CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Apr 11

Faculty candidate talk right after this!

- Recap working in log-space and posterior mean
- Finish human evolution overview
- Begin: PCA

- Office hours TODAY 1-2:30pm
- Interviewing faculty candidate 2:30-3:30pm, available after that
- Email me if you would like to meet for your project
- Optional PCA lab released on Thursday, PCA reading posted
- Graded labs and midterm 2 study guide coming out soon!

Main takeaways from feedback forms

- Needs more work: DP, population genetics, coalescent, Tajima's D, HMMs, tree algorithms

- Lot of diversity in terms of in-class style (some prefer board, some slides, some group work) and lab style (some like problem sets, some don't, etc)

- Overall: recent labs long, too much I/O formatting, lots of material too quickly, lots of math, need more biological motivation + big picture, still like worksheets

- <u>Modifications:</u>
  - *More time for Lab 8, more code for reading files*
  - *Lab 9 optional + human evolution motivation for PCA*
  - *Midterm review session (Monday in-class)*
  - *Project scope small (data X, algorithm Y, results Z)*

Topics to cover in the last three weeks

- Machine learning, deep learning, and neural networks
- Ethics (HeLa, prenatal testing)
- Evolution (natural selection, disprove evolution?)
- More statistics, mathematical models, Bayesian computation and inference
- What is going on in bioinformatics industry
- Conservation genetics, other species (plants)
- Structures (i.e. RNA folding, protein structure) vs. sequences
- Algorithms with broad applications, how do our algorithms fit together
- Epigenetics
- More population genetics
- More runtime analysis

Order roughly corresponds to how many people mentioned a topic

Recap: working in log-space and posterior mean

Why do we need to work in log-space?

- Multiplying (or adding) many small probabilities will result in underflow
- Underflow: at some point the computer will represent very small numbers as 0
- Overflow: at some point very large numbers can no longer be represented
- Multiplication and division are okay:

$$\log(ab) = \log(a) + \log(b)$$

$$\log(a/b) = \log(a) - \log(b)$$

- The difficulty arises with adding:

$$\log(a + b) = ???$$

In Python3 on my system:

$e^{709} = 8 \times 10^{307}$
e^{710} = overflow

$e^{-745} = 5 \times 10^{324}$
e^{-746} = 0.0 (underflow)

working in log-space

$\tilde{p} = \log p$

forward recursion

$$\log\left(f_k(i) = e_k(x_i) \cdot \sum_{\ell} f_\ell(i-1)\, a_{\ell k}\right)$$

$$\tilde{f}_k(i) = \log(e_k(x_i)) + \log\Big(\underbrace{f_0(i-1)\,a_{0,k}}_{p} + \underbrace{f_1(i-1)\,a_{1,k}}_{q} + \underbrace{f_2(i-1)\,a_{2,k}}_{r} + \underbrace{f_3(i-1)\,a_{3,k}}_{s}\Big)$$

$\log \swarrow \approx -1600$ $\qquad$ $\log \swarrow -1700$ $\qquad$ $\rightarrow \tilde{q} - \tilde{p} = -100$

$$\rightarrow \log(\underbrace{\underbrace{\underbrace{p+q}_{t}+r}_{u}+s})$$

$$\log(p+q)$$
$$= \log\left(e^{\tilde{p}} + e^{\tilde{q}}\right)$$
$$= \log\left(e^{\tilde{p}} \cdot \left(1 + \frac{e^{\tilde{q}}}{e^{\tilde{p}}}\right)\right)$$
$$= \log\left(e^{\tilde{p}} \cdot \left(1 + e^{\tilde{q}-\tilde{p}}\right)\right)$$
$$= \tilde{p} + \log\left(1 + e^{\tilde{q}-\tilde{p}}\right)$$

def log-add($\tilde{p}$, $\tilde{q}$):

$\boxed{}$

$\log(p+q)$

log-add-all(array)

$\tilde{t} = \text{log-add}(\tilde{p}, \tilde{q})$

$\tilde{u} = \text{log-add}(\tilde{t}, \tilde{r})$

result $= \text{log-add}(\tilde{u}, \tilde{s})$

loop!

variance

$\bar{x}$

high variance

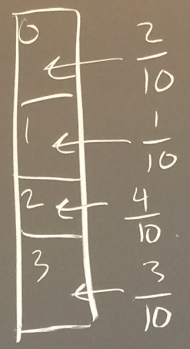
$\bar{x}$

$\boxed{-\bar{f})^2}$

posterior mean

times = [0.32, 1.75, 4.54, 9.40]

$$\bar{z}_i = \sum_k P(z_i = k \mid \vec{x}) \cdot g(k)$$



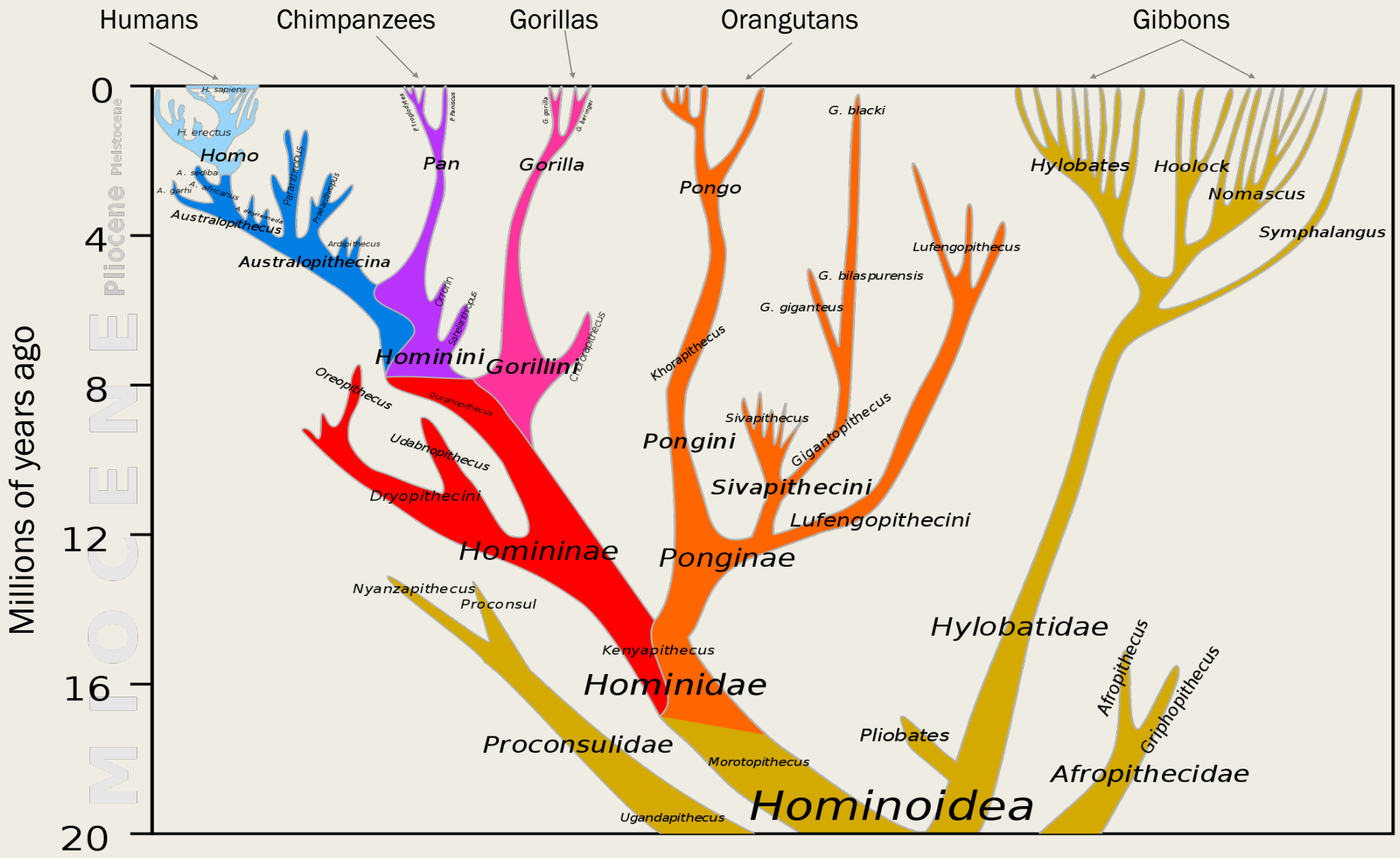
$$\bar{z}_i = \underbrace{\frac{2}{10}(0.32) + \frac{1}{10}(1.75) + \frac{4}{10}(4.54) + \frac{3}{10}(9.40)}_{\text{weighted avg.}}$$

$$\frac{f_k(i)\, b_k(i)}{P(\vec{x})}$$

$$\underbrace{\sum_k P(z_i = k \mid \vec{x})}_{\text{transform back into non-log space}} = 1 \approx \pm \boxed{10^{-6}}$$

Continue: human evolution overview

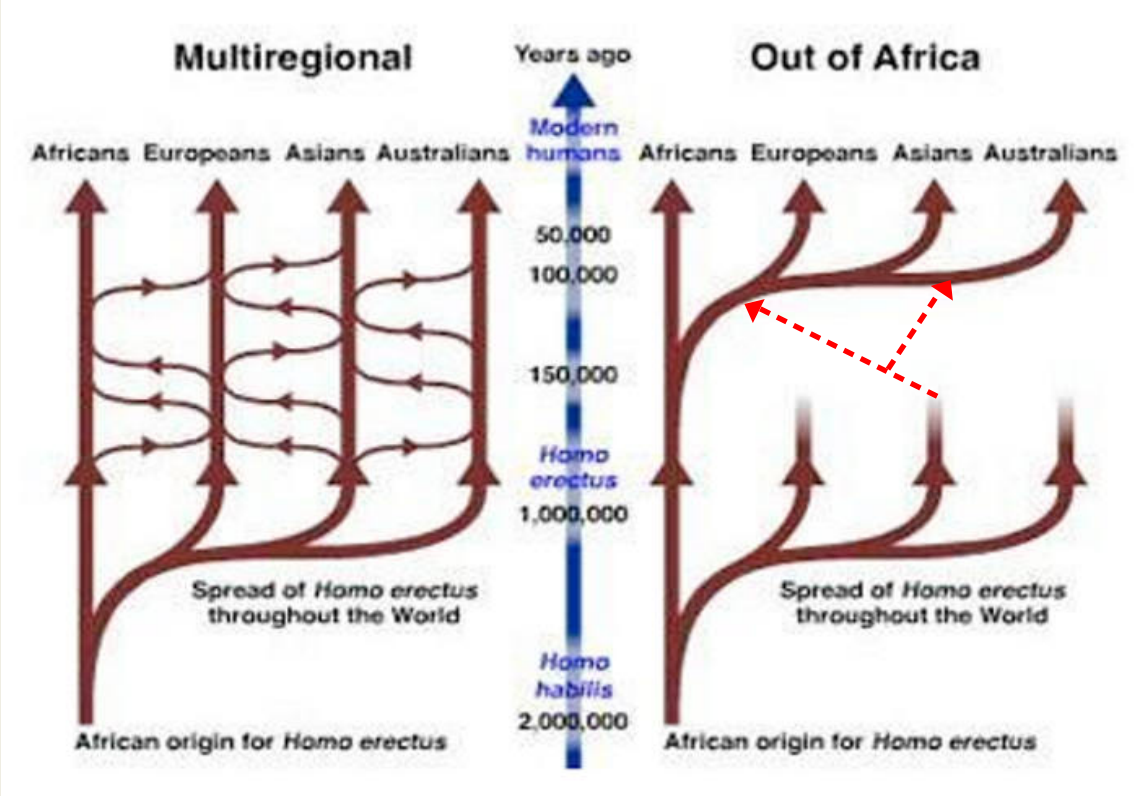
20 million years of Apes



Slide: Modified from Iain Mathieson

Wikipedia/Dbachmann

Multi-regionalism vs Recent African Origin



An old debate about human origins

Genetic data largely resolved this question in favor of RAO, starting in the late 1980's

Recent data , particularly ancient DNA slightly modifies this model.

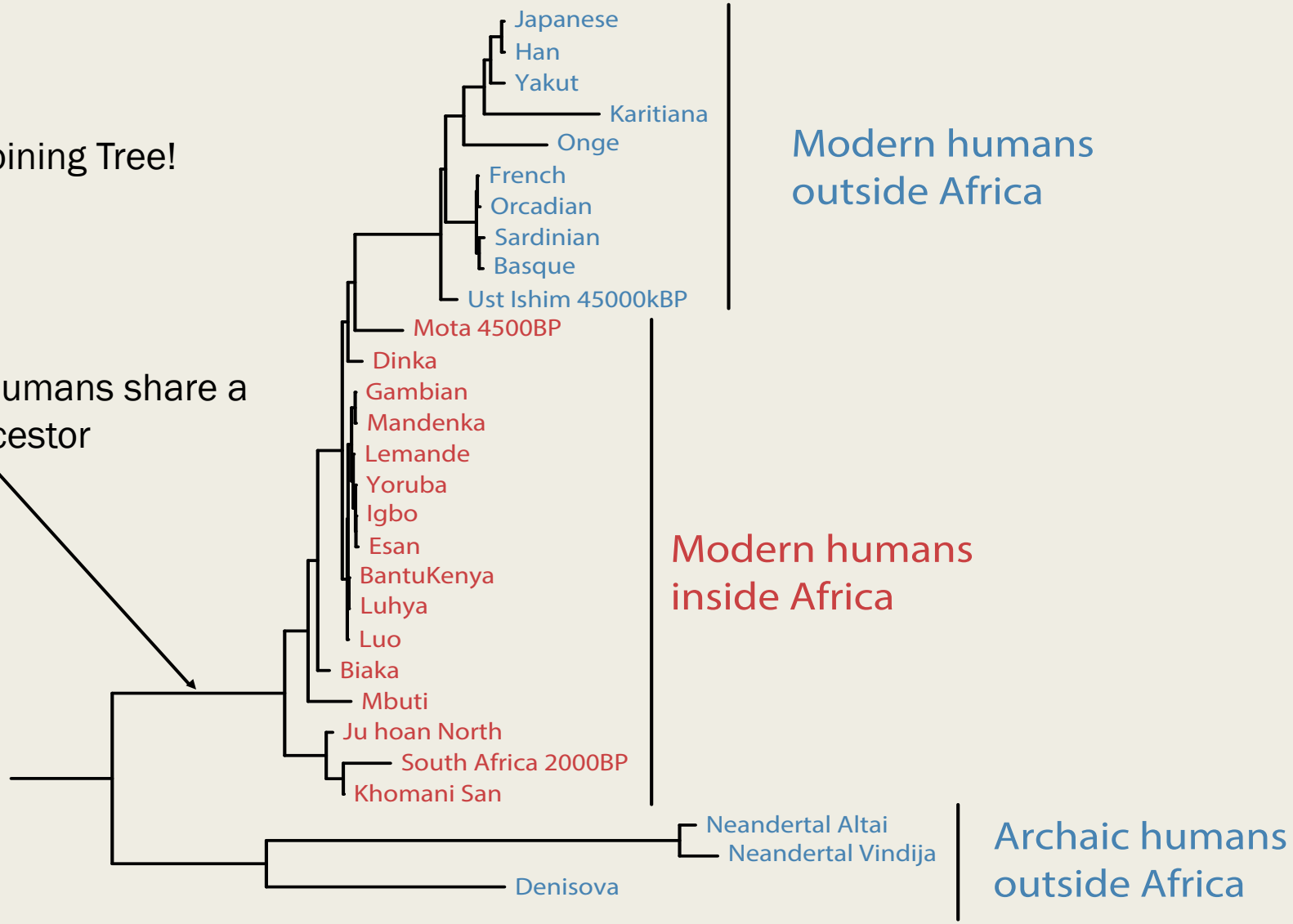
"Leaky replacement" – Svante Pääbo

Slide: Modified from Iain Mathieson

DNA data confirms RAO (Recent African Origin)

Neighbor Joining Tree!

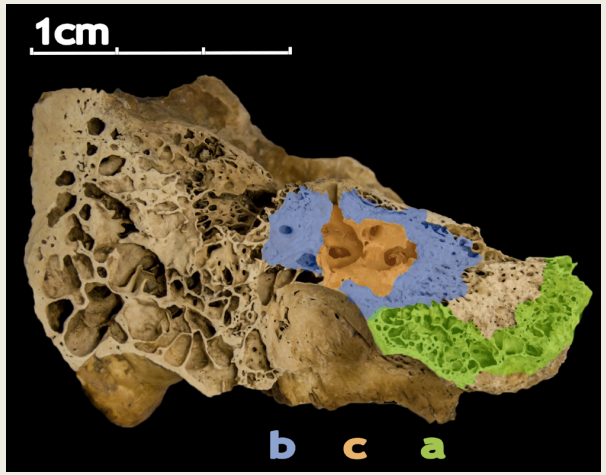
All modern humans share a common ancestor



Skoglund & Mathieson 2018

Ancient DNA

- Best preserved in cold and dry conditions, and dense parts of the skeleton (teeth, inner ear)



- Most DNA recovered is bacterial contamination

- When recovered:
  - fragmented to <200 bp molecules
  - cytosine deamination: sequence errors
  - Be paranoid about modern human contamination

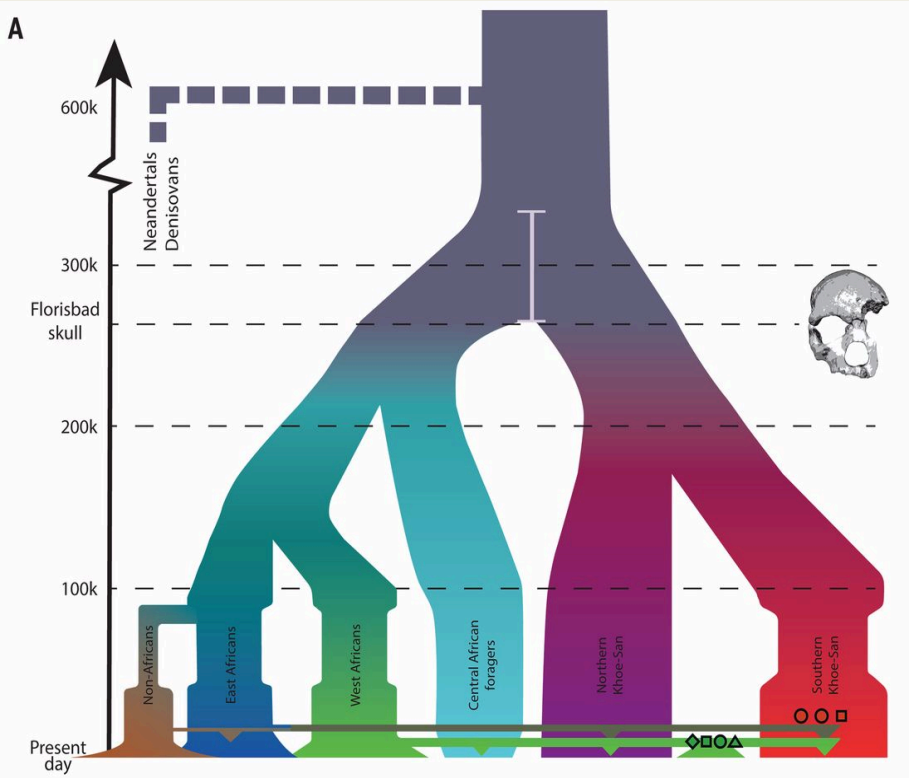
- Upper limit?
  - ~700,000 year-old horse from permafrost in Alaska (Orlando et al. 2010)
  - Oldest human DNA – Sima Los Huesos (Spain). 430,000 year-old proto-Neanderthal (Meyer et al. 2016)

- As of today: 2 high coverage (5 low coverage) Neanderthal genomes, 1 high-coverage Denisovan genome, ~1500 ancient modern humans from 45,000 BP to present.

Slide: Modified from Iain Mathieson

See. Orlando et al. 2015, *Nature Reviews Genetics*

Ancient divergence in Africa



C

Split method	Human-Neandertal (Nean-BBayA)	Human-Neandertal (Nean-San)	Human-Neandertal (Nean-Dinka)	Deep Human (Dinka-BBayA)	Deep Human (Dinka-San)	Deep Human (Mandenka-BBayA)	Deep Human (Mandenka-San)	NKSP-SKSP (San-BBayA)	Out of AFR (Dinka-Sardinian)
G-PhoCS	545 ± 9	534 ± 8	535 ± 9	336 ± 7	282 ± 7	356 ± 7	298 ± 7	185 ± 6	115 ± 6
TT-method	660 ± 33	639 ± 26	632 ± 28	265 ± 5	255 ± 5	256 ± 6	261 ± 5	156 ± 5	76 ± 6

Oldest divergences among human populations found between present-day African populations. >250 KYA separation between Khoe-san and others

Other lineages are also deeply diverged, including those separating East-central and West-central African populations.

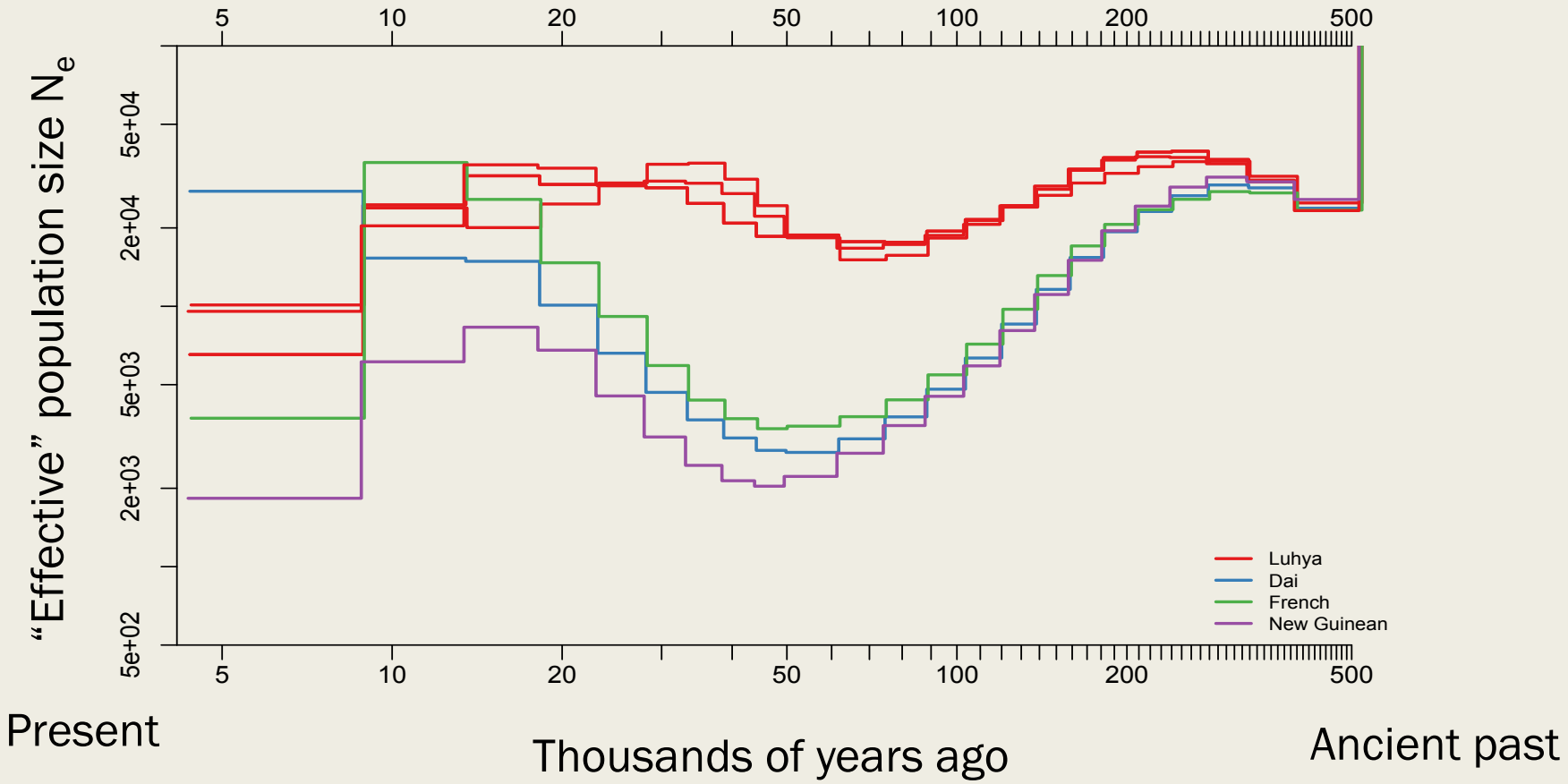
Non-Africans diverge ~70-120 KYA, from a population most closely related to present-day East Africans.

Recent (past 10KYA) population movements have affected this structure

Slide: Modified from Iain Mathieson

Schlebusch et al. *Science* 2017

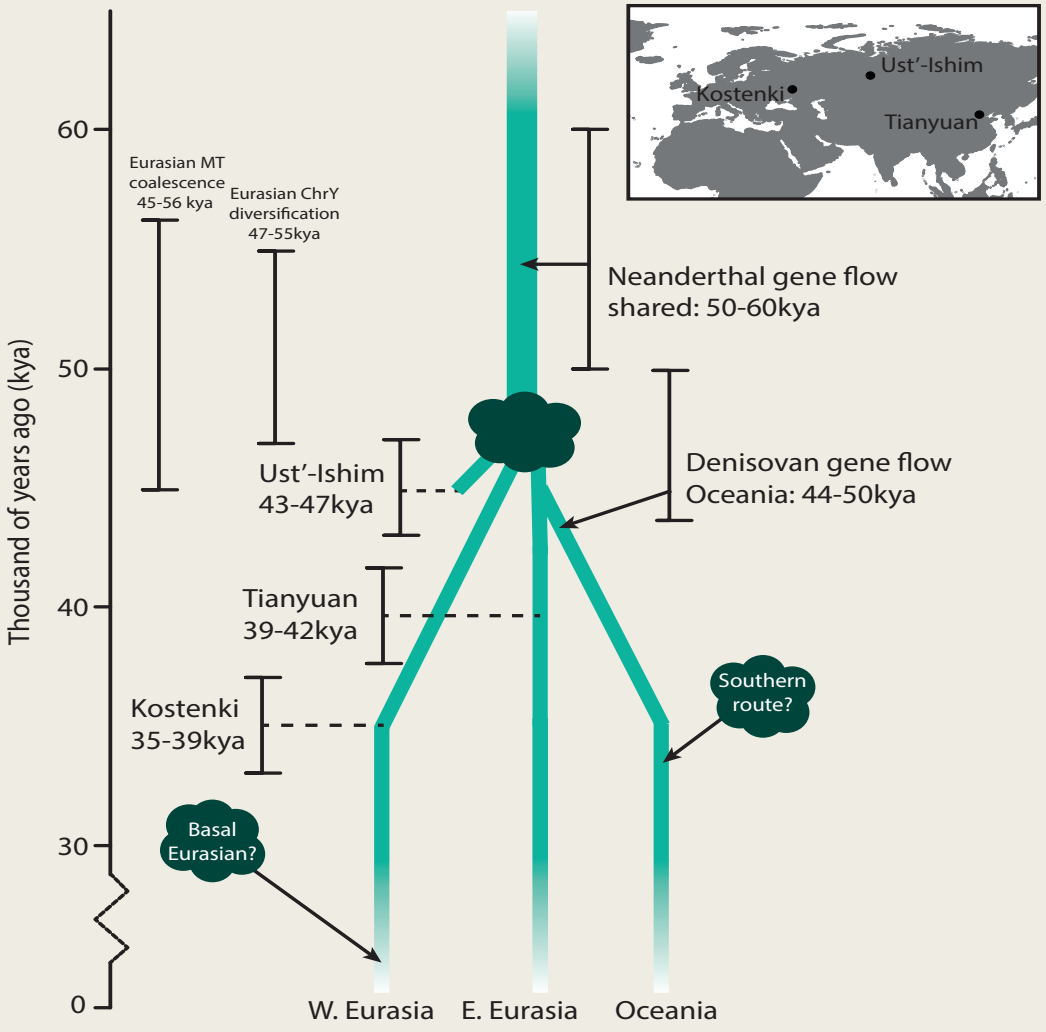
Out-of-Africa bottleneck ~ 50 KYA



Slide: Modified from Iain Mathieson

Method: "PSMC" See https://www.nature.com/articles/nature10231

Relationship between present-day non-Africans



All present-day non-Africans share a common ancestor 45-60 KYA

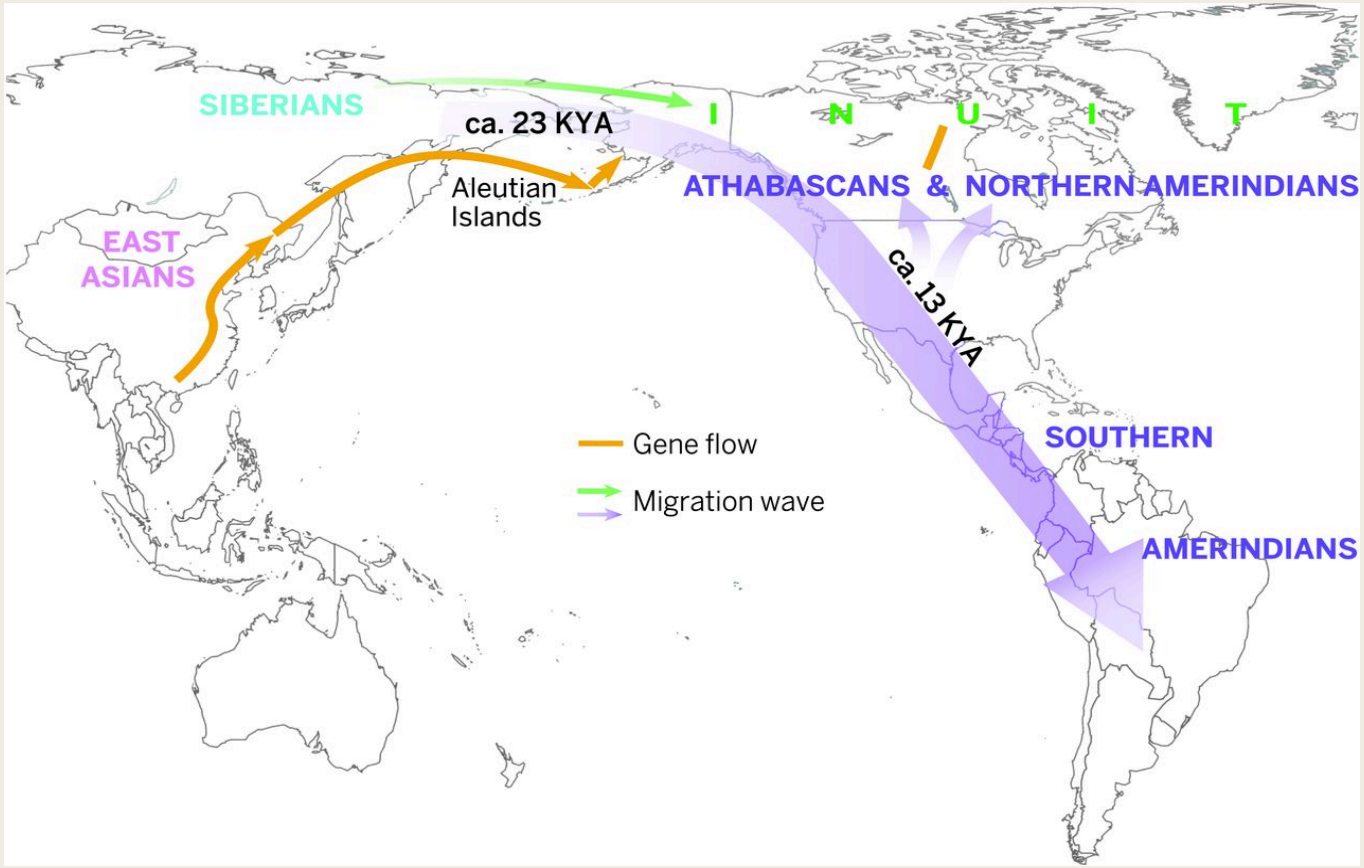
Oceanian lineage diverged by 45 KYA

East and West Eurasian Lineages diverged by 40 KYA

Slide: Modified from Iain Mathieson

Skoglund & Mathieson 2018

The peopling of America



Raghavan et al. 2015

Last major landmass inhabited by humans. At least 14,500 BP

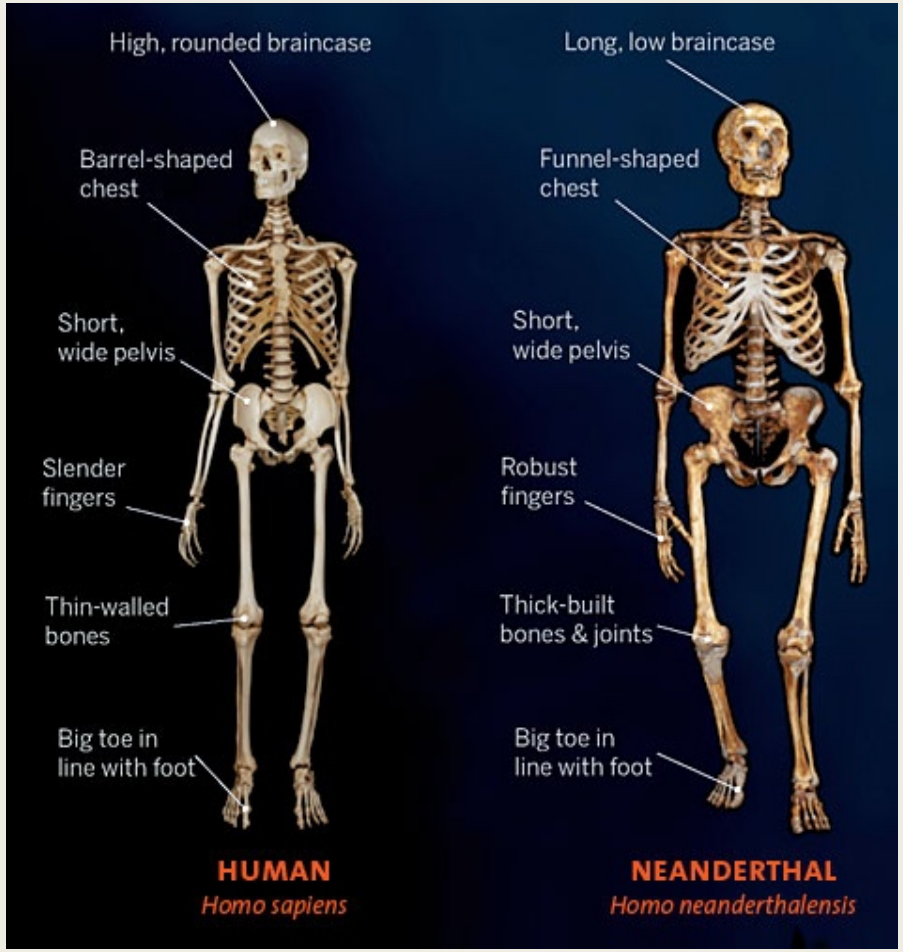
Current model is that Native Americans derive from an admixture between populations related to present-day East Asians and Siberians, likely 10-25 KYA.

Later gene flow in the Arctic and farther south (Paleo-Eskimo, present-day Inuit, Athabascan-speakers).

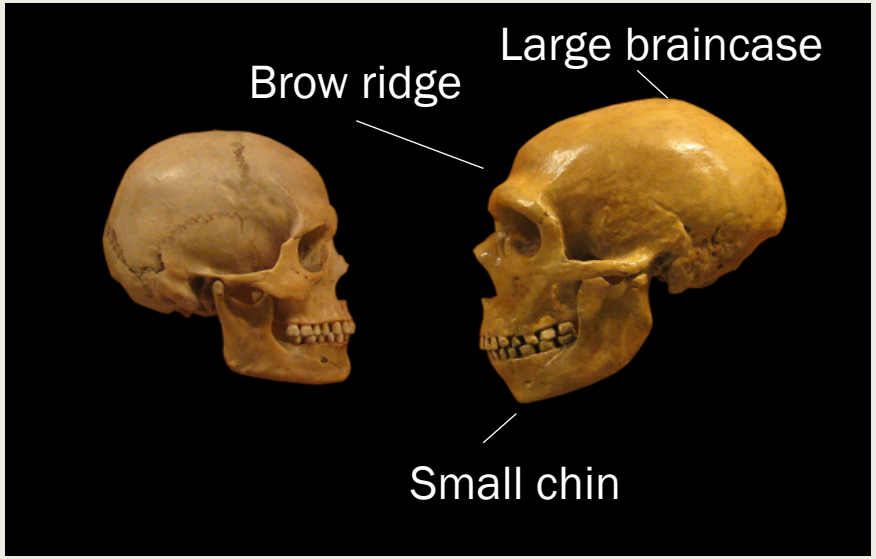
Slide: Modified from Iain Mathieson

Human-Neanderthal divergence

Neanderthals



Smithsonian



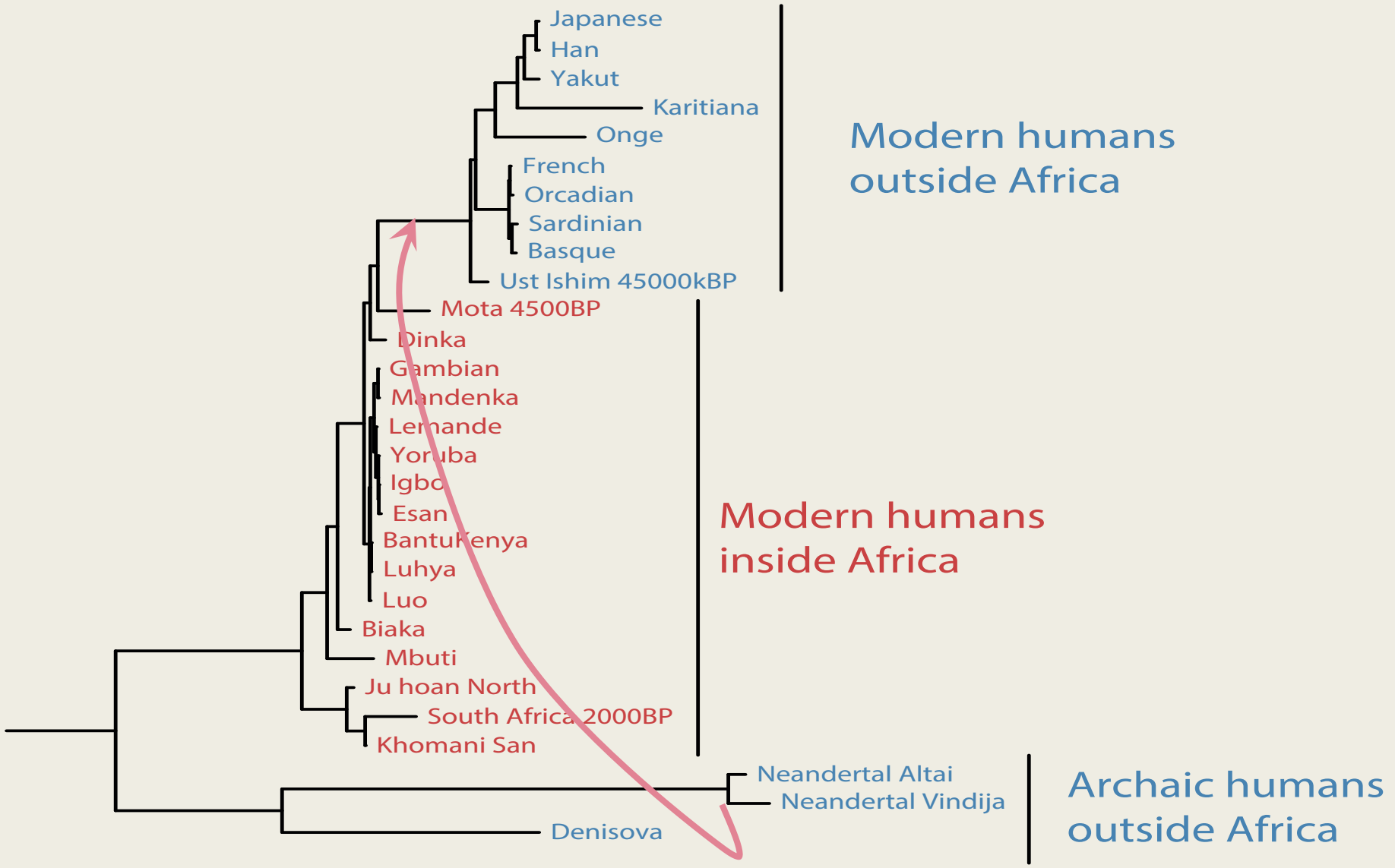
Remains of ~400 Neanderthals identified

Dated from ~250 KYA to ~40 KYA

Estimated divergence from modern humans 400-800 KYA

Slide: Modified from Iain Mathieson

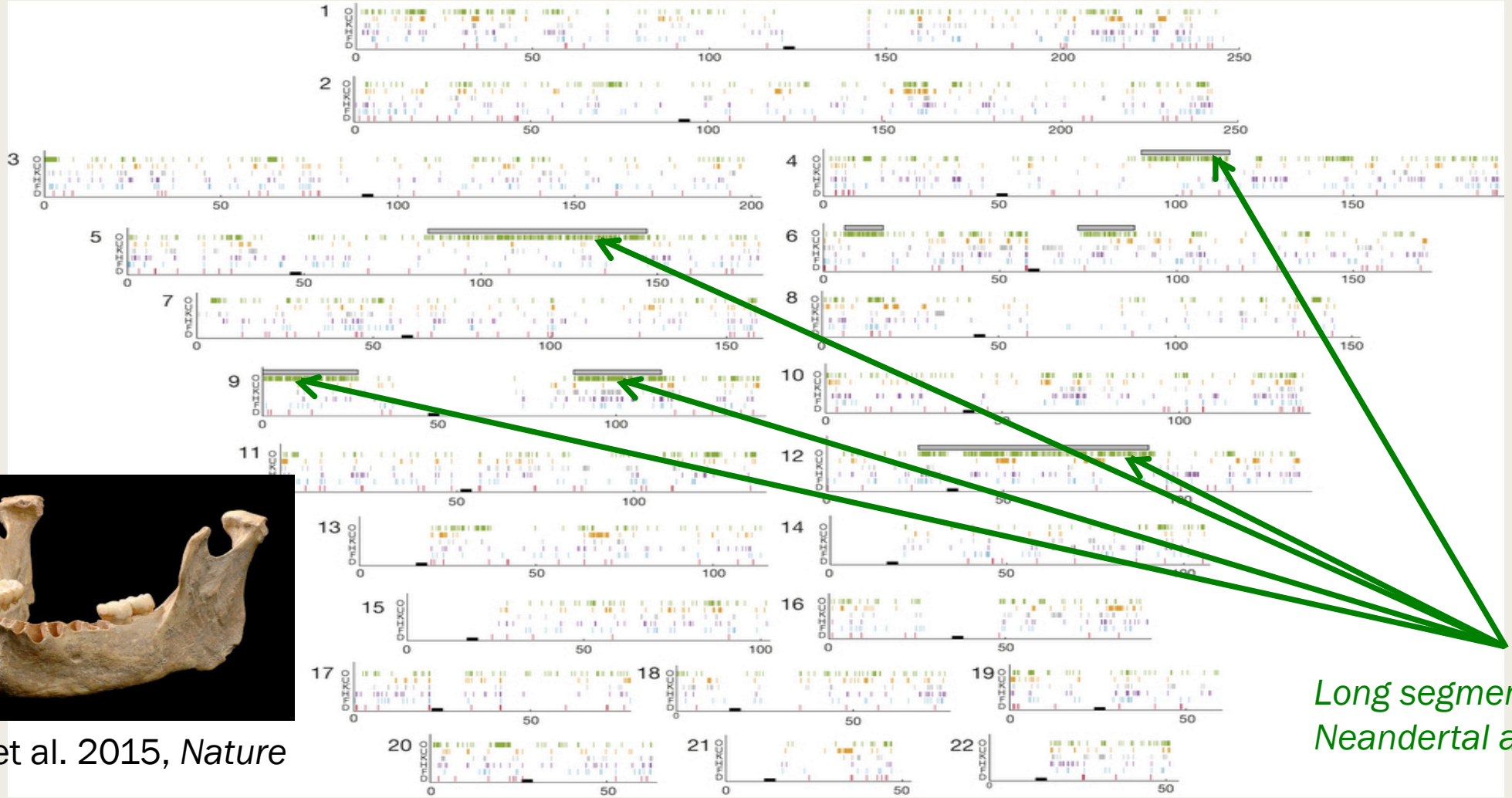
Neanderthal variation falls outside modern humans



Slide: Modified from Iain Mathieson

Skoglund & Mathieson 2018

Oase 1: 40,000 year old European has a recent Neanderthal ancestor

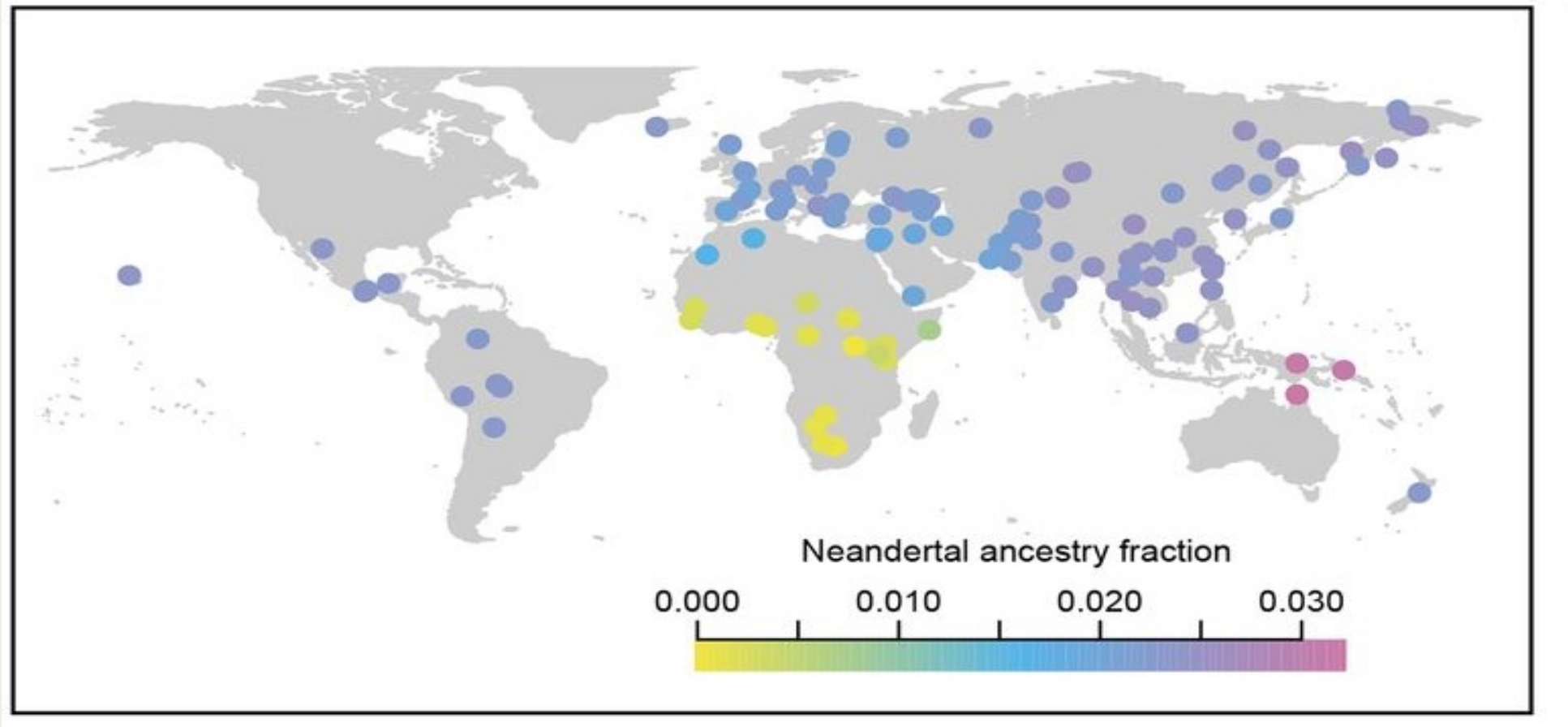


Fu et al. 2015, *Nature*

Long segments of Neandertal ancestry

Slide: Modified from Iain Mathieson

Neanderthal ancestry varies across populations



Western Eurasia: 1.8-2.4%
East Asia: 2.3-2.6%
Oceania - > 3% But, possibly confounded with Denisovan ancestry

Slide: Modified from Iain Mathieson

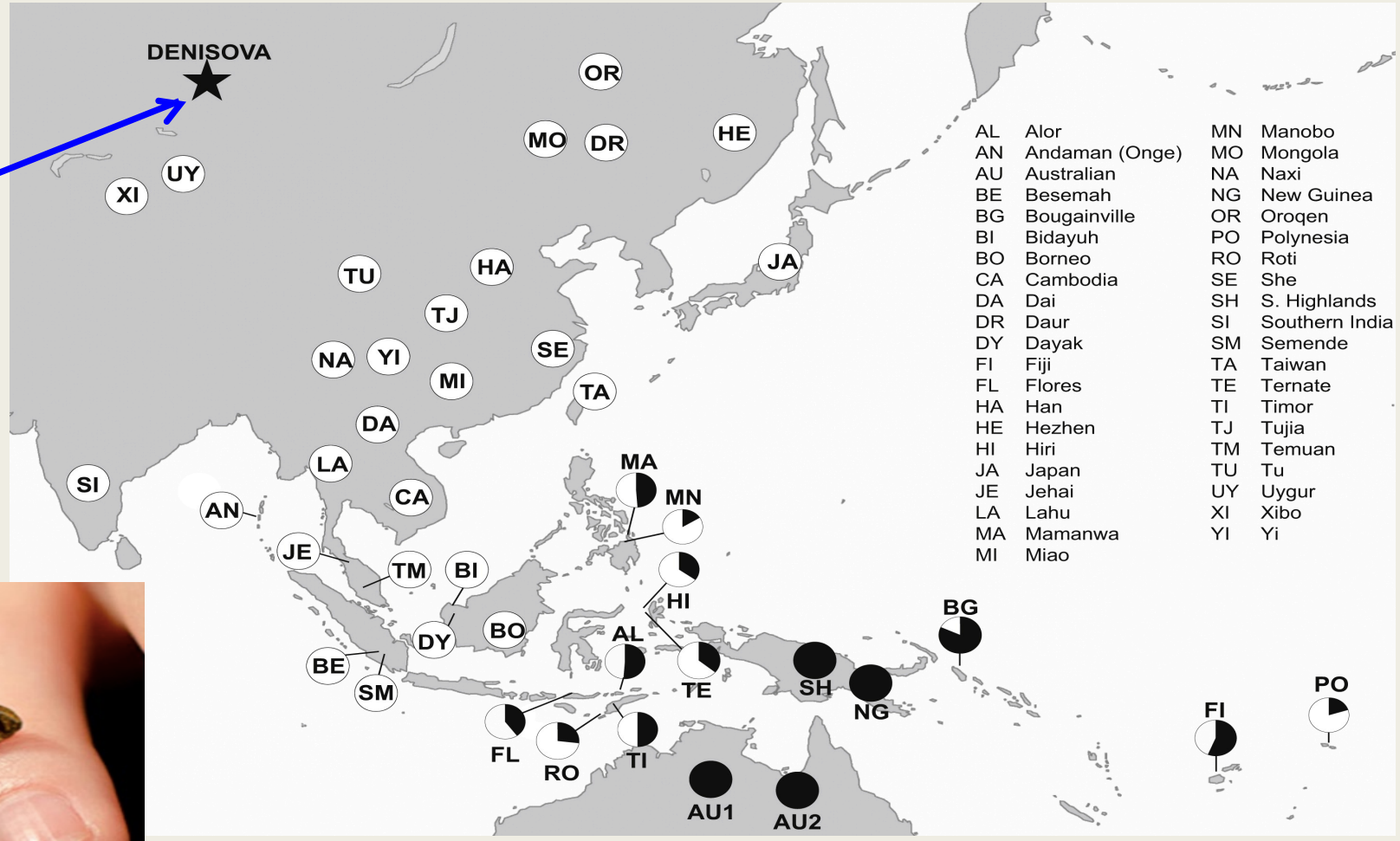
Prüfer et al. 2017, *Nature*

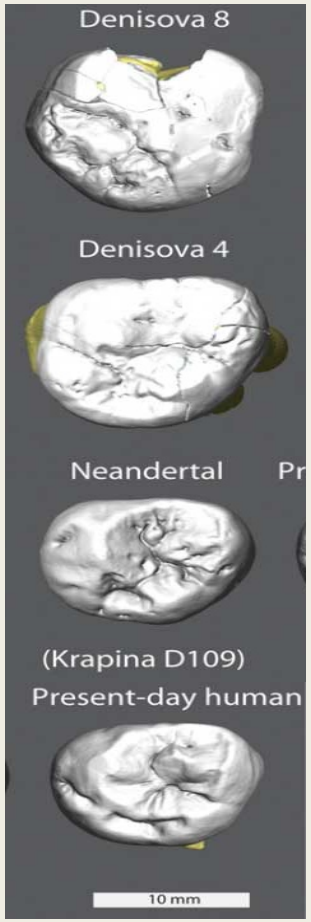
Denisova Cave

MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY

A new human group – Denisovans

Denisova Cave

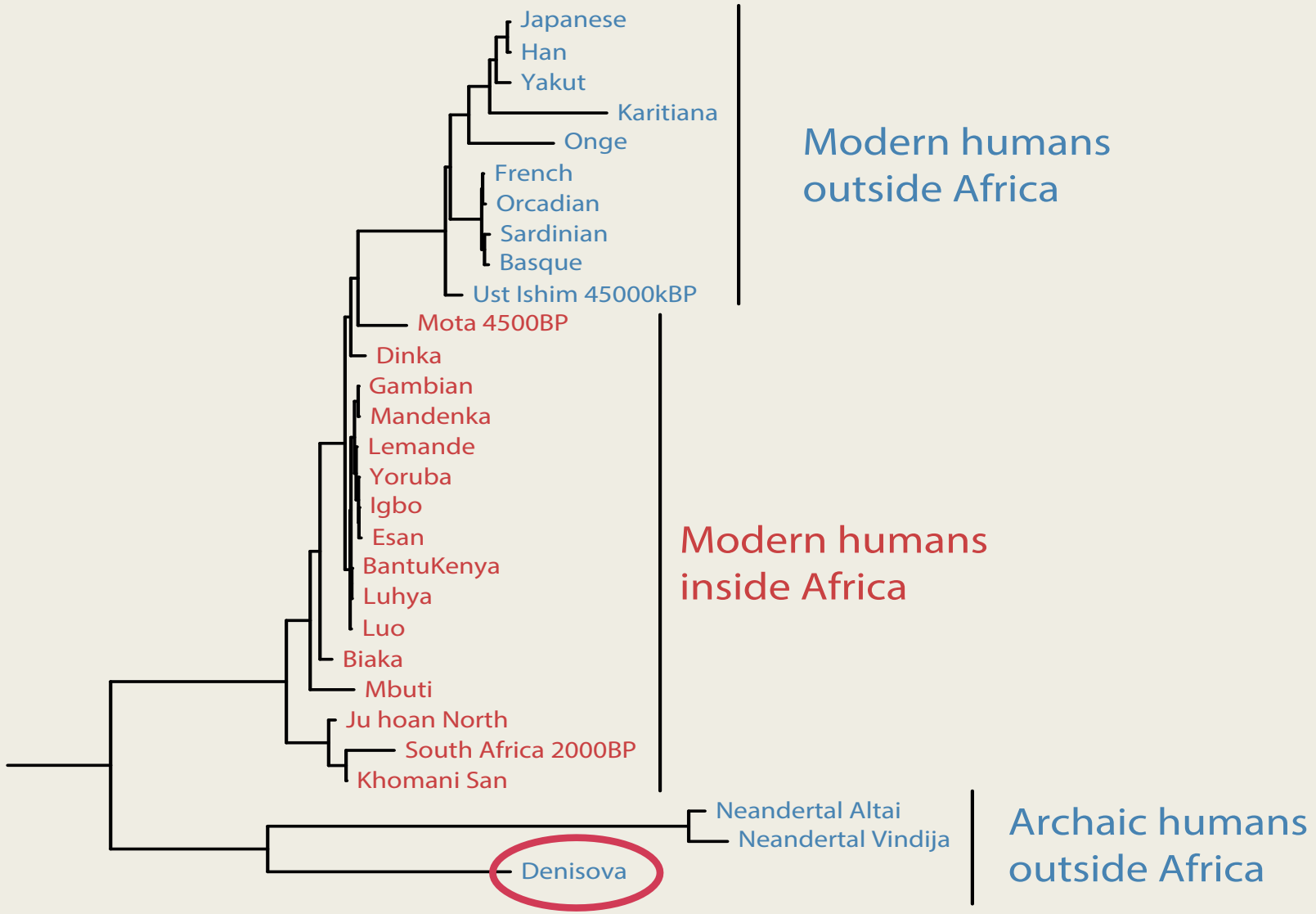




Reich et al. 2010, *Nature*; Reich et al. 2011 *AJHG,* Meyer et al. 2012 *Nature.* Sawyer *et al. PNAS* 2015

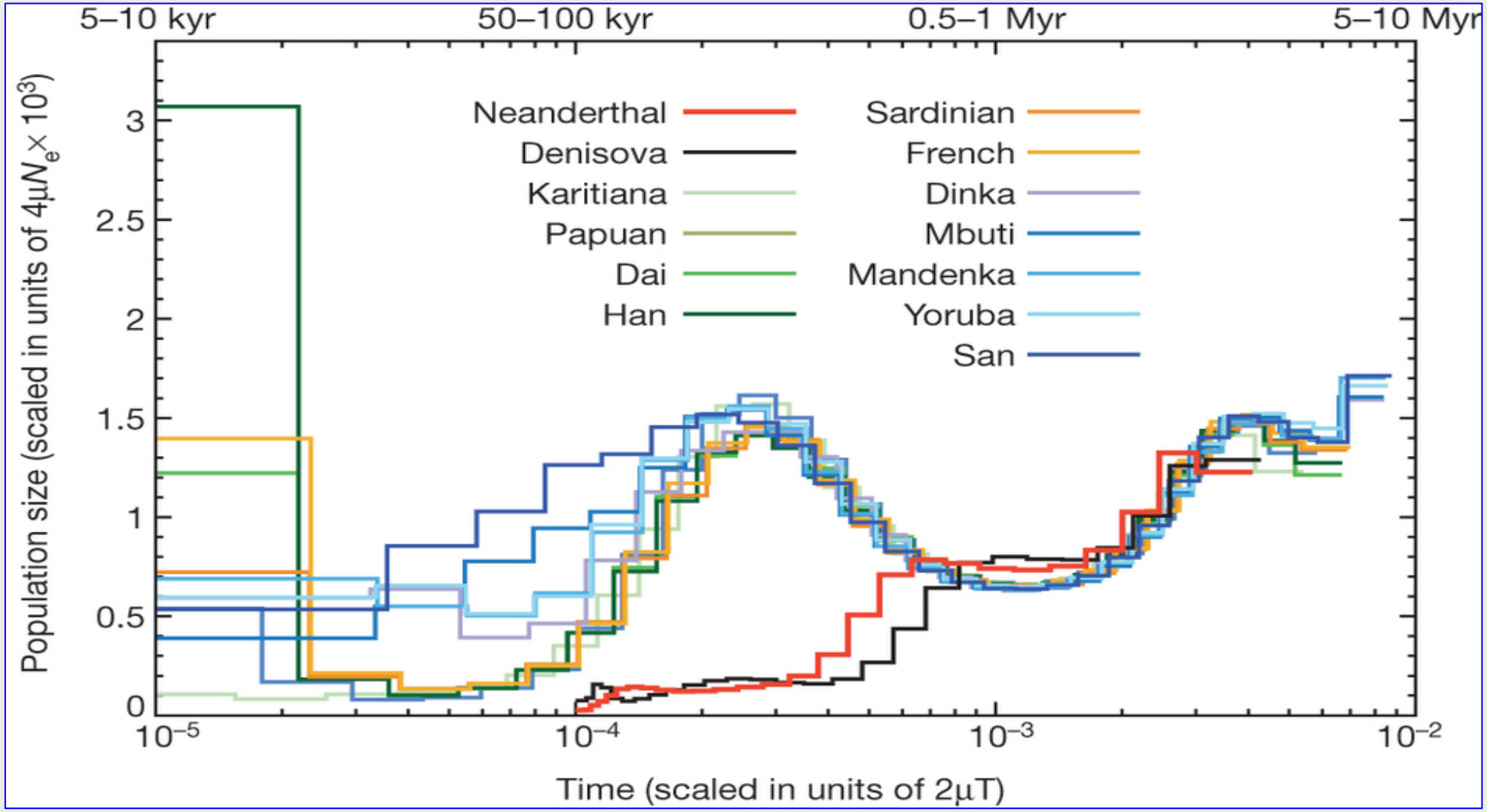
Slide: Modified from Iain Mathieson

Denisovans are a sister group to Neanderthals



Skoglund & Mathieson 2018

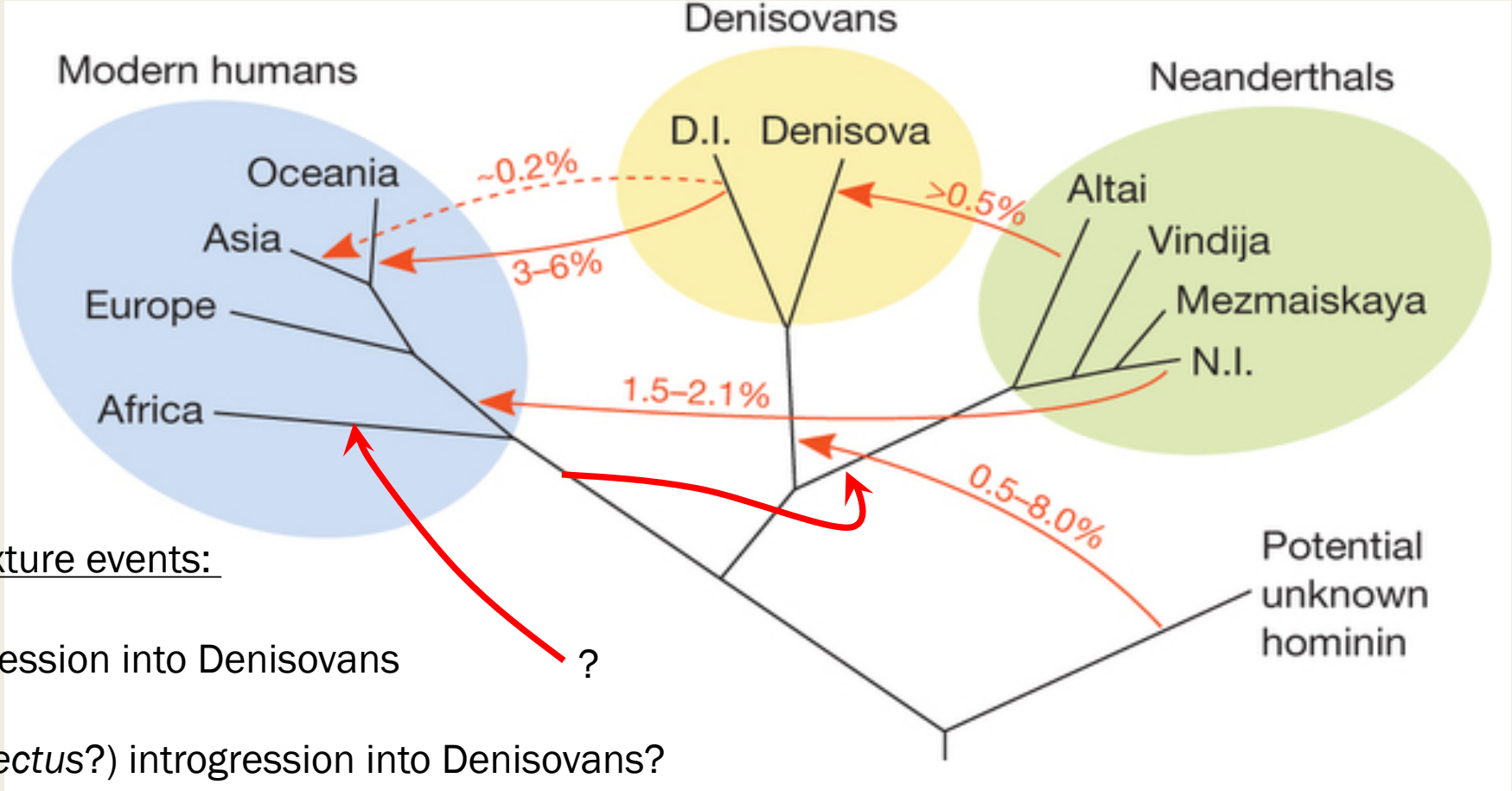
Neanderthals and Denisovans had very small population sizes



- Prüfer *et al* 2014
- PSMC: Li & Durbin 2011

Slide: Modified from Iain Mathieson

Ghost populations and additional admixture



<u>More complex admixture events:</u>

-Neanderthal introgression into Denisovans

-Unknown (*Homo erectus*?) introgression into Denisovans?

-Unknown archaic introgression into west Africans?

-Proto-modern human introgression into Neanderthals??

Slide: Modified from Iain Mathieson

Modified from Meyer et al. 2012

Further reading

Archaic humans:

Prüfer et al. "**The complete genome sequence of a Neanderthal from the Altai Mountains**"
Nature, 2014

Meyer et al. "**A High-Coverage Genome Sequence from an Archaic Denisovan Individual**"
Science, 2012

Human history:

Nielsen et al. "**Tracing the peopling of the world through genomics**"
Nature Reviews Genetics, 2017

Skoglund & Mathieson. "**Ancient genomics: a new view into human prehistory and evolution**"
https://t.co/YrZS45Q0xt, 2017

Natural selection:

Fu & Akey. "**Selection and Adaptation in the Human Genome**"
Annual Review of Genomics and Human Genetics, 2013

Slide: Modified from Iain Mathieson

Begin: PCA

PCA

input $n \times p$ matrix, X_{orig}

n = # samples (2 × # individuals)

p = # features (SNPs, sites)

$$X_{orig} = \begin{array}{c} \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \\ \begin{array}{cc} f & g \end{array} \end{array}$$

$n = 6$, $p = 2$

output:

- feature combinations that best explain observed variation in the data

Step 1: get data

step 2

subtract off column-wise mean

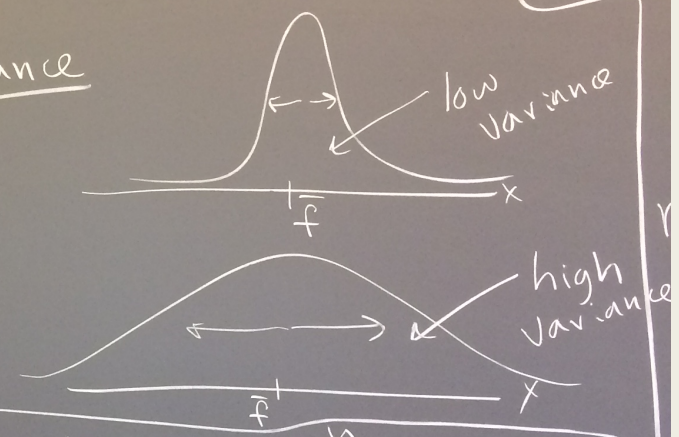
$$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \,, \quad \underline{ex}: \quad \bar{f} = \frac{1}{2}, \quad \bar{g} = \frac{1}{2}$$

$$X = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ -1/2 & 1/2 \\ -1/2 & \frac{1}{2} \\ 1/2 & -1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix}$$

step 3: compute covariance

log-

Variance



$$\boxed{Var(f) = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})^2}$$