Lesson 9: qpWave

Wednesday July 25, 2018

9:00 – 11:30 am

July - August 2018

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1	2	3	4	5	6	7
8	9	10	11	12	13	14
Preparation: Read two studies for Monday journal club.	**9:00-11:30**,QF, Introduction to ancient DNA research; **2:00-4:00 pm**: QF+MY: Journal Club	**Preparation:** Follow instructions to install python/access server; Read Haak et al. 2015 paper.	**9:00-11:30**, MY: Introduction to Haak et al 2015/Data sets; **2:00-4:30 pm**, MY, QF: Lab preparation and data processing		**9:00-11:30**, AK: Uniparental markers, mtDNA/chrY	
15	16	17	18	19	20	21
	9:00-11:30, MY: D-statistics (Test of Treeness) **2:00-4:00 pm:** HW: Journal Club		**9:00-11:30**, HW: PCA		**9:00-11:30**, MY: Outgroup f3-statistic/Introduce Mini-Project	
22	23	24	25	26	27	28
	9:00-11:30, HS: D-statistics (Test of Admixture); **2:00-4:00 pm:** AK: Journal Club		**9:00-11:30**, MY: qpWave		**9:00-11:30**, MY: qpAdm	
29	30	31	1	2	3	4
	9:00-11:30, HW: ADMIXTURE/PLINK; **2:00-4:00 pm**: HS: Journal Club		**9:00-11:30**, MY: ADMIXTUREGRAPH/ Modeling		**9:00-11:30**, MY: Test!	
5	6	**Notes:**				
	9:00-11:30: HS/AK/HS/MY/QF: Final Presentation	Lecturer/Discussion Leaders are listed by initials, as follows: QF=Qiaomei Fu; MY=Melinda Yang; AK=Albert Ko; HW=Hongru Wang; HS=Hassan Shafiey				

Test: Friday August 3

- Study lecture and exercises
- No computer scripting
- Focus on how to interpret results

So far, what can each of these tell us?

- PCA
- f-statistics
  - F3
    - Outgroup
    - Admixture
  - F4/D
    - Treeness
    - Admixture

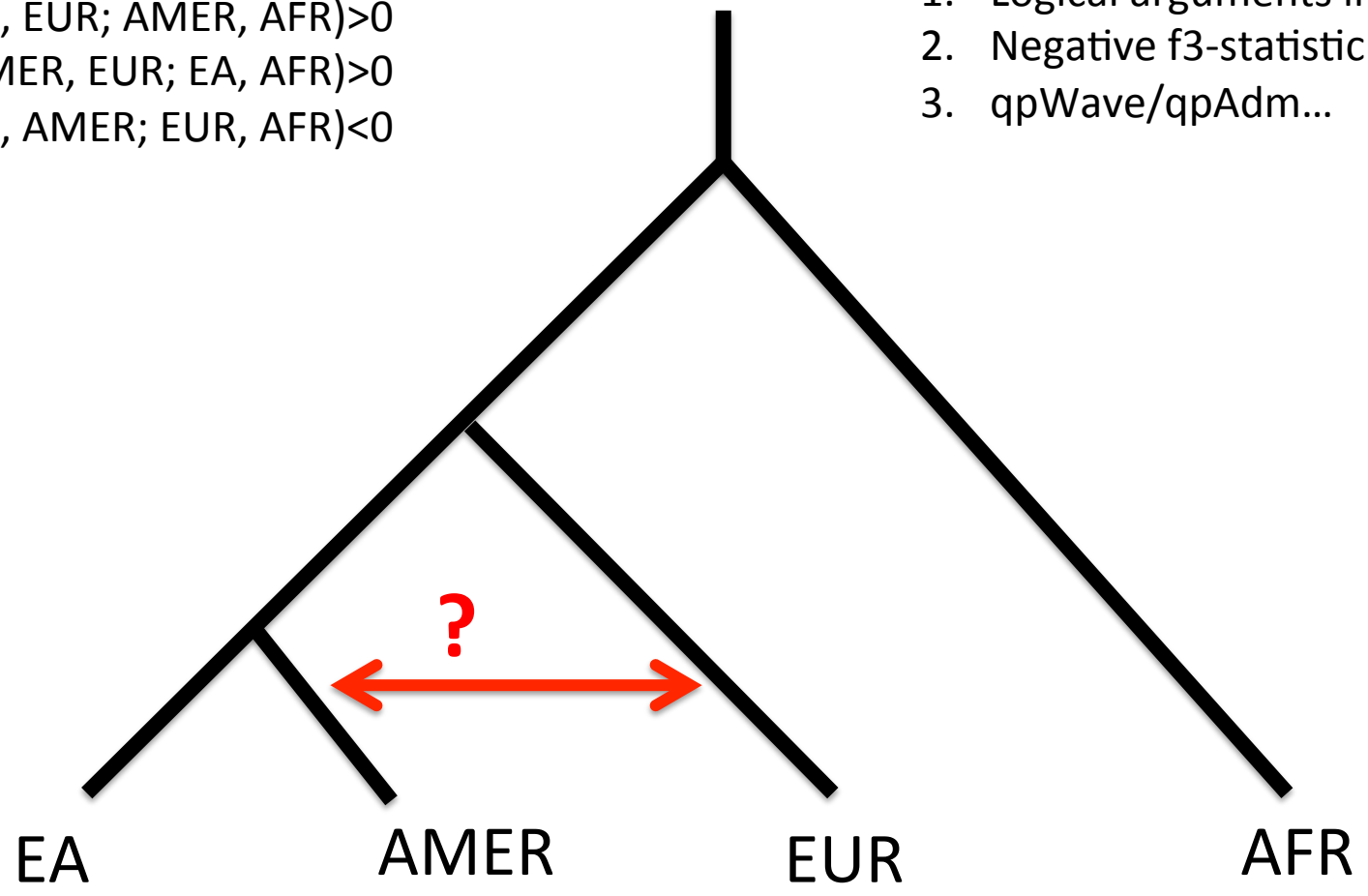
Admixture

- D-statistics show a connection, but not easy to argue direction of gene flow.
- F3-statistics can show an admixed population, but does not always show when true, and it doesn't show how much.
  - Can only consider two source populations

How do we know the direction of gene flow?

D(EA, EUR; AMER, AFR)>0
D(AMER, EUR; EA, AFR)>0
D(EA, AMER; EUR, AFR)<0

1. Logical arguments in D
2. Negative f3-statistic
3. qpWave/qpAdm…



qpWave and qpAdm

- Expansion on f-statistics, allowing study of direction of gene flow from N potential sources, even estimating the amount of admixture.
  - qpWave: number of sources
  - qpAdm: estimate amount of admixture

qpWave and qpAdm

- Expansion on f4-statistics, allowing study of direction of gene flow from N potential sources, even estimating the amount of admixture.
  - **qpWave: number of sources (Today)**
  - qpAdm: estimate amount of admixture (Friday)

qpWave

- Use Reich et al. (2012), where first introduced, to explore how qpWave works (Note S6)
- https://reich.hms.harvard.edu/publications

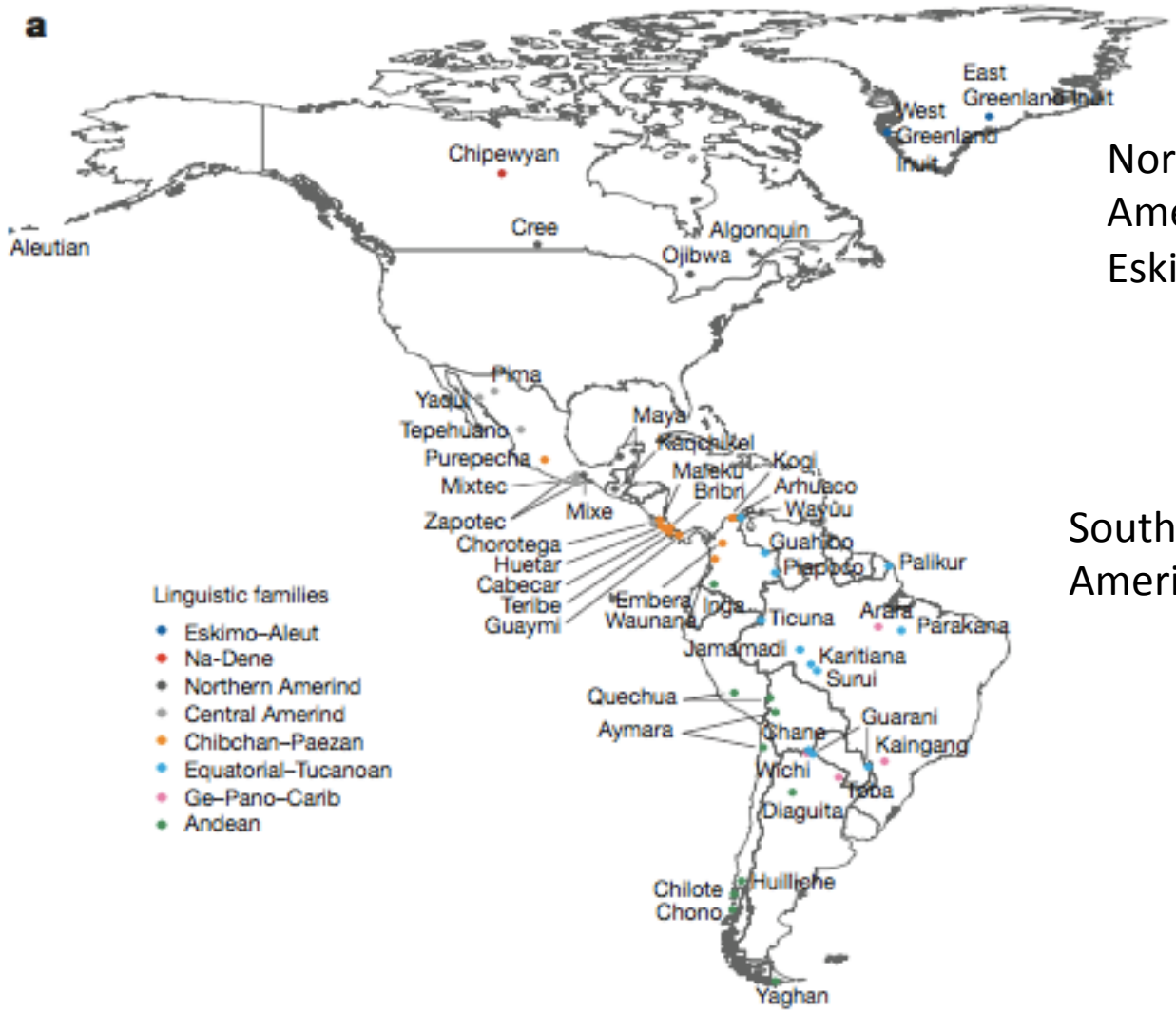
LETTER

doi:10.1038/nature11258

Reconstructing Native American population history

David Reich[1,2], Nick Patterson[2], Desmond Campbell[3,4], Arti Tandon[1,2], Stéphane Mazieres[3,5], Nicolas Ray[6], Maria V. Parra[3,7], Winston Rojas[3,7], Constanza Duque[3,7], Natalia Mesa[3,7], Luis F. García[7], Omar Triana[7], Silvia Blair[7], Amanda Maestre[7], Juan C. Dib[8], Claudio M. Bravi[3,9], Graciela Bailliet[9], Daniel Corach[10], Tábita Hünemeier[3,11], Maria Cátira Bortolini[11], Francisco M. Salzano[11], María Luiza Petzl-Erler[12], Victor Acuña-Alonzo[13], Carlos Aguilar-Salinas[14], Samuel Canizales-Quinteros[15,16], Teresa Tusié-Luna[15], Laura Riba[15], Maricela Rodríguez-Cruz[17], Mardia Lopez-Alarcón[17], Ramón Coral-Vazquez[18], Thelma Canto-Cetina[19], Irma Silva-Zolezzi[20]†, Juan Carlos Fernandez-Lopez[20], Alejandra V. Contreras[20], Gerardo Jimenez-Sanchez[20]†, Maria José Gómez-Vázquez[21], Julio Molina[22], Ángel Carracedo[23], Antonio Salas[23], Carla Gallo[24], Giovanni Poletti[24], David B. Witonsky[25], Gorka Alkorta-Aranburu[25], Rem I. Sukernik[26], Ludmila Osipova[27], Sardana A. Fedorova[28], René Vasquez[29], Mercedes Villena[29], Claudia Moreau[30], Ramiro Barrantes[31], David Pauls[32], Laurent Excoffier[33,34], Gabriel Bedoya[7], Francisco Rothhammer[35], Jean-Michel Dugoujon[36], Georges Larrouy[36], William Klitz[37], Damian Labuda[30], Judith Kidd[38], Kenneth Kidd[38], Anna Di Rienzo[25], Nelson B. Freimer[39], Alkes L. Price[2,40] & Andrés Ruiz-Linares[3]

What do we know about Native Americans?



Northern Native Americans, Eskimo-Aleut

Southern Native Americans

What do we know about Native Americans?

How many streams of ancestries make up Native Americans?

Northern Native Americans, Eskimo-Aleut

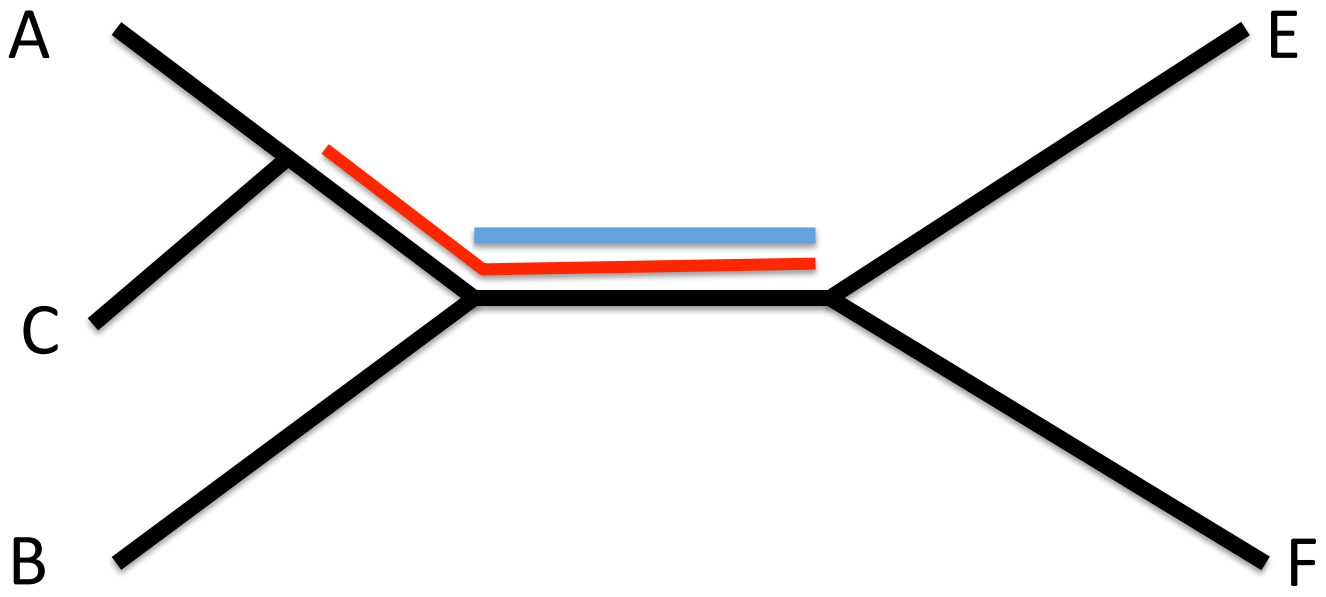
Southern Native Americans

Outgroups and Related

Consider populations A, B, and C as outgroup to populations E and F, such that Out={A, B, C} and Main={E, F}. Then, if we focus on E and A,
We would expect f4(E, F; A, B)=f4(E, F; A, C)=0.

Note:

1. We are no longer considering treating one position as an outgroup.
2. Because E and F both show no special relationship to A or B, still expect symmetry.

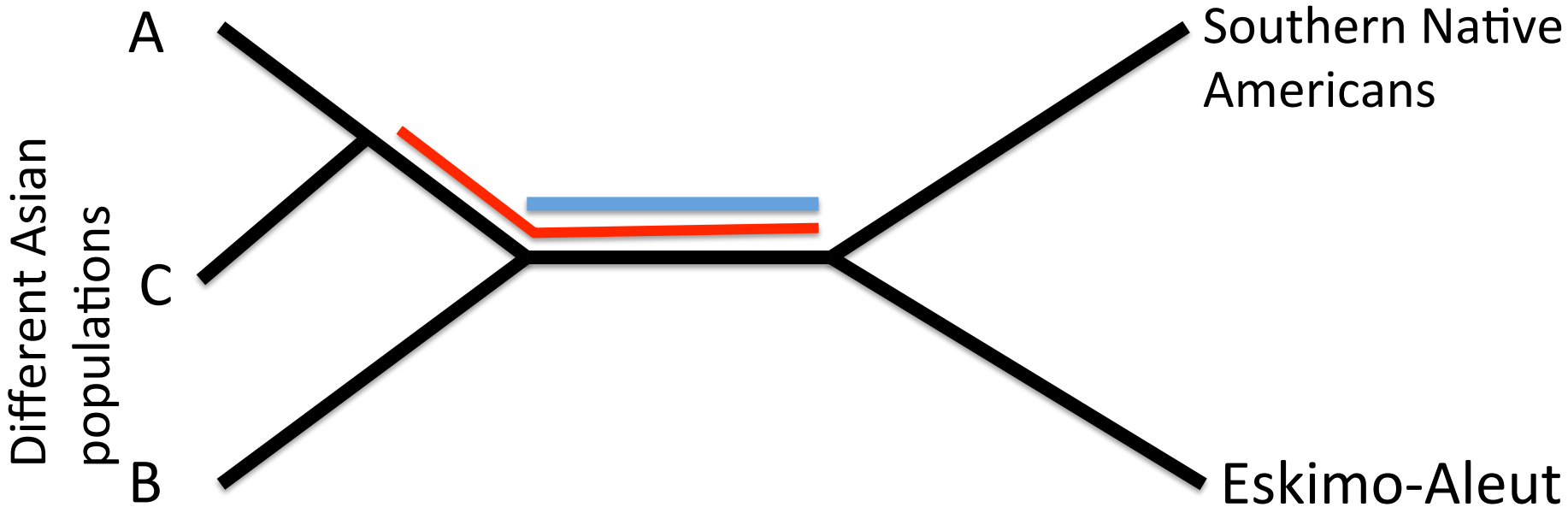


Outgroups and Related

Consider populations A, B, and C as outgroup to populations E and F, such that Out={A, B, C} and Main={E, F}. Then, if we focus on E and A,
We would expect f4(E, F; A, B)=f4(E, F; A, C)=0.

Note:

1. We are no longer considering treating one position as an outgroup.
2. Because E and F both show no special relationship to A or B, still expect symmetry.



F4(AMER1, AMER2; Out1, Out2)

- Consider each Native American in AMER, and each outgroup in OUT
  - F4(Southern, Northern; Asian, Asian) – deviations from zero indicate different streams of ancestry.
  - Contrast different northerns to each other:
    - F4(Southern, Northern1; Asian, Asian)
    - F4(Southern, Northern2; Asian, Asian)

See three different patterns – two depicted below.

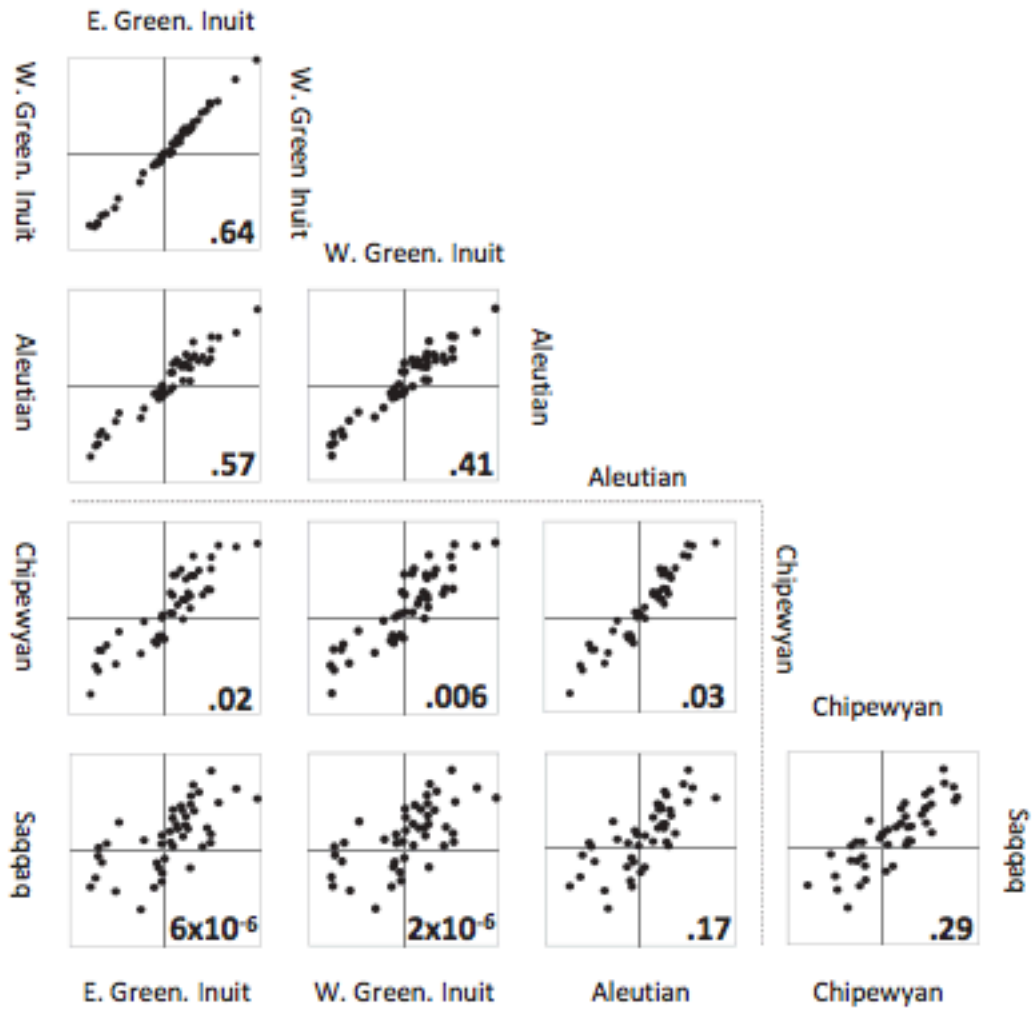


Figure S6.1: Qualitative evidence of 3 different patterns of relatedness to Asians among Native Americans. We plot f_4 statistics for all possible pairs of northern Native American populations with the strongest evidence of a distinct relationship to Asians compared with more southern Native Americans. Two groupings of populations harbor significantly different historical relationships with a panel of 10 Asian outgroups. Within groups, f_4 statistics are highly correlated, whereas across groups they are significantly different: the P-value in each panel is from Table 1: the Hotelling T-test for whether the vectors of f_4 statistics are consistent with being scalar multiples of each other. The dashed line is added to highlight the difference in within-group and across-group comparison.

Statistically…

- "If two Native American populations derive all their non-First American ancestry from the same ancestral stream of Asian gene flow, **their vectors of f4-statistics are expected to be scalar multiples of each other**, and [Reich et al. 2012] developed a formal statistical test for whether this is the case." (p. 21, Note S6)

X(Q, S) = F4(Karitiana, Q; Han, S)

S

Q

$$\begin{bmatrix} X(Q_1, S_1) & X(Q_1, S_2) & \cdots & X(Q_1, S_n) \\ X(Q_2, S_1) & \cdots & & \vdots \\ \vdots & & & \\ & & & \\ & & & \\ & & & \\ X(Q_m, S_1) & \cdots & & X(Q_m, S_n) \end{bmatrix}$$

X(Q, S)

- S6.4 (p. 21): Let r be the rank of X, and n the number of independent gene flows into the Americas. Then,

$$r+1 \leq n$$

- Thus, the number of streams of ancestry at minimum must be greater than one plus the rank of the matrix.
- A likelihood ratio test can be used to compare rank r=k to rank r=k+1

Table 1 in Reich et al. (2012)

Table 1 | Native Americans descend from at least three streams of Asian gene flow

Population groupings tested	*P* value for this many Asian streams being enough to explain the data			Minimum number of streams of Asian gene flow needed to explain the data
	1	2	3	
East Greenland Inuit/West Greenland Inuit/First American	$<10^{-9}$	0.64	1	2
East Greenland Inuit/Aleutian/First American	$<10^{-9}$	0.57	1	2
West Greenland Inuit/Aleutian/First American	$<10^{-9}$	0.41	1	2
Chipewyan/East Greenland Inuit/First American	$<10^{-9}$	0.02	1	3
Chipewyan/West Greenland Inuit/First American	$<10^{-9}$	0.006	1	3
Chipewyan/Aleutian/First American	$<10^{-9}$	0.03	1	3
Saqqaq/East Greenland Inuit/First American	$<10^{-9}$	6×10^{-6}	1	3
Saqqaq/West Greenland Inuit/First American	$<10^{-9}$	2×10^{-6}	1	3
Saqqaq/Aleutian/First American	$<10^{-9}$	0.17	1	2
Saqqaq/Chipewyan/First American	$<10^{-9}$	0.29	1	2
Saqqaq/Eskimo–Aleut/Chipewyan/First American	$<10^{-9}$	8×10^{-6}	0.27	3

We use the method described in Supplementary Notes to test formally whether specified groupings of Native American populations are consistent with descending from one, two or three streams of gene flow from Asia. We use 'First American' to refer to a pool of 43 populations from Meso-America southward, and 'Eskimo–Aleut' to refer to a pool of East and West Greenland Inuit and Aleuts. We test either three or four population groupings (when there are three groupings, the maximum number of streams we can reject is two, and so the *P* value for three streams is always 1). At least two streams of Asian gene flow are required to explain all rows ($P < 10^{-9}$). The Chipewyan, Eskimo–Aleut and First Americans can only be jointly explained by at least three streams. Analysis of the Saqqaq Palaeo-Eskimo (using about sixfold fewer SNPs than for the other analyses) show that the Asian ancestry in this individual has a component that is different from that in First Americans and Greenland Inuit, but indistinguishable from the Chipewyan.

Summary

- Given a set of populations and a set of ‘outgroup’ populations, qpWave asks the minimum streams of ancestry required to explain the included populations.
- If you expect three populations to have three distinct ancestries, but you find less than that, then perhaps finding admixture (stay tuned for qpAdm).

How to calculate qpWave?

- Requires:
  - PAR file
  - Leftpop file (Main populations of interest)
  - Rightpop file (Outgroup populations)
  - Geno/snp/ind

Populations

- Leftpop
  - Loschbour, LBK_EN, Yamnaya
- Rightpop
  - Han, Eskimo, Mbuti, Karitiana, Kharia, Onge, Ulchi

Results of qpWave

- How many waves of ancestry are needed for these three European populations?
  - Which column do we look at below?

```
[mel_yang@comput14 lesson9]$ less example.log | grep f4rank
f4rank: 0 dof:     12 chisq:   454.955 tail:       8.37514316e-90 dofdiff:      0 chisqdiff:     0.000 taildiff:                   1
f4rank: 1 dof:      5 chisq:    29.151 tail:       2.16611578e-05 dofdiff:      7 chisqdiff:   425.804 taildiff:       6.94783958e-88
f4rank: 2 dof:      0 chisq:     0.000 tail:                    1 dofdiff:      5 chisqdiff:    29.151 taildiff:       2.16611578e-05
```

Results of qpWave

- How many waves of ancestry are needed for these three European populations?
  - Which column do we look at below?

```
[mel_yang@comput14 lesson9]$ less example.log | grep f4rank
f4rank: 0 dof:     12 chisq:   454.955 tail:       8.37514316e-90 dofdiff:      0 chisqdiff:     0.000 taildiff:                  1
f4rank: 1 dof:      5 chisq:    29.151 tail:       2.16611578e-05 dofdiff:      7 chisqdiff:   425.804 taildiff:       6.94783958e-88
f4rank: 2 dof:      0 chisq:     0.000 tail:                    1 dofdiff:      5 chisqdiff:    29.151 taildiff:       2.16611578e-05
```

r=2, so n≥3 – therefore, at least three streams of ancestry to describe Loschbour, LBK_EN, and Yamnaya

Populations

- Leftpop
  - Spain_EN, Corded_Ware_LN, Yamnaya
- Rightpop (SI 9 of Haak et al. 2015, p. 94)

In this section, we develop a method that can be used (i) to identify reference populations that may have contributed ancestry to a *Test* population, and (ii) to estimate mixture proportions from these reference populations for *Test*. The method uses the intuition that the reference populations are not identically related to a panel of "focal" or "outgroup" populations, but share different amounts of genetic drift with them as a result of their deep evolutionary history (which is, however, not explicitly modeled). These "outgroups" must be devoid of recent gene flow with either the *Test* or the candidate reference population, as such gene flow introduces additional common genetic drift. One way to identify them is to pick a varied set of world populations that are (i) geographically remote from the area under study, and (ii) do not show evidence of admixture from that area using an algorithm such as ADMIXTURE[1] that can identify recently admixed populations and individuals. In practice, we will use the following set *O* of 15 previously identified[2] outgroups for West Eurasia:

"World Foci 15" set of outgroups *O*: Ami, Biaka, Bougainville, Chukchi, Eskimo, Han, Ju_hoan_North, Karitiana, ~~Kharia,~~ Mbuti, ~~Onge,~~ Papuan, She, Ulchi, Yoruba

Why do I kick some out?

Activity

- Make the leftpop and rightpop files, as well as the par file.
- Run qpWave
- What is the minimum streams of ancestry that you find?
  - What are the p-values for each rank?
  - Does this make sense to you?