Demographic inference using summary statistics

Ben Peter
Max Planck Institute for Evolutionary Anthropology
Leipzig, Germany

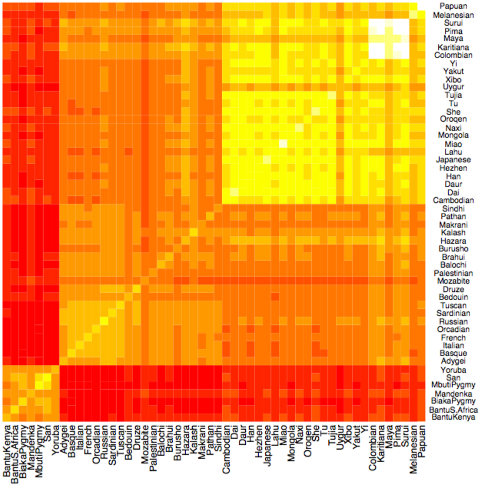
Motivation

Motivation

- Archaeological and linguistic sources of data give us alternative sources of data with which to confirm/contrast genetic inferences regarding population history
- Observational studies only possible - so statistical methods are key for inference

Similarity matrices

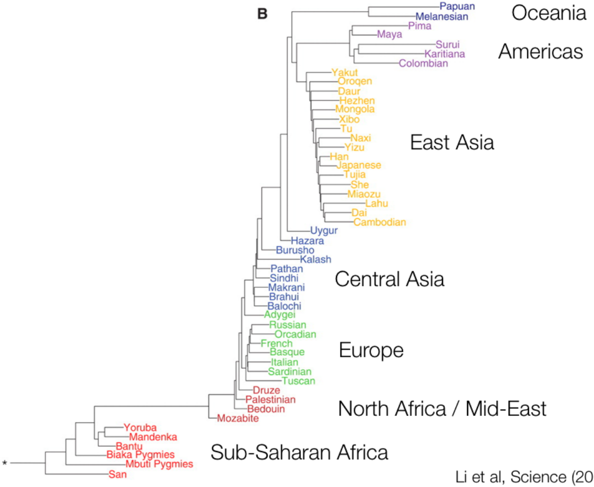
Covariance matrix of allele frequencies across HGDP populations



Coop et al (2010) Genetics

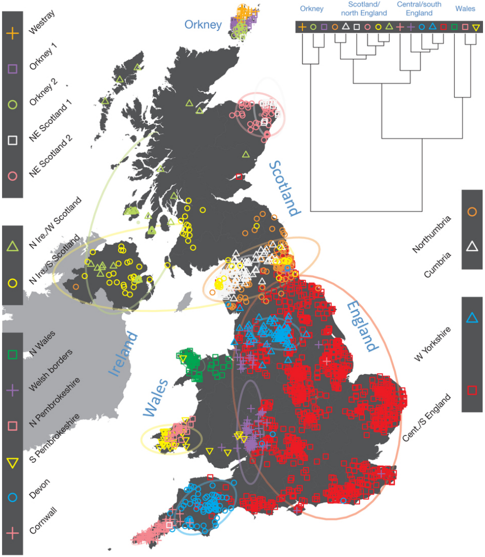
Phylogenetic trees

Neighbor-joining tree built with `PHYLIP` on the basis of similarity in allele frequencies:



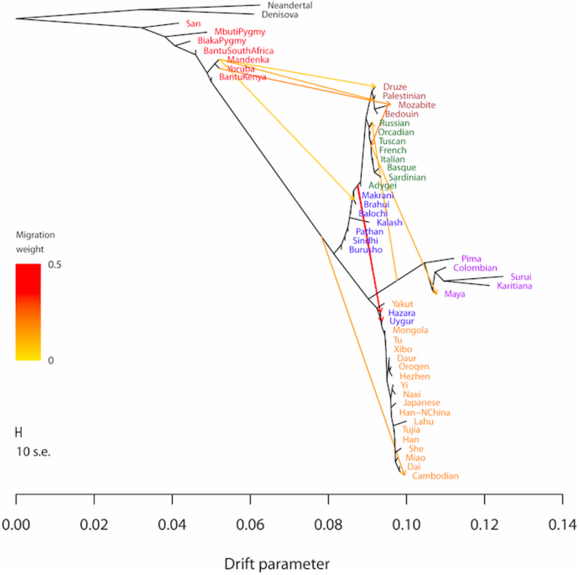
Li et al, Science (2008)

Finestructure algorithm (Leslie et al. 2015)

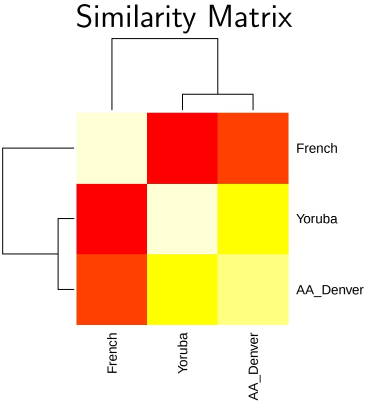


Phylogenetic trees

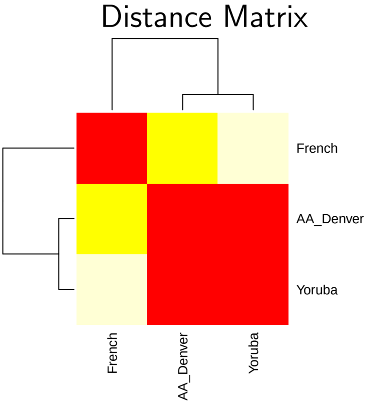
Population tree with admixture events inferred using `TreeMix` software on the basis of allele frequencies:



Measuring Similarity vs measuring distance

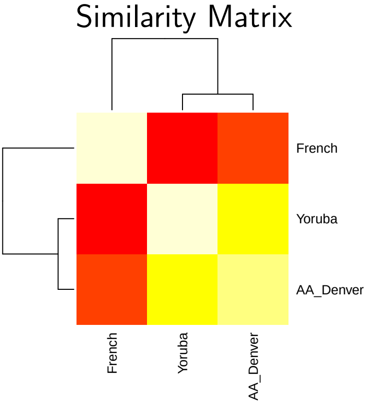


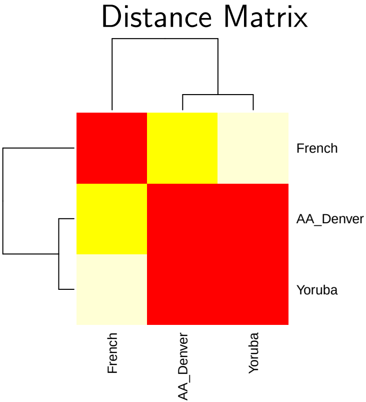
Similarity Matrix



Distance Matrix

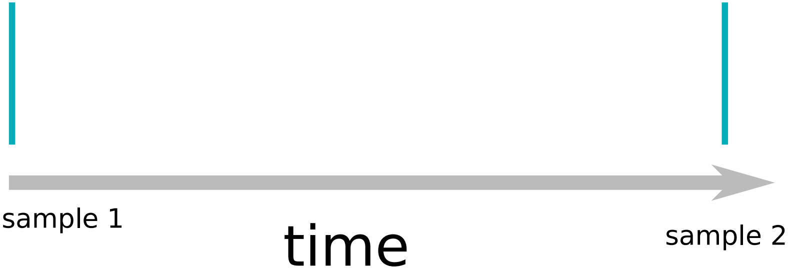
Measuring Similarity vs measuring distance


Similarity Matrix

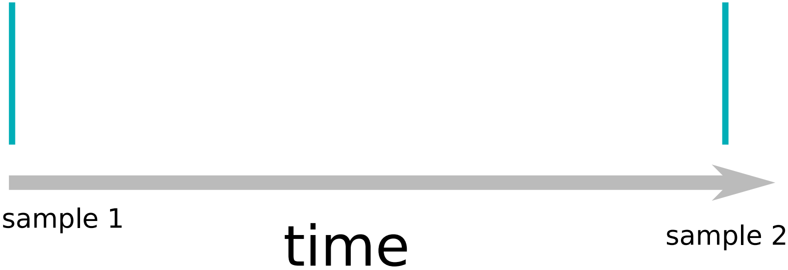

Distance Matrix

Difference: $D_{i,i} = 0$

How could we measure genetic similarity/dissimilarity in a population?



How could we measure genetic similarity/dissimilarity in a population?



sample 1

time

sample 2

- change in allele frequency
- loss of heterozygosity
- probability of coalescence

Introducing today's superhero

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

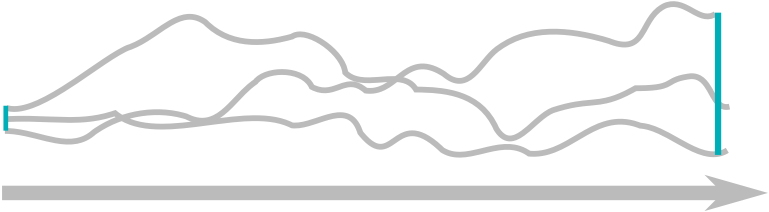
Introducing today's superhero

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

- change in allele frequency
- loss of heterozygosity
- probability of coalescence

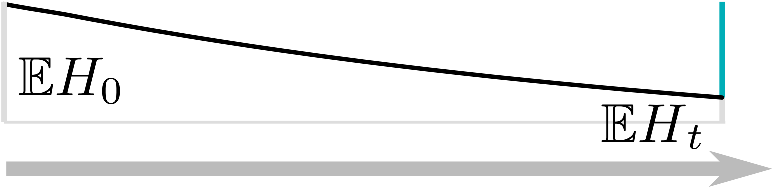
F_2: Changing allele frequencies

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



F_2: Loss of heterozygosity

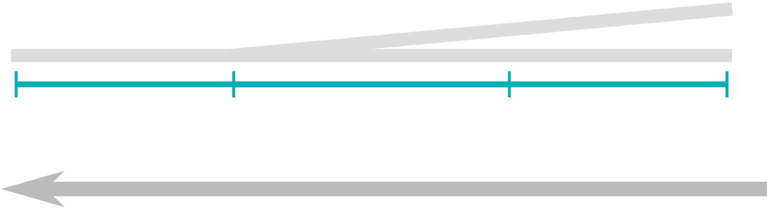
$$F_2 = \frac{\mathbb{E}H_0 - \mathbb{E}H_t}{2}$$



F_2: Probability of coalescence

C

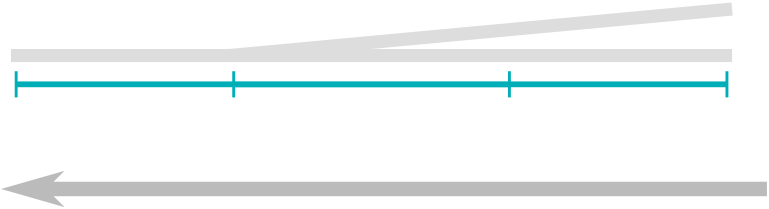
$$F_2 = \frac{1}{2} f \mathbb{E} H_0$$



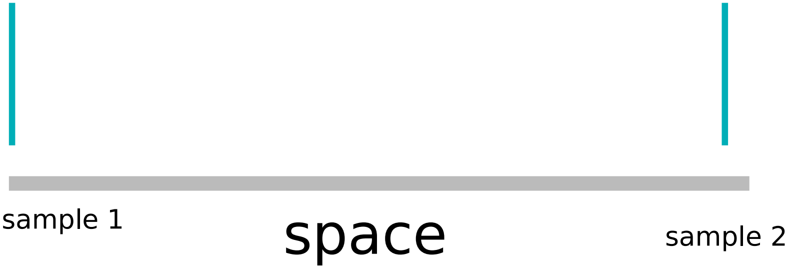
F_2: Probability of coalescence

C

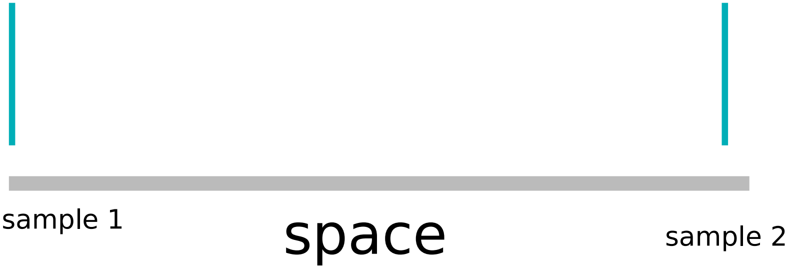
$$F_2 = \frac{1}{2} f \mathbb{E} H_0$$



How could we measure genetic similarity/dissimilarity between populations?



How could we measure genetic similarity/dissimilarity between populations?



- difference in allele frequency
- Heterozygosity: H_{between} vs H_{within}
- Coalescence: T_{between} vs T_{within}

How could we measure genetic similarity/dissimilarity between populations?

sample 1
space
sample 2

- difference in allele frequency
- Heterozygosity: H_{between} vs H_{within}
- Coalescence: T_{between} vs T_{within}

Conveniently, F_2, measures difference equivalently in this scenario

From differences to trees

sample 1

space

sample 2

- difference in allele frequency
- Heterozygosity: H_{between} vs H_{within}
- Coalescence: T_{between} vs T_{within}

Conveniently, F_2, measures difference equivalently in this scenario

F_2: Time to coalescence

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

F_2 vs F_{ST}

$$F_{ST}(P_1, P_2) = \frac{2F_2(P_1, P_2)}{\mathbb{E}H}$$

F_2 vs F_{ST}

$$F_{ST}(P_1, P_2) = \frac{2F_2(P_1, P_2)}{\mathbb{E}H}$$

Main difference is normalization:

- $F_{ST} = 0$: no differentiation
- $F_{ST} = 1$: maximum differentiation

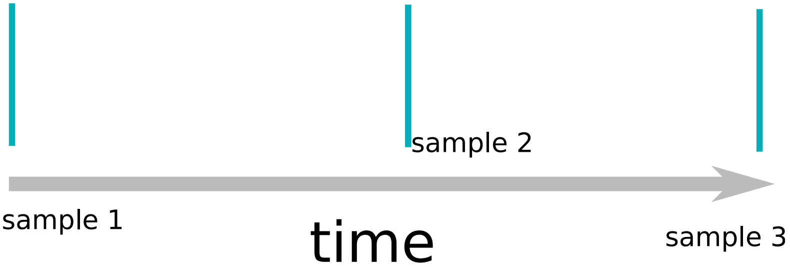
F_2 vs F_{ST}

$$F_{ST}(P_1, P_2) = \frac{2F_2(P_1, P_2)}{\mathbb{E}H}$$

Main difference is normalization:

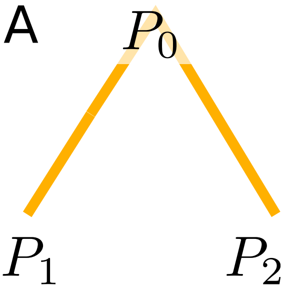
- $F_{ST} = 0$: no differentiation
- $F_{ST} = 1$: maximum differentiation
- $F_2 = 0$: no differentiation
- $F_2 = ???$: maximum differentiation

F_2 is additive



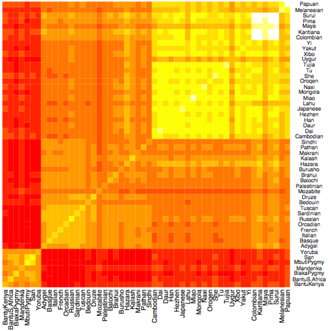
$$F_2(P_1, P_3) = F_2(P_1, P_2) + F_2(P_2, P_3)$$

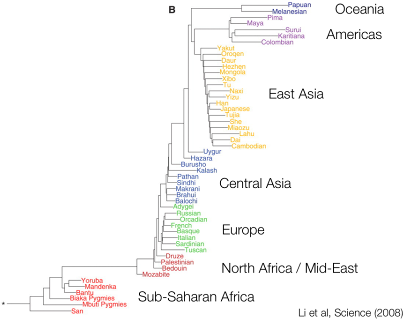
F_2 is tree-additive



$$F_2(P_1, P_2) = F_2(P_0, P_1) + F_2(P_0, P_2)$$

Dissimilarity matrices vs Tree





Li et al, Science (2008)

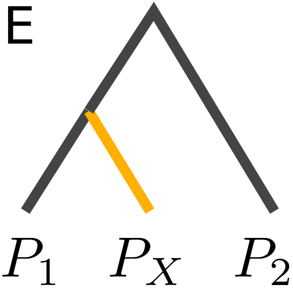
$$F_2(P_1, P_2) = F_2(P_0, P_1) + F_2(P_0, P_2)$$

testing treeness

$$2F_3(P_X; P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)$$

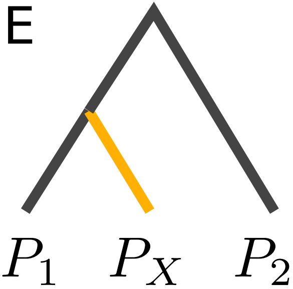
testing treeness

$$2F_3(P_X; P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)$$



testing treeness

$$2F_3(P_X; P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)$$



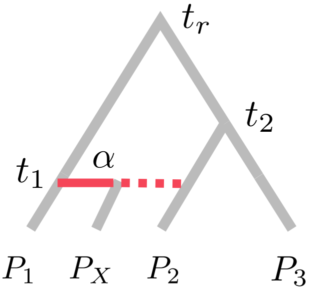
In a tree, $F_3 \geq 0$!

example when this is violated

$$2F_3(P_X; P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)$$

example when this is violated

$$2F_3(P_X; P_1, P_2) = F_2(P_X, P_1) + F_2(P_X, P_2) - F_2(P_1, P_2)$$



$$F_3 = t_1 - 2\alpha(1-\alpha)(1-c_x)(t_r - t_1)$$

alternative interpretation

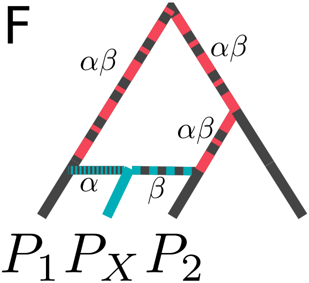
overlap between paths:

$$F_2(P_1, P_2) = \mathbb{E}(p_1 - p_2)(p_1 - p_2)$$
$$F_3(P_X; P_1, P_2) = \mathbb{E}(p_x - p_1)(p_x - p_2)$$

alternative interpretation

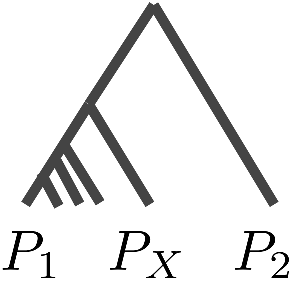
overlap between paths:

$$F_2(P_1, P_2) = \mathbb{E}(p_1 - p_2)(p_1 - p_2)$$
$$F_3(P_X; P_1, P_2) = \mathbb{E}(p_x - p_1)(p_x - p_2)$$



outgroup-F_3

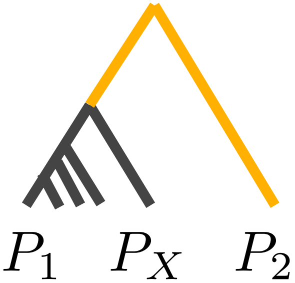
Assume we have an unknown sample, and would like to know which potential population P_1 it is closest to:



What statistic would you calculate?

outgroup-F_3

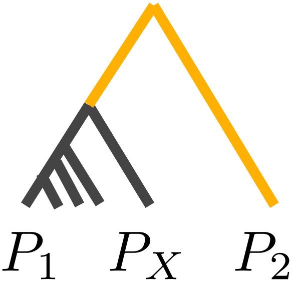
Assume we have an unknown sample, and would like to know which potential population P_1 it is closest to:



$F_3(P_2; P_X, P_1)$ will be larger the closer P_X and P_1 are!

outgroup-F_3

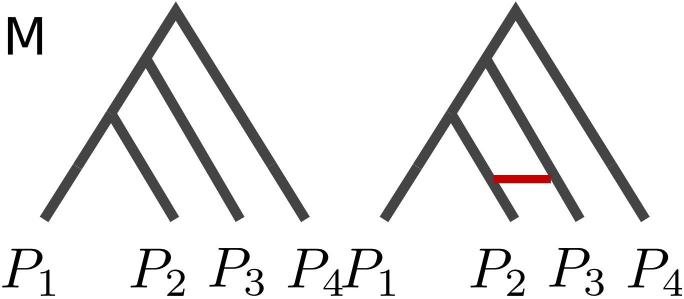
Assume we have an unknown sample, and would like to know which potential population P_1 it is closest to:



$F_3(P_2; P_X, P_1)$ will be larger the closer P_X and P_1 are!
Advantage over direct measures of differentiation if sampling times of P_1 are different.

D-statistic / (*F*$_4$-statistic)

Imagine you sequence a Neandertal for the first time. How do you test for gene flow?

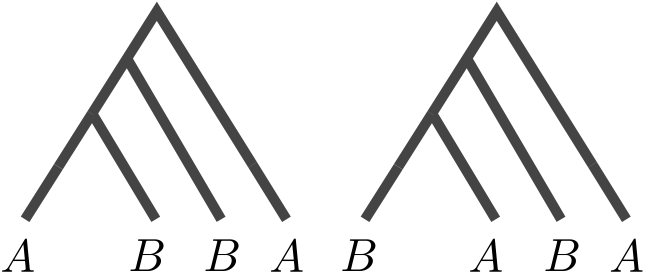


D-statistic

Imagine you sequence a Neandertal for the first time. How do you test for gene flow?

D-statistic

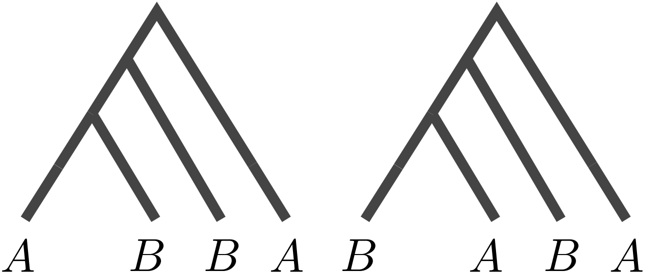
Imagine you sequence a Neandertal for the first time. How do you test for gene flow?



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

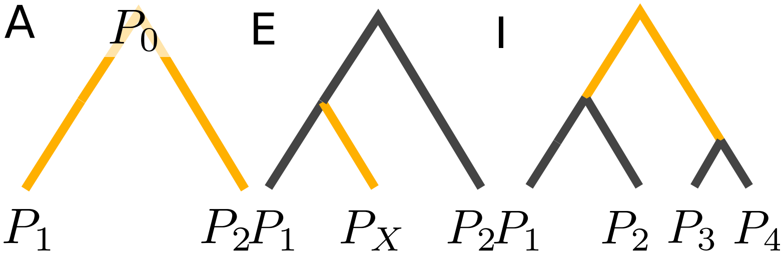
D-statistic

Imagine you sequence a Neandertal for the first time. How do you test for gene flow?



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

What does D/F_4 actually measure?

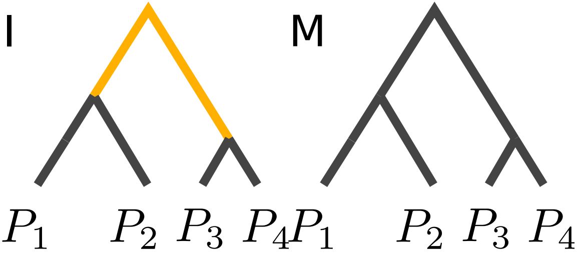


$$D = \frac{\text{ABBA} - \text{BABA}}{\text{ABBA} + \text{BABA}}$$

$$F4 = F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_2) - F_2(P_3, P_4)$$

What does D/F_4 actually measure?

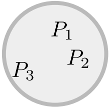
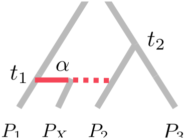
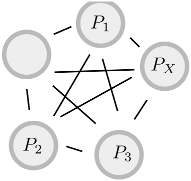
Two possibilities:



$$2F_4 = F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_2) - F_2(P_3, P_4)$$

$$2F_4 = F_2(P_1, P_3) + F_2(P_2, P_4) - F_2(P_1, P_4) - F_2(P_2, P_3)$$

how do these statistics behave under other demographic models?

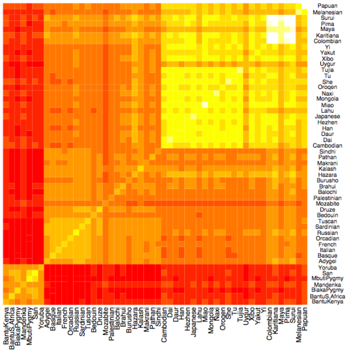
Model		$F_3(P_X; P_1, P_2)$	$F_4(P_1; P_X; P_2, P_3)$
Panmictic		0	0
Admixture Graph	 t_2 t_1 α P_1 P_X P_2 P_3	$t_1 - 2\alpha(1-\alpha)\times$ $(1-c_x)t_r$	$(1-\alpha)(t_2-t_1)$
Island Model		$\frac{1}{M}$	0

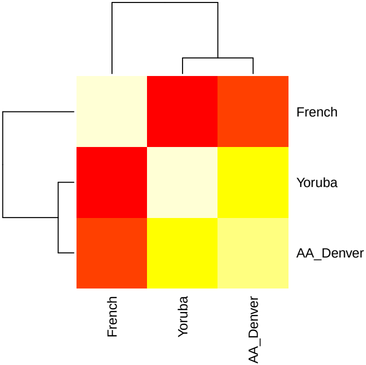
how do these statistics behave under other demographic models?

Stepping stone	$P_1 - P_X - P_2 - P_3$	$\frac{2}{7M}$	$-\frac{8}{7M}$
Hierarchical stepping stone	$P_1\,P_1\,P_X\,P_X\,P_2\,P_2$	$-\frac{\mathbf{0.06}}{\mathbf{M}}$	$\frac{14}{55M}$
Serial founder model	$P_1 \rightarrow P_X \rightarrow P_2 \rightarrow P_3$	t_x	0

Recap

1. F_3 and F_4 are simple statistics that test for admixture
2. F_3 requires just 3 populations, and is most useful for recent admixture at approximately equal proportions
3. F_4 is suitable to more ancient admixture, but more sensitive


Papuan
Melanesian
Surui
Pima
Maya
Karitiana
Colombian
Yi
Yakut
Xibo
Uygur
Tujia
Tu
She
Oroqen
Naxi
Mongola
Miao
Lahu
Japanese
Hezhen
Han
Daur
Dai
Cambodian
Sindhi
Pathan
Makrani
Kalash
Hazara
Burusho
Brahui
Balochi
Palestinian
Mozabite
Druze
Bedouin
Tuscan
Sardinian
Russian
Orcadian
French
Italian
Basque
Adygei
Yoruba
San
MbutiPygmy
Mandenka
BiakaPygmy
BantuS.Africa
BantuKenya
BantuKenya
BantuS.Africa
BiakaPygmy
Mandenka
MbutiPygmy
San
Yoruba
Adygei
Basque
Italian
French
Orcadian
Russian
Sardinian
Tuscan
Bedouin
Druze
Mozabite
Palestinian
Balochi
Brahui
Burusho
Hazara
Kalash
Makrani
Pathan
Sindhi
Cambodian
Dai
Daur
Han
Hezhen
Japanese
Lahu
Miao
Mongola
Naxi
Oroqen
She
Tu
Tujia
Uygur
Xibo
Yakut
Yi
Colombian
Karitiana
Maya
Pima
Surui
Melanesian
Papuan


French
Yoruba
AA_Denver
French
Yoruba
AA_Denver