CAAP Biology 2019 Computational Lab Handout

Author: Chris Porras

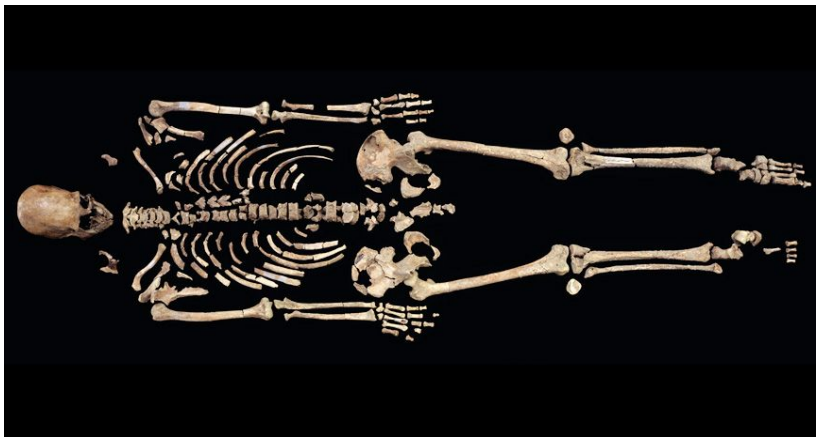
What can genetic data teach us?

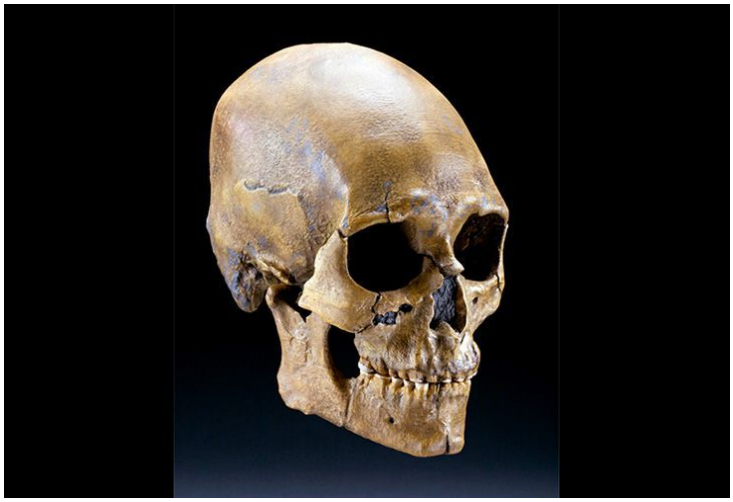
The Story of Kennewick Man:

BACKGROUND:
In 1996, human skeletal remains were discovered near the Columbia River in the city of Kennewick, Washington. Dubbed “The Kennewick Man,” this sample was found to be almost 9,000 years old and its discovery sparked a heated debate about its ancestry. The Umatilla people and other Columbia Basin tribes claimed that the Kennewick Man was their ancestor and demanded he be returned to them for burial. However, questionable conclusions made by anthropologists inferred a possible link to Pacific peoples and imprecise terminology convinced some to conclude that the sample was more closely related to Caucasians. Following decades of legal battles, geneticists stepped in and their analyses of ancient DNA from the Kennewick Man confirmed his ties to the modern Columbia Basin tribes. In 2017, these tribes were finally able to bury the remains of the Kennewick Man at an undisclosed location.

INTRODUCTION:
This lab will walk through the various methods used to infer the Kennewick Man’s relatedness to modern populations. We will examine real data collected from the skeleton beginning with morphometric and leading into genetic. Employing computational tools, we will recreate the methods used by researchers who studied this problem. Our ultimate goal is to examine the flaws and promises of each approach and gain a first-hand account of the messiness in ancestral inference and the importance of ethical science.

Kennewick man skeleton (Smithsonian mag)





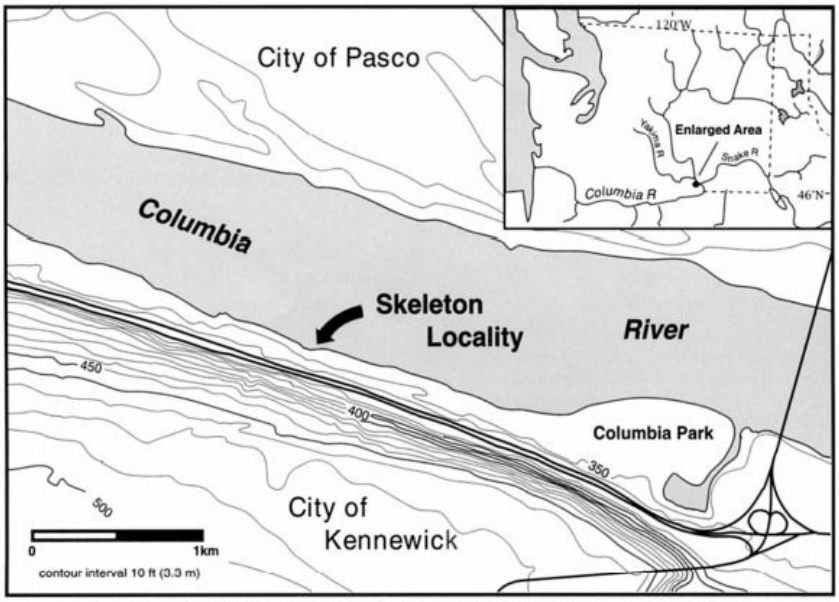


Figure 1. Topographic map showing the locality where the Kennewick skeleton was found.

Table 7. Comparison of the Kennewick Skeleton with Other Paleoamerican Males based on Metric Indices, Stature, and Age.

Measure	Browns Valley[d]	Horn Shelter	Hourglass Cave	Spirit Cave[d]	Wizards Beach[d]	Gore Creek	Mean	Kennewick
Cranial index	71.0[e]	73.8	—	70.2	72.8	—	71.9 ± 1.6	73.5
Orbital index	83.0	78.0	—	78.6	85.3	—	81.2 ± 3.5	80.4
Upper face index	46.4	51.2	—	47.8	54.7	—	50.2 ± 3.8	55.6
Nasal index	48.0	55.0	—	52.1	47.3	—	50.5 ± 3.5	47.3
Palatal index	81.2	90.6	—	86.2	83.9	—	85.5 ± 4.0	84.8
Height index[b]	91.4	93.8	—	91.8	91.3	—	92.3 ± 1.1	90.5
Facial forwardness[c]	2.88	2.79	—	2.73	2.82	—	2.80 ± 0.06	3.06[a]
Crural index	—	85.1	—	81.6	84.7	83.6	83.8 ± 1.5	85.1
Platymeric index	75.0	68.8	—	81.2	77.1	66.0	73.6 ± 6.2	82.5
Platycnemic index	—	52.6	—	65.7	59.0	57.5	58.7 ± 5.4	61.0
Humeral robusticity	—	18.9	—	17.4	18.9	—	18.4 ± 0.9	18.5[o]
Femoral robusticity	12.1	12.5	—	12.4	13.3	13.3	12.7 ± 0.5	13.5
Stature (all ± 4) cm	165.1[f]	165.4	161.6[i]	164.2	171.8	166.5[m]	165.8 ± 3.4	173.1[a]
Age	25-35[g]	35-44[h]	35-45[j]	40-44[k]	32-42[l]	27-35[n]	—	35-45

[a]Significantly different values.
[b]Calculated as bregma radius/size factor (geometric mean of cranial length, max. breadth, bregma radius, nasion radius, and prosthion radius).
[c]Sum of nasion radius, subspinale radius, prosthion radius, zygomaxillary radius/size factor.
[d]Postcranial values by David R. Hunt, used courtesy of Richard L Jantz; Browns Valley skull measurements from Owsley and Jantz 1999:89.
[e]Values by David Hunt, used courtesy of Richard L. Jantz., based on original skull; no confidence in cast reconstruction of length and breadth.
[f]Based on partial femur length after Steele and Bramblett 1988, using the formula for segments 1 and 2.
[g]Barbara O'Connell, personal communication 1999.
[h]Young 1988.
[i]Value reported by Mosch and Watson 1996 is 162.5 cm, although they do not specify the formula used, I presume it to be the Mongoloid equation of Trotter and Gleser 1958, back computed for femur length, and recalculated stature using Mesoamerican formulae as described in the text.
[j]Mosch and Watson 1996.
[k]Jantz and Owsley 1997.
[l]Edgar 1997.
[m]Value reported by Cybulski et al. 1981 is 168 cm, based on Mongoloid formula of Trotter and Gleser (1958).
[n]Cybulski et al. 1981.
[o]Robusticity indices of right humeri for Kennewick and Horn Shelter are both 20.6.

1. The figure above was displayed in a 2000 paper written by anthropologist James Chatters, the first academic to examine the Kennewick Man skeleton. Chatters compares morphometric data collected from the Kennewick skeleton to measurements from other ancient American male skeletons.

   a) Which Kennewick man measurements differ significantly from other paleoamerican samples? Facial forwardness and stature

   b) From this table, what can you infer about the Kennewick man's relationship with paleoamerican populations? Mostly similar, some significant differences

   c) Is this analysis sufficient for making claims about ancestry? Mostly no

PCA overview:

Principal Component Analysis (PCA) is a statistical method used to reduce the size of a large data set by displaying only axes that capture the greatest variance. In a PC plot, points that are closer together are considered to be more similar to one another. Each

PC explains a fraction of the total variance in the data. It's important to report exactly how much variance is explained by each PC used, because a comparably low fraction of variance explained implies a weak capacity to draw inferences of relatedness.

PCA with morphometric data:

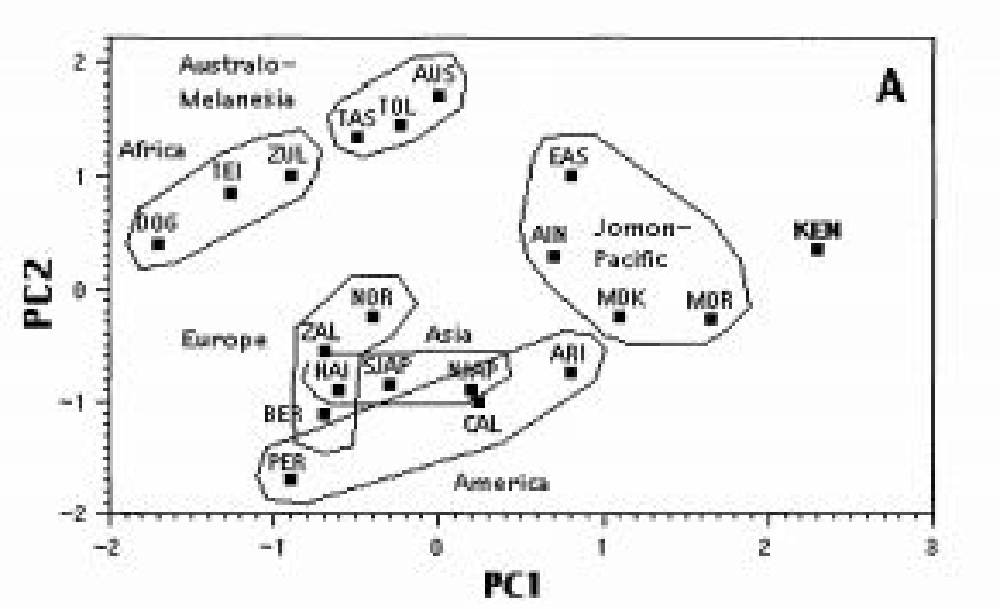


Figure 11. A plot of the first two principal components in a comparison with 19 populations in the Howells database. This comparison considers both size and shape, and the Kennewick skull (KEN) clusters outside modern populations, but near the Jomon-Pacific Cluster (EAS-Easter Island, MOR-Moriori, MOK-Mokapu, and AIN-Ainu.) Other populations are as follows: Asian (NJAP-North Japan, SJAP-South Japan, HAI-Hainan Chinese.) European (NOR-Old Norse, BER-Berg, ZAL-Zalavar), African (ZUL-Zulu, DOG-Dogon, TEI-Tieta), Australo-melanesian (AUS-Australian, TAS-Tasmanian, TOL-Tolai), and American (ARI-Arikara, PER-Peru, CAL-Santa Cruz). From Chatters et al. 1999: Fig. 1a.

2. This PC plot is also from the early Kennewick Man analysis done by James Chatters. It was made comparing morphometric data from a set of 19 ancient human populations.

a) The Kennewick man skeleton was first described as "Caucosoid," an ambiguous anthropological term that some interpreted to mean "Caucasian," or related to modern

Europeans. Does this figure support the hypothesis that the Kennewick man is closely related to modern Europeans? Why or why not? No, KEN is far away from Europeans in plot

b) Which populations are shown to be most closely related to the Kennewick man? Jomon-Pacific cluster and Arikara

c) The first two principal components should explain the largest fraction of the total variance in the data. How much of the variance is explained by PC1 and PC2? Trick question. The variance explained isn't even reported in the original publication…

d) What does your answer for part c) suggest about the interpretability of this data? Not very interpretable…

e) Combining conclusions from questions 1 and 2, what can you infer about the ancestral identity of the Kennewick man? KEN is both similar to Native Americans and not similar to them. Possible Jomon-Pacific origin.

PCA with genetic data: Switching gears to elaborate on our focal method

Come up with scenario to motivate simulation-- OPTIONS

1) Identifying the ancestral country of an international criminal
2) Forensics identification of human remains
3) Optimized marketing based on country from genetic inference

Simulation reflection:

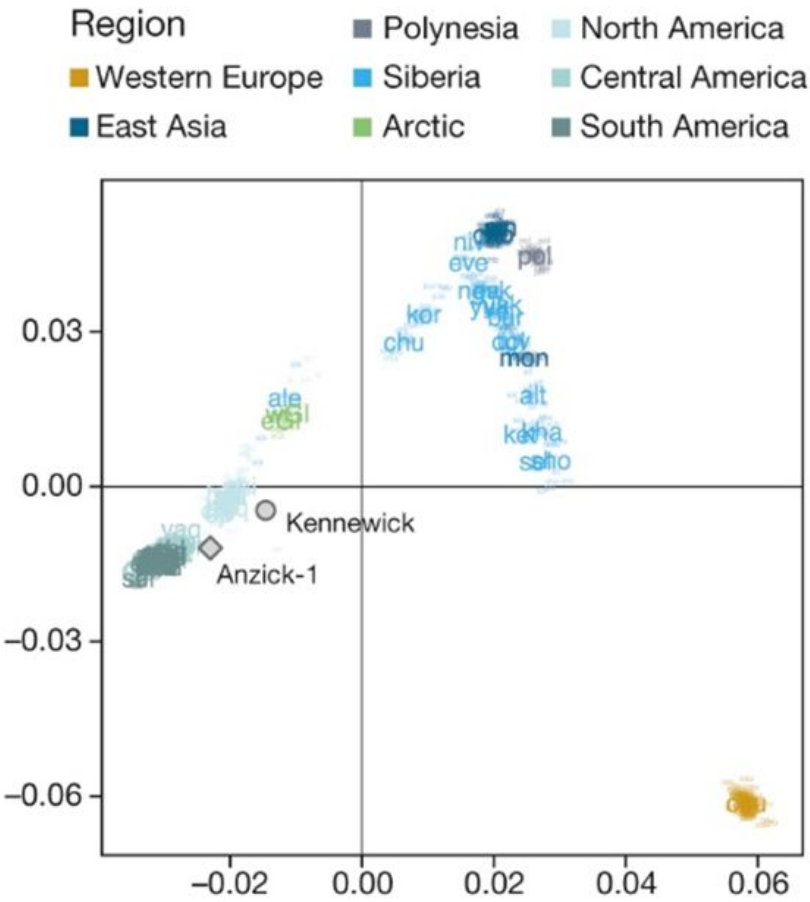
3. A. What is the identity of the unknown? European from Great Britain

B. What other genetic information might you need to narrow down the search region? Regional and familial data

C. What might be some problems with our methodology? Analysis of PCA plots is mostly qualitative, need additional stats, unclear separation of admixed pops, other ideas...

D. Do you think the utility of this type of analysis outweighs privacy concerns? Why or why not? Student opinion

E. In what other scenarios can you imagine this type of analysis being performed? Student opinion



4. The Kennewick Man's genome was sequenced in 2015, equipping geneticists to tackle the nearly 20-year long controversy. PCA was performed to compare the Kennewick man to a known ancient Native American sample (Anzick-1) and modern populations.

a. What can we interpret as our "control" in this analysis? Anzick-1 should cluster near Americans
b. Which population is Kennewick Man most related to? North American
c. How does this compare with the Chatters study? Chatters claimed East Asian and non-Native American ancestry
d. Is there strong evidence for a European ancestry of Kennewick Man? No, Europe clusters far away from KEN

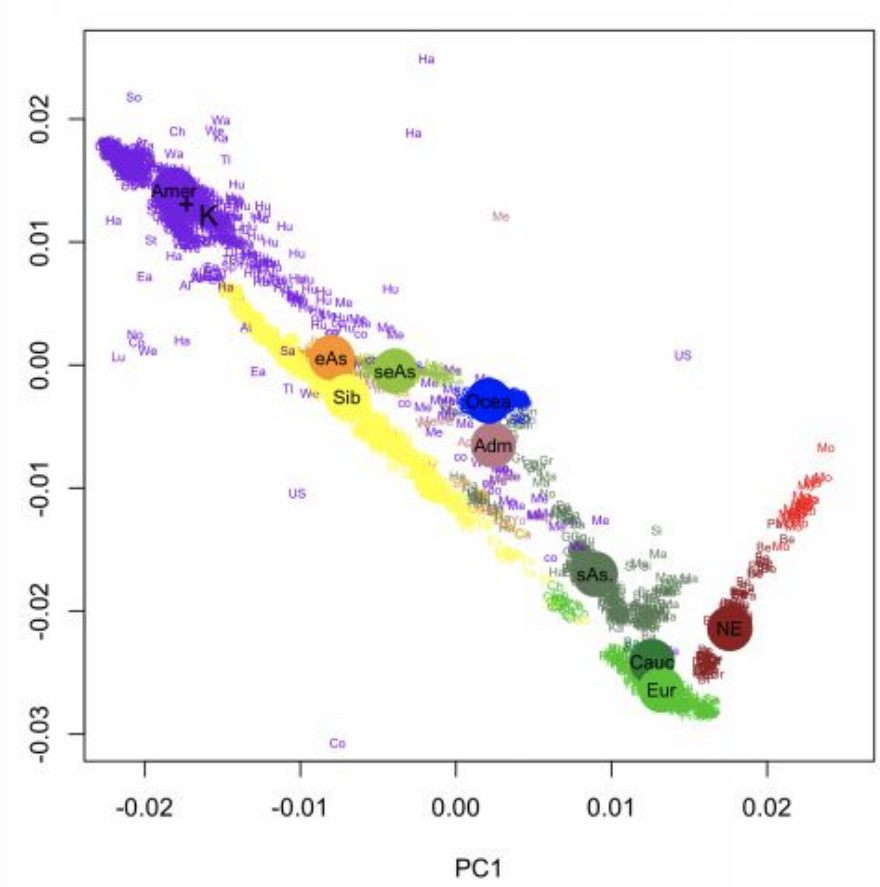


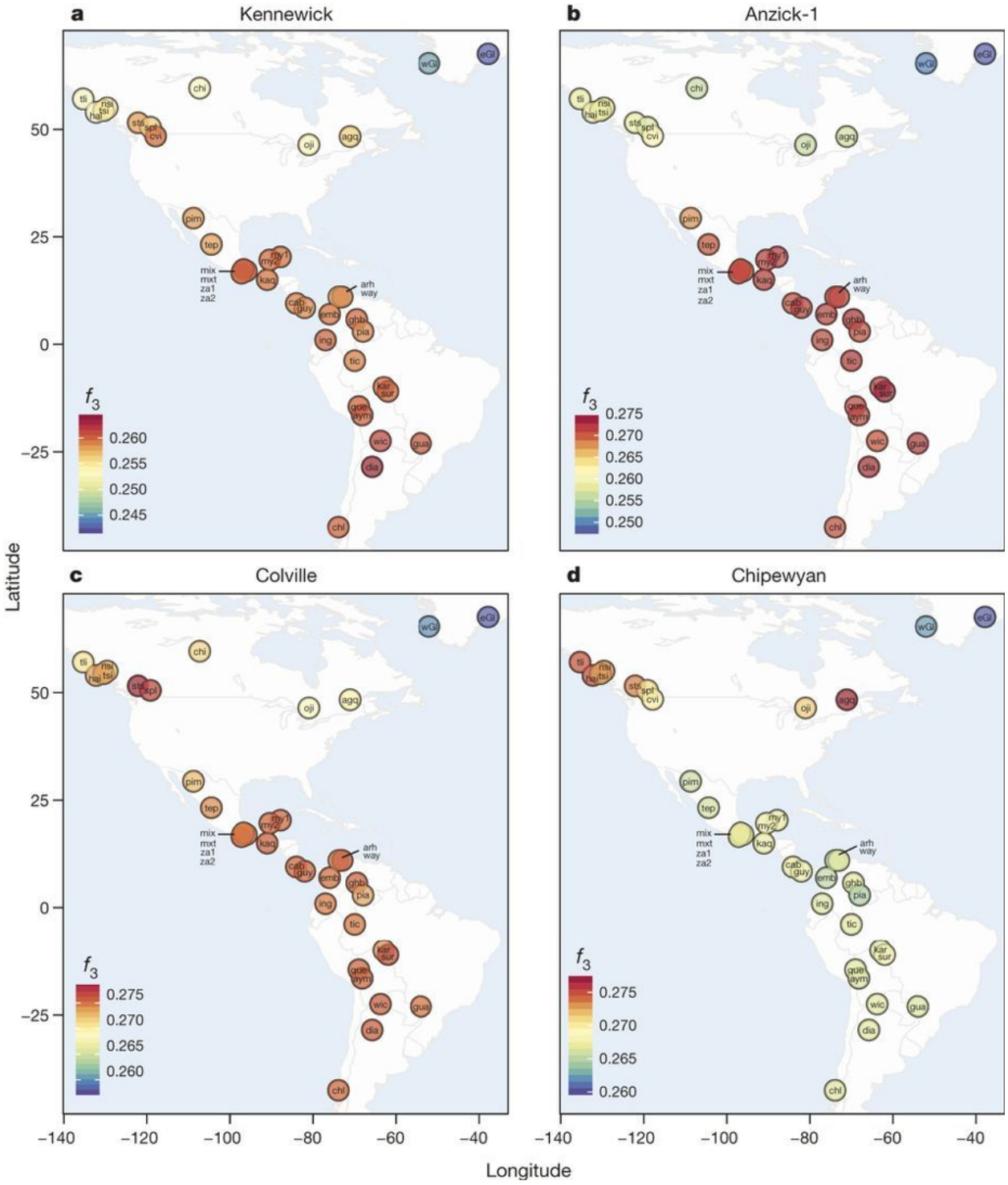
Figure 4: PCA of global samples with ancient samples projected to show ancestry. 'K' indicates the genetic position of the Kennewick sample, + indicates Anzick-1. For the reference set, each two letter abbreviation denotes the position of an individual and the abbreviation denotes the population label given by Rhagavan et al. (2015). The large points represent median positions of individuals from across different geographic regions or groupings. Regional abbreviations: Afr=Africa, Amer=America, Adm=Admixed from the Americas, Cauc=Caucasus, eAs=East Asia, Eur=Europe, NE=Near East, Ocea=Oceania, Sib=Siberia, sAs=South Asia, seAs=Southeast Asia. Samples from Africa are to the right along the PC1 axis but are not visible due to the plotted region.

5. In 2016, researchers from the University of Chicago published a report assessing the genetic analyses of the 2015 paper from question (4). The report aimed to re-examine the genetic evidence surrounding the Kennewick Man case. A PCA plot from this report is shown above.

a. Why might it be important to repeat the analyses of previous publications? Peer review, independent verification
b. Are there any populations that appear more closely related to Kennewick Man here than in the previous PCA plot? Siberians and East Asians appear more closely related to KEN than before
c. What does your answer to part c) tell you about the reliability of PCA for genetic inference? Not an exact science, representation can vary with construction of data set
d. Are you convinced of Kennewick Man's ancestry? Why or why not? Student opinion

6. Short survey:
Name 1 thing you learned:
Name 1 thing you'd like to know more about:
?Rate engagement out of 10?

EXTRA FIGURES TO POSSIBLY INCLUDE


a
Kennewick
b
Anzick-1
c
Colville
d
Chipewyan
Latitude
Longitude
50
25
0
−25
−140
−120
−100
−80
−60
−40
f_3
0.260
0.255
0.250
0.245
0.275
0.270
0.265
0.260
wGl
eGl
chi
tli
nsi
haj
tsi
st6
spt
cvi
oji
agq
pim
tep
my1
my2
mix
mxt
za1
za2
kaq
arh
way
cab
guy
emb
ghb
pia
ing
tic
kar
sur
que
aym
wic
gua
dia
chl

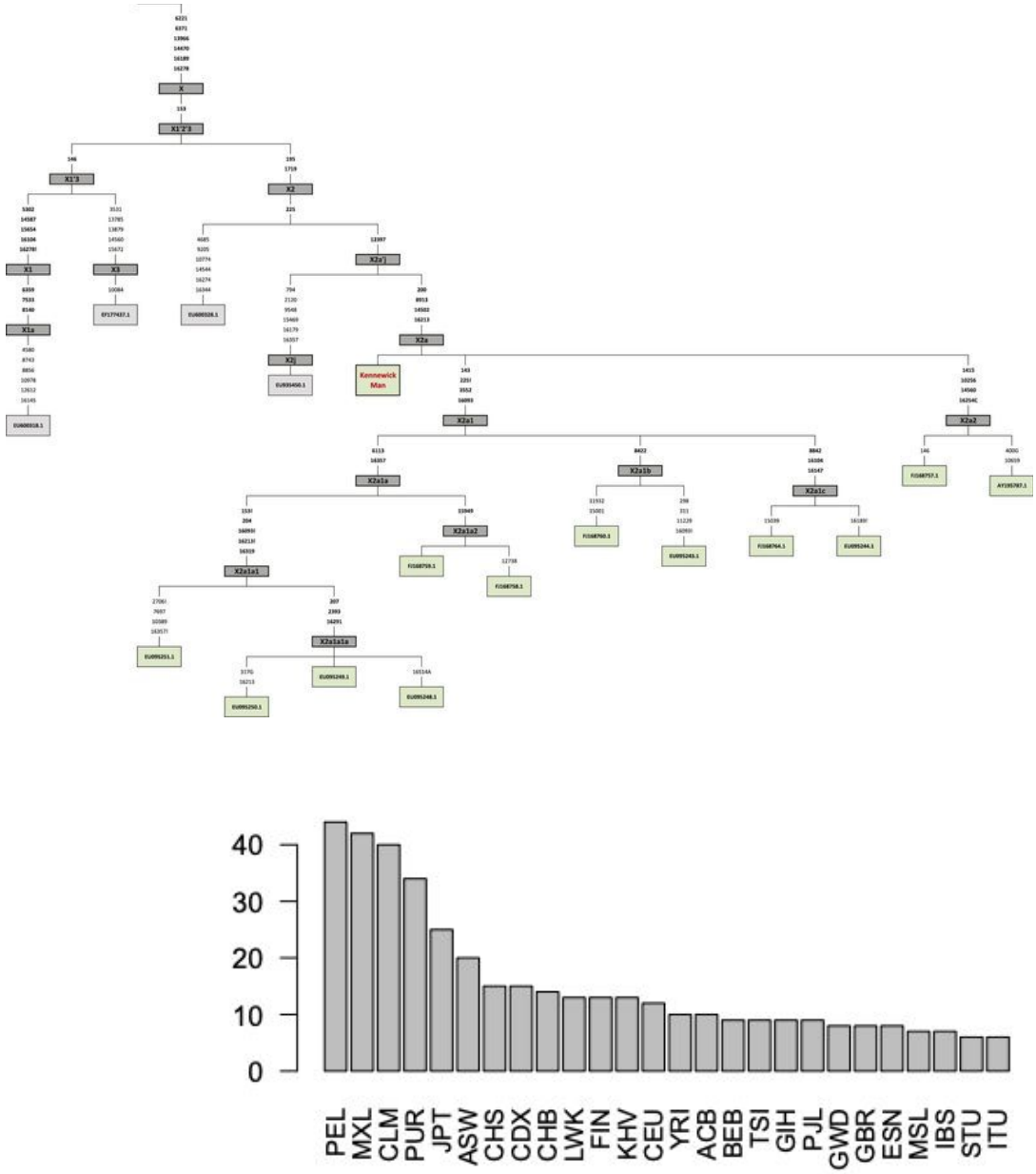


Figure 8: **Rare variant sharing profile of the Kennewick sample.** The y-axis counts the number of globally rare variants in the 1000 Genomes Project found in each population and carried by the Kennewick sample. The highest levels of sharing for Kennewick are with populations from the Americas: PEL = Peruvian from Lima; MXL = Mexican from Los Angeles; CLM = Colombian from Medellin; PUR = Puerto Rican, full list of codes available in the phase 3 1000 Genomes Project Consortium paper (2015).