Human Evolution

Michael Schatz

October 25 – Lecture 16

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research

Assignment 2: Genome Assembly

Assignment Date: Wednesday, October 25, 2017
Due Date: Monday, November 6, 2017 @ 11:59pm

Assignment Overview

In this assignment, you will explore a few properties of the sequencing data. You can either submit your results in a jupiter notebook, or as a single PDF document. Feel free to sketch the figures by hand, and then include a photograph of your solution. I encourage you to discuss your solutions with other members of the class, but everyone should submit their own write up. You are allowed to use the notes from class, and notes found online to help you work through the problem.

Here are a few helpful resources:

- Python 2 reference
- Jupyter notebooks
- Matplotlib and Gallery
- Numpy and Scipy

Question 1. Read coverage [10pts]

1a. The cichlid fish genome is 1 Gbp. Approximately how many 100bp reads should we sequence so that we expect at least 99.85% of the genome will be sequenced at least 40 times? (Hint: show your work)

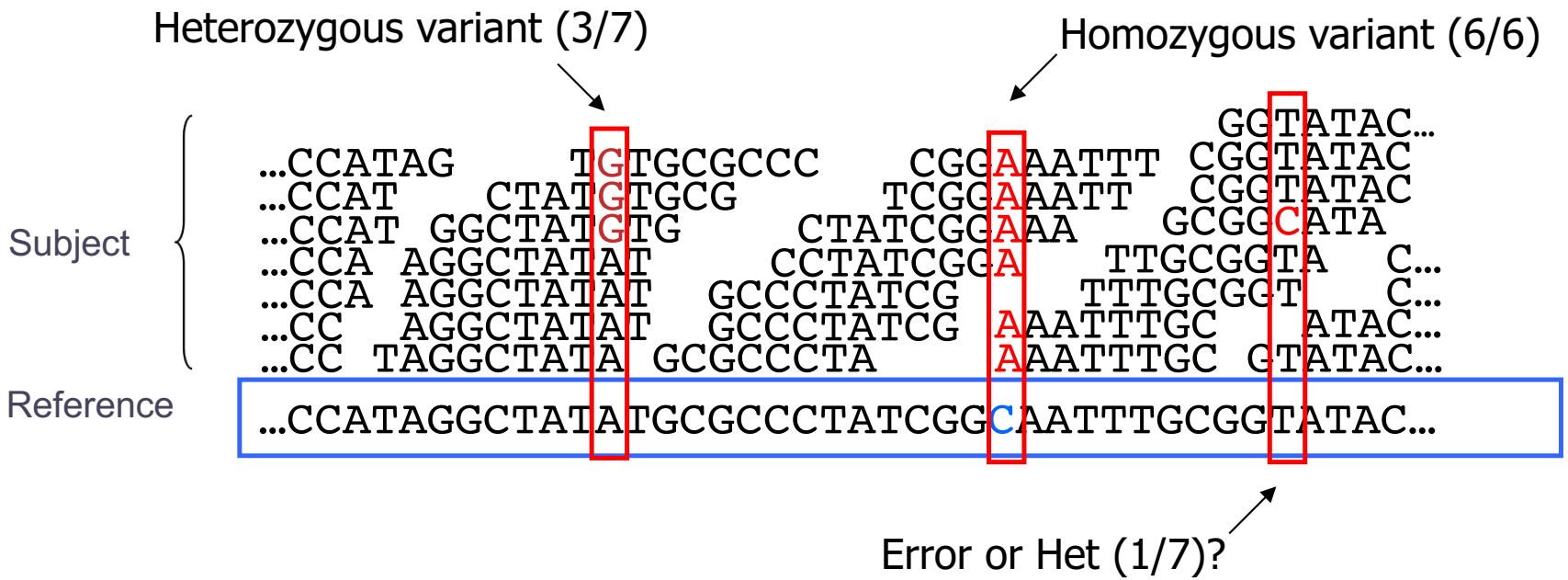
1b. Sketch the expected coverage distribution for this number of reads; be sure to clearly label the mean coverage, and how 40 fold coverage relates to the mean. (Hint: In a normal distribution, 68.2% of the data will be within 1 standard deviation, 95.4% within 2, 99.7% within 3, and 99.9% within 4)

Question 2. de Bruijn graph construction [10pts]

2a. Draw the de Bruijn graph for the following reads using k=3 (assume all reads are from the forward strand, no sequencing errors, complete coverage of the genome)

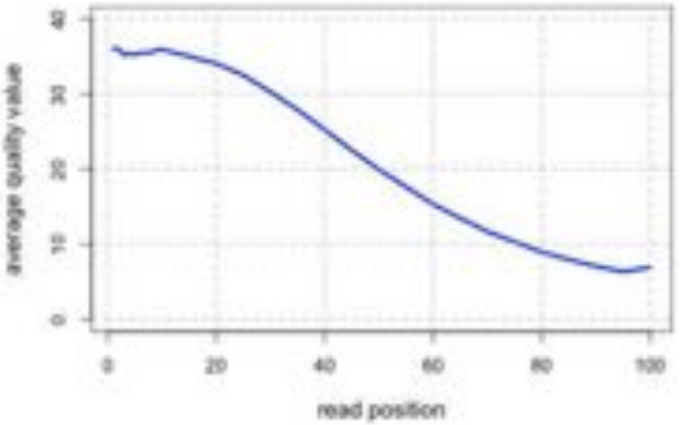
```
ATTC
ATTG
GATT
CTTA
GATT
TATT
TCAT
```

Genotyping Theory



- If there were no sequencing errors, identifying SNPs would be very easy: any time a read disagrees with the reference, it must be a variant!

- Sequencing instruments make mistakes
  - Quality of read decreases over the read length



- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times

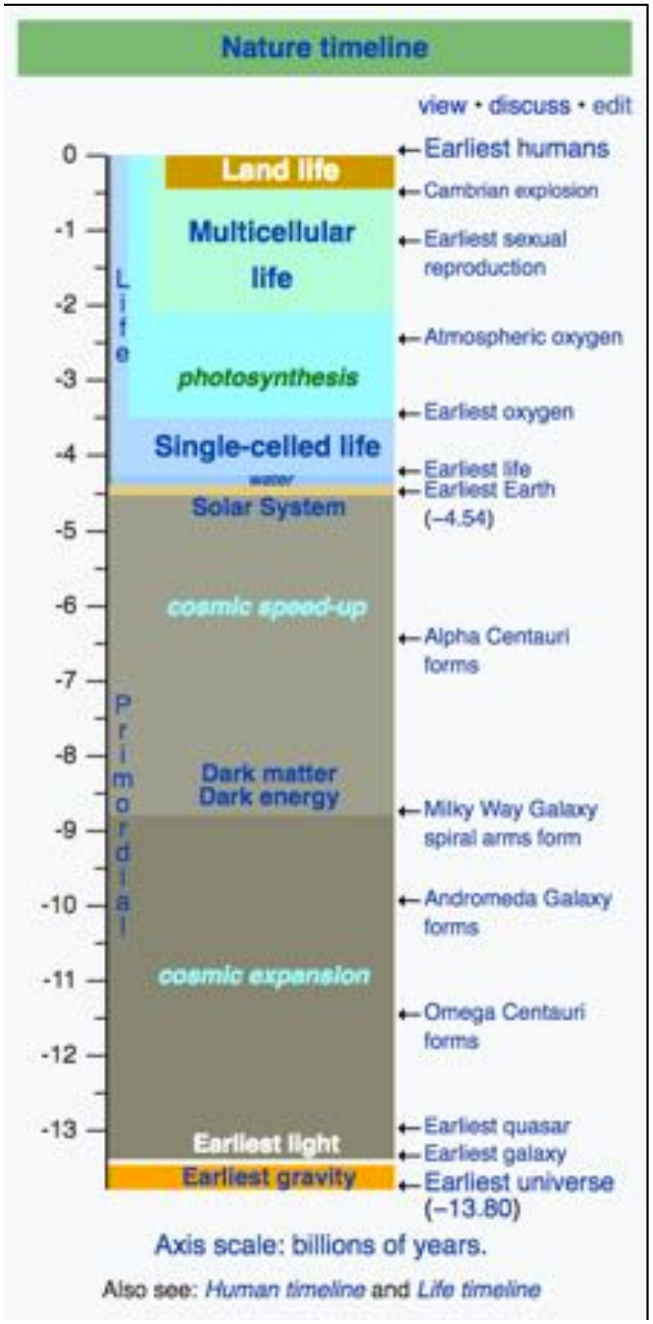
The Binomial Distribution: Adventures in Coin Flipping

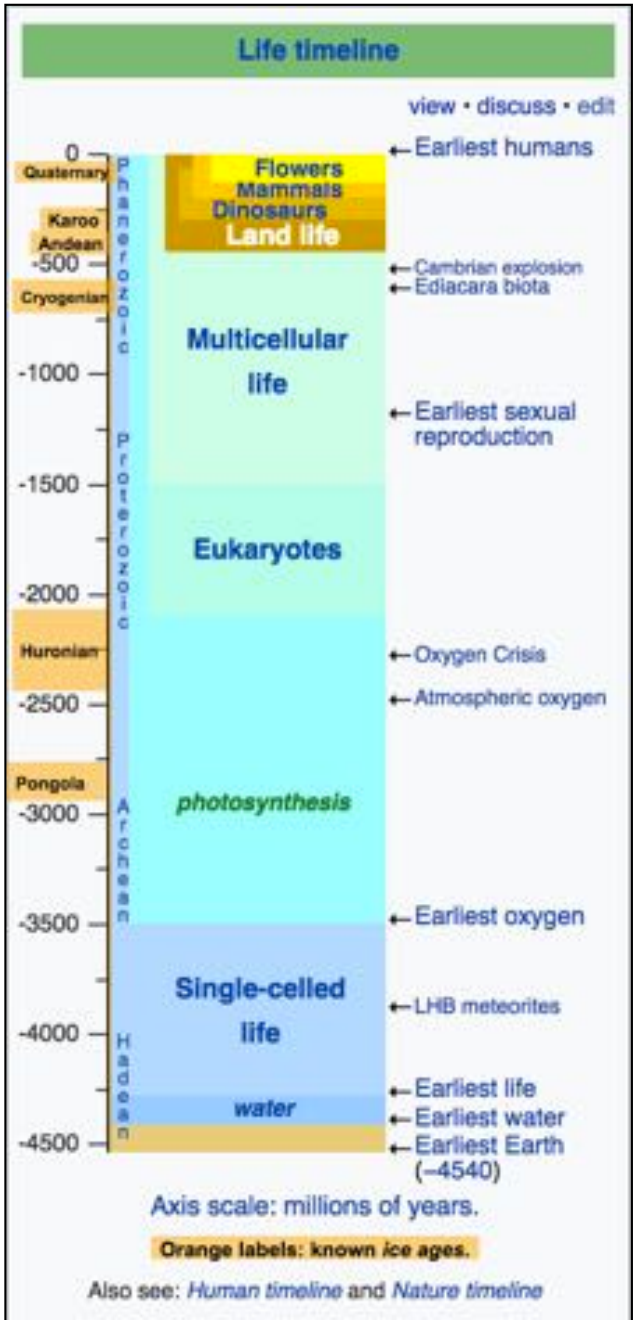
P(heads) = 0.5

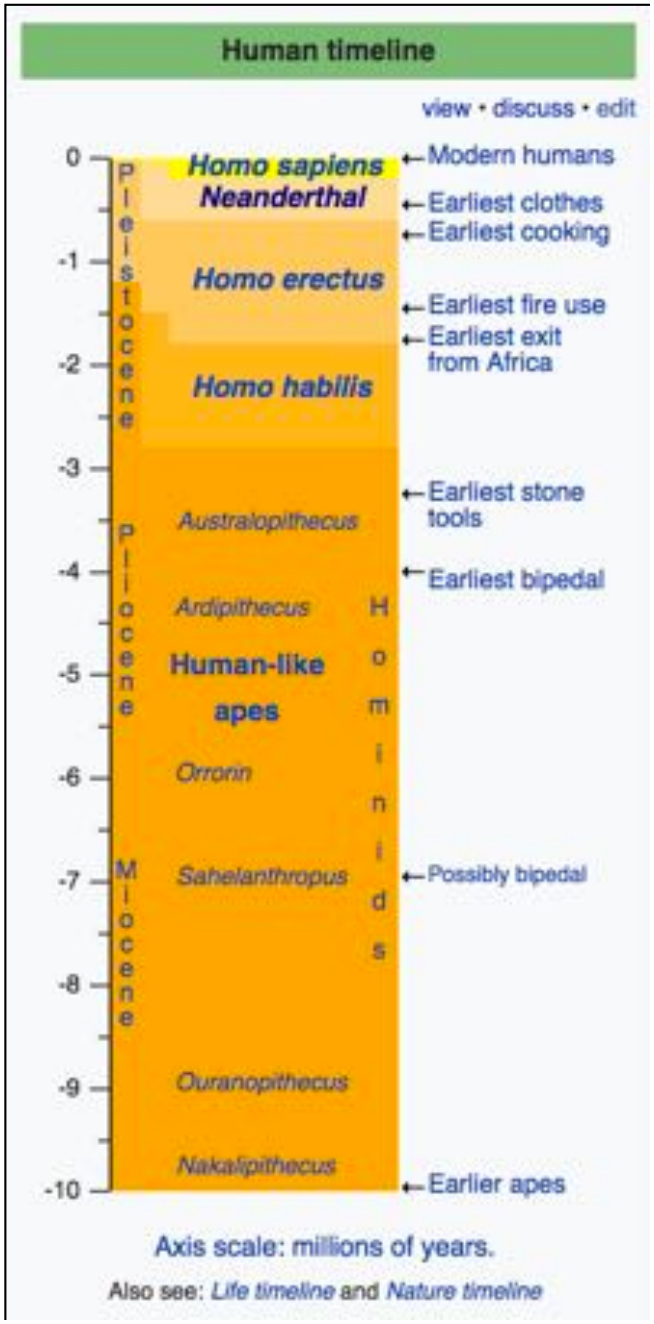
P(tails) = 0.5

Aaron Quinlan

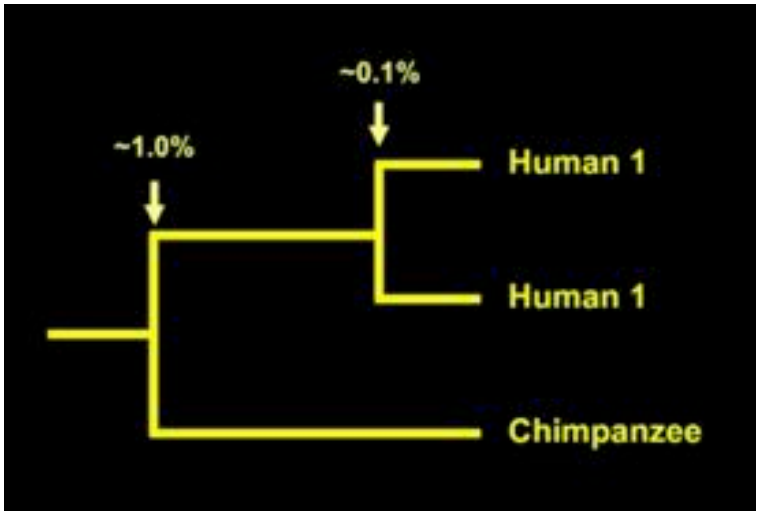
Our Origins

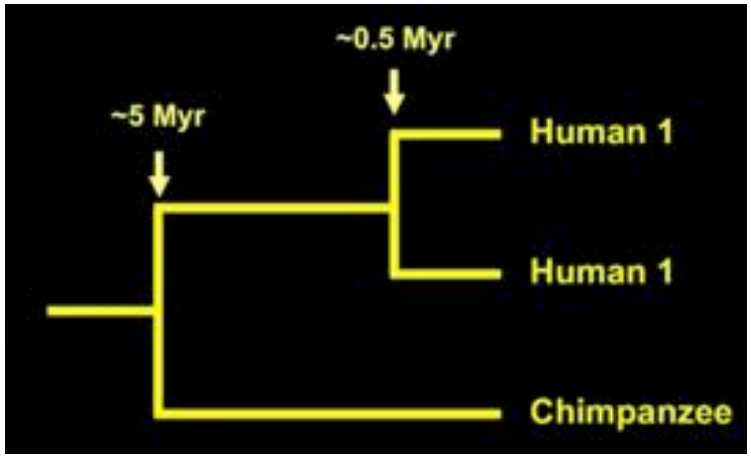

Nature timeline
view • discuss • edit
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-11
-12
-13
Life
Primordial
Land life
Multicellular life
photosynthesis
Single-celled life
water
Solar System
cosmic speed-up
Dark matter
Dark energy
cosmic expansion
Earliest light
Earliest gravity
Earliest humans
Cambrian explosion
Earliest sexual reproduction
Atmospheric oxygen
Earliest oxygen
Earliest life
Earliest Earth (–4.54)
Alpha Centauri forms
Milky Way Galaxy spiral arms form
Andromeda Galaxy forms
Omega Centauri forms
Earliest quasar
Earliest galaxy
Earliest universe (–13.80)
Axis scale: billions of years.
Also see: Human timeline and Life timeline


Life timeline
view • discuss • edit
0
-500
-1000
-1500
-2000
-2500
-3000
-3500
-4000
-4500
Quaternary
Karoo
Andean
Cryogenian
Huronian
Pongola
Phanerozoic
Proterozoic
Archean
Hadean
Flowers
Mammals
Dinosaurs
Land life
Multicellular life
Eukaryotes
photosynthesis
Single-celled life
water
Earliest humans
Cambrian explosion
Ediacara biota
Earliest sexual reproduction
Oxygen Crisis
Atmospheric oxygen
Earliest oxygen
LHB meteorites
Earliest life
Earliest water
Earliest Earth (–4540)
Axis scale: millions of years.
Orange labels: known ice ages.
Also see: Human timeline and Nature timeline


Human timeline
view • discuss • edit
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
Pleistocene
Pliocene
Miocene
Hominids
Homo sapiens
Neanderthal
Homo erectus
Homo habilis
Australopithecus
Ardipithecus
Human-like apes
Orrorin
Sahelanthropus
Ouranopithecus
Nakalipithecus
Modern humans
Earliest clothes
Earliest cooking
Earliest fire use
Earliest exit from Africa
Earliest stone tools
Earliest bipedal
Possibly bipedal
Earlier apes
Axis scale: millions of years.
Also see: Life timeline and Nature timeline

Mutation Rates and Evolutionary Time





Since mutation occur as a function of time we can use the number of mutation to age when different populations split

Interestingly, there is much more variability within Africa than outside of Africa despite the much smaller population

We see "African" alleles all around the world
- Zero SNPs occur exclusively in Africa
- Only 12 SNPs across the entire genome 'unique' to Africa (allowing 95% tolerance)
- We are all African (either currently living in Africa or recent exiles)!

Open question if/how early modern humans interacted with earlier hominid

DNA clues to our inner neanderthal
Svante Pääbo (2011). *TED Global.*
https://www.ted.com/talks/svante_paeaebo_dna_clues_to_our_inner_neanderthal

Sequencing ancient genomes

Janet Kelso
Max-Planck Institute

Homo neanderthalensis

•Proto-Neanderthals emerge around 600k years ago

•“True” Neanderthals emerge around 200k years ago

•Died out approximately 40,000 years ago

•Known for their robust physique

•Made advanced tools, probably had a language (the nature of which is debated and likely unknowable) and lived in complex social groups

Homo sapiens sapiens

- Apparently emerged from earlier hominids in Africa around 50k years ago
- Capable of amazing intellectual and social behaviors
- Mostly Harmless ☺


Engis
Neandertal
Spy
Arcy-sur-Cure
La Quina
Saint-Césaire
Le Moustier
La Ferrassie
Châtelperron
Steinheim
La Chapelle-aux-Saints
Moula
Figueira Brava
Zaferraya
Forbe's Quarry
Saccopastore
Guattari
Külna
Sipka
Tata
Erd
Krapina
Vindija
Sukhaya Mechetka
Molodova
Kiyik-Koba
Starosillya
Shanidar
Amud
Tabun
sites ayant livré des fossiles de Neandertaliens classiques
(les lignes de rivages et l'extension des glaciers correspondent à un maximum glaciaire)

Science
AAAS

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*

Science **328**, 710 (2010);
DOI: 10.1126/science.1188021

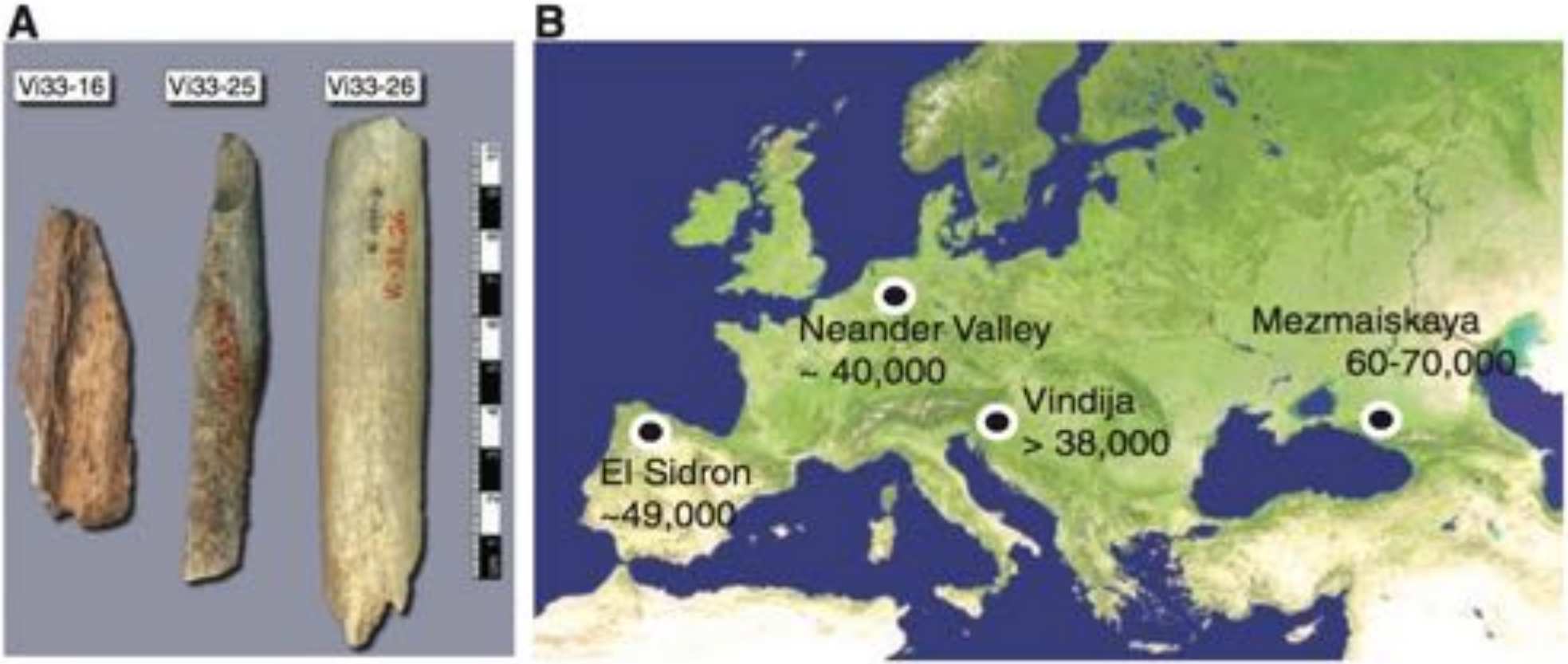
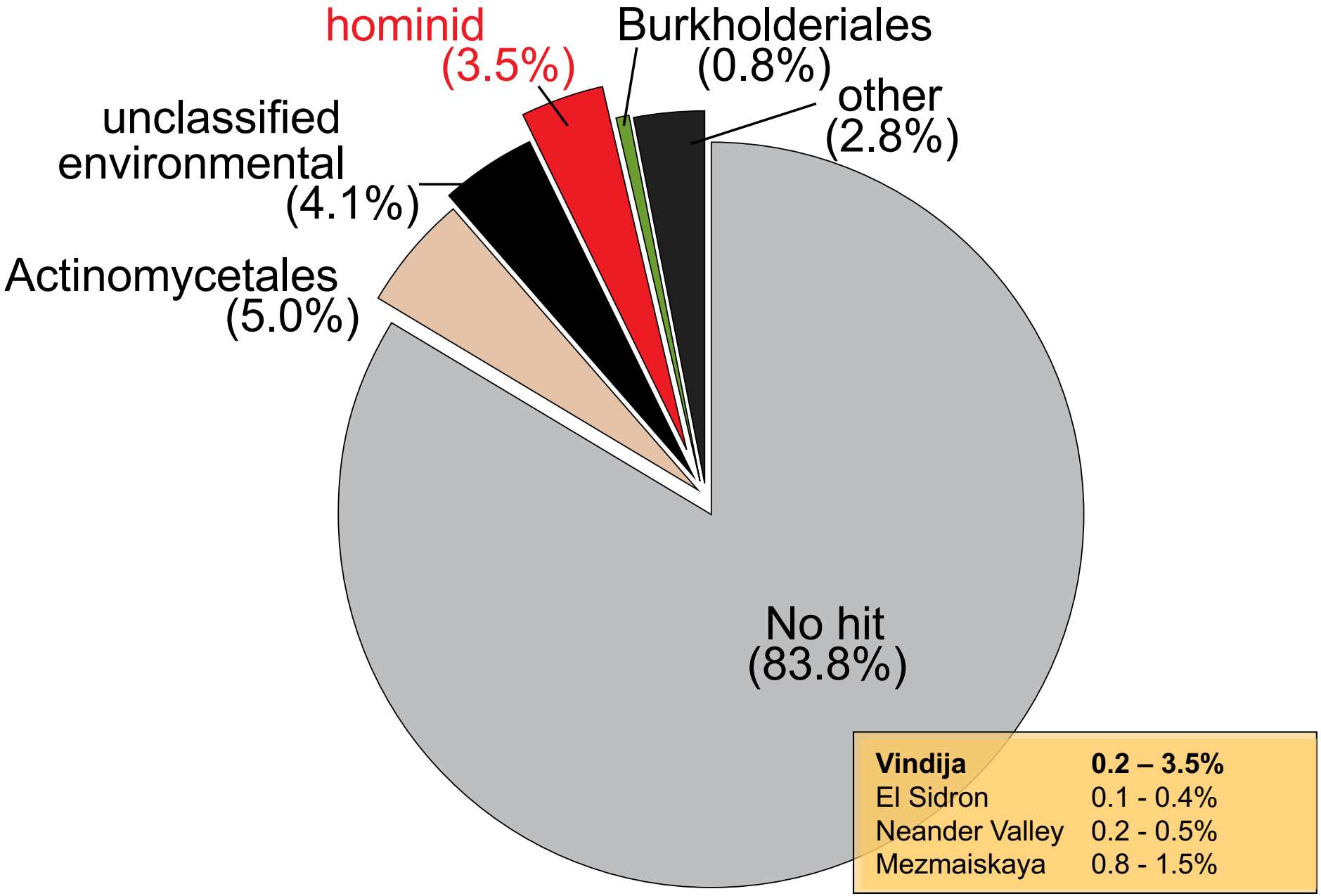


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

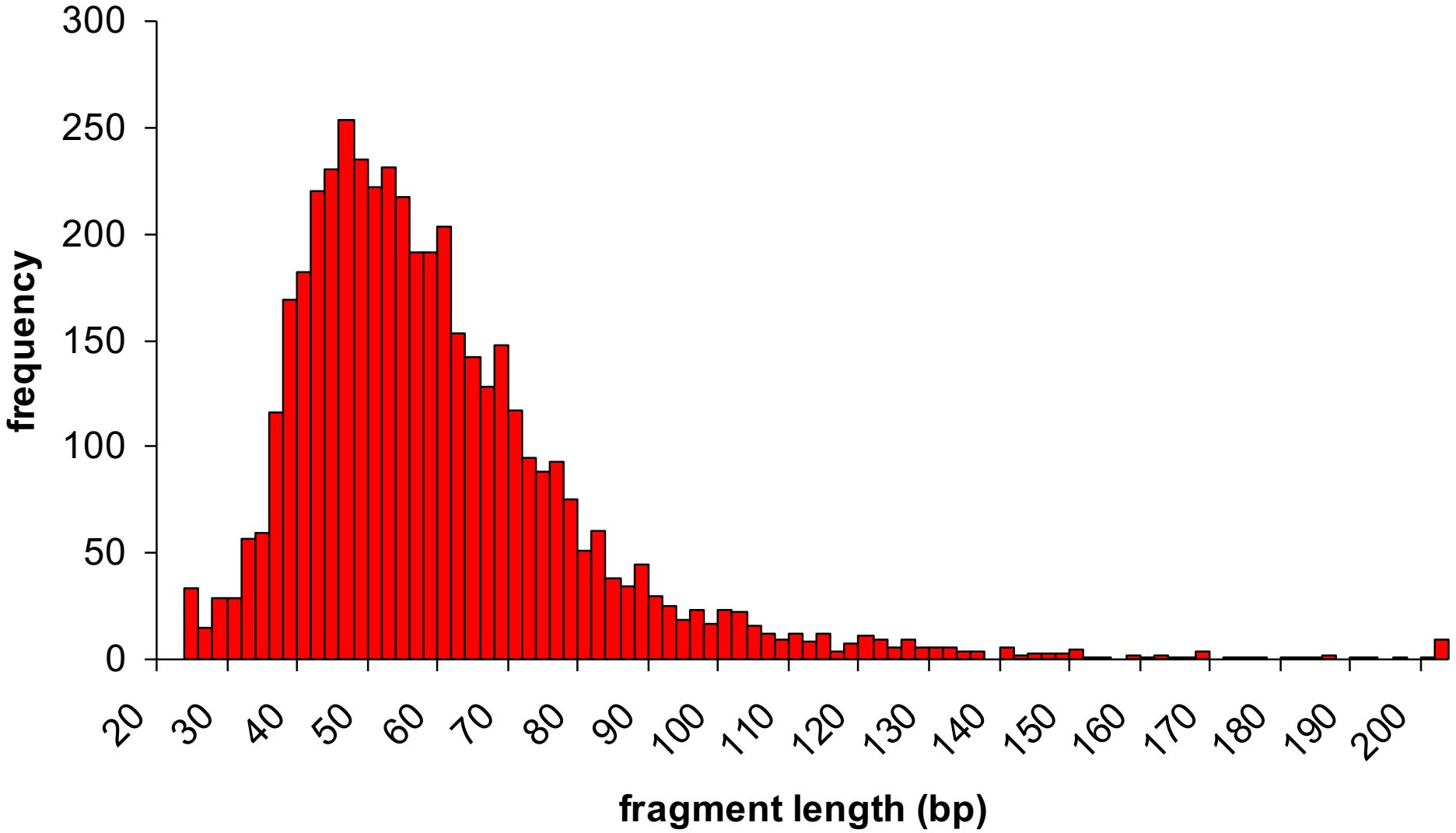
Extracting Ancient DNA

DNA is from mixed sources

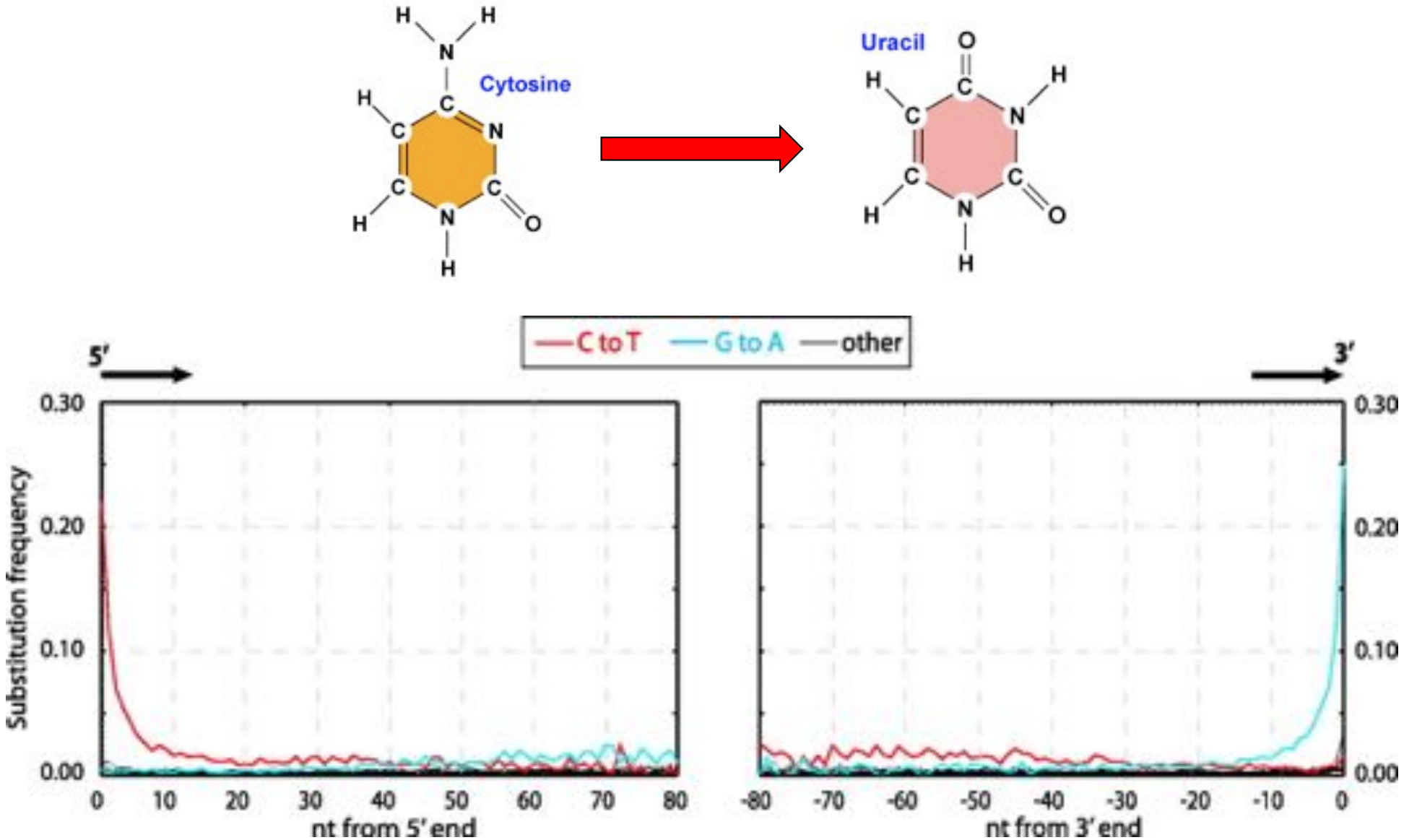


Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged



Briggs A W et al. PNAS 2007;104:14616-14621

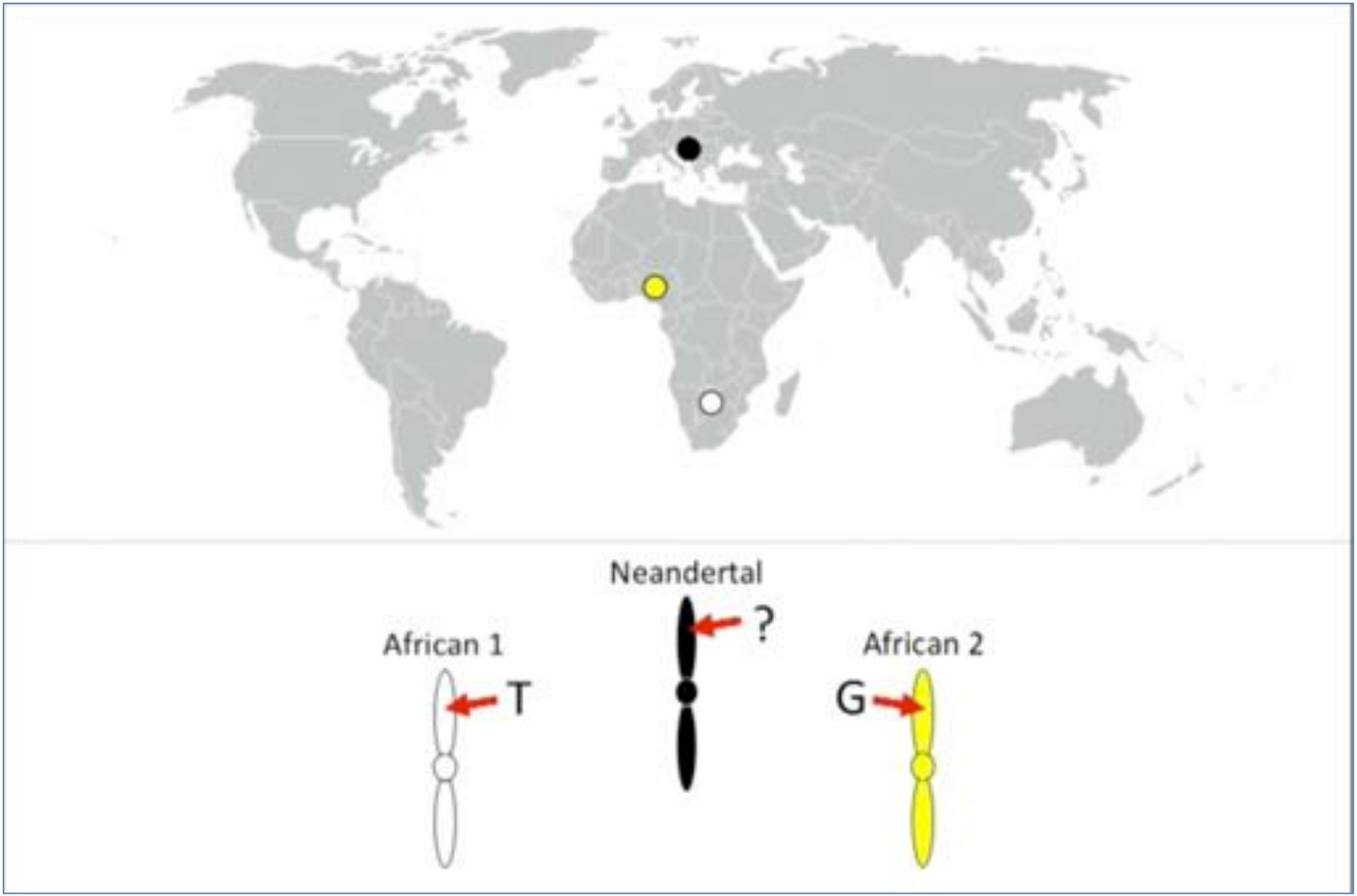
Green et al. 2010

Vindija	33.16	~1.2 Gb
	33.25	~1.3 Gb
	33.26	~1.5 Gb
El Sidron (1253)		~2.2 Mb
Feldhofer 1		~2.2 Mb
Mezmaiskaya 1		~56.4 Mb

~35 Illumina flow cells

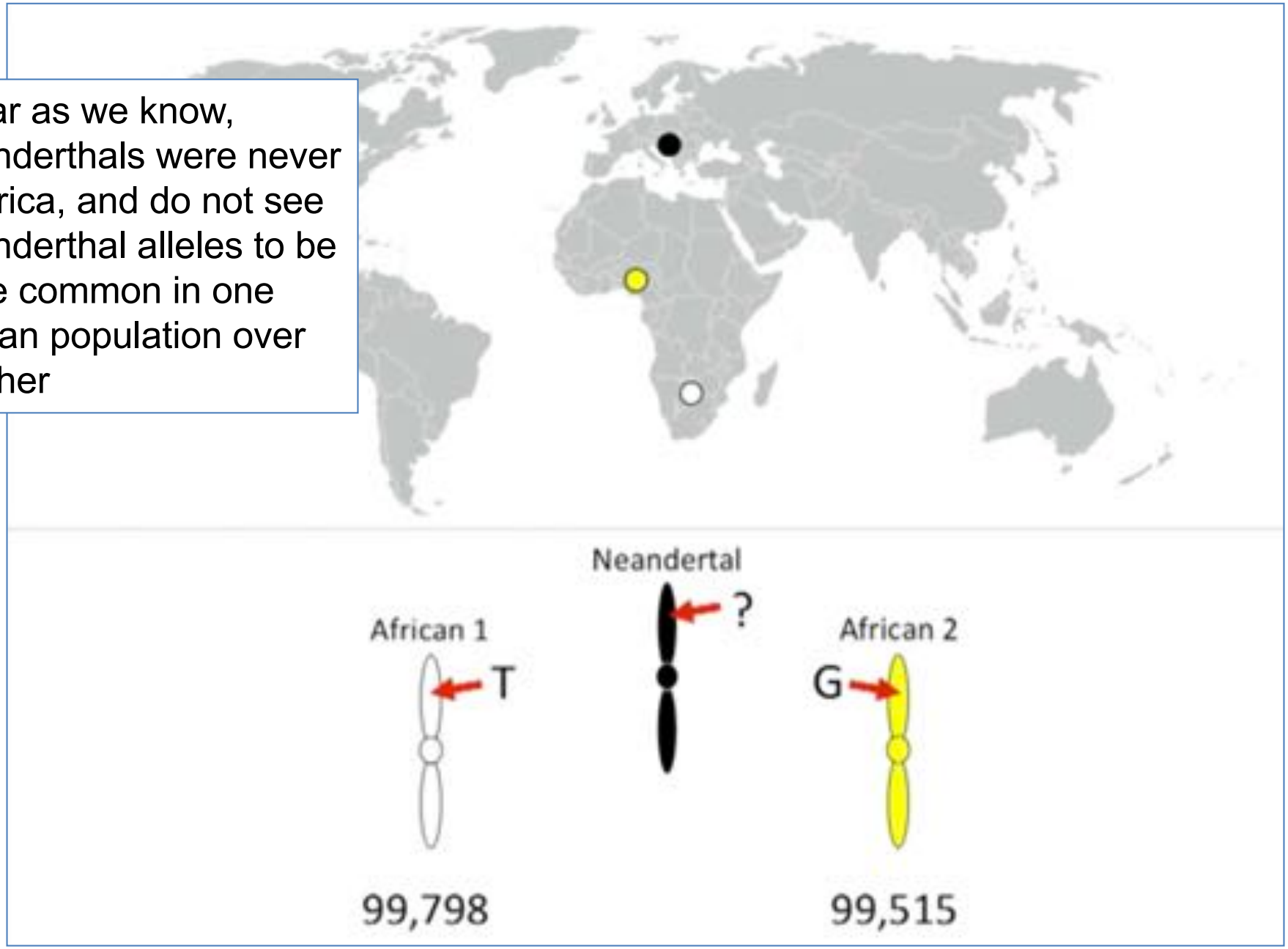
Genome coverage ~1.3 X

Did we mix?



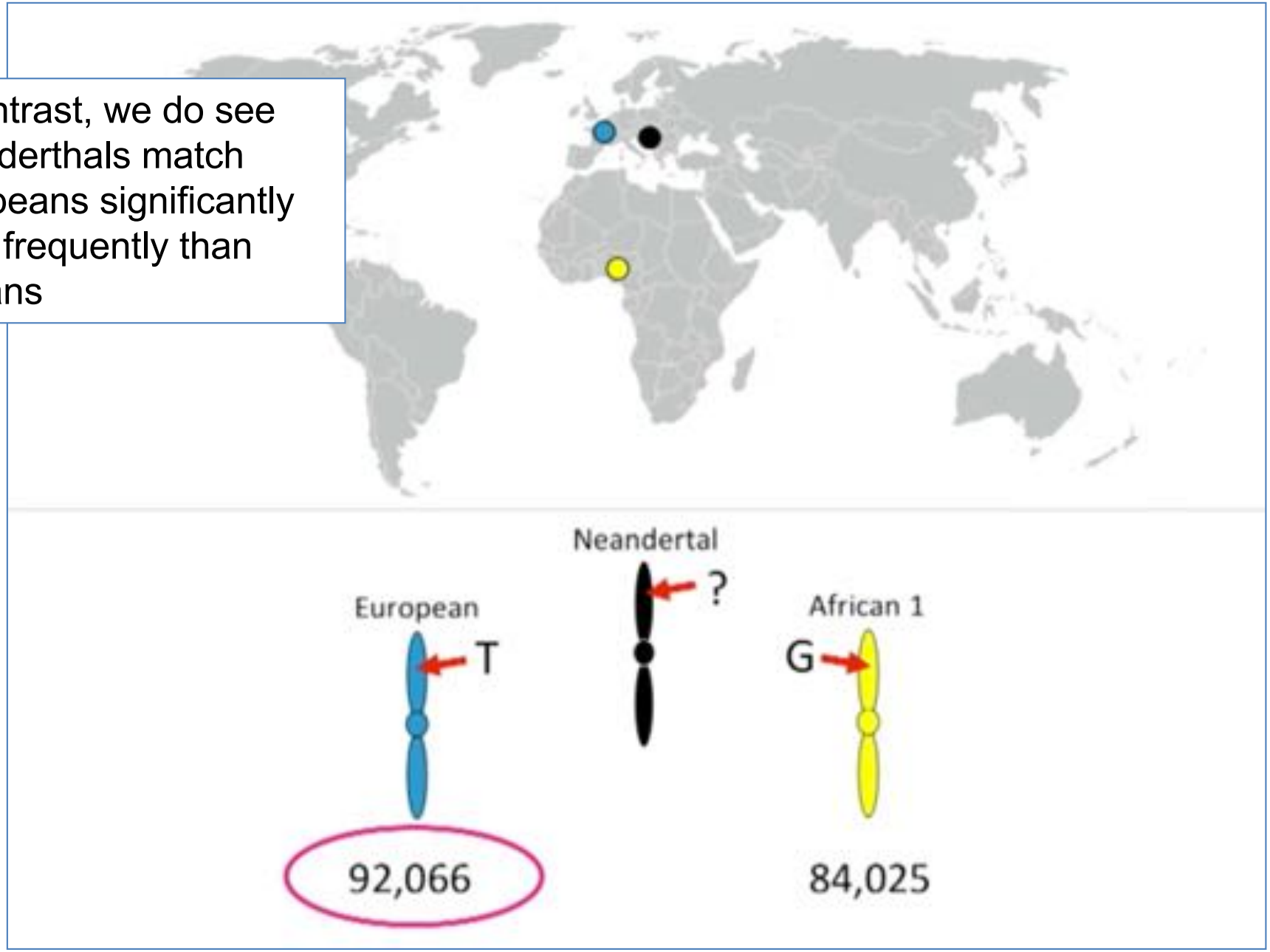
Did we mix?

As far as we know, Neanderthals were never in Africa, and do not see Neanderthal alleles to be more common in one African population over another



Did we mix?

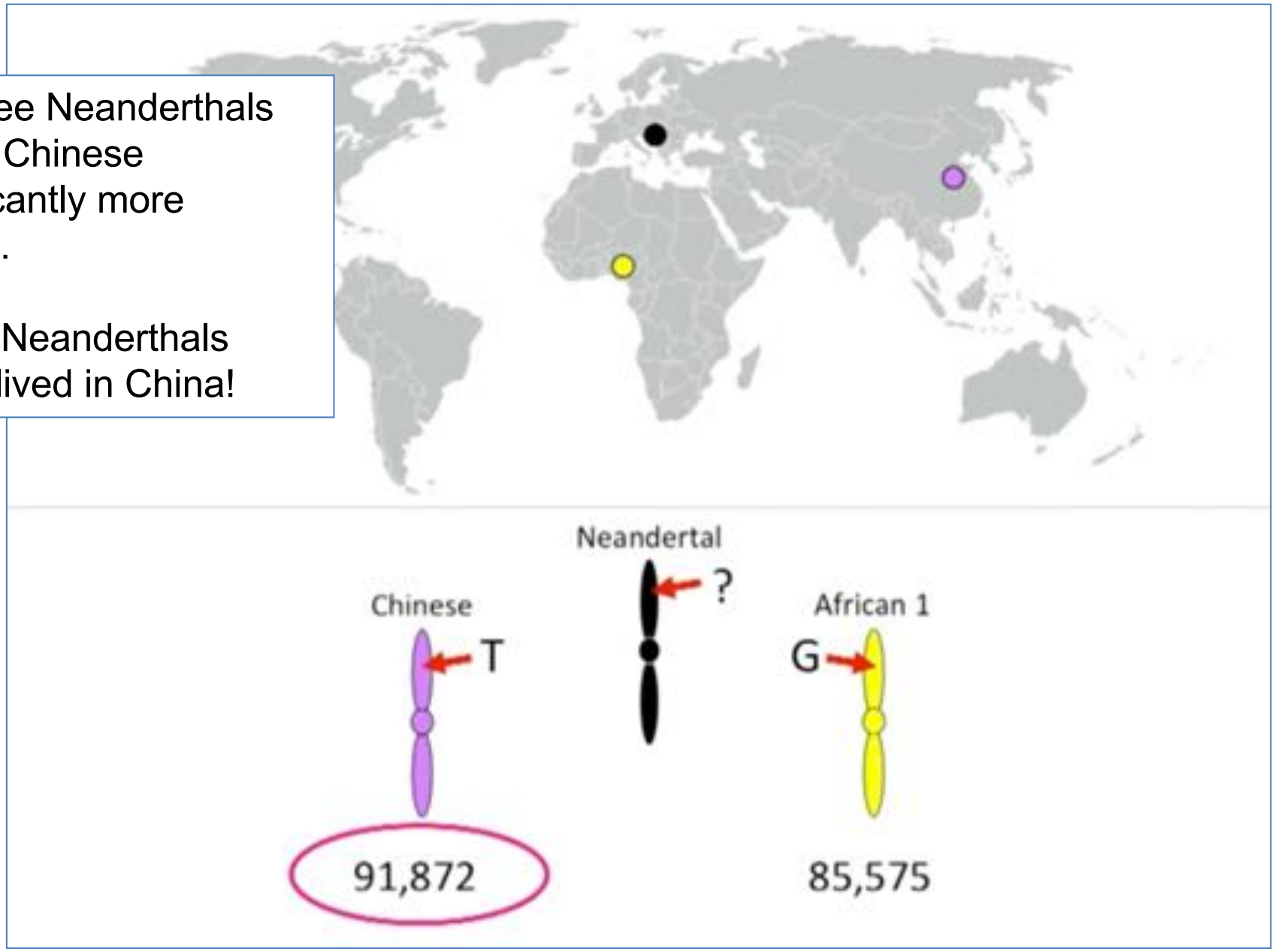
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



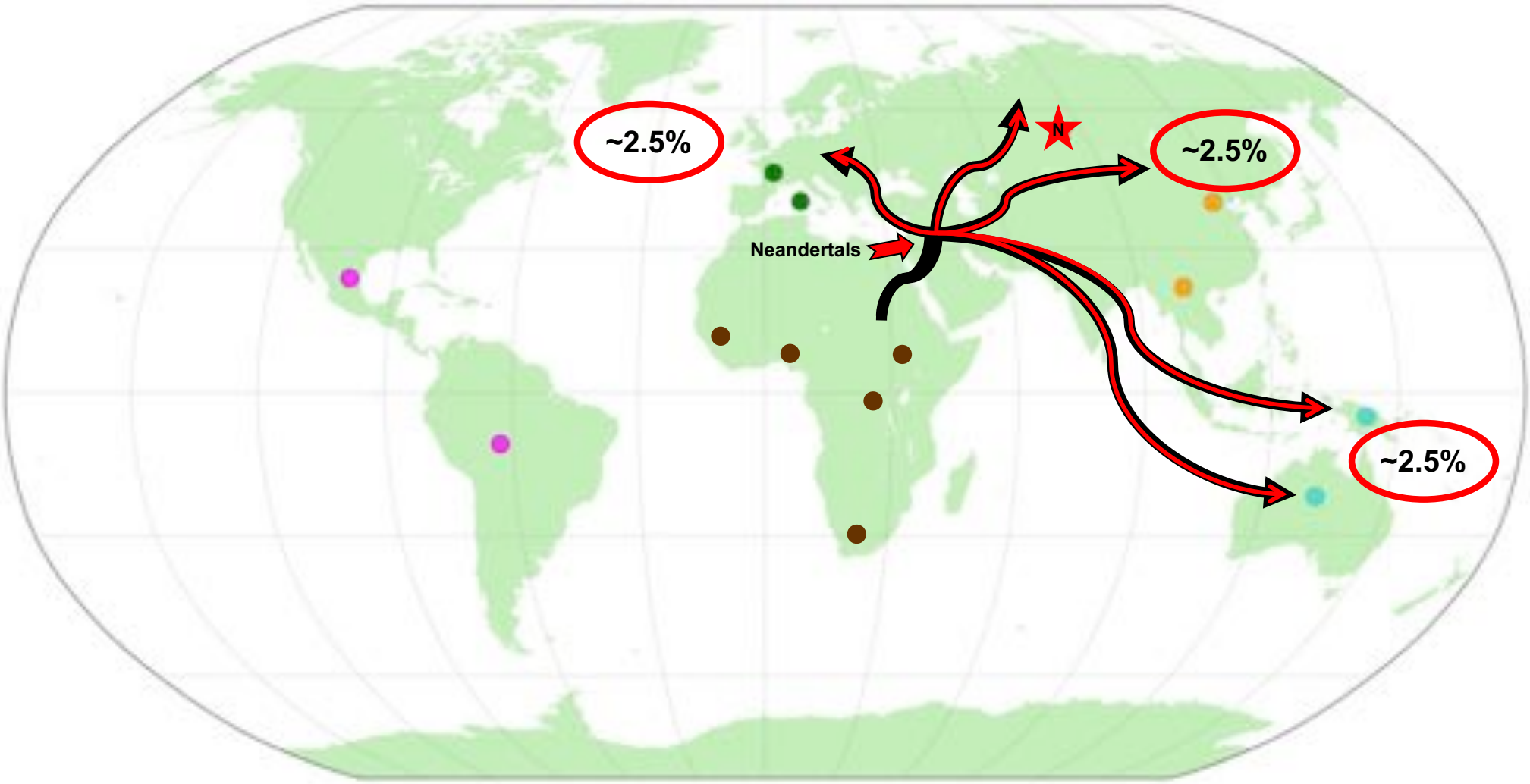
Did we mix?

Also see Neanderthals match Chinese significantly more often…

… but Neanderthals never lived in China!



Neanderthal Interbreeding

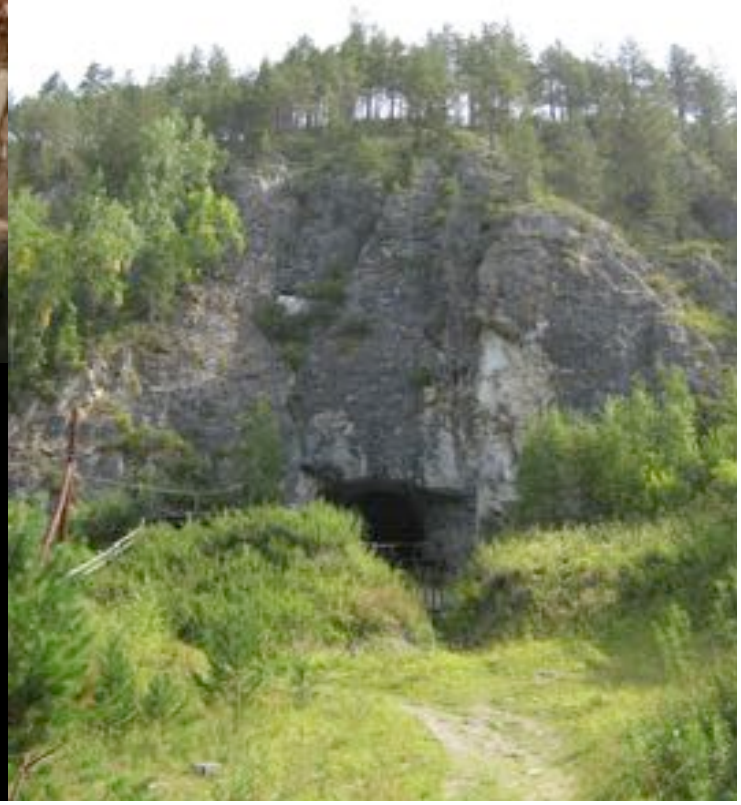


As modern humans migrated out of Africa, they apparently interbred with Neanderthal's so we see their alleles across the rest of the world and carry about 2.5% of their genome with us!

What about other ancient hominids?

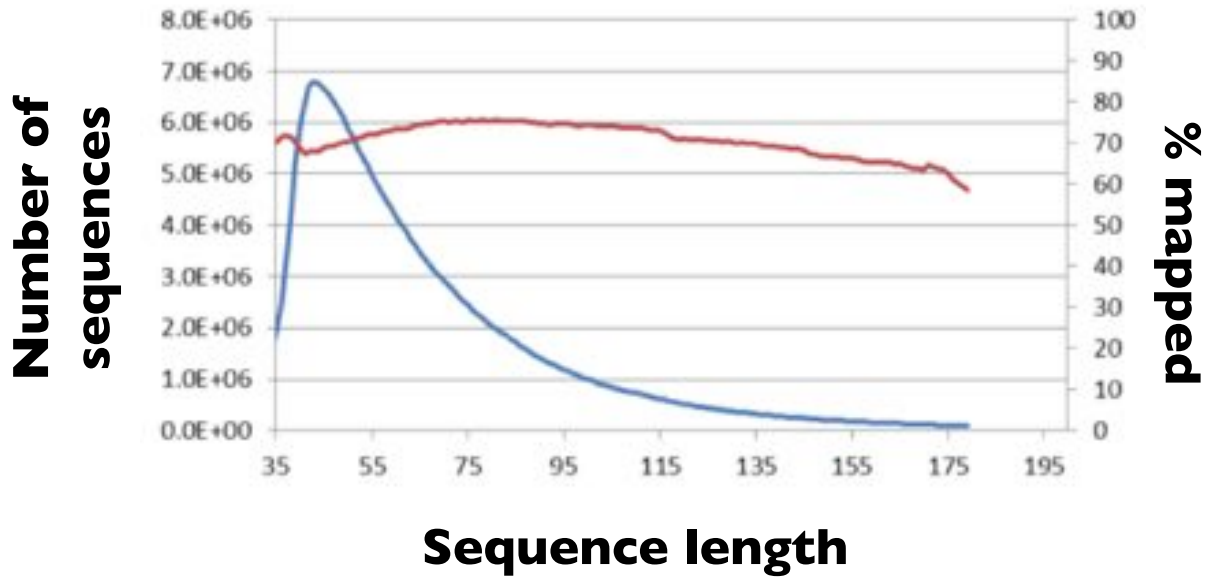
Denisova cave
Altai mountains
Russia

Academician A.P. Derevianko



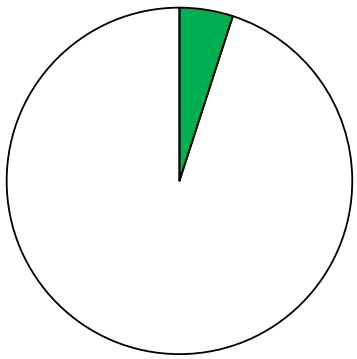

UNITED STATES OF AMERICA
E PLURIBUS UNUM
ONE CENT

Extraordinary preservation

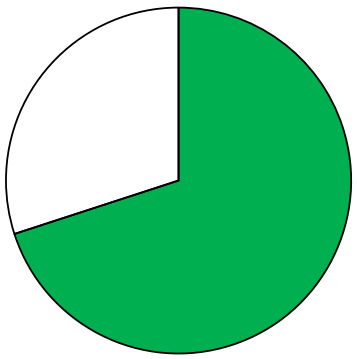


5% endogenous DNA

Best Neandertal bone

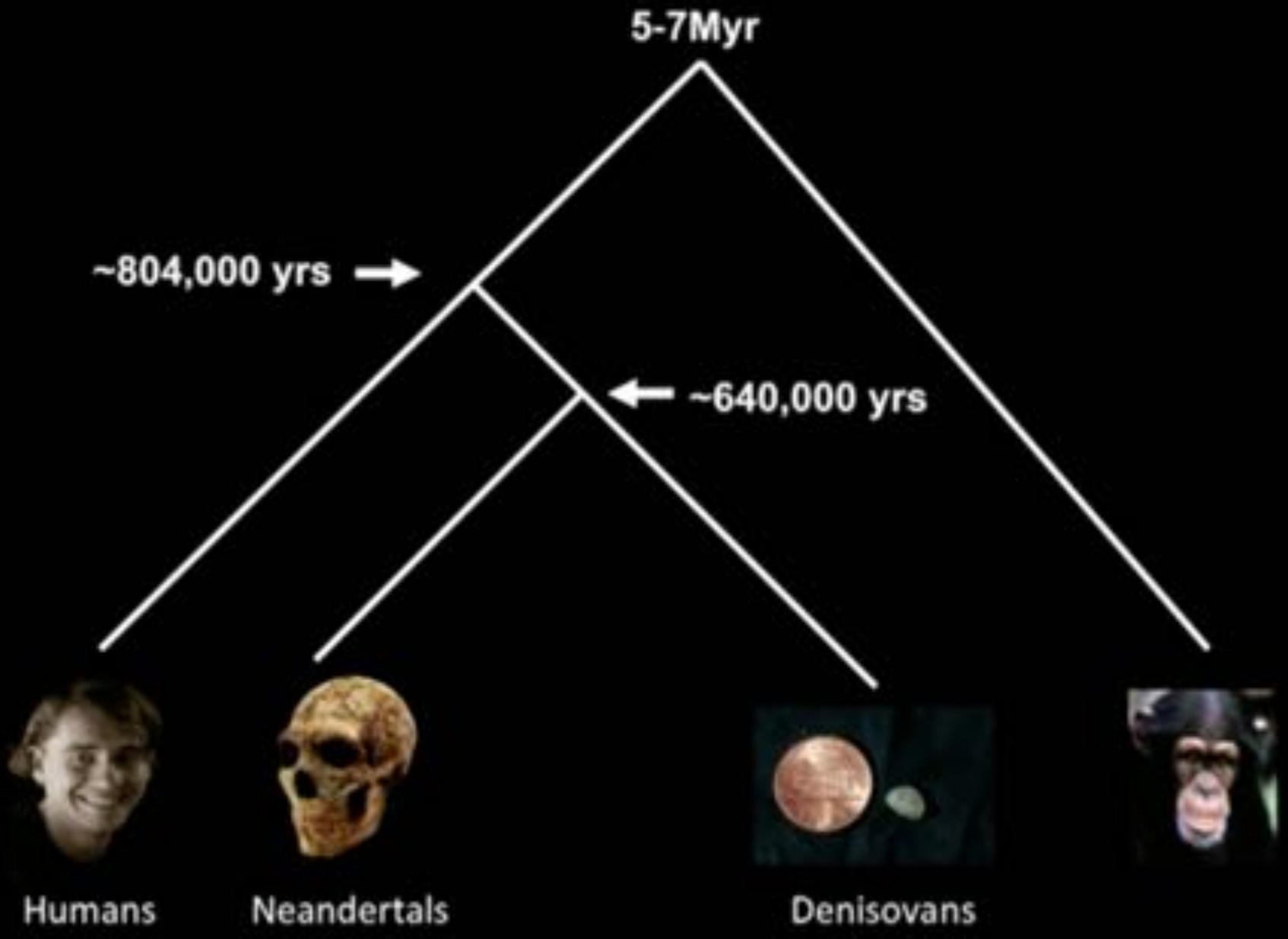


>70% endogenous DNA



Denisova bone

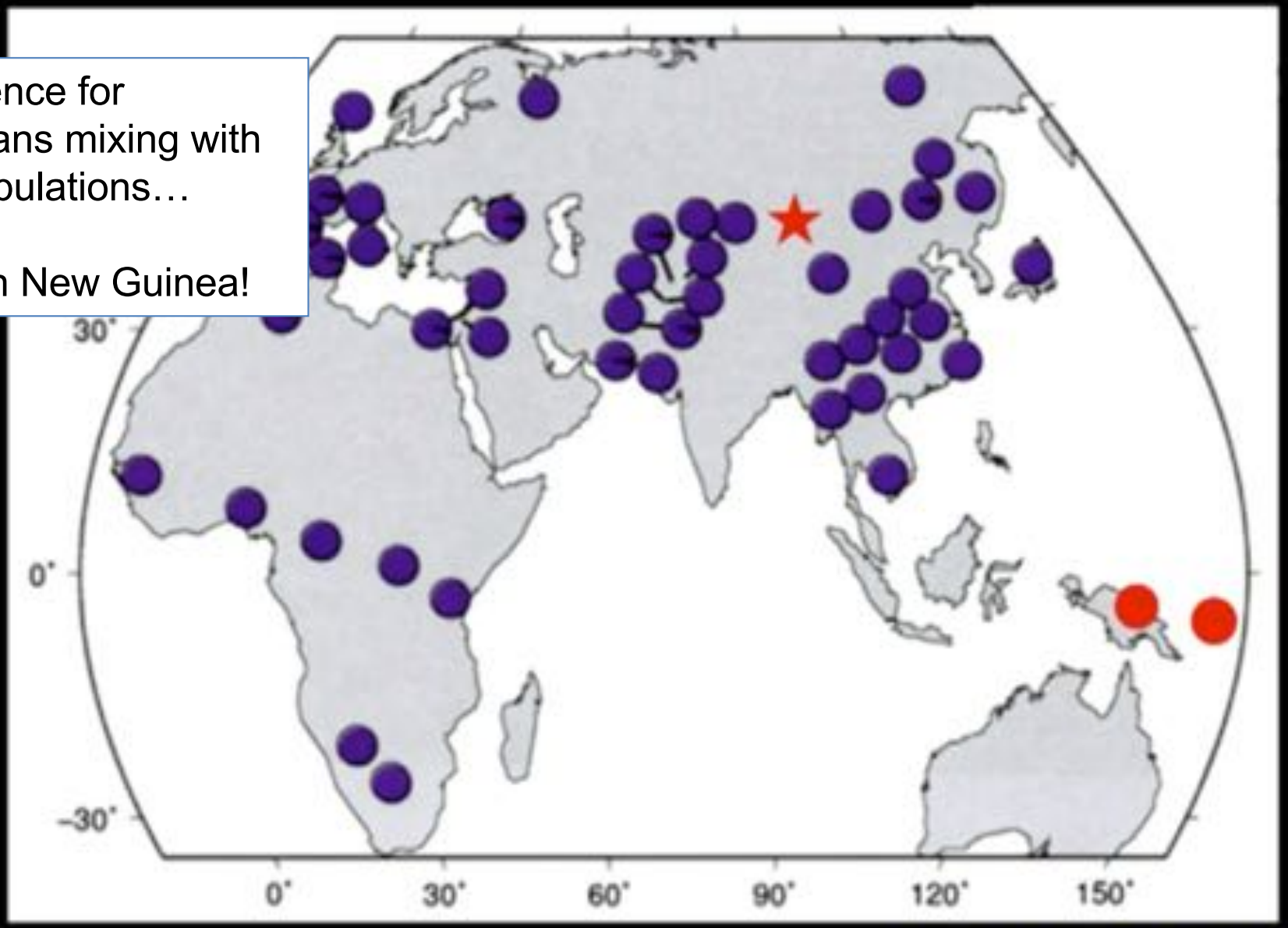
Denisovans & Neandertals



Did we mix?

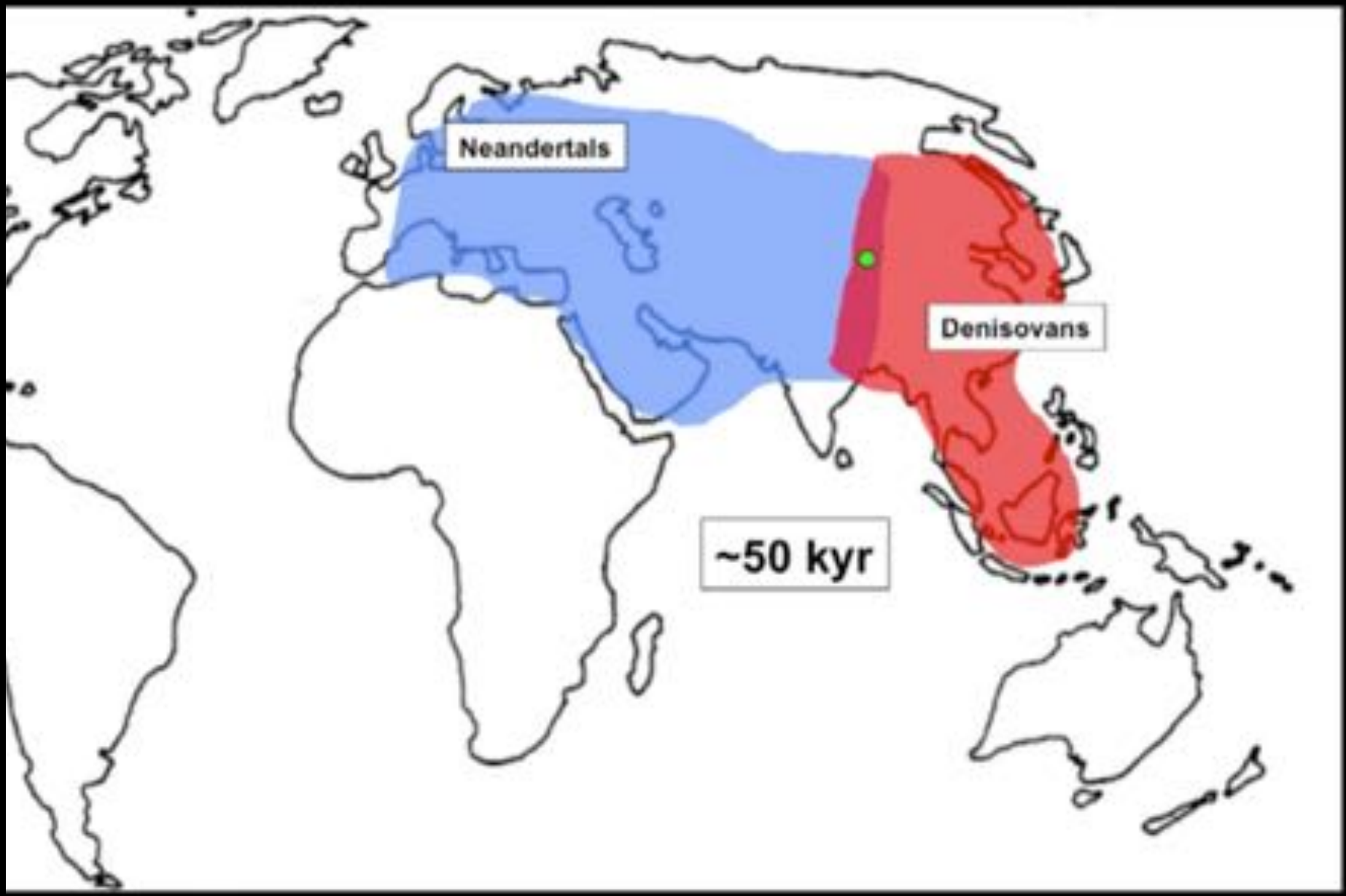
No evidence for Denisovans mixing with other populations…

Except in New Guinea!

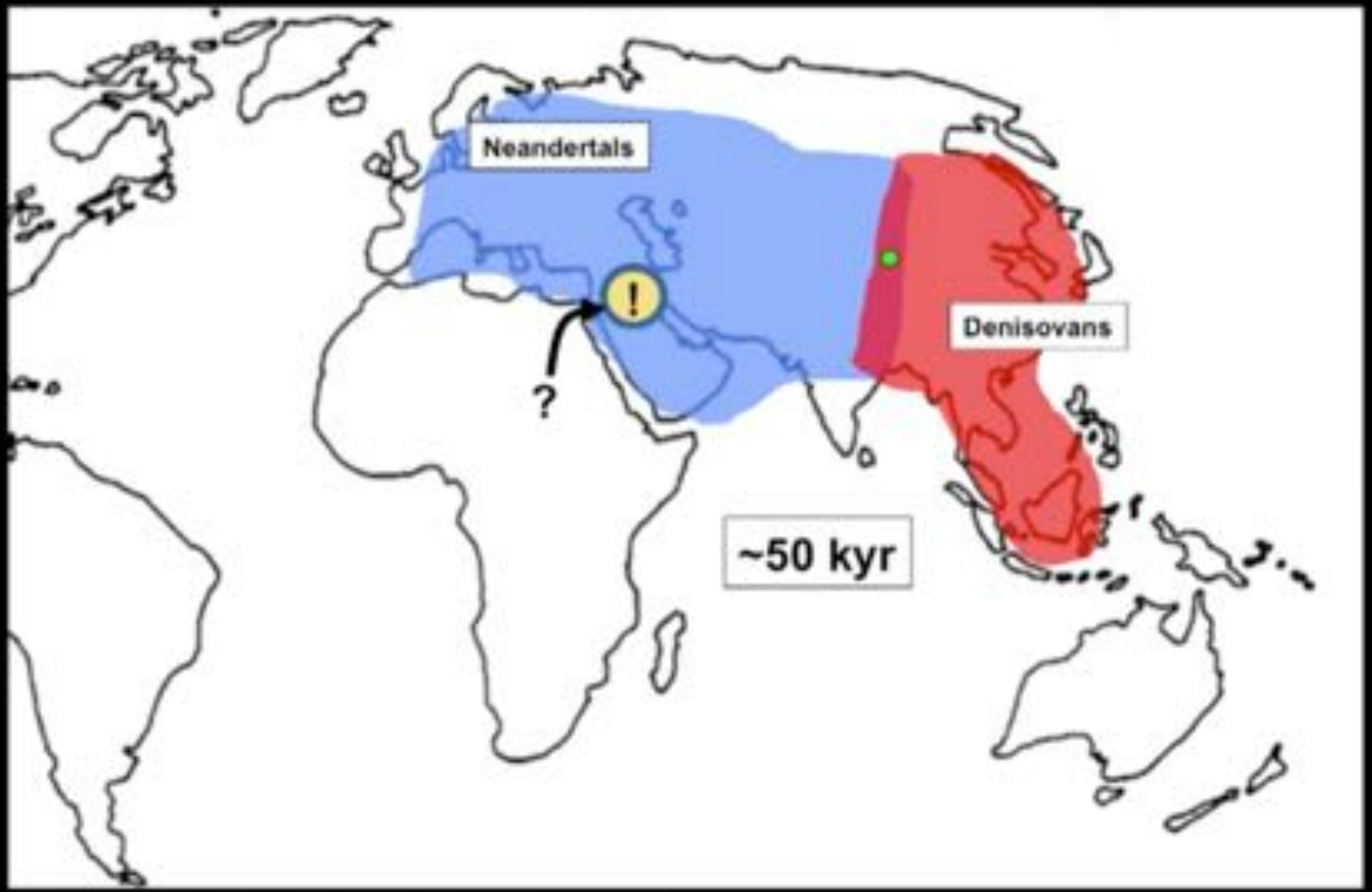


Map after Pickrell et al., 2009

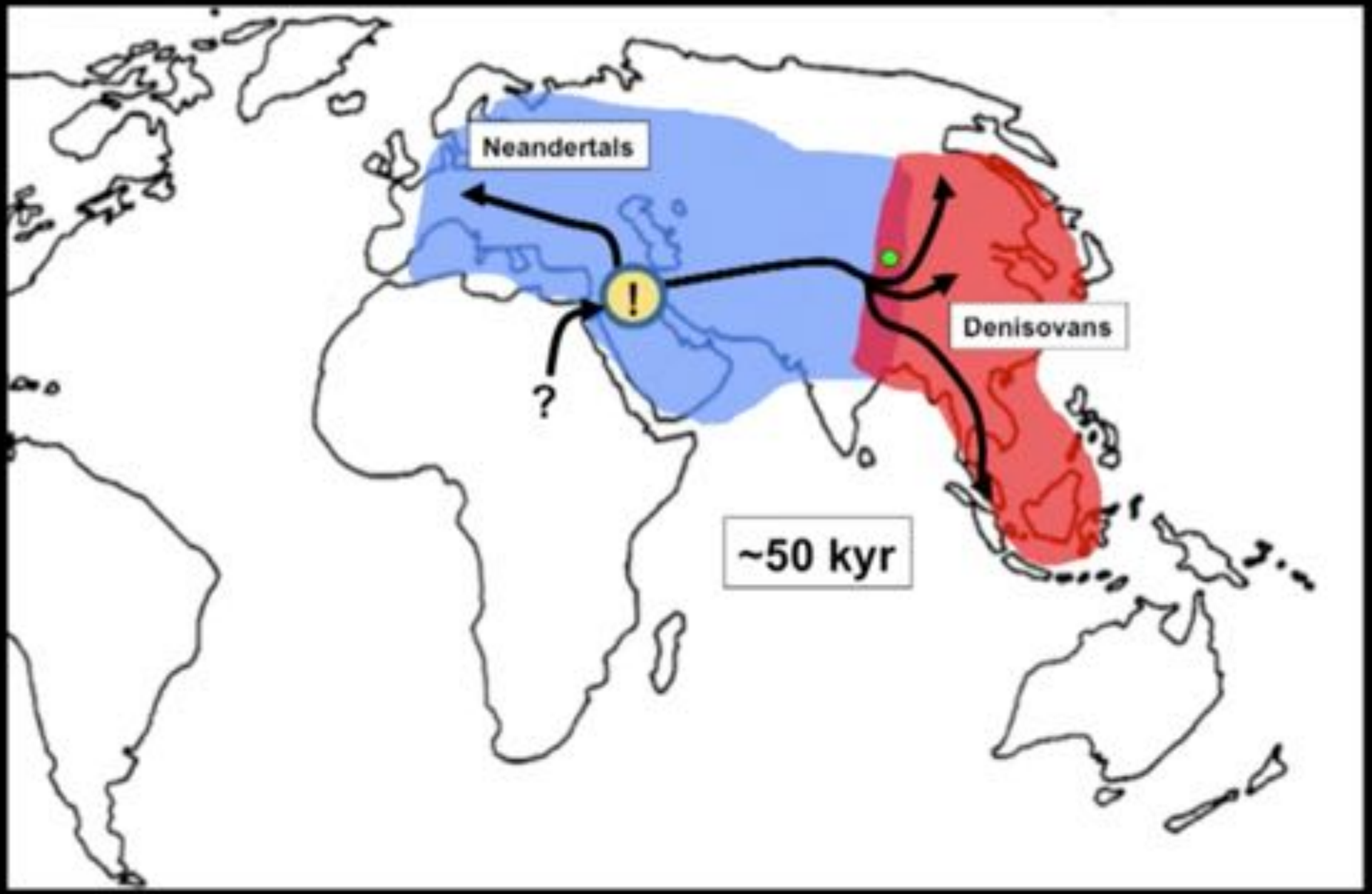
Timeline of ancient hominids



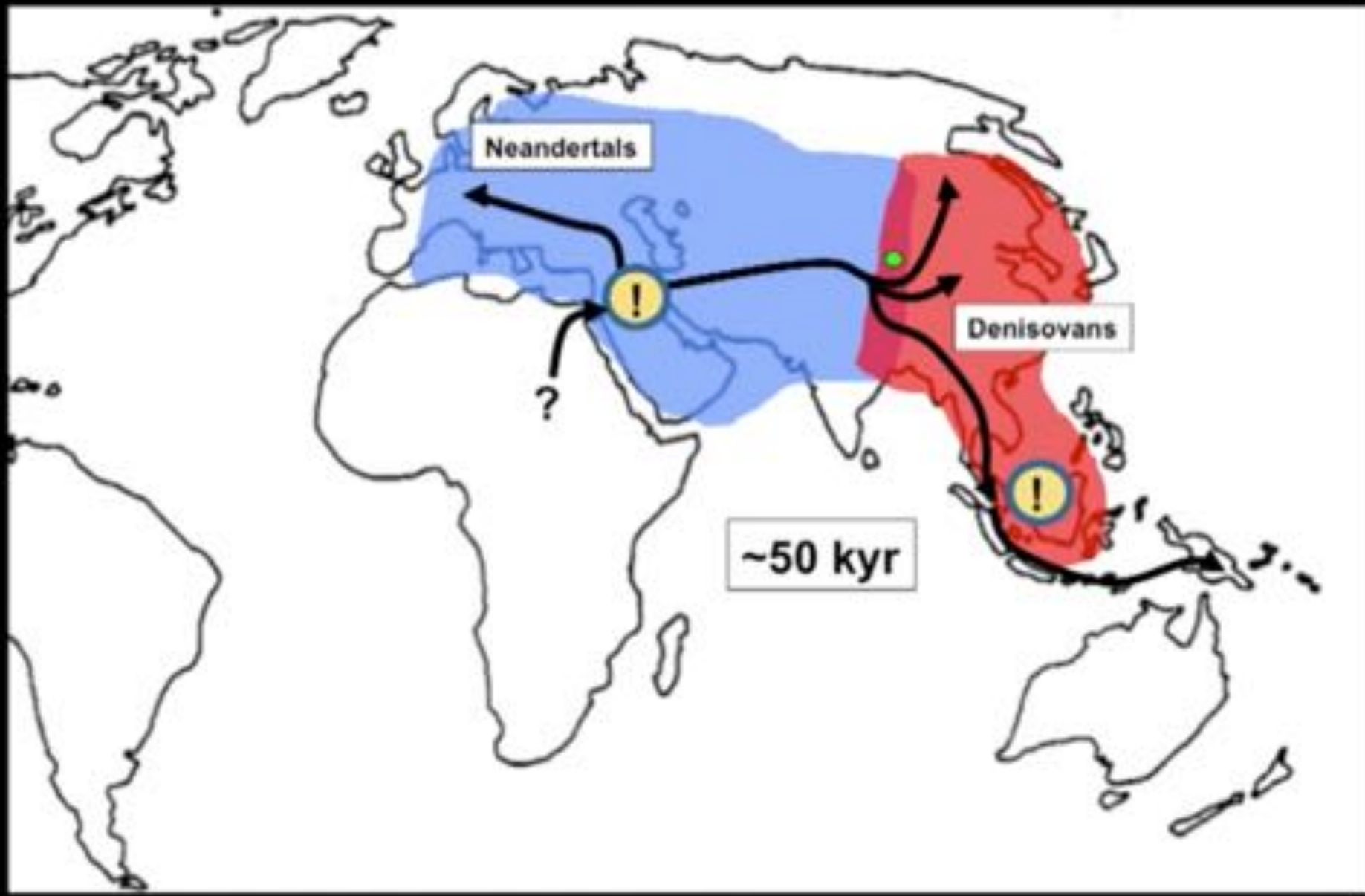
Timeline of ancient hominids



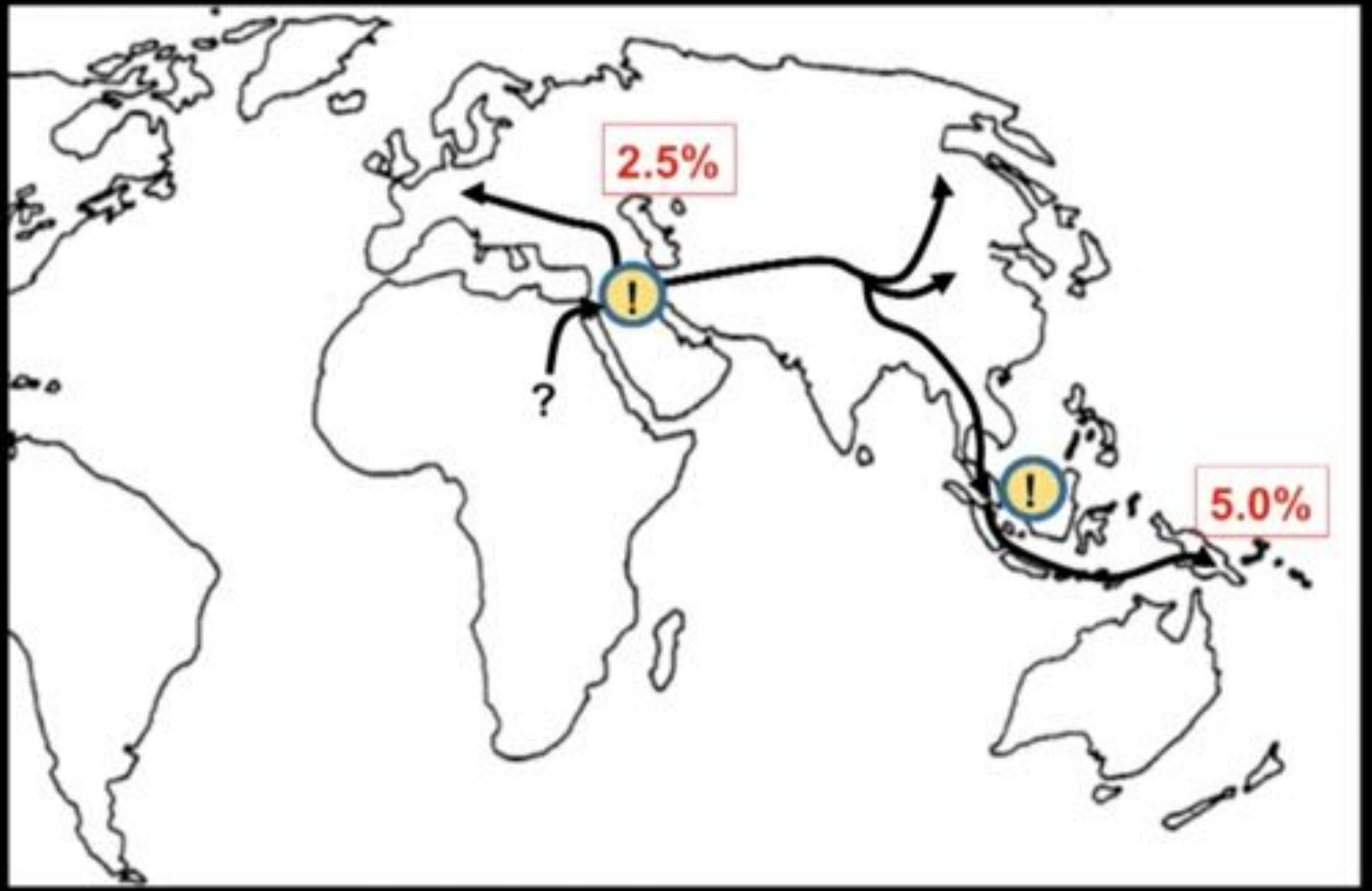
Timeline of ancient hominids



Timeline of ancient hominids



Timeline of ancient hominids



We have always mixed!

Science

REPORTS

Cite as: B. Vernot *et al.*, *Science* 10.1126/science.aad9416 (2016).

Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals

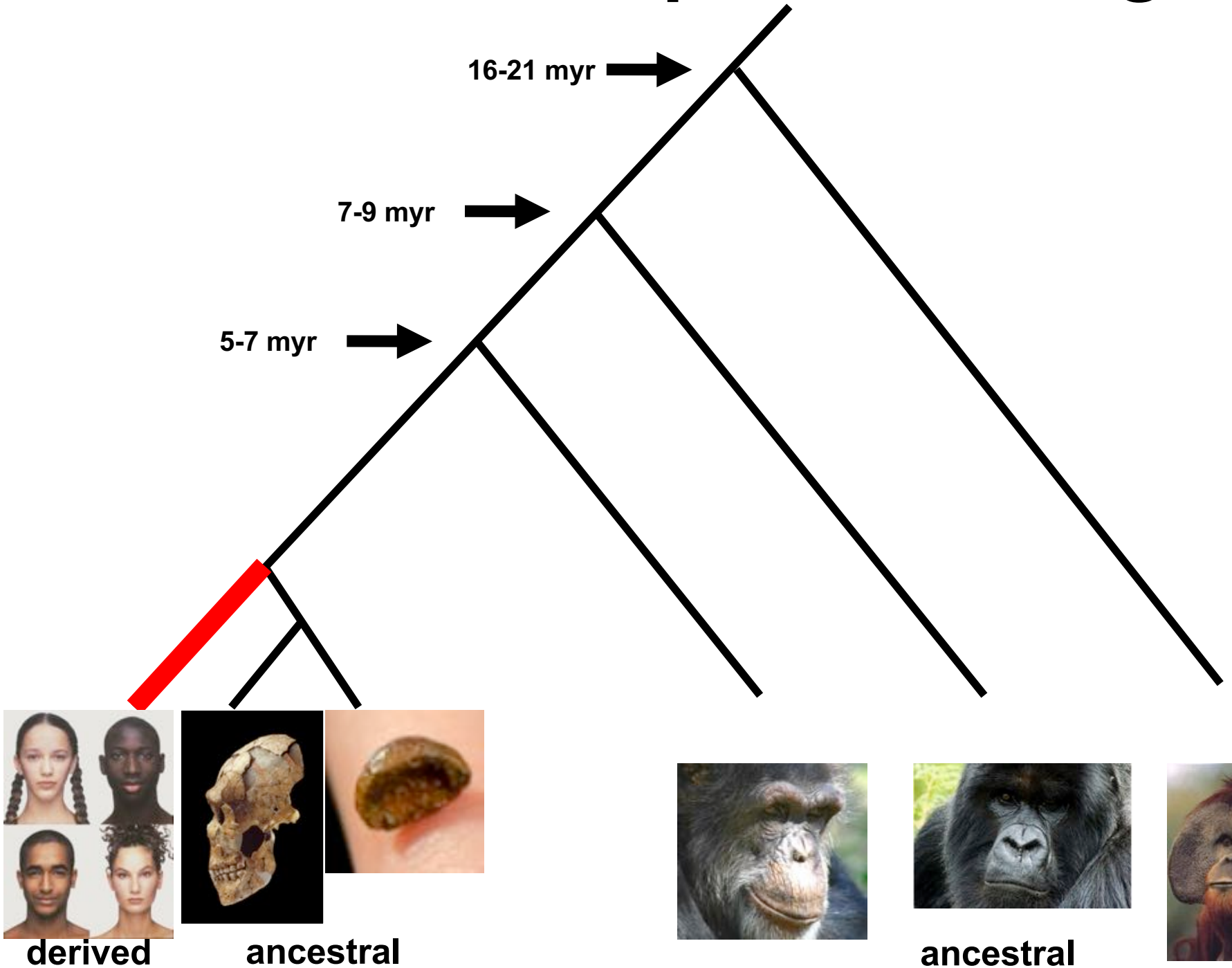
Benjamin Vernot,[1] Serena Tucci,[1,2] Janet Kelso,[3] Joshua G. Schraiber,[1] Aaron B. Wolf,[1] Rachel M. Gittelman,[1] Michael Dannemann,[3] Steffi Grote,[3] Rajiv C. McCoy,[1] Heather Norton,[4] Laura B. Scheinfeldt,[5] David A. Merriwether,[6] George Koki,[7] Jonathan S. Friedlaender,[8] Jon Wakefield,[9] Svante Pääbo,[3]* Joshua M. Akey[1]*

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Department of Life Sciences and Biotechnology, University of Ferrara, Italy. [3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [4]Department of Anthropology, University of Cincinnati, Cincinnati, OH, USA. [5]Coriell Institute for Medical Research, Camden, NJ, USA. [6]Department of Anthropology, Binghamton University, Binghamton, NY, USA. [7]Institute for Medical Research, Goroka, Eastern Highlands Province, Papua New Guinea. [8]Department of Anthropology, Temple University, Philadelphia PA, USA. [9]Department of Statistics, University of Washington, Seattle, Washington, USA.

*Corresponding author. E-mail: paabo@eva.mpg.de (S.P.); akey@uw.edu (J.M.A.)

Although Neandertal sequences that persist in the genomes of modern humans have been identified in Eurasians, comparable studies in people whose ancestors hybridized with both Neandertals and Denisovans are lacking. We developed an approach to identify DNA inherited from multiple archaic hominin ancestors and applied it to whole-genome sequences from 1523 geographically diverse individuals, including 35 new Island Melanesian genomes. In aggregate, we recovered 1.34 Gb and 303 Mb of the Neandertal and Denisovan genome, respectively. We leverage these maps of archaic sequence to show that Neandertal admixture occurred multiple times in different non-African populations, characterize genomic regions that are significantly depleted of archaic sequence, and identify signatures of adaptive introgression.

Modern human-specific changes



Recipe for a modern human

109,295 single nucleotide changes (SNCs)

7,944 insertions and deletions

Changes in protein coding genes

277 cause fixed amino acid substitutions

87 affect splice sites

Changes in Non-coding & regulatory sequences

26 affect well-defined motifs inside regulatory regions

Enrichment analysis

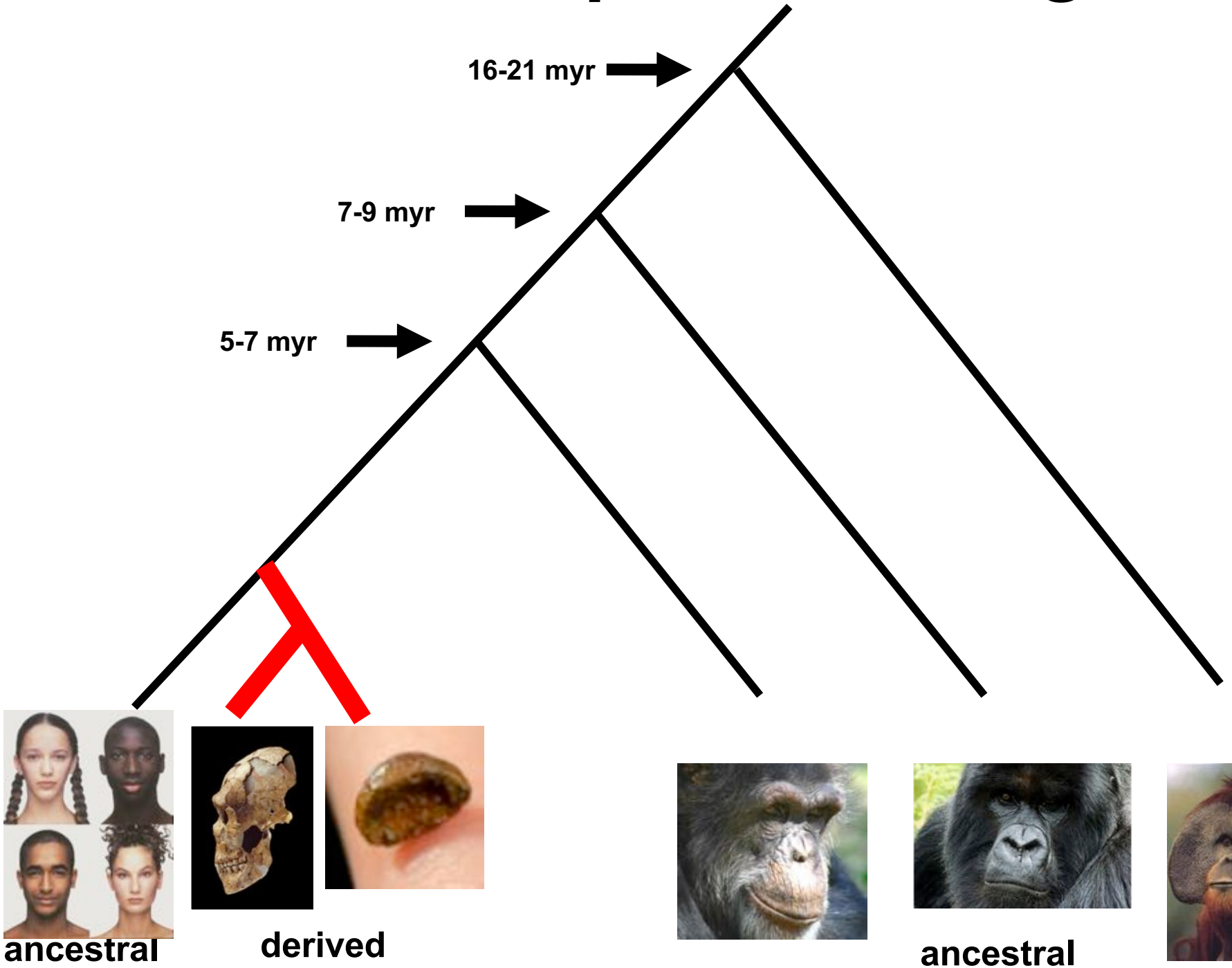
skin pigmentation

skeletal morphologies (limb length, digit development)

morphologies of the larynx and the epiglottis

Nonsynonymous	None	- Giant melanosomes in melanocytes (p-6.77e-6; FWER=0.091;
Splice sites		
3' UTR	None	- 1-3 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - 1-5 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Aplasia/Hypoplasia of the distal phalanx of the thumb (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Bifid or hypoplastic epiglottis (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Central polydactyly (feet) (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Distal urethral duplication (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Dysplastic distal thumb phalanges with a central hole (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Laryngeal cleft (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Midline facial capillary hemangioma (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Preductal coarctation of the aorta (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Radial head subluxation (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Short distal phalanx of the thumb (p=1.34288e-05; FWER=0.538; FDR=0.0887928)

Neandertal-specific changes



Enrichment analysis

Nonsynonymous	None	- Abnormality of the thumb (p=3.01e-5; FWER=0.025; FDR=0.02) - Aplasia/Hypoplasia of the thumb (p=6.31-5; FWER=0.054; FDR=0.024) - Facial cleft (p=0.0004; FWER=0.36; FDR=0.098) - Wide pubic symphysis (p=0.0004; FWER=0.36; FDR=0.098) - Abnormality of the frontal hairline (p=0.00042; FWER=0.39; FDR=0.096) - Abnormality of the finger (p=0.0005; FWER=0.44; FDR=0.08) - Brachydactyly syndrome (p=0.00062; FWER=0.48; FDR=0.088)

Skeletal and hair morphology

Protein	Ensembl ID	Protein position	Ancestral amino acid	Derived amino acid	Description
ABCA12	ENSP00000272895	199	W	C	ATP-binding cassette, sub-family A (ABC1)
FRAS1	ENSP00000264895	209	P	S	Fraser syndrome 1
GLI3	ENSP00000379258	1537	R	C	GLI family zinc finger 3
LAMB3	ENSP00000355997	926	A	D	Laminin, beta 3
MOGS	ENSP00000233616	495	R	Q	Mannosyl-oligosaccharide glucosidase

FOXP2 Analysis

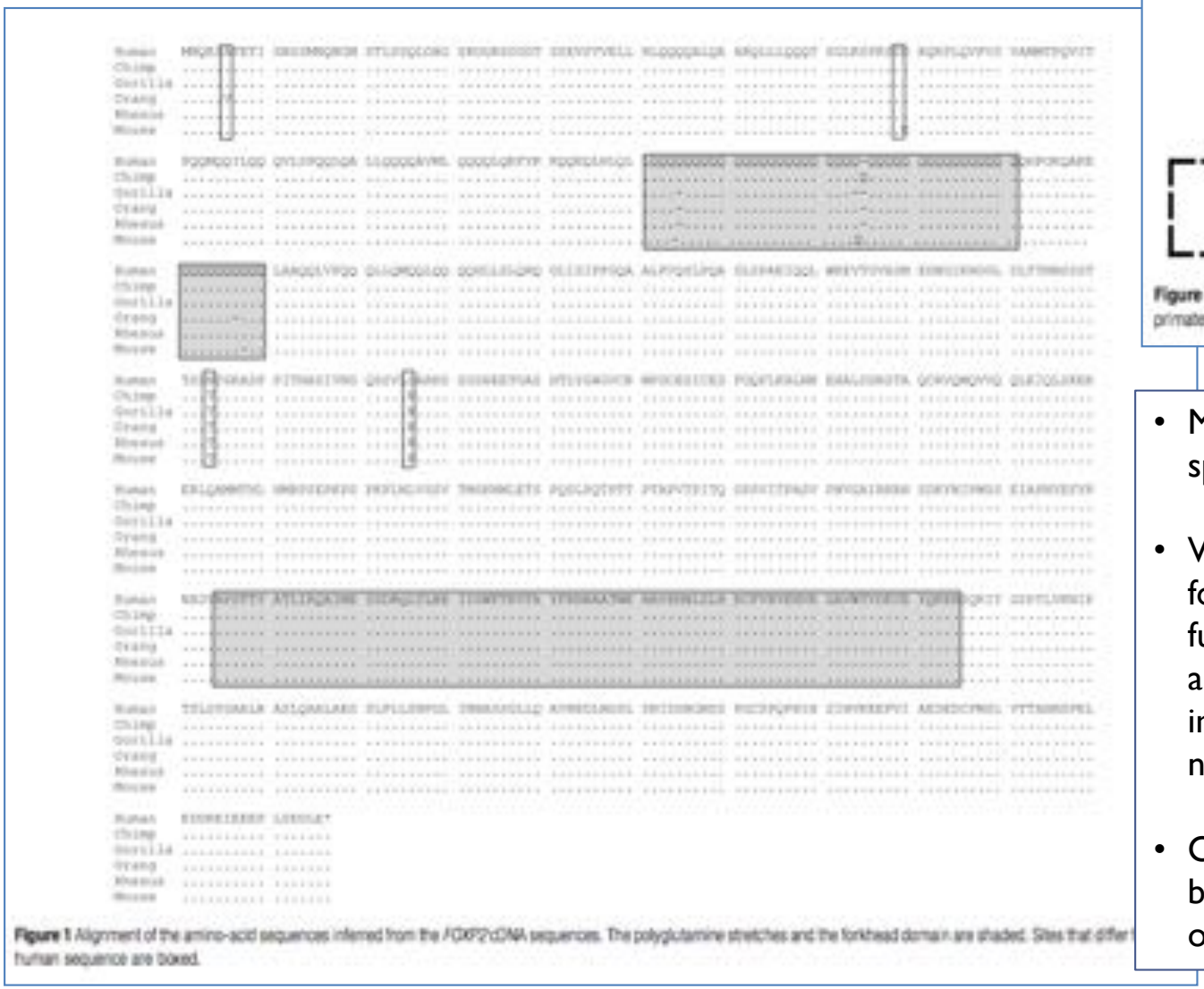


Figure 1 Alignment of the amino-acid sequences inferred from the FOXP2 cDNA sequences. The polyglutamine stretches and the forkhead domain are shaded. Sites that differ from human sequence are boxed.

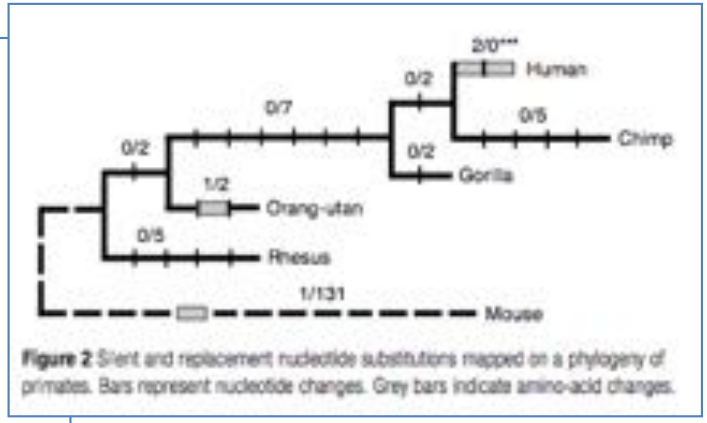


Figure 2 Silent and replacement nucleotide substitutions mapped on a phylogeny of primates. Bars represent nucleotide changes. Grey bars indicate amino-acid changes.

- Mutations of FOXP2 cause a severe speech and language disorder in people
- Versions of FOXP2 exist in similar forms in distantly related vertebrates; functional studies of the gene in mice and in songbirds indicate that it is important for modulating plasticity of neural circuits.
- Outside the brain FOXP2 has also been implicated in development of other tissues such as the lung and gut.

Molecular evolution of FOXP2, a gene involved in speech and language

Enard *et al* (2002) *Nature*. doi:10.1038/nature01025