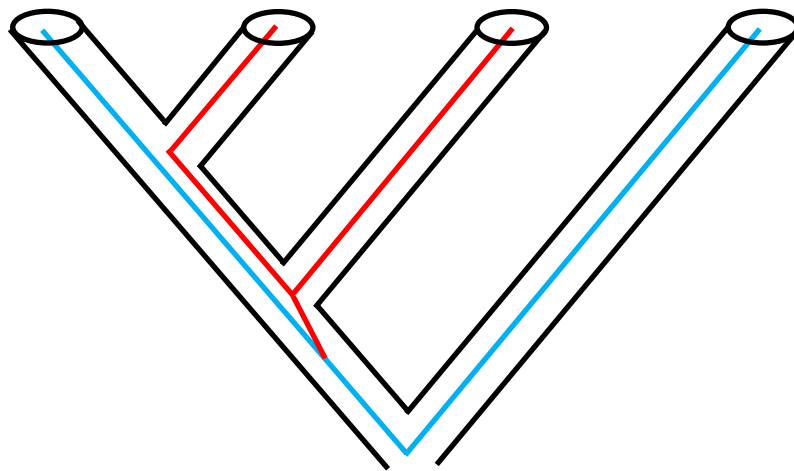
Expect  $f_4$  to be zero if the two pairs form clades with respect to each other

# D-statistics (ABBA-BABA)



$$D = \frac{\#ABBA - \#BABA}{\#ABBA + \#BABA}$$

$$f_4 = \frac{\#ABBA - \#BABA}{\#sites}$$



$D(O, P_1; P_2, P_3)$

$D = 0$  true phylogeny

$D < 0$  gene flow  $P_1 \& P_2$

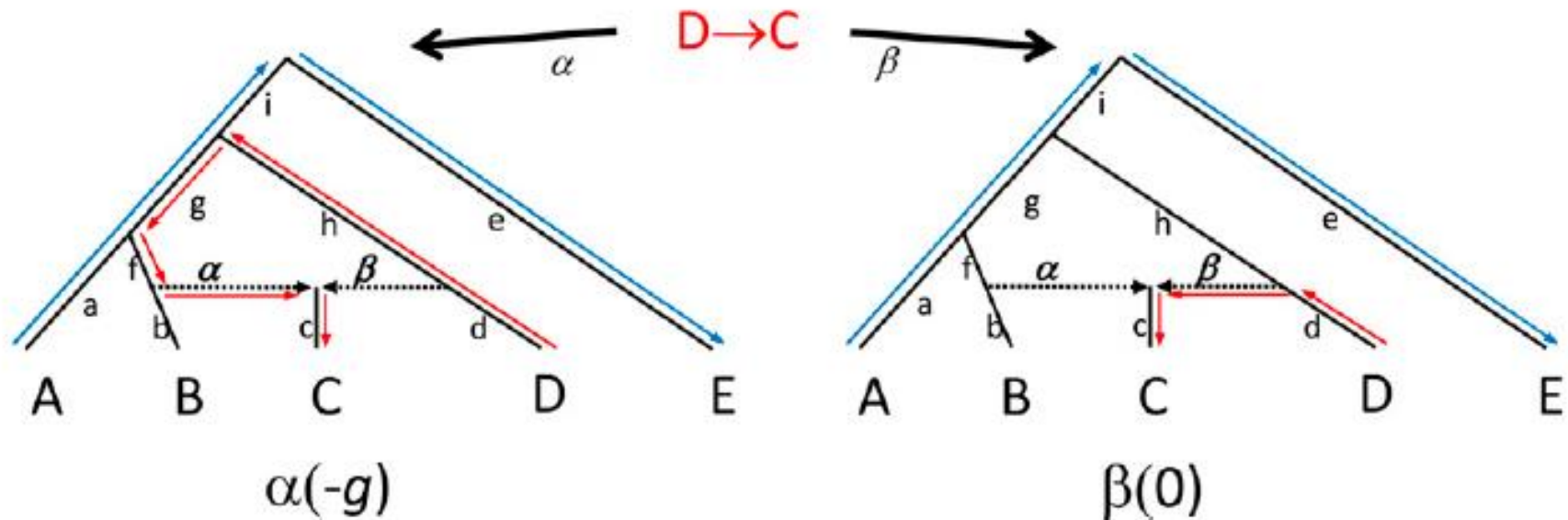
$D > 0$  gene flow  $P_1 \& P_3$



$f_4$  if topology is not a simple tree

$$F_4(A,E;D,C) = -\alpha g$$

$$F_4 \text{ ratio} = \frac{F_4(A,E;D,C)}{F_4(A,E;D,B)} = \frac{-\alpha g}{-g} = \alpha$$



A & E set the background, B & D are the sources

## Testing significance of relationships

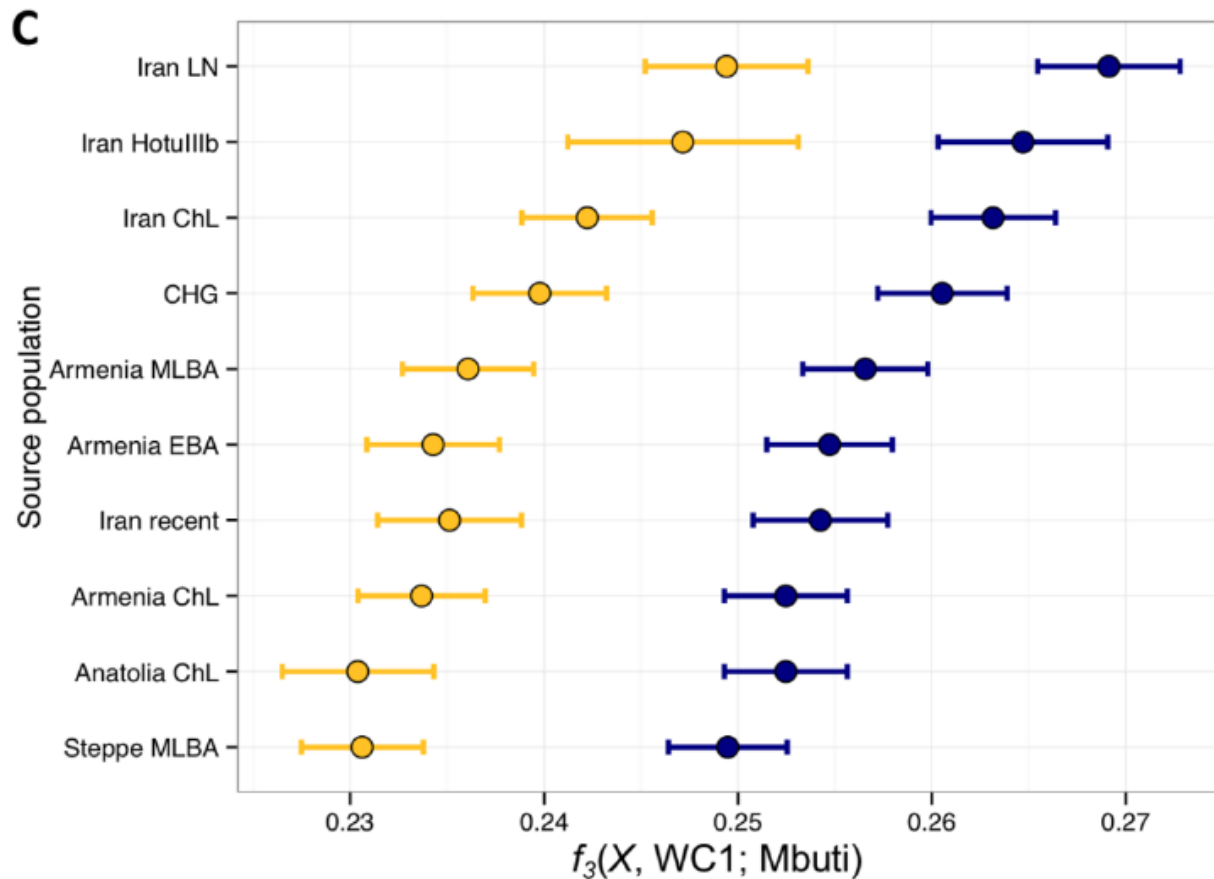
We can test significant by block jackknifing (usually using 5cM)

Estimate  $f_x$  and its standard error, and then compute a standardized  $Z$

$|Z| > 2$  or  $3$  considered significant (but note that the critical value has been changing... not always for the right reasons)

Be careful that you are not comparing apples and oranges (SNPs from the same panel)

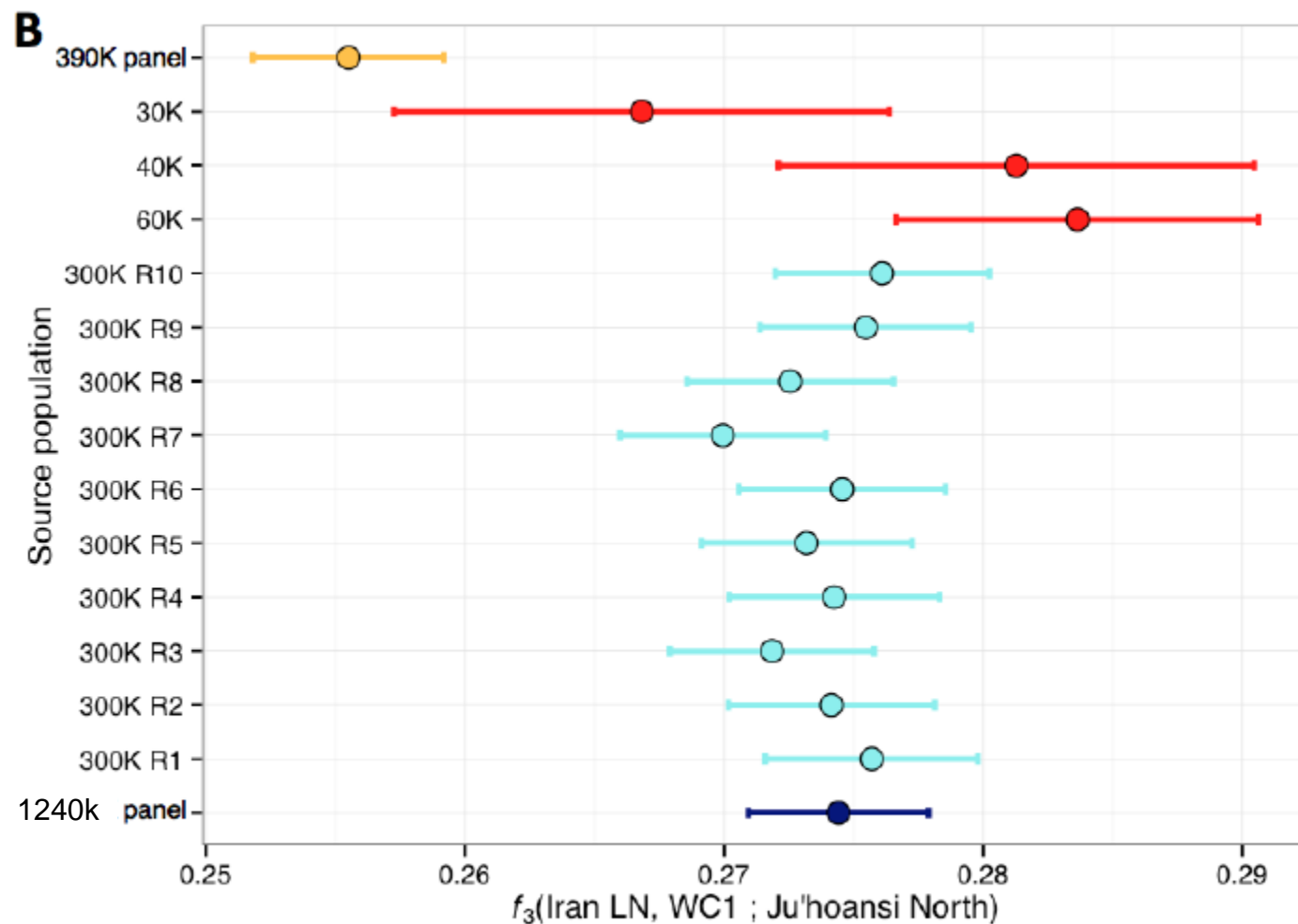
# Ascertainment and $f$ stats



Blue: 1240k SNP panel

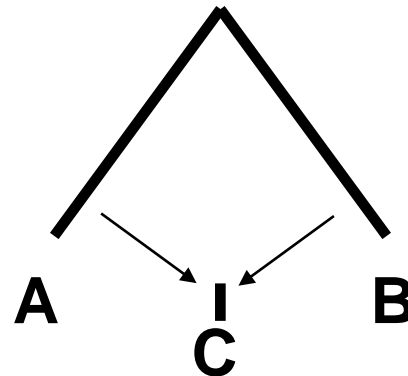
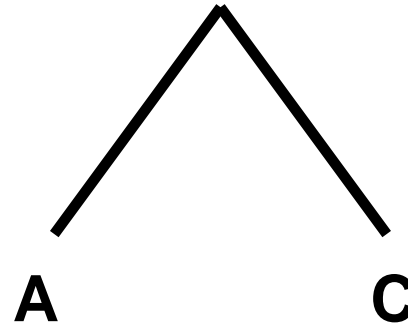
Yellow: 390k SNP panel (subset of 1240k)

# Ascertainment and $f$ stats



# Summary

- Using F statistics (measure of drift) to compare populations
- Relationships among populations on a simple tree
- Admixture



## Practical

- Use *admixturetools* to compute different  $f$  stats and interpret them
- Human dataset with modern and ancient
- Commands and questions in  
2024\_Manica\_practical1\_worksheet.pdf

## References

- Peter, B.M. (2016), Admixture, Population Structure, and F-Statistics, *Genetics* 202(4): 1485–1501 doi:10.1534/genetics.115.183913
- Patterson N, et al. (2012) Ancient admixture in human history. *Genetics* 192(3):1065-93. doi: 10.1534/genetics.112.145037.