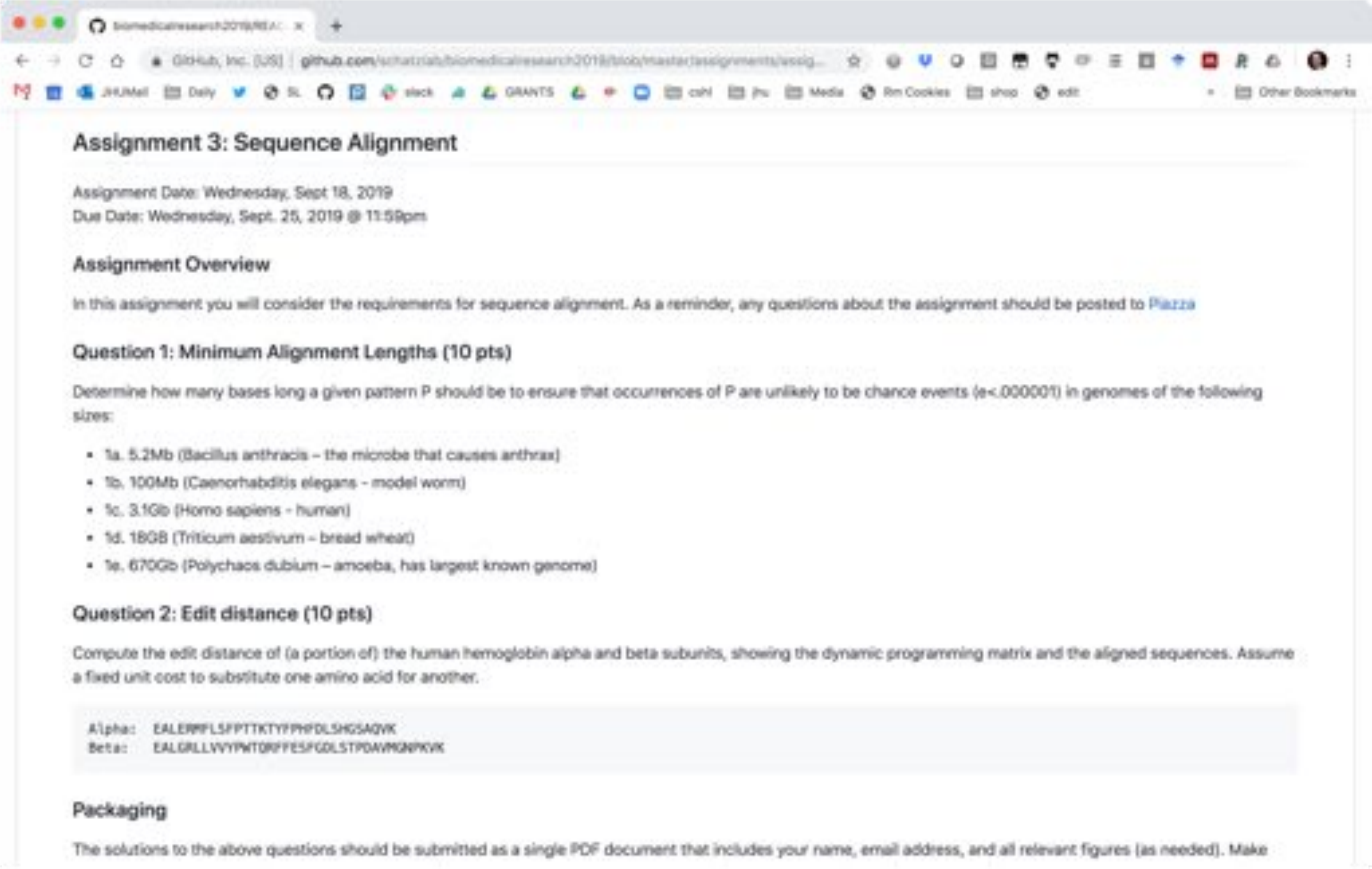
Human Evolution

Michael Schatz

Sept 25, 2019

Lecture 8: Computational Biomedical Research

Assignment 3: Sequence Alignment
Due Monday Sept 30 @ 11:59pm


biomedicalresearch2019/REA
GitHub, Inc. [US] | github.com/schatzlab/biomedicalresearch2019/blob/master/assignments/assig...
JHUMail
Daily
SL
slack
GRANTS
cshl
jhu
Media
Rm Cookies
shop
edit
Other Bookmarks

Assignment 3: Sequence Alignment

Assignment Date: Wednesday, Sept 18, 2019
Due Date: Wednesday, Sept. 25, 2019 @ 11:59pm

Assignment Overview

In this assignment you will consider the requirements for sequence alignment. As a reminder, any questions about the assignment should be posted to Piazza

Question 1: Minimum Alignment Lengths (10 pts)

Determine how many bases long a given pattern P should be to ensure that occurrences of P are unlikely to be chance events (e<.000001) in genomes of the following sizes:

- 1a. 5.2Mb (Bacillus anthracis – the microbe that causes anthrax)
- 1b. 100Mb (Caenorhabditis elegans - model worm)
- 1c. 3.1Gb (Homo sapiens - human)
- 1d. 18GB (Triticum aestivum – bread wheat)
- 1e. 670Gb (Polychaos dubium – amoeba, has largest known genome)

Question 2: Edit distance (10 pts)

Compute the edit distance of (a portion of) the human hemoglobin alpha and beta subunits, showing the dynamic programming matrix and the aligned sequences. Assume a fixed unit cost to substitute one amino acid for another.

```
Alpha:   EALERMFLSFPTTKTYFPHFDLSHGSAQVK
Beta:    EALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVK
```

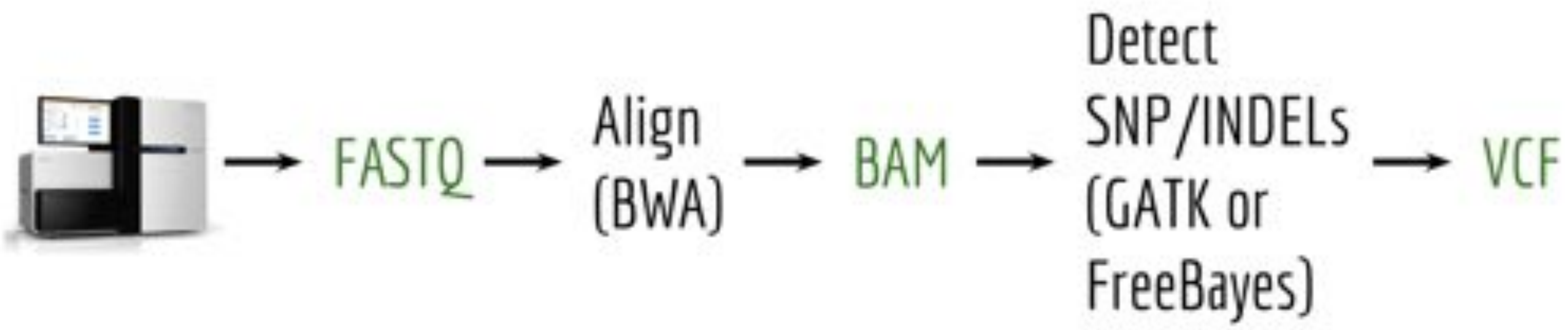
Packaging

The solutions to the above questions should be submitted as a single PDF document that includes your name, email address, and all relevant figures (as needed). Make

https://github.com/schatzlab/biomedicalresearch2019

Part 1: Recap

Variant Calling Overview



→ FASTQ → Align (BWA) → BAM → Detect SNP/INDELs (GATK or FreeBayes) → VCF

Similarity metrics

- Hamming distance
  - Count the number of substitutions to transform one string into another

```
MIKESCHATZ
||x||xxxx|
MICESHATZZ
     5
```

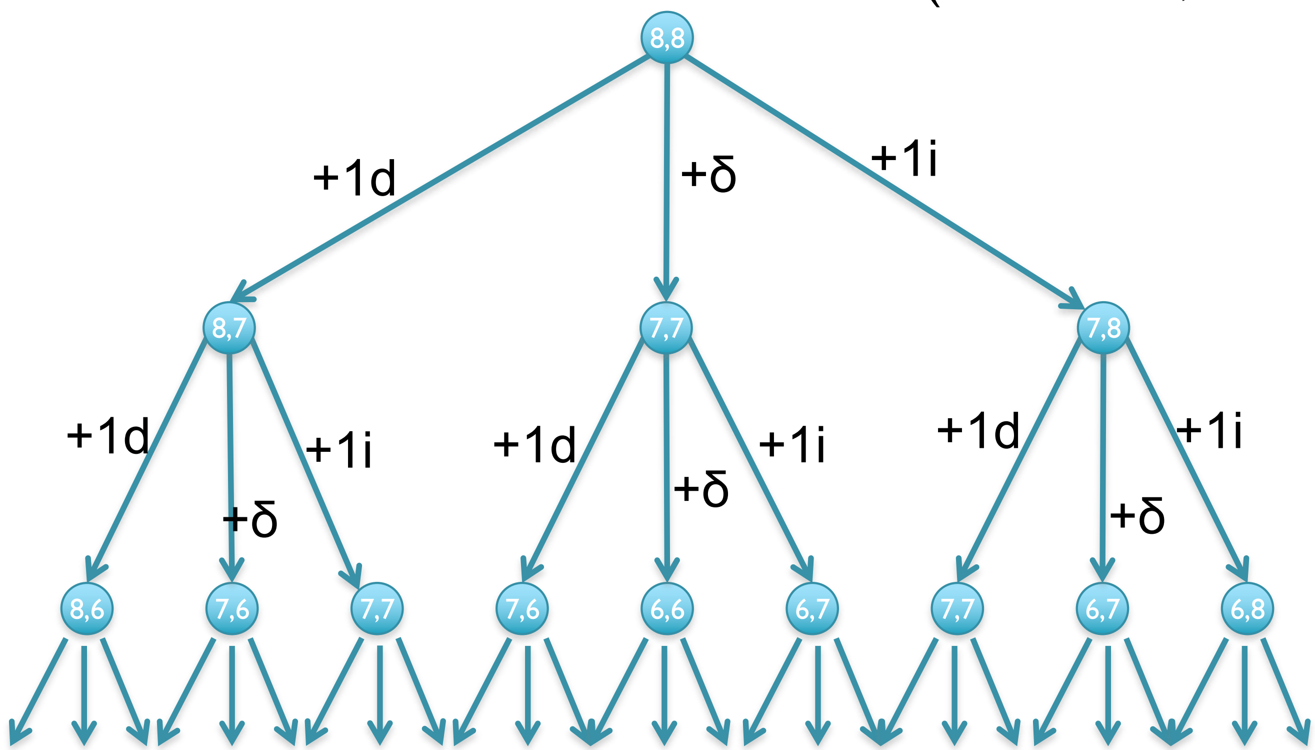
- Edit distance
  - The minimum number of substitutions, insertions, or deletions to transform one string into another

```
MIKESCHAT-Z
||x||x|||x|
MICES-HATZZ
     3
```

Recursive solution

- Computation of D is a recursive process.
  - At each step, we only allow matches, substitutions, and indels
  - D(i,j) in terms of D(i' ,j') for i' ≤ i and j' ≤ j.

D(AGCACACA, ACACACTA) = min{D(AGCACACA, ACACACT) + 1,
D(AGCACAC, ACACACTA) + 1,
D(AGCACAC, ACACACT) +δ(A, A)}



[What is the running time?]

Dynamic Programming Matrix

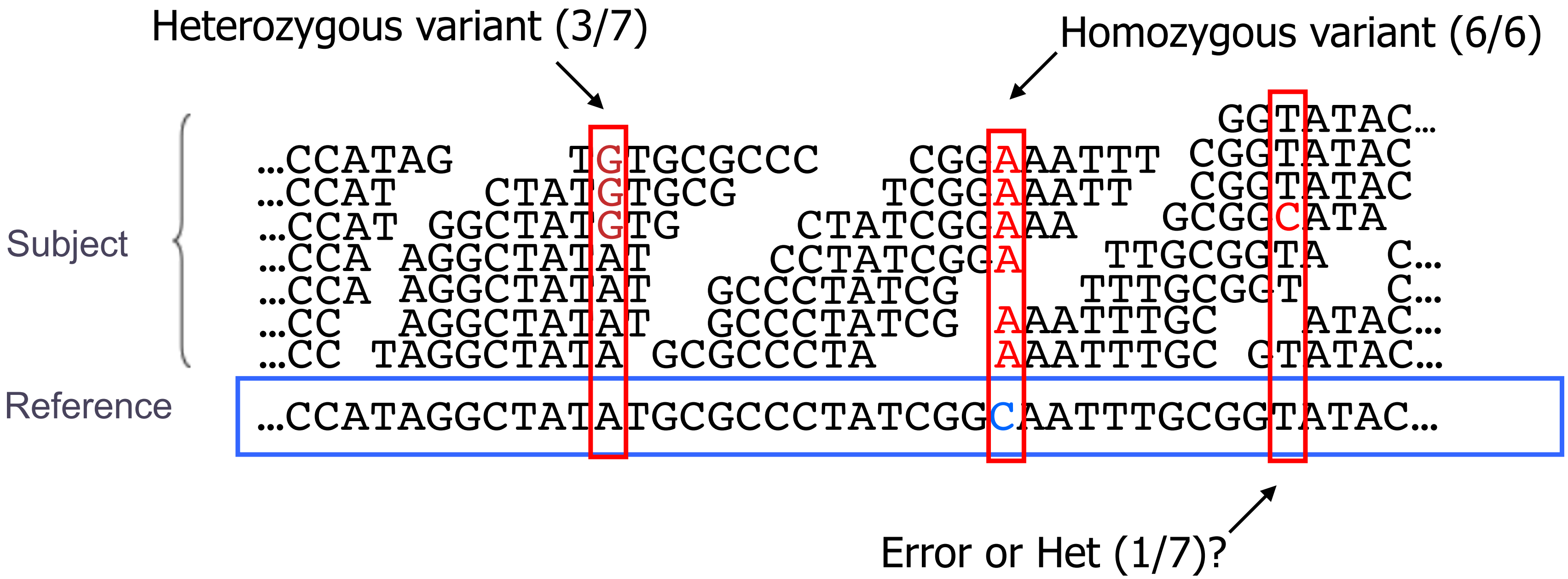
		A	C	A	C	A	C	T	A
	0	1	2	3	4	5	6	7	8
A	1	0	1	2	3	4	5	6	7
G	2	1	1	2	3	4	5	6	7
C	3	2	1	2	2	3	4	5	6
A	4	3	2	1	2	2	3	4	5
C	5	4	3	2	1	2	2	3	4
A	6	5	4	3	2	1	2	3	3
C	7	6	5	4	3	2	1	2	3
A	8	7	6	5	4	3	2	2	2

D[AGCACACA,ACACACTA] = 2

```
AGCACAC-A
|*|||||*|
A-CACACTA
```

[Can we do it any better?]

Genotyping Theory



Heterozygous variant (3/7)

Homozygous variant (6/6)

Subject

```
                                    GGTATAC...
...CCATAG    TGTGCGCCC   CGGAAATTT CGGTATAC
...CCAT   CTATGTGCG     TCGGAAATT  CGGTATAC
...CCAT GGCTATGTG    CTATCGGAAA   GCGGCATA
...CCA AGGCTATAT    CCTATCGGA   TTGCGGTA  C...
...CCA AGGCTATAT  GCCCTATCG    TTTGCGGT   C...
...CC  AGGCTATAT  GCCCTATCG AAATTTGC   ATAC...
...CC TAGGCTATA GCGCCCTA    AAATTTGC GTATAC...
```

Reference

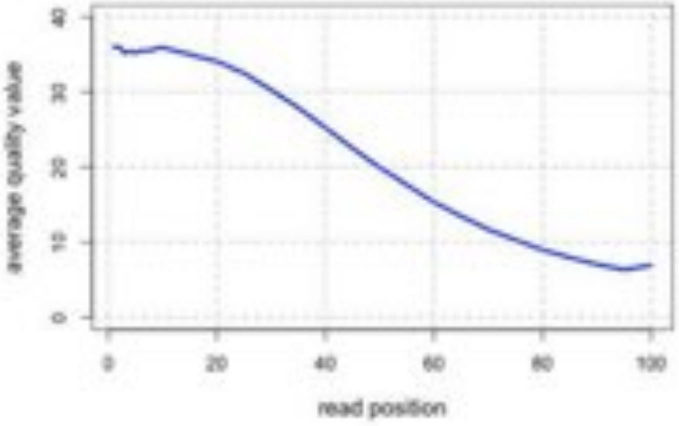
```
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

Error or Het (1/7)?

- 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



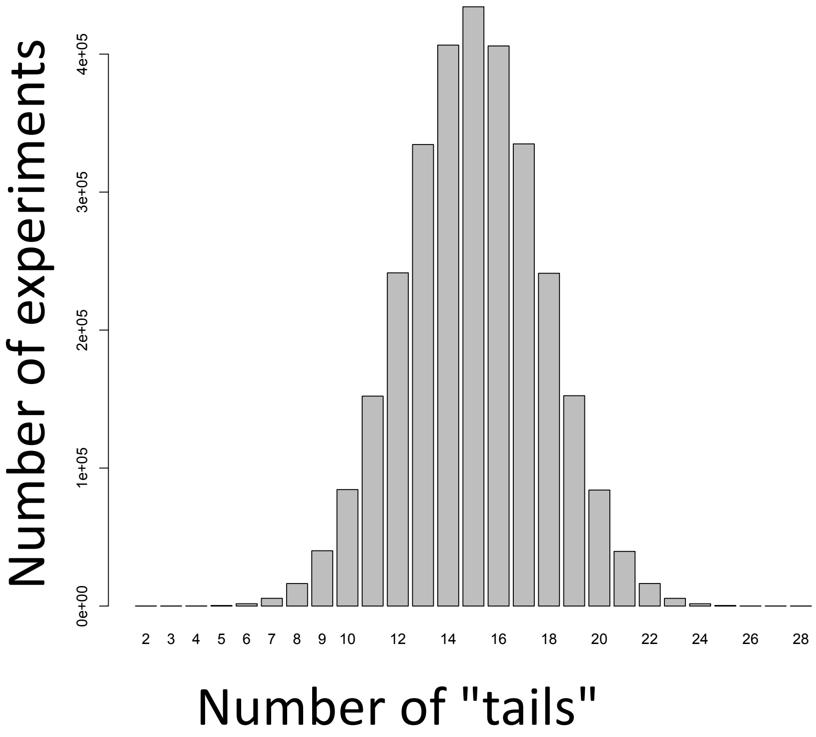
The Binomial Distribution: Adventures in Coin Flipping

P(heads) = 0.5

P(tails) = 0.5

Aaron Quinlan

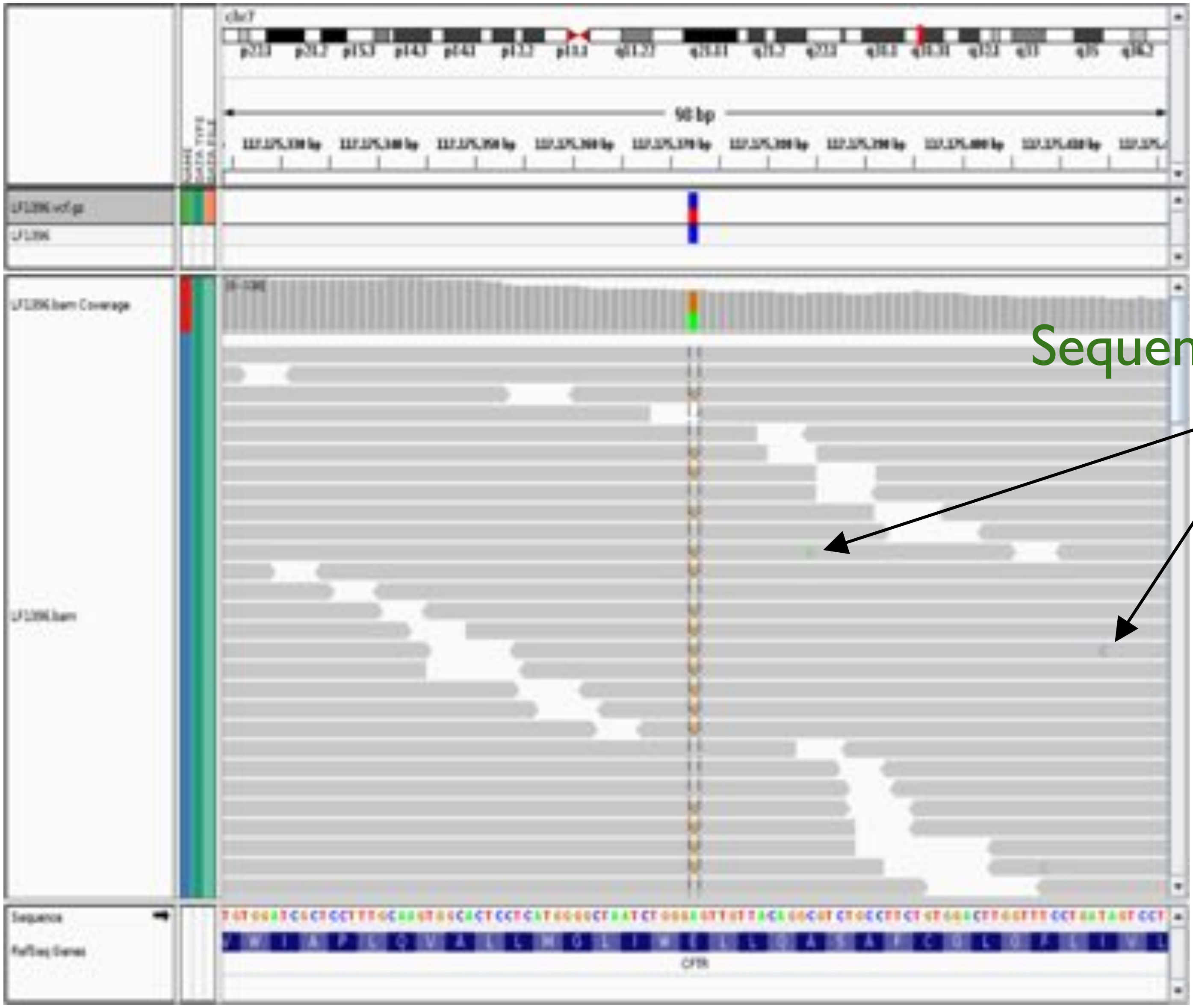
So, with 30 tosses (reads), we are much more likely to see an even mix of alternate and reference alleles at a heterozygous locus in a genome



This is why at least a "30X" (30 fold sequence coverage) genome is recommended: it confers sufficient power to distinguish heterozygous alleles and from mere sequencing errors

P(3/30 het) <?> P(3/30 err)

Sequencing errors fall out as noise (most of the time)

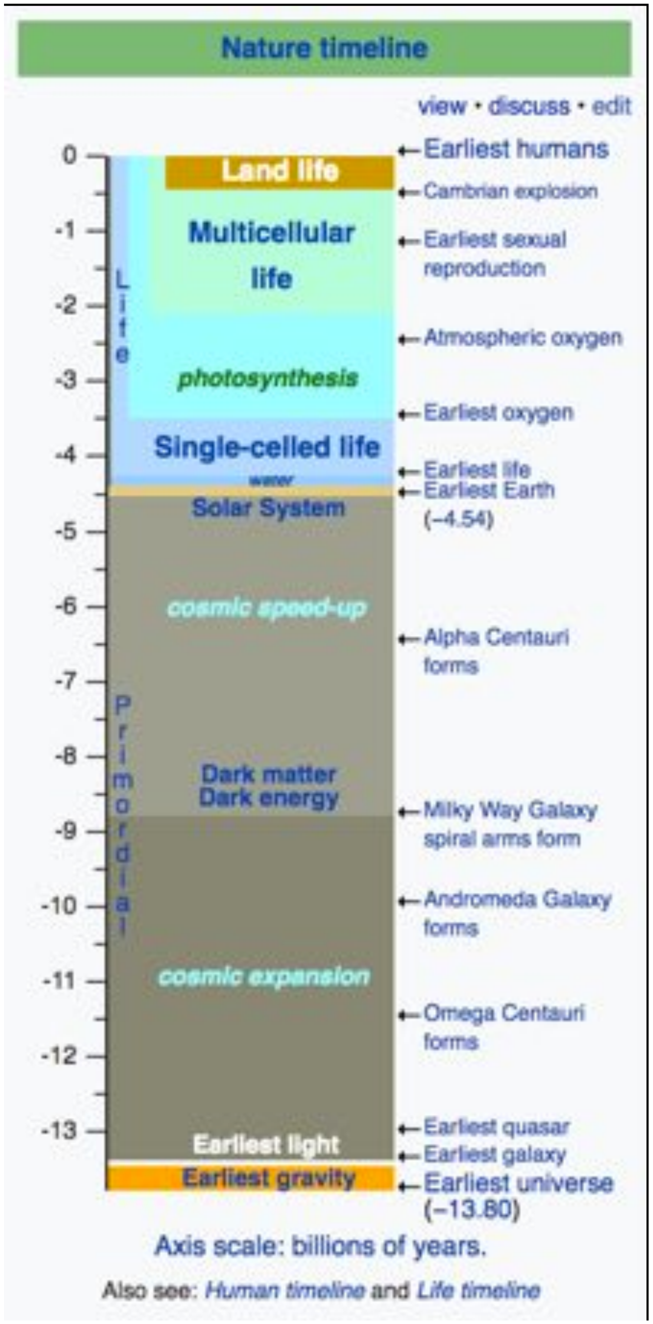


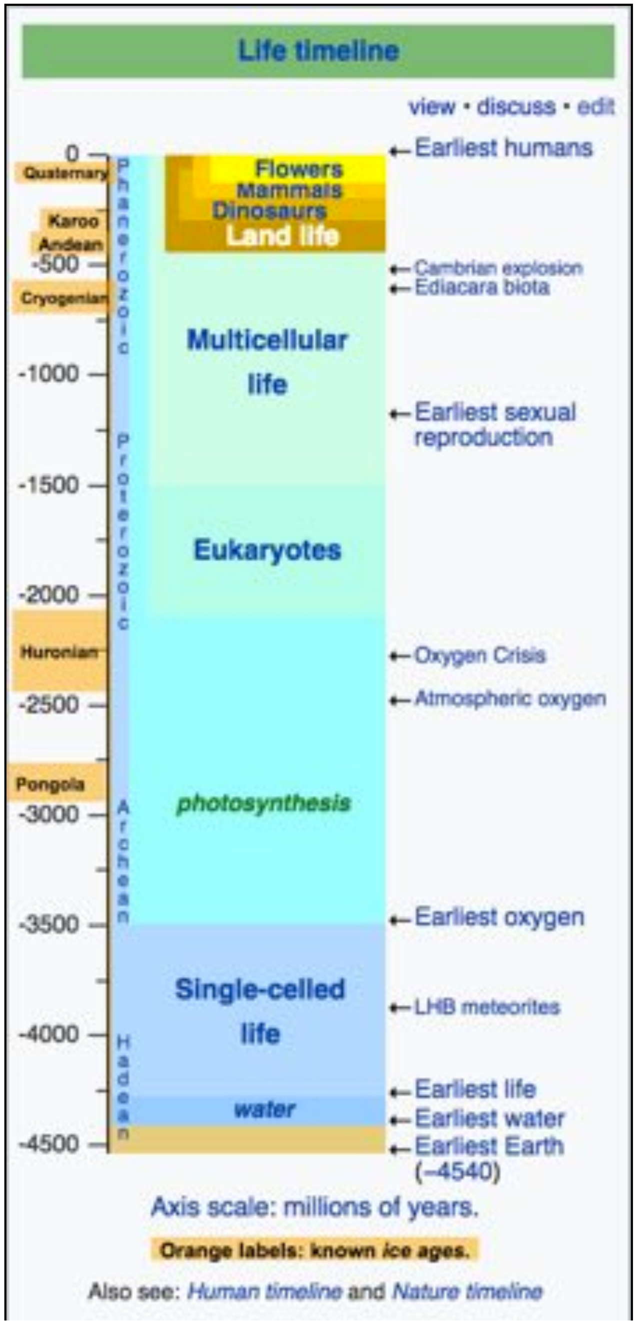
Sequencing errors

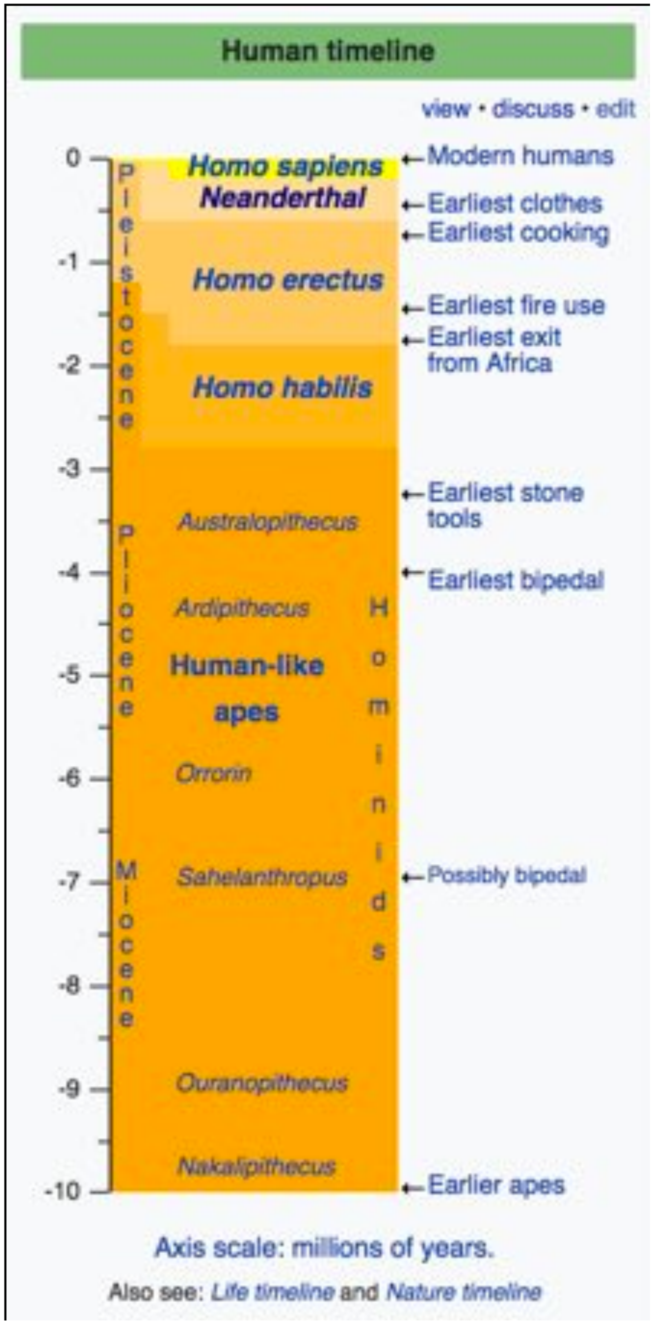
https://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

Part 2: Ancient Hominds

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

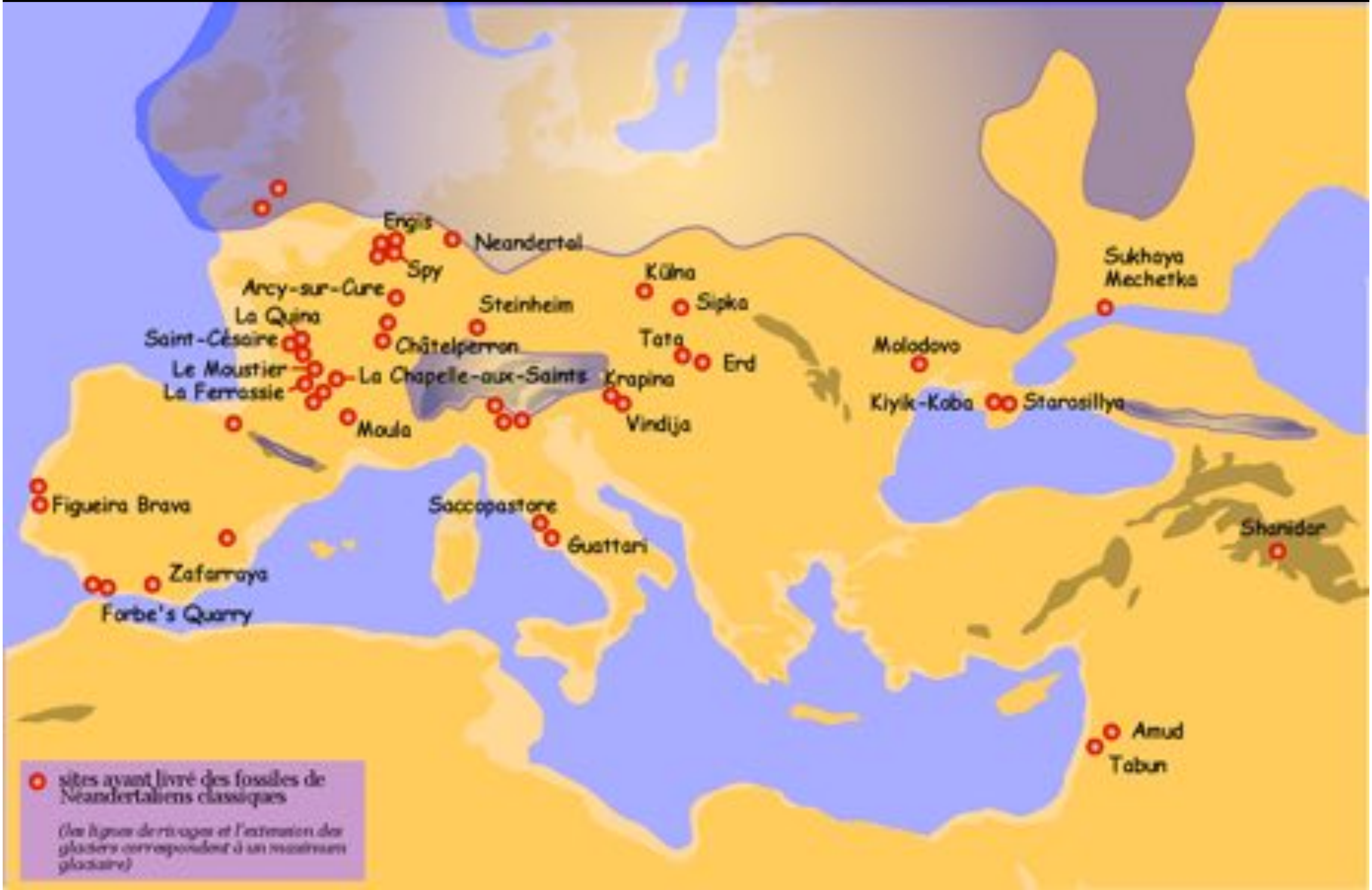

Life timeline
view • discuss • edit
0
-500
-1000
-1500
-2000
-2500
-3000
-3500
-4000
-4500
Quaternary
Karoo
Andean
Cryogenian
Huronian
Pongola
Phanerozoic
Proterozoic
Archean
Hadean
Flowers
Mammals
Dinosaurs
Land life
Multicellular life
Eukaryotes
photosynthesis
Single-celled life
water
Earliest humans
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
Homo sapiens
Neanderthal
Homo erectus
Homo habilis
Australopithecus
Ardipithecus
Human-like apes
Orrorin
Sahelanthropus
Ouranopithecus
Nakalipithecus
Hominids
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

Sequencing ancient genomes

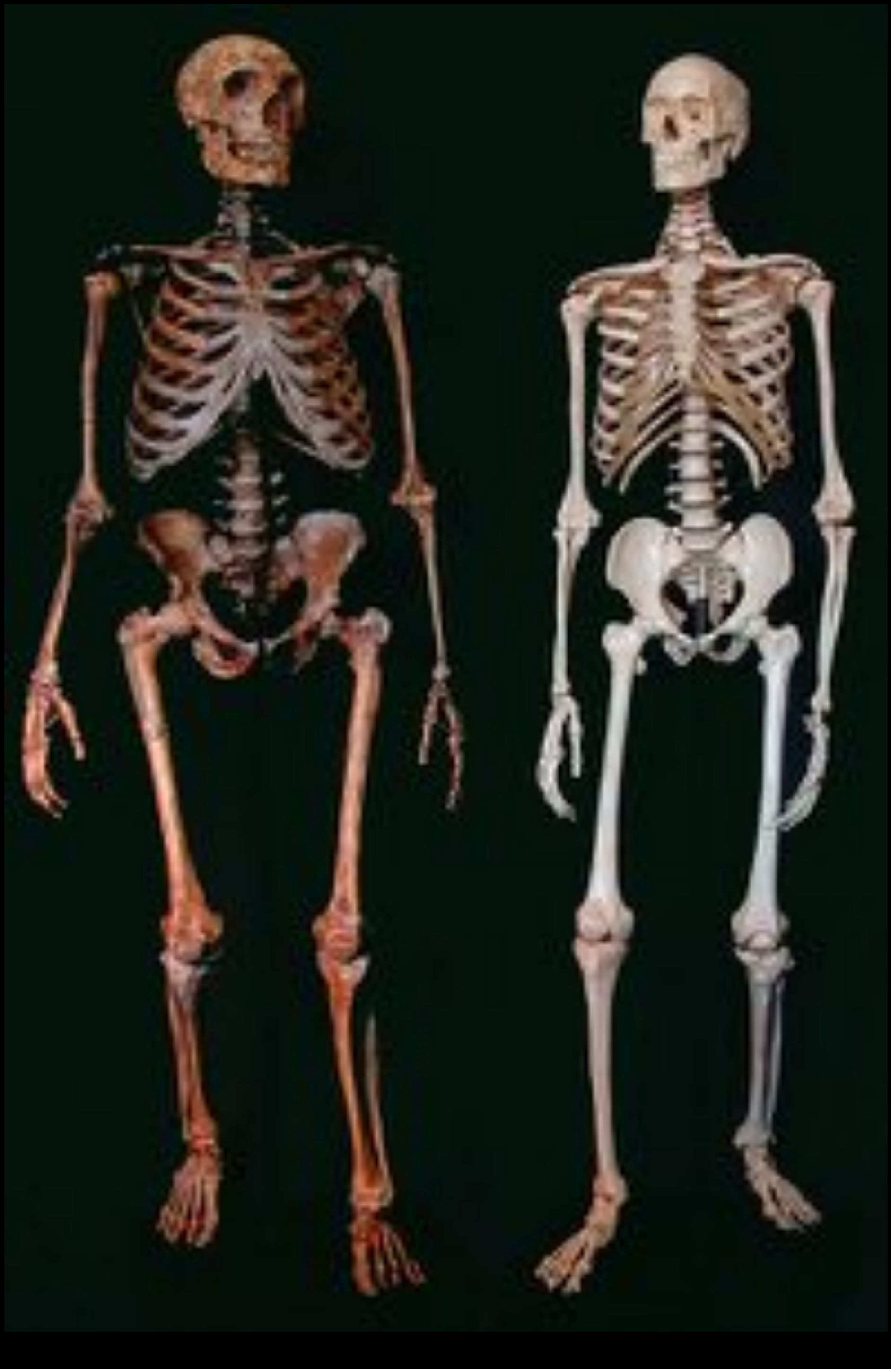
Janet Kelso

Max-Planck Institute


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

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 ☺

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*
Science **328**, 710 (2010);
DOI: 10.1126/science.1188021

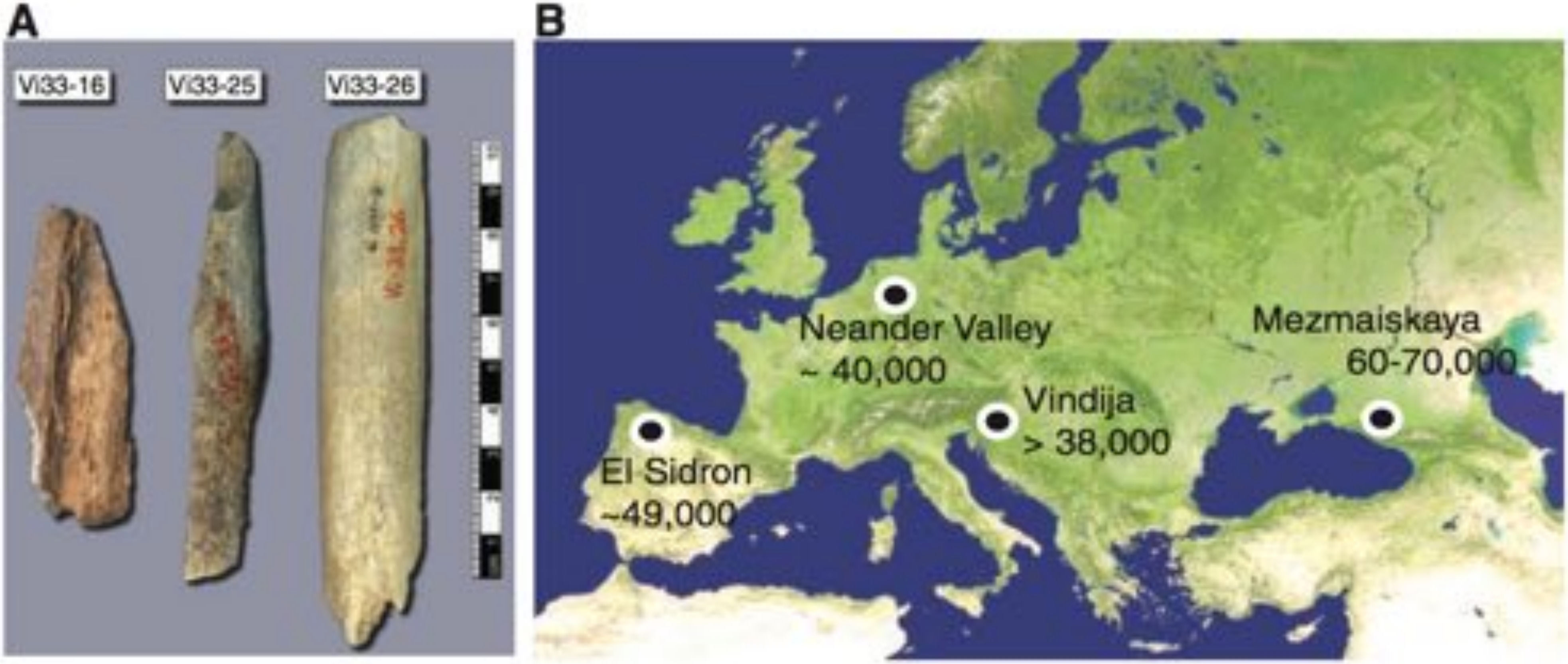
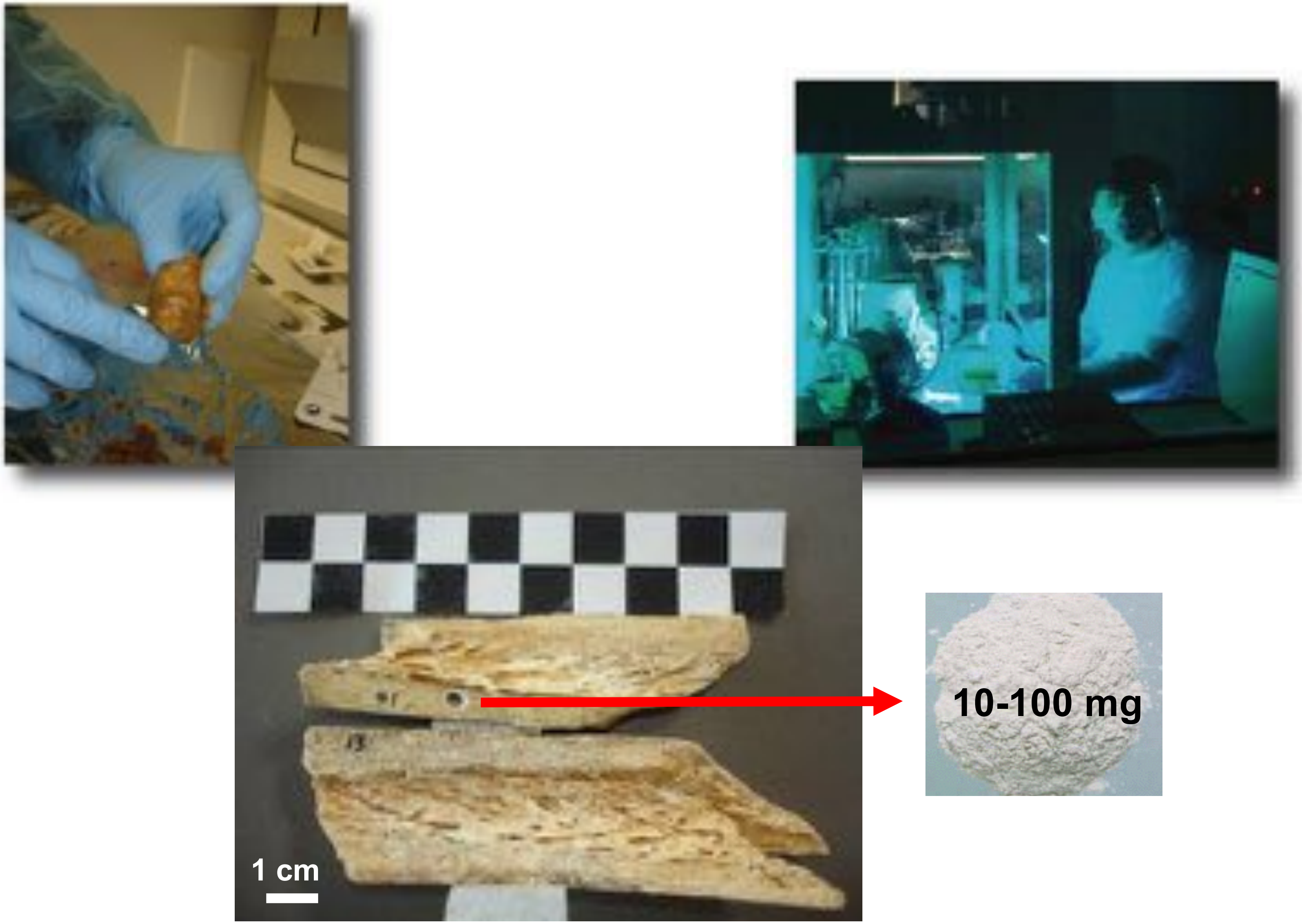
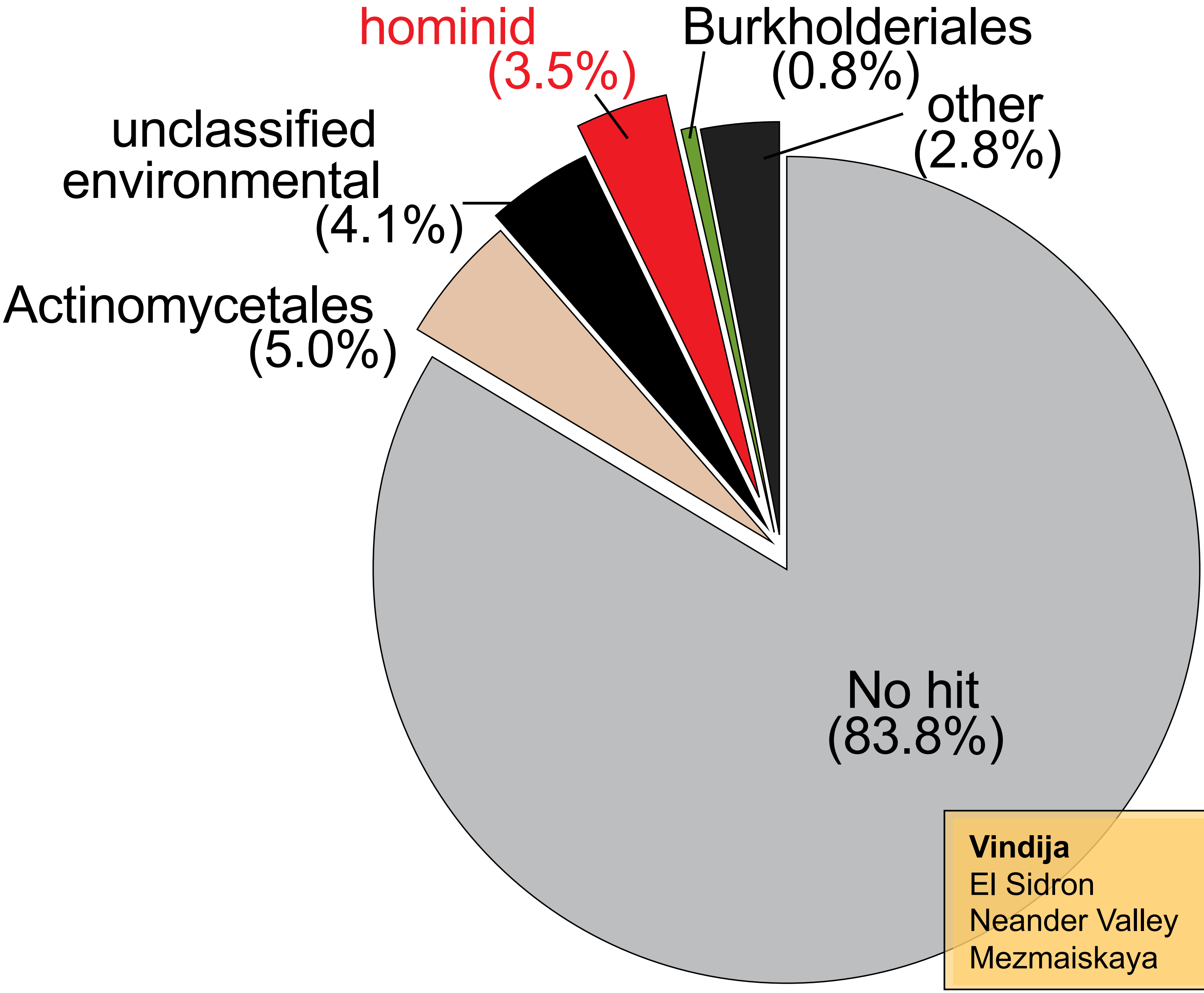


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

Extracting Ancient DNA

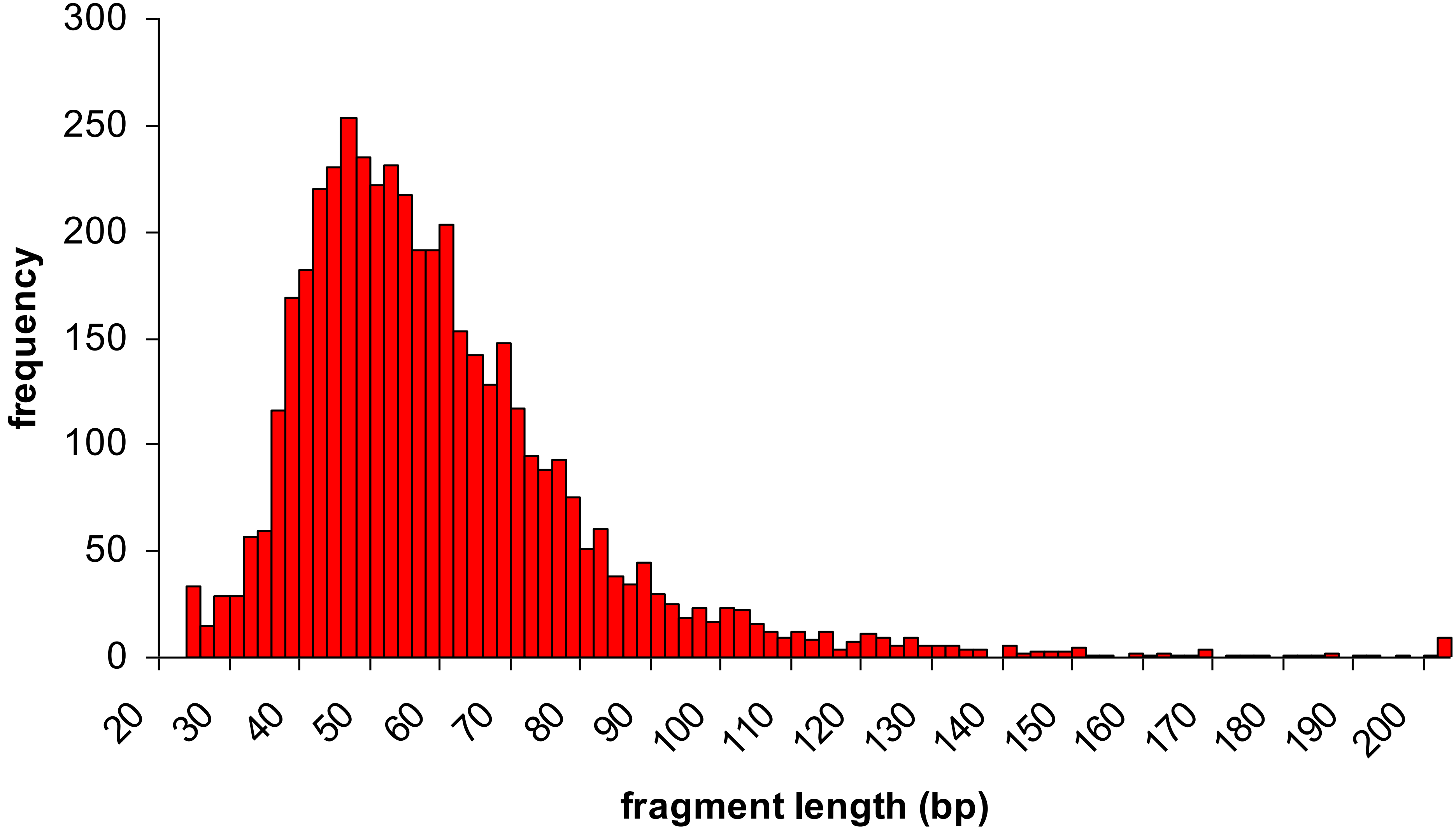


DNA is from mixed sources

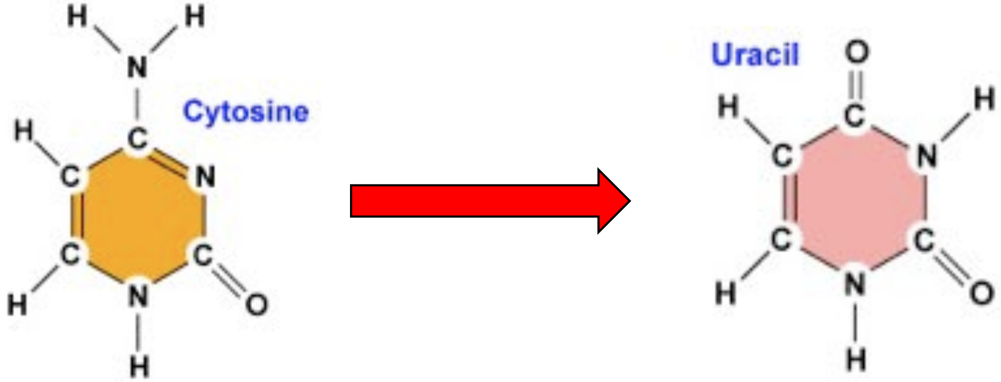


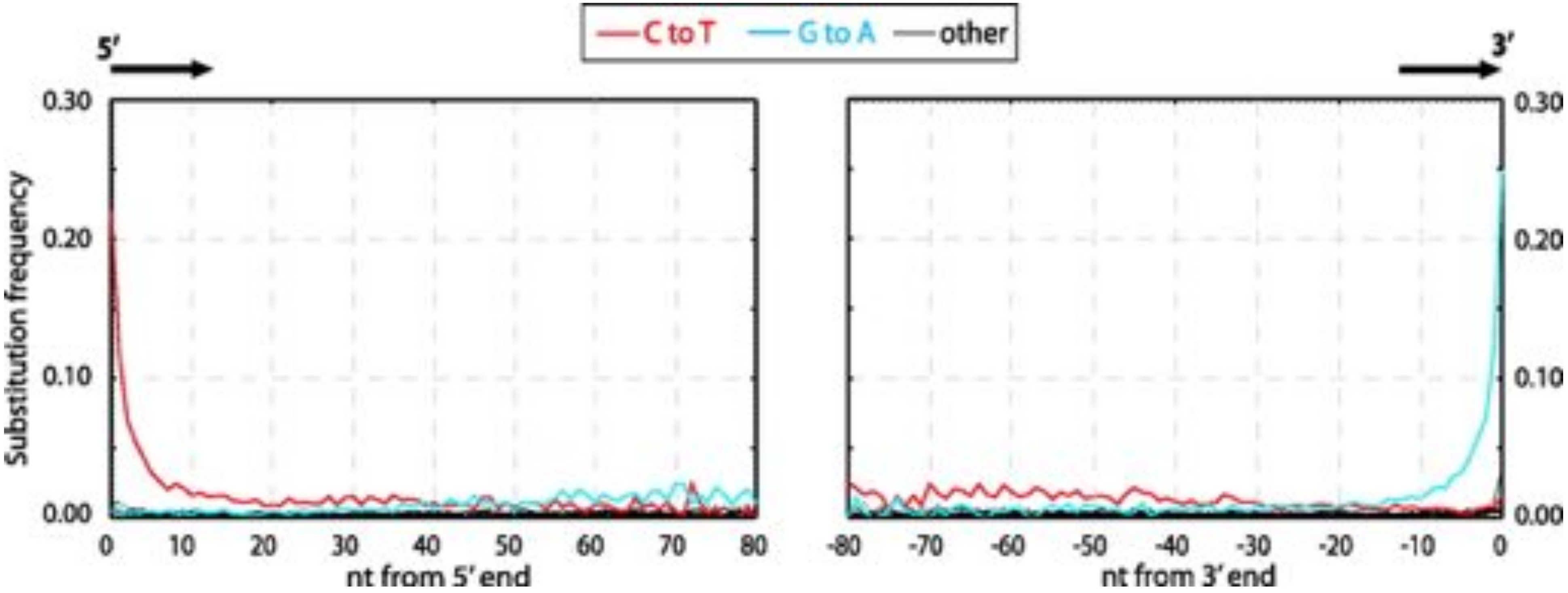
Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged





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

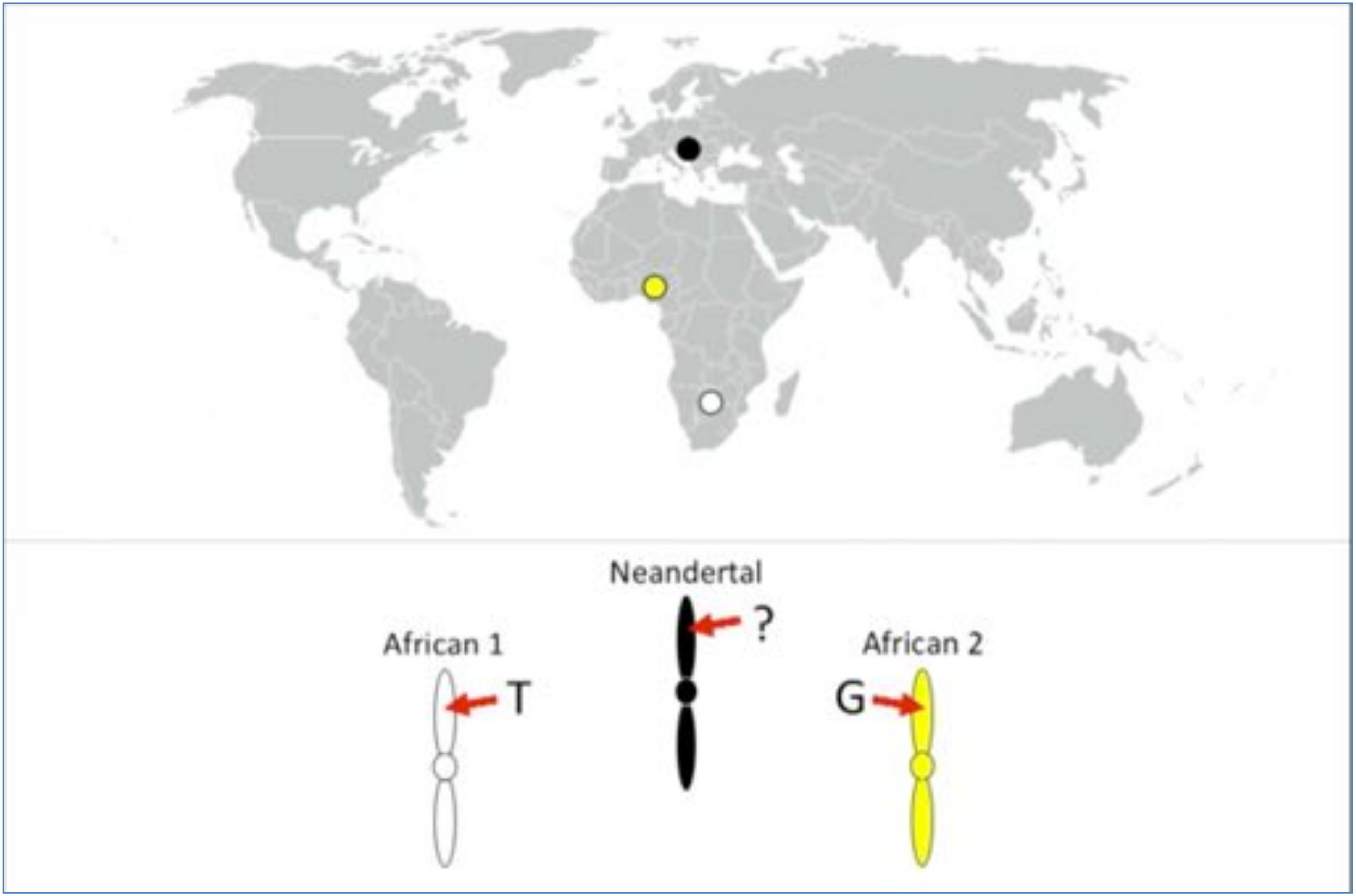
Green et al. 2010

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

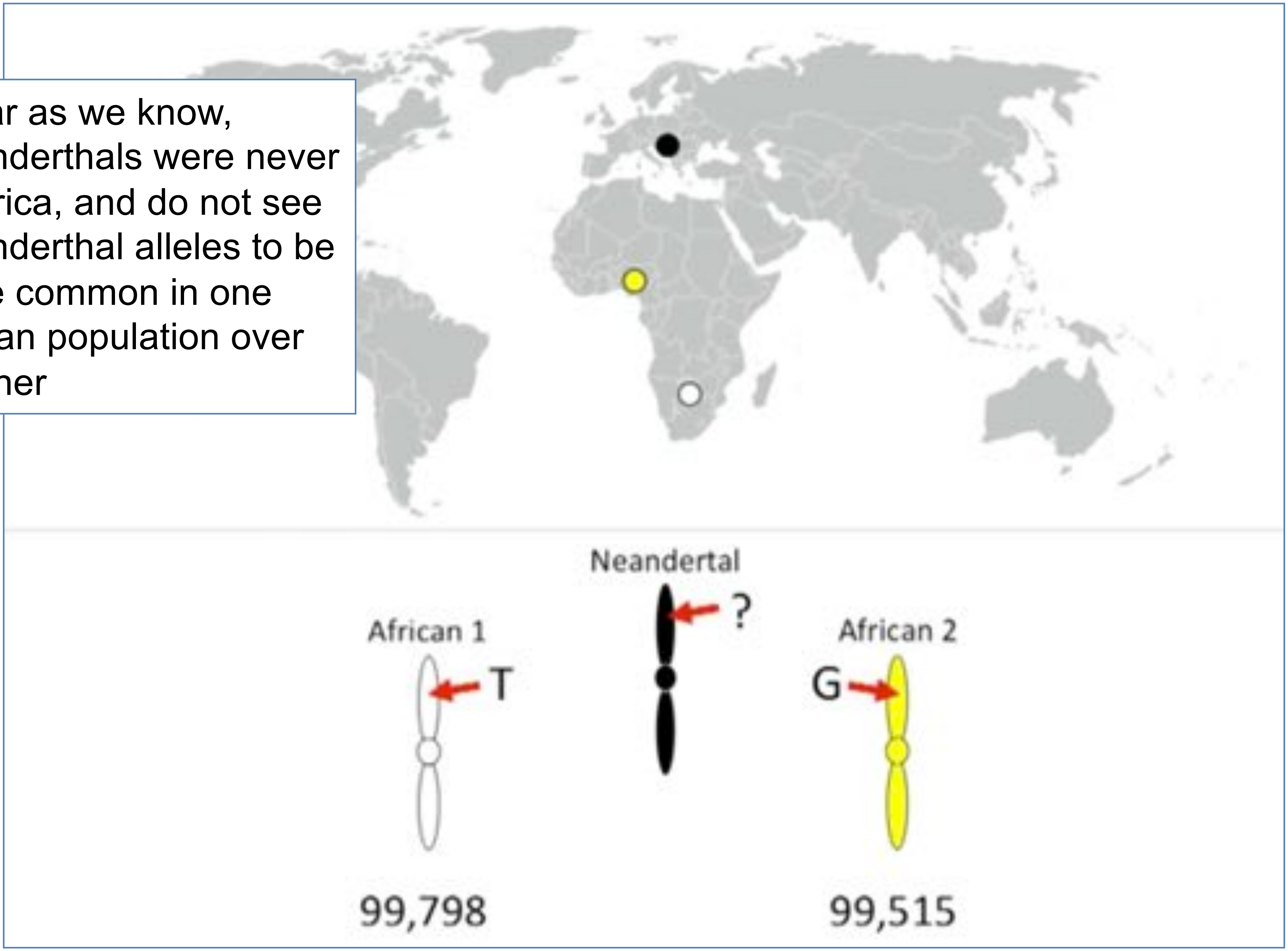
Genome coverage ~1.3 X

Did we mix?



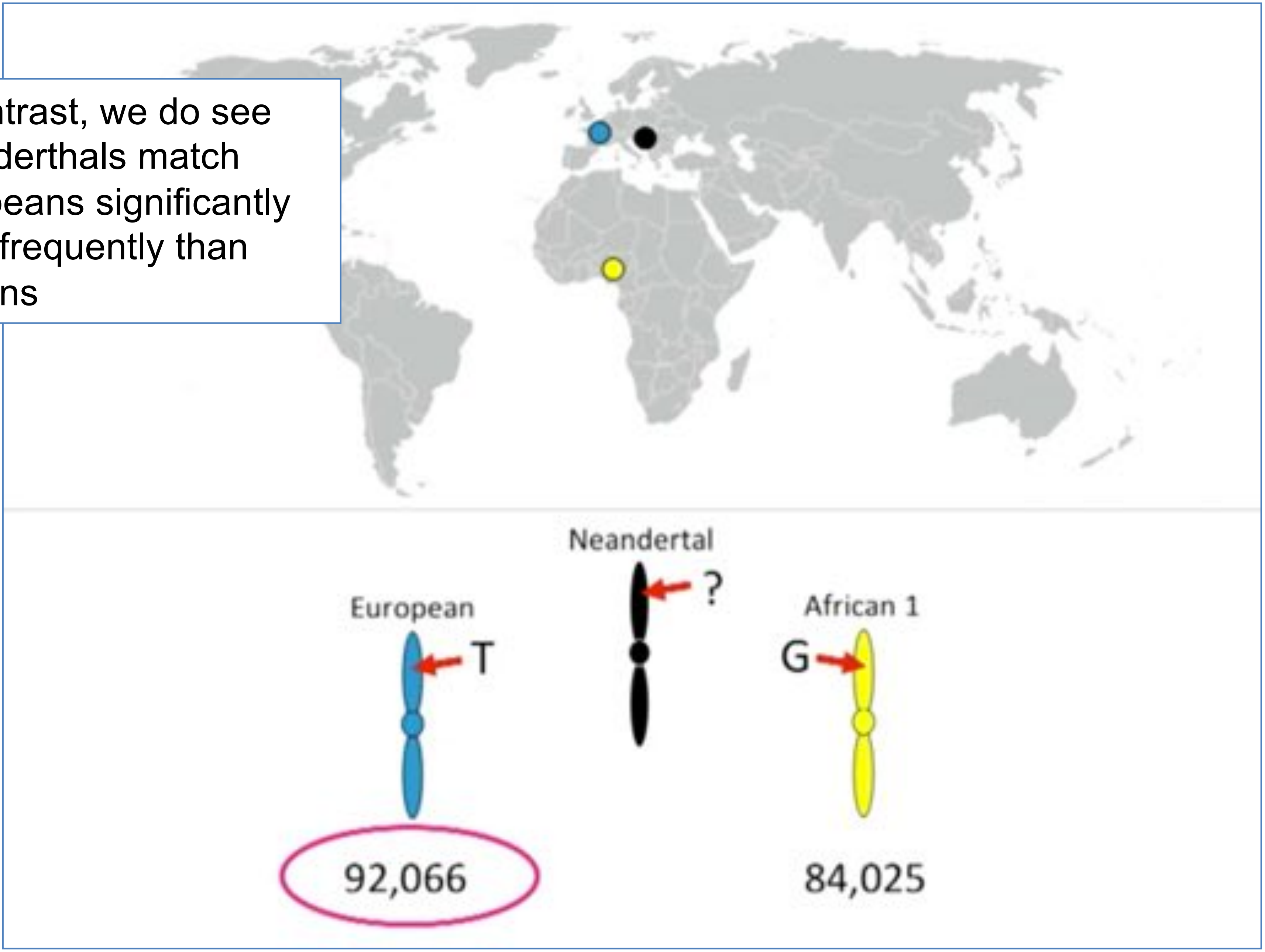
Did we mix?

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?

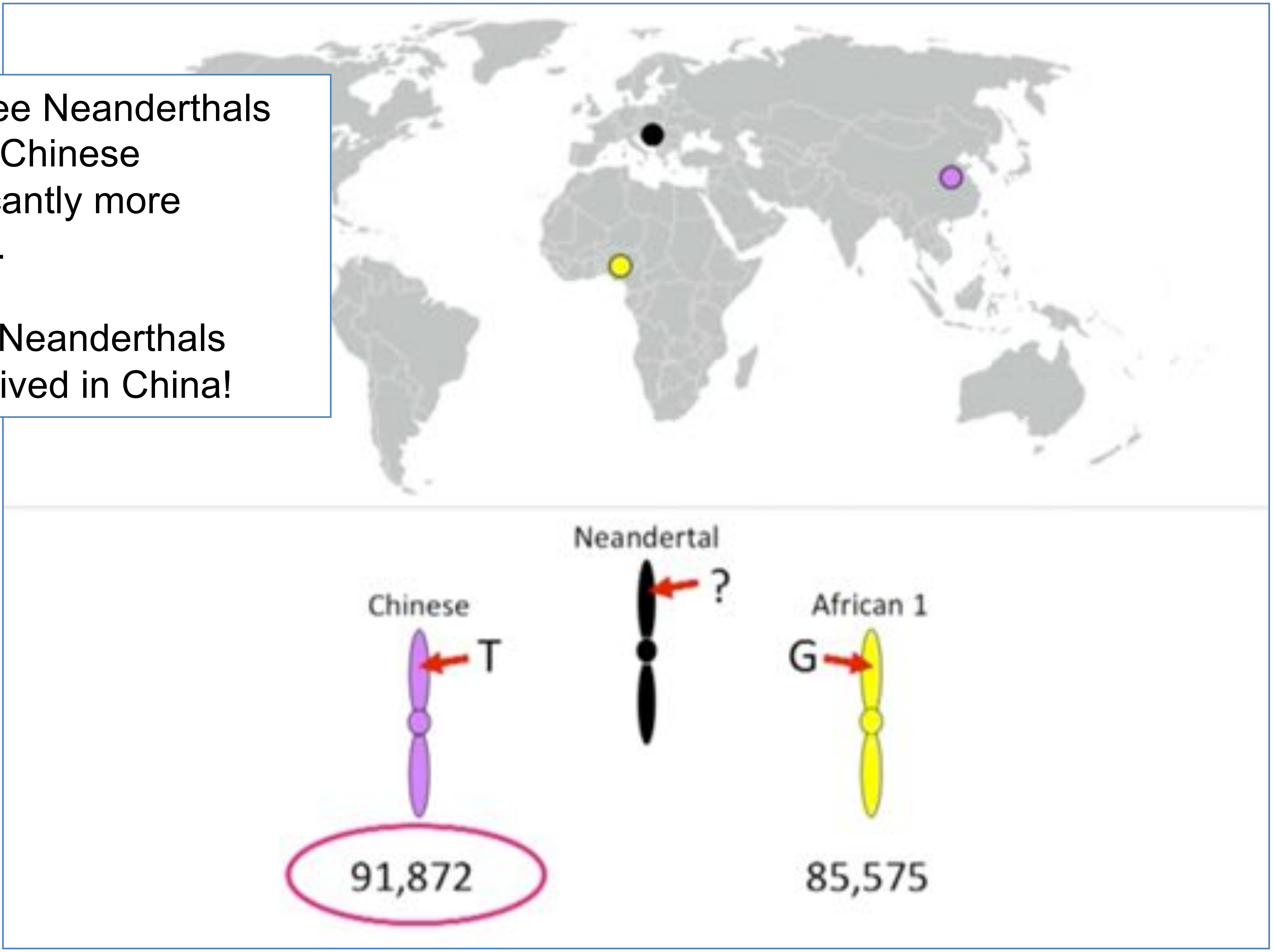
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



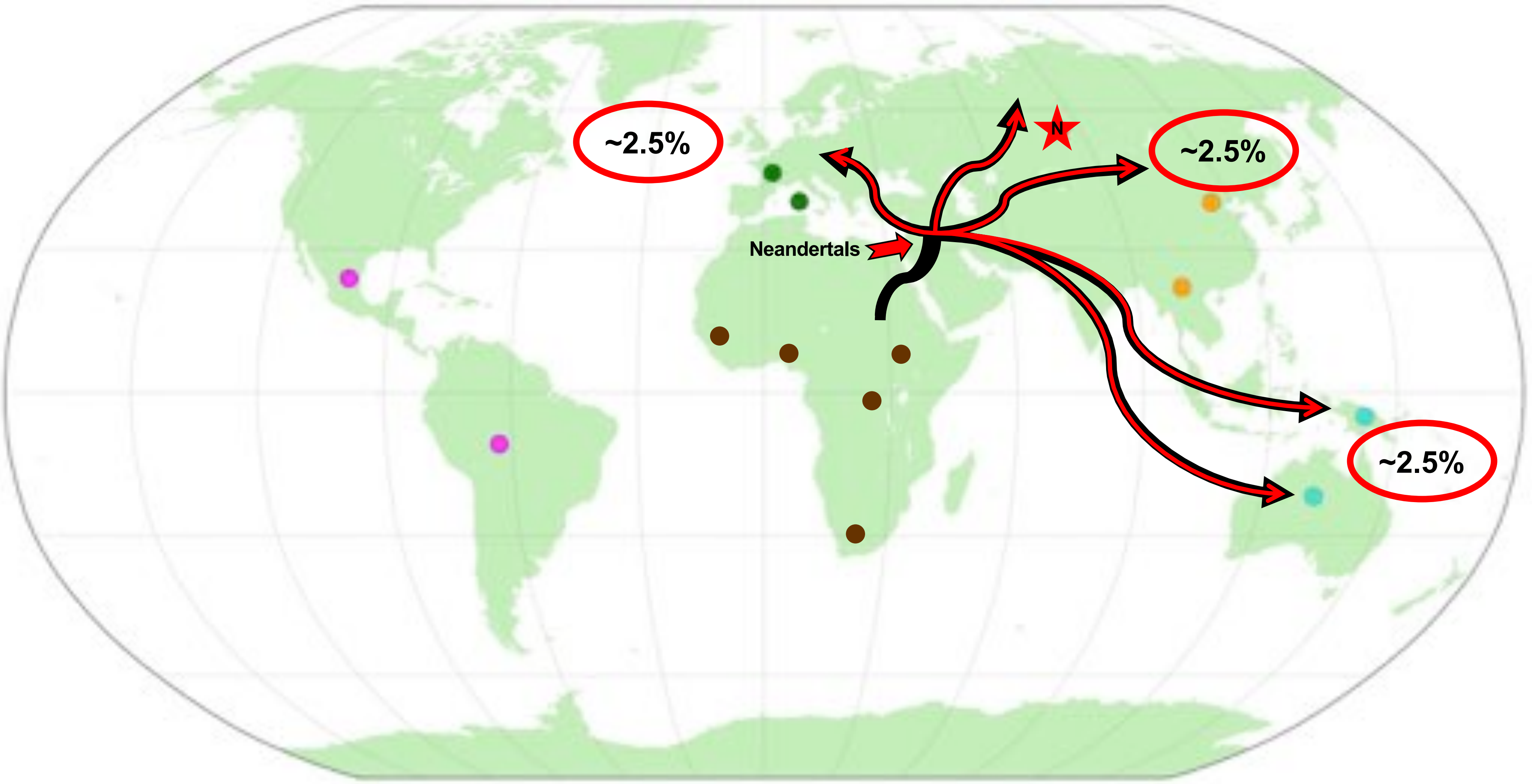
Did we mix?

Also see Neanderthals match Chinese significantly more often…

… but Neanderthals never lived in China!

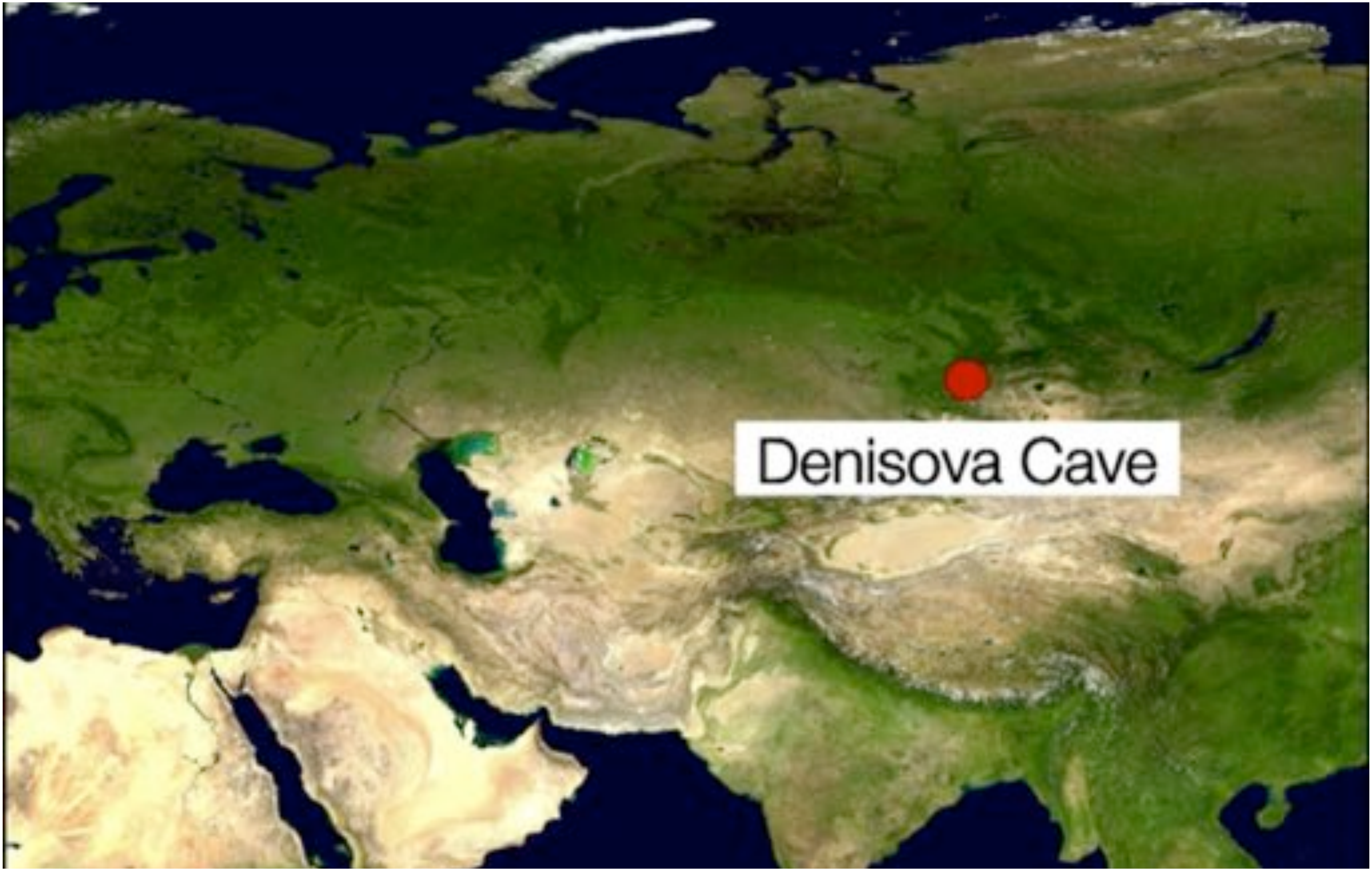


Neanderthal Interbreeding



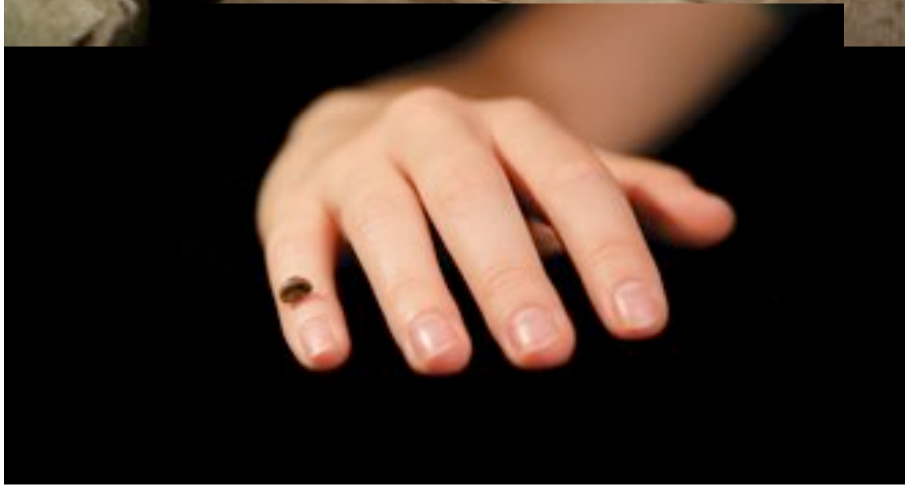
As modern humans migrated out of Africa, they apparently interbred with Neanderthal's so we see their alleles across the rest of the world and carry about 2.5% of their genome with us!

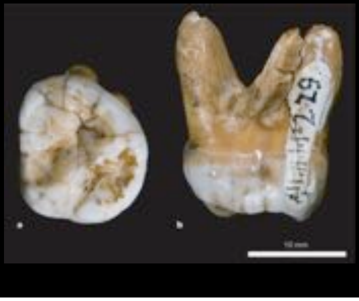
What about other ancient hominids?



Denisova cave
Altai mountains
Russia

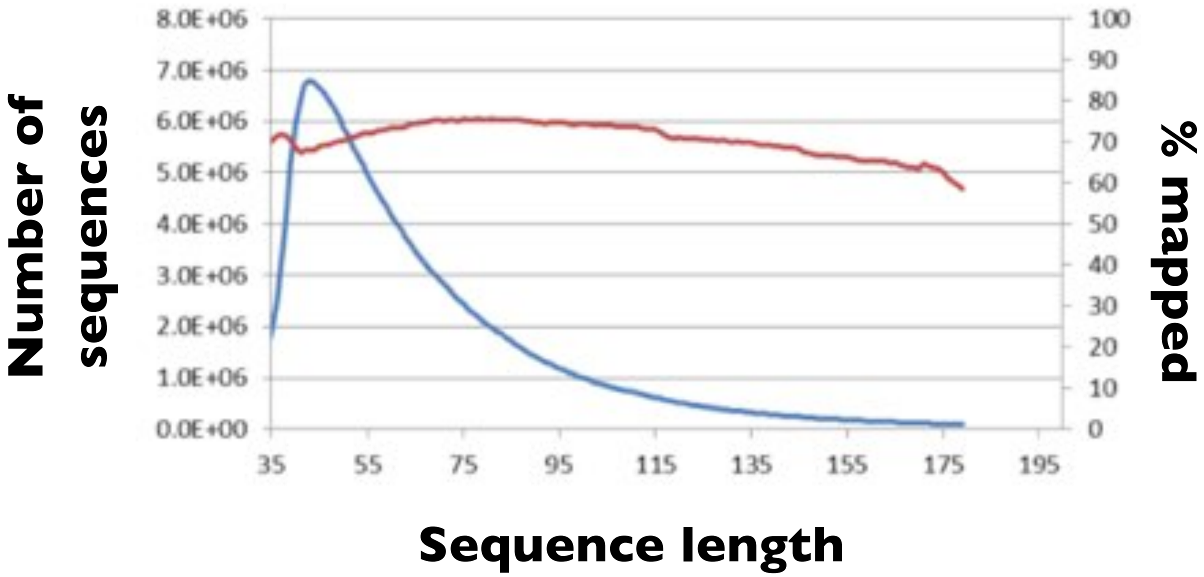
Academician A.P. Derevianko





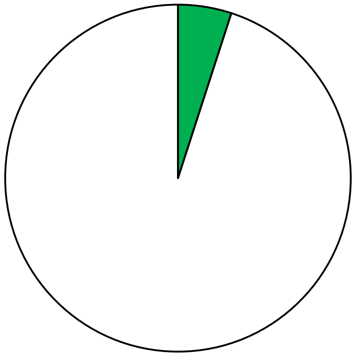
UNITED STATES OF AMERICA
E PLURIBUS UNUM
ONE CENT

Extraordinary preservation

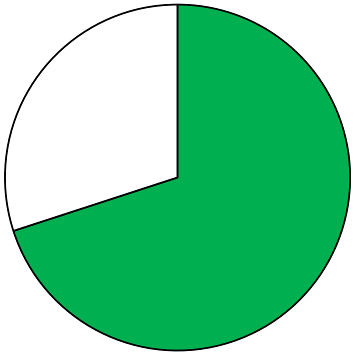


5%
endogenous
DNA

Best
Neandertal
bone

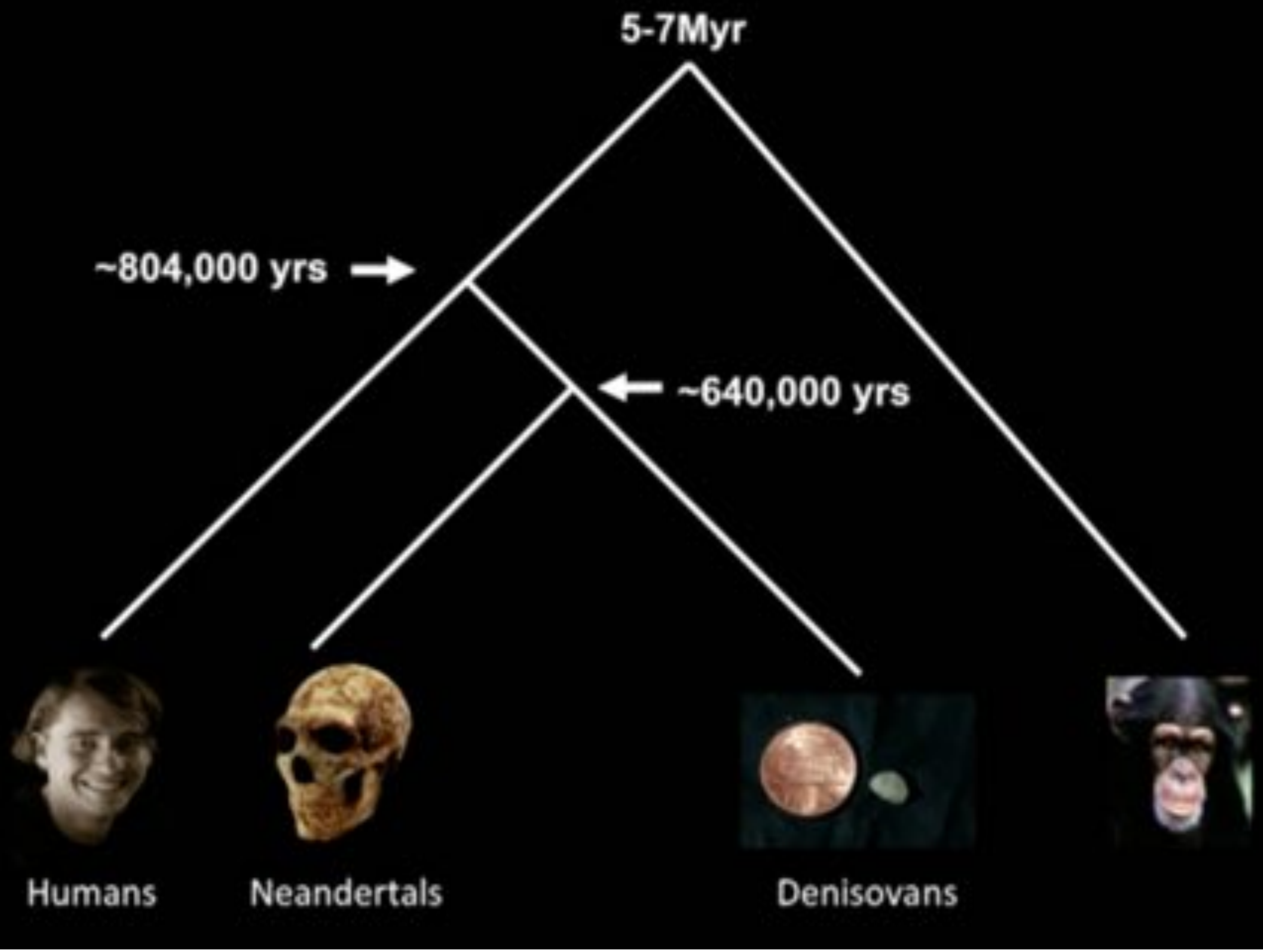


>70%
endogenous
DNA



Denisova
bone

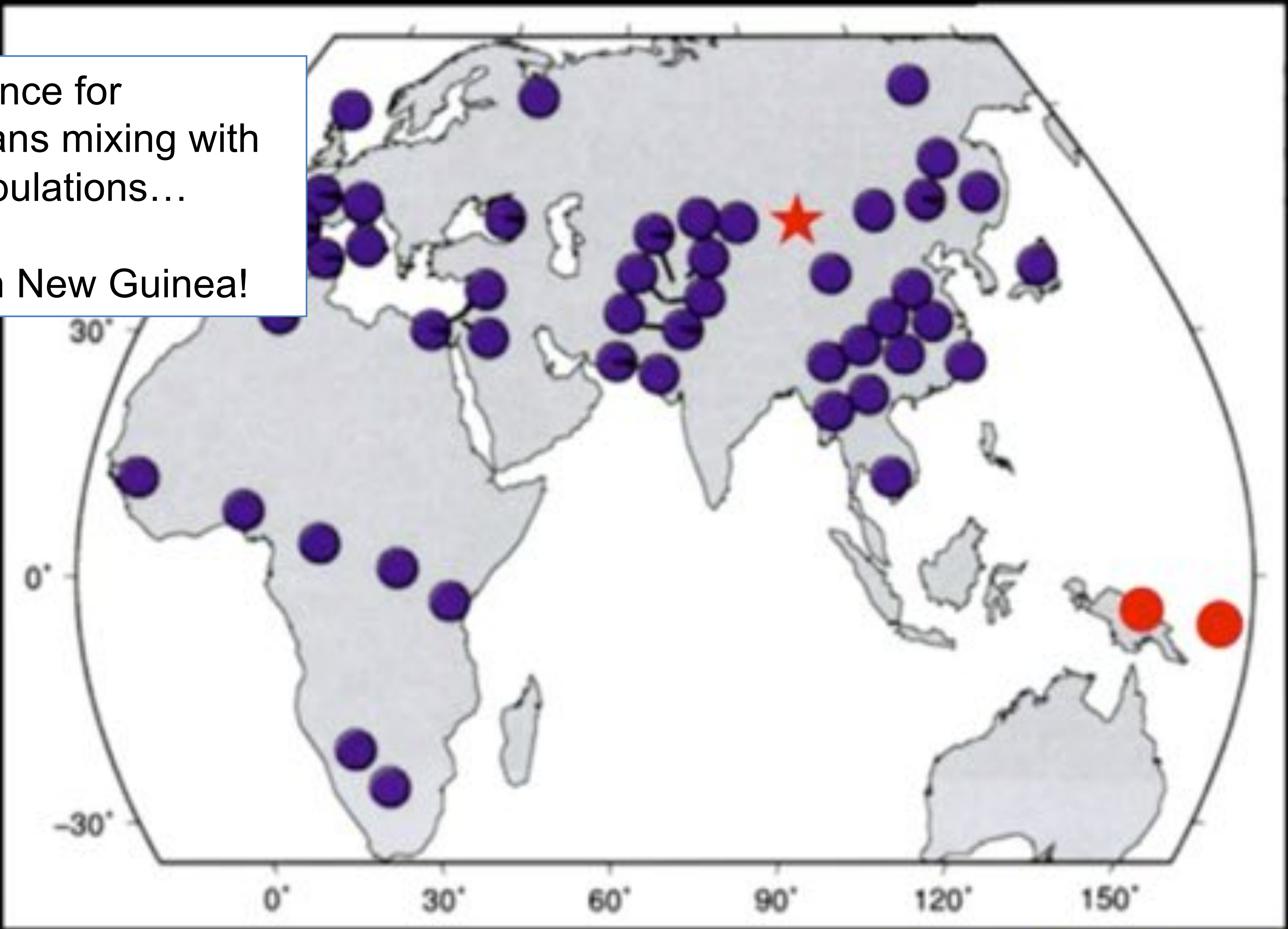
Denisovans & Neandertals



Did we mix?

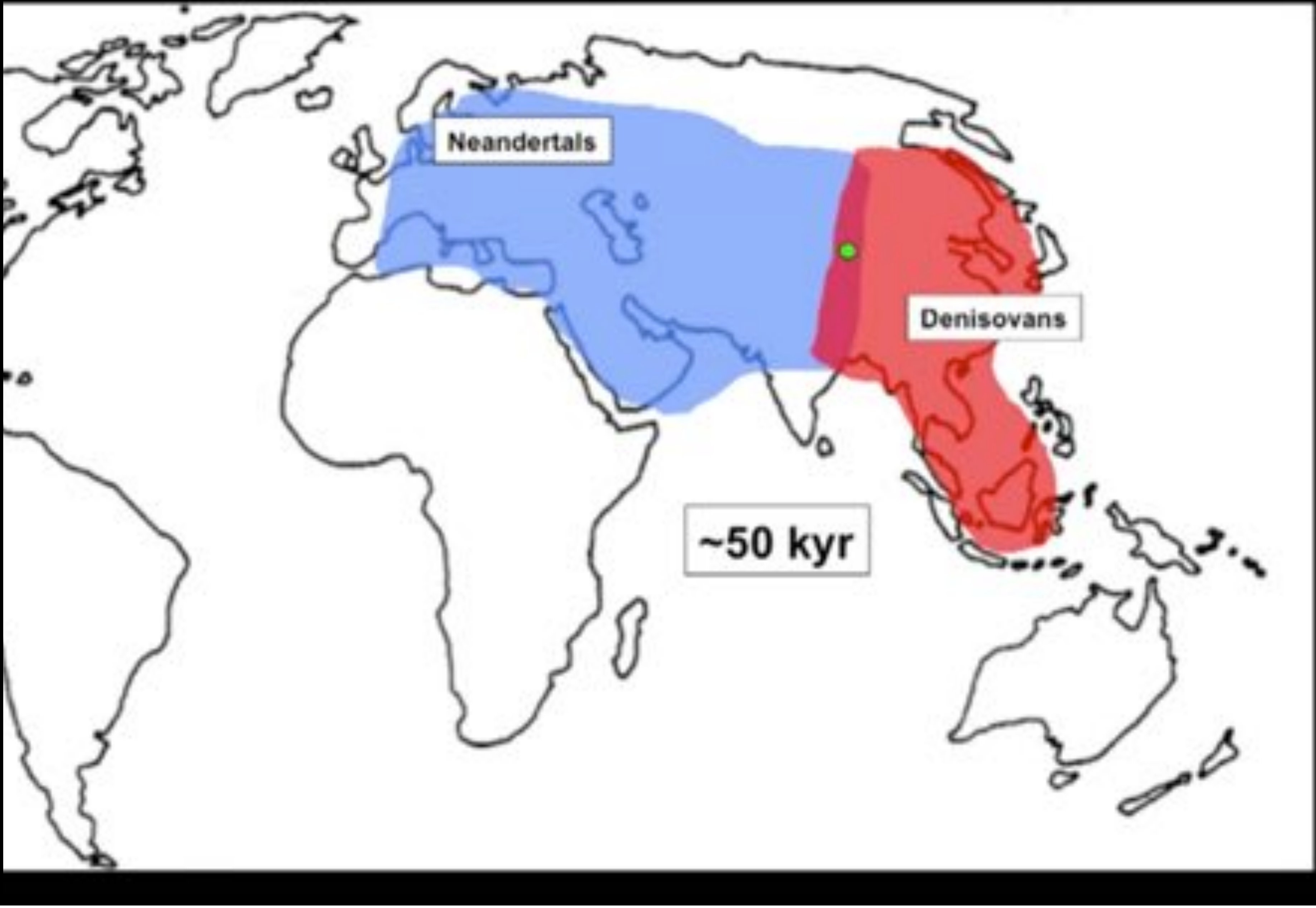
No evidence for Denisovans mixing with other populations…

Except in New Guinea!

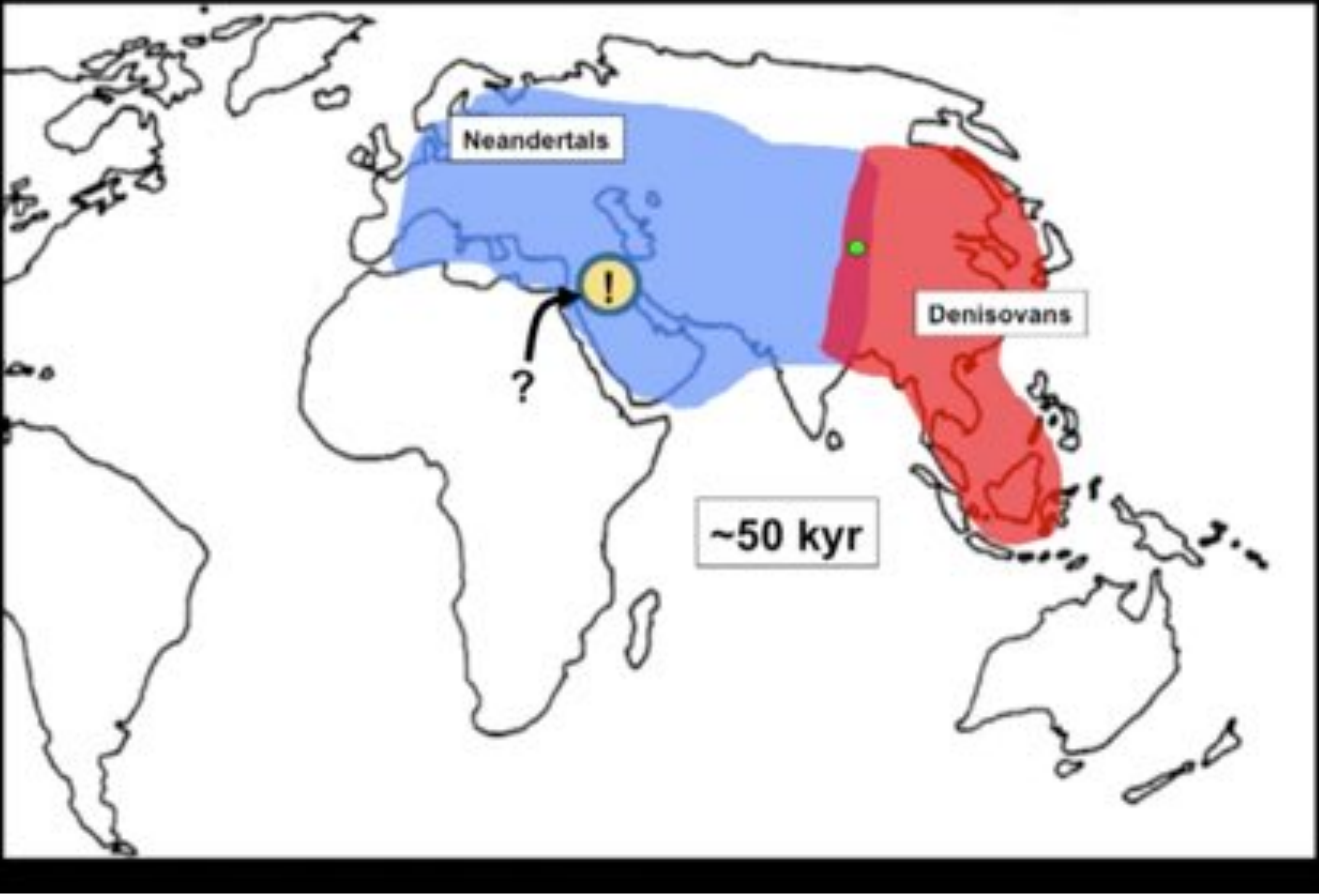


Map after Pickrell et al., 2009

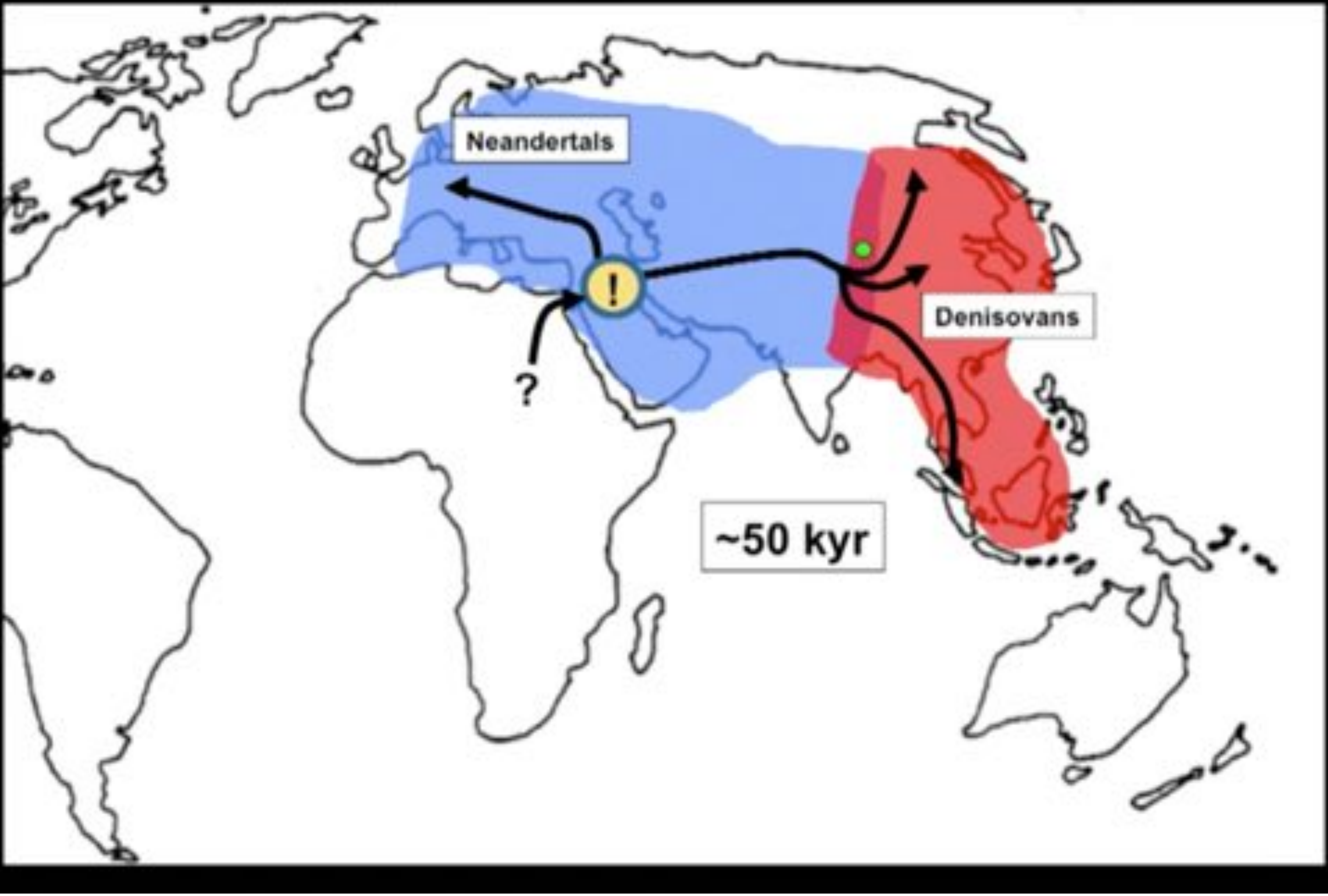
Timeline of ancient hominids



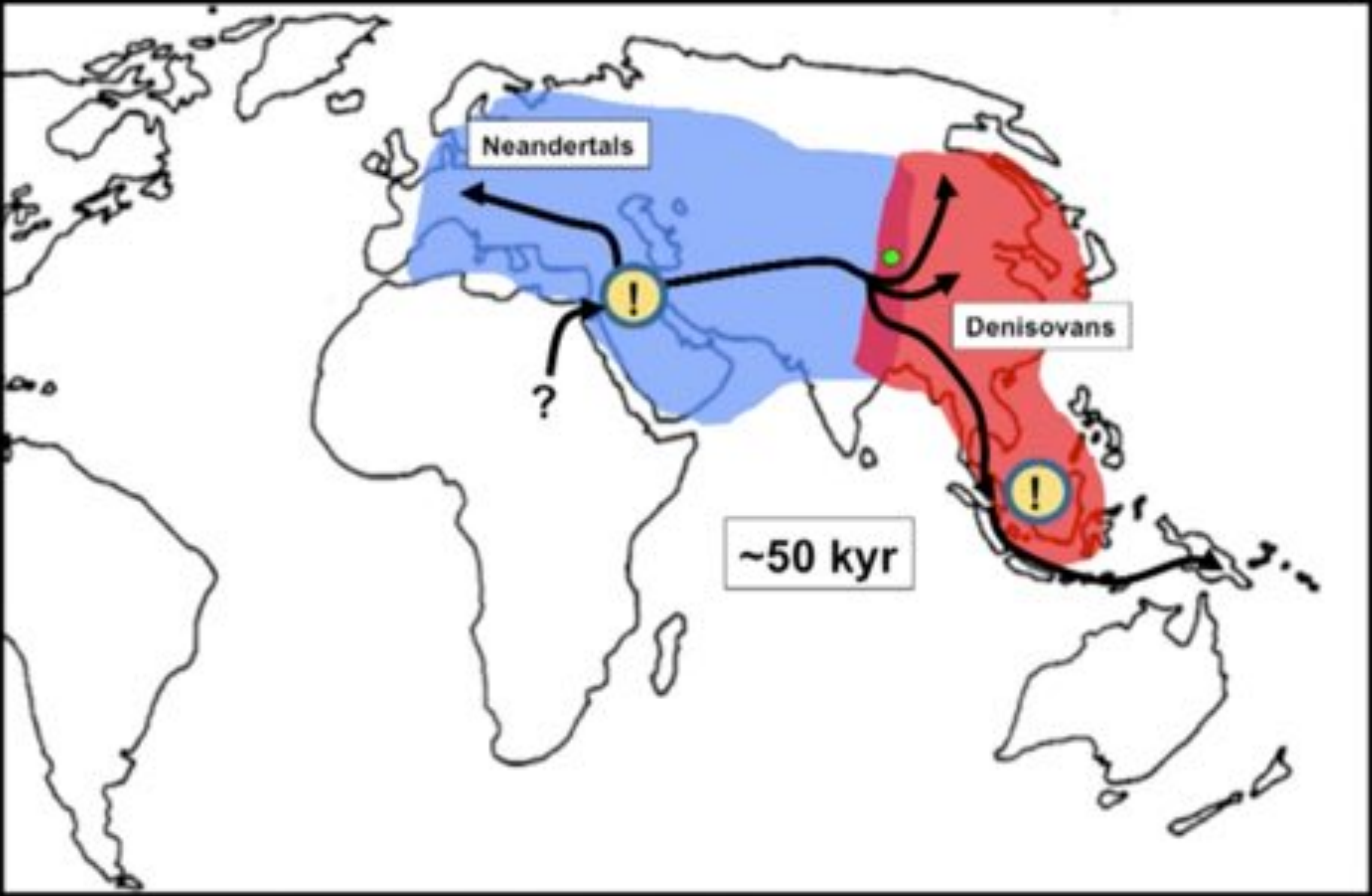
Timeline of ancient hominids



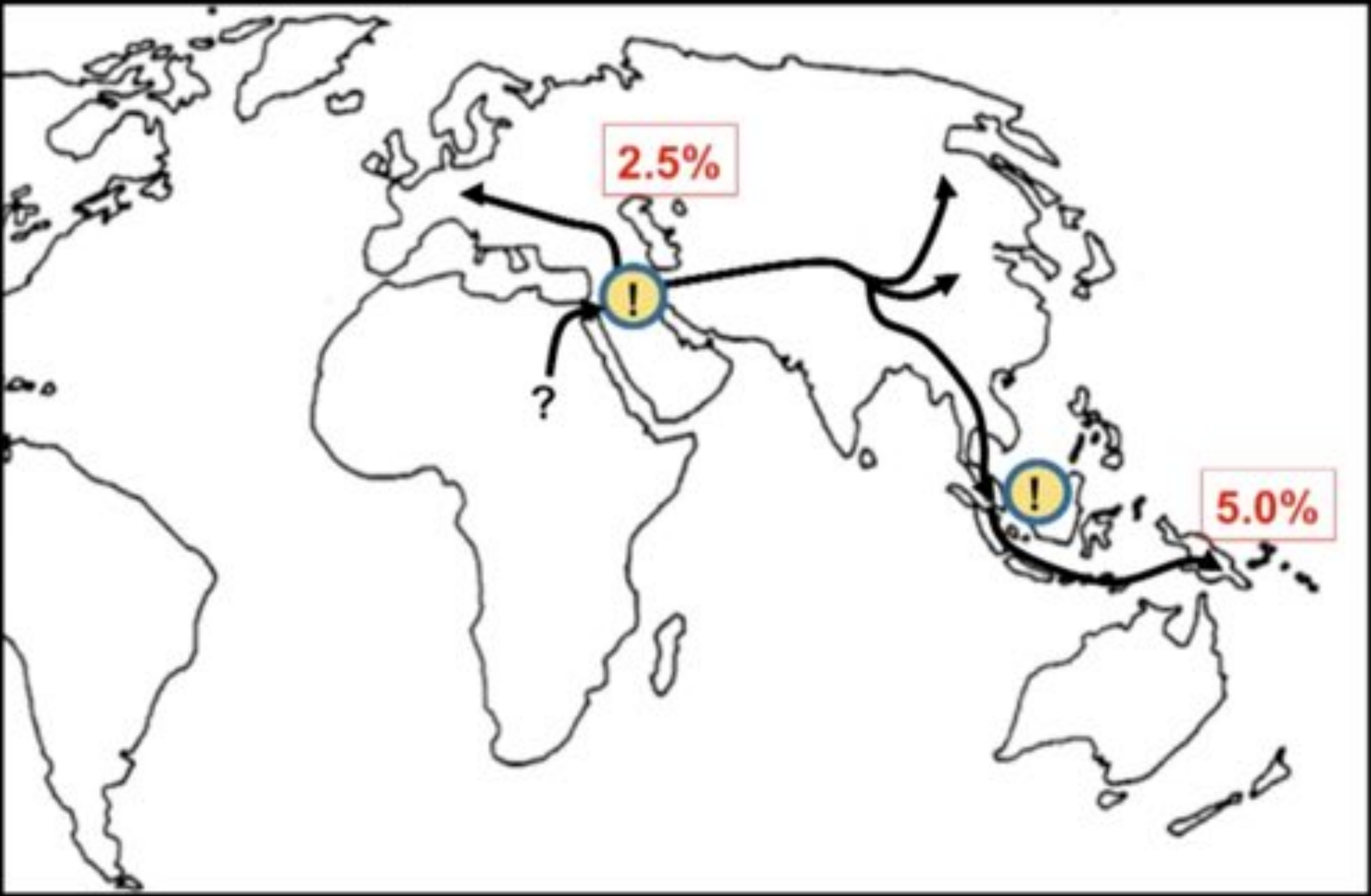
Timeline of ancient hominids

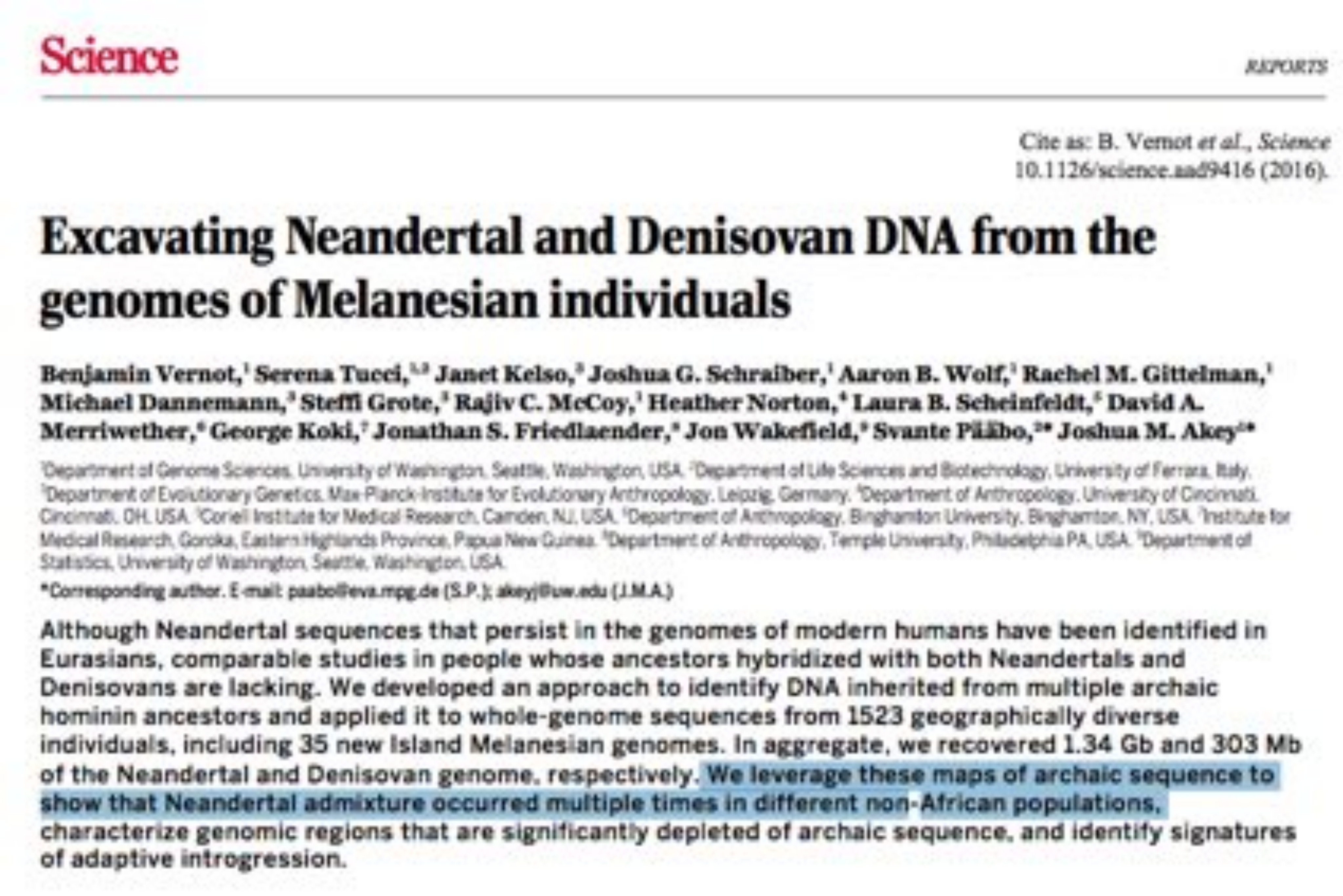


Timeline of ancient hominids



Timeline of ancient hominids





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.); akeyj@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.

Recipe for a modern human

109,295	single nucleotide changes (SNCs)
7,944	insertions and deletions

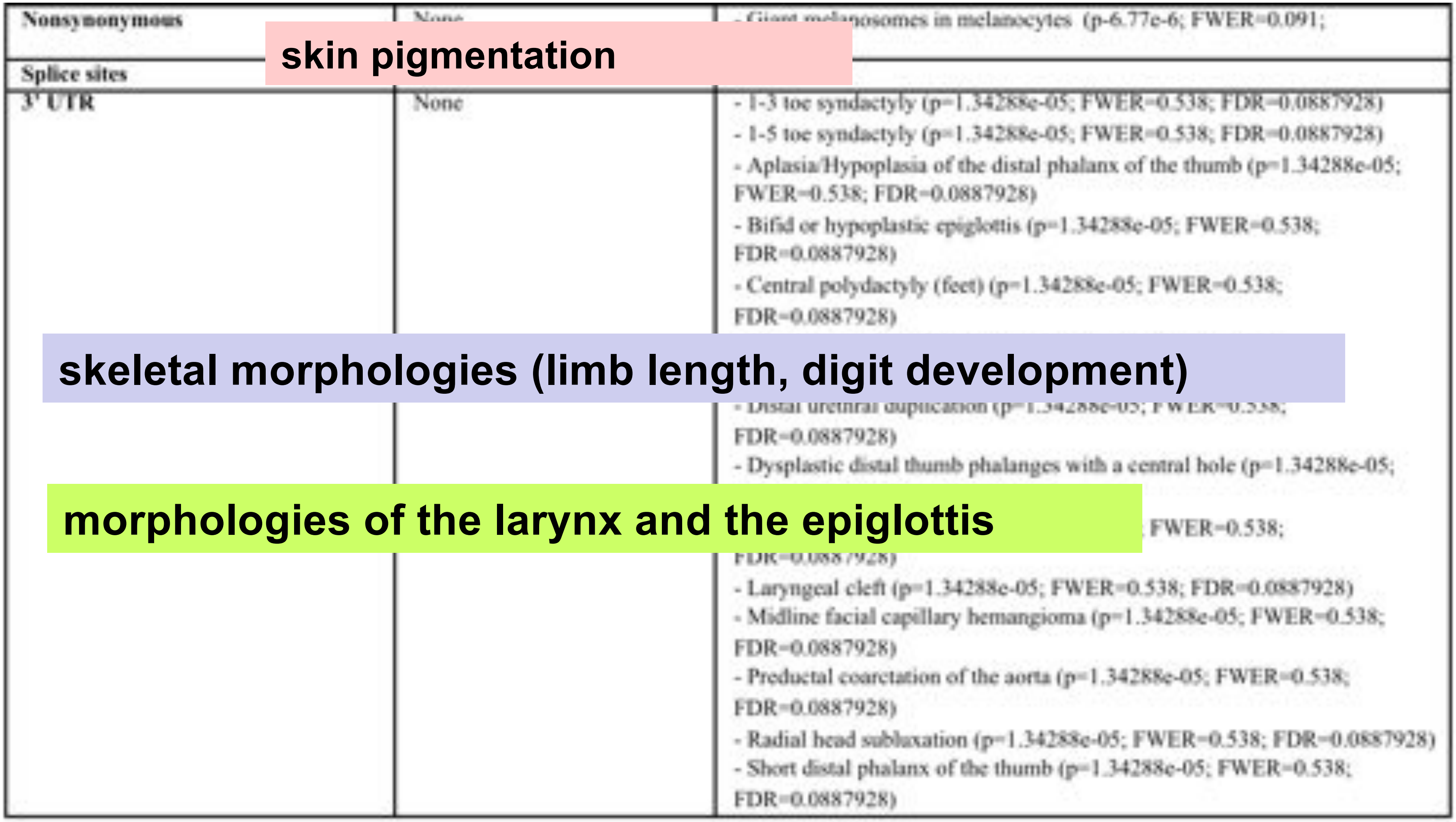
Changes in protein coding genes

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



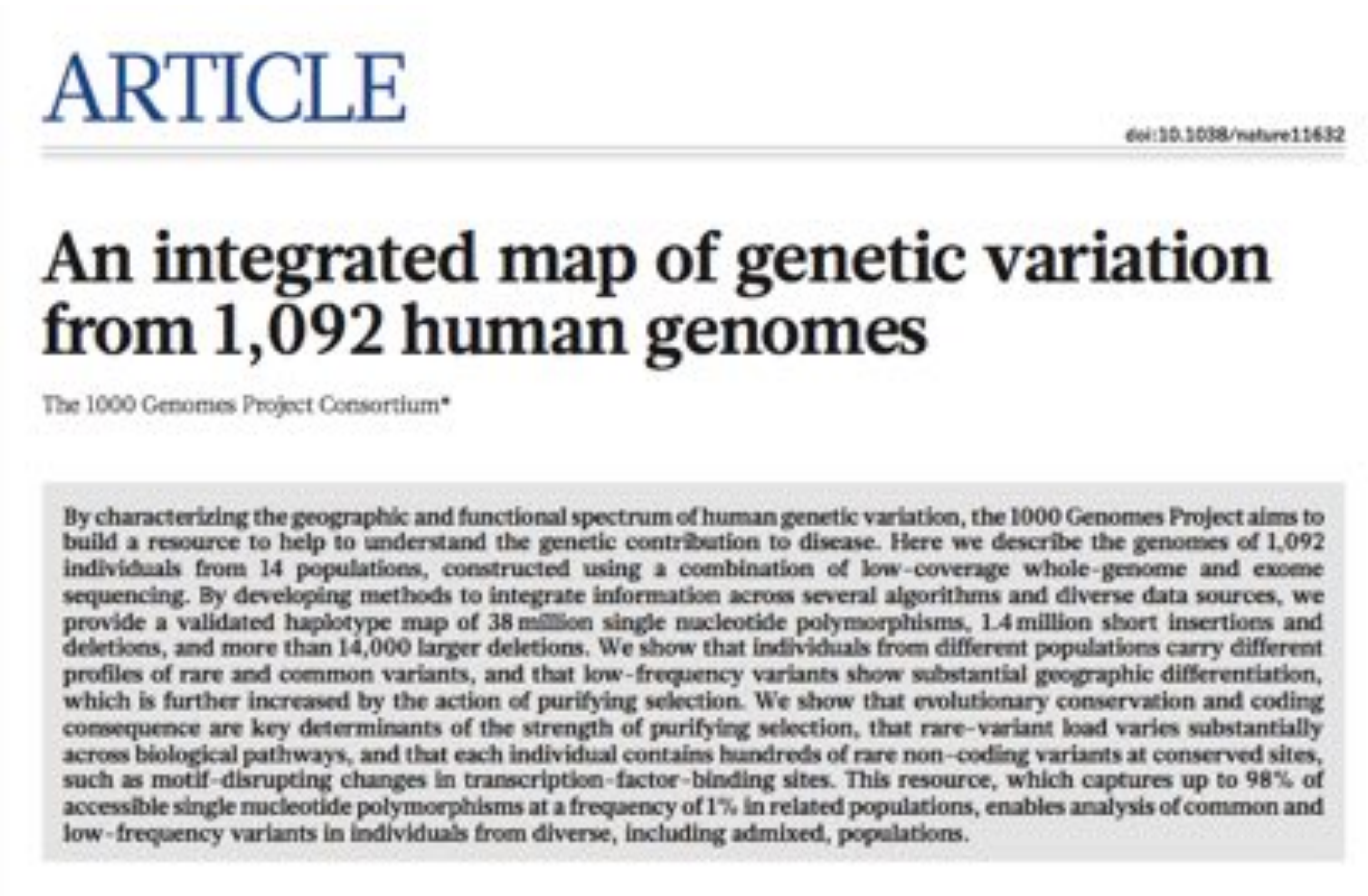
Nonsynonymous	None	- Giant melanosomes in melanocytes (p-6.77e-6; FWER=0.091;
Splice sites		
3' UTR	None	- 1-3 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - 1-5 toe syndactyly (p=1.34288e-05; FWER=0.538; FDR=0.0887928) - Aplasia/Hypoplasia of the distal phalanx of the thumb (p=1.34288e-05; 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)

skin pigmentation

skeletal morphologies (limb length, digit development)

morphologies of the larynx and the epiglottis

Part 3: Modern Humans



ARTICLE

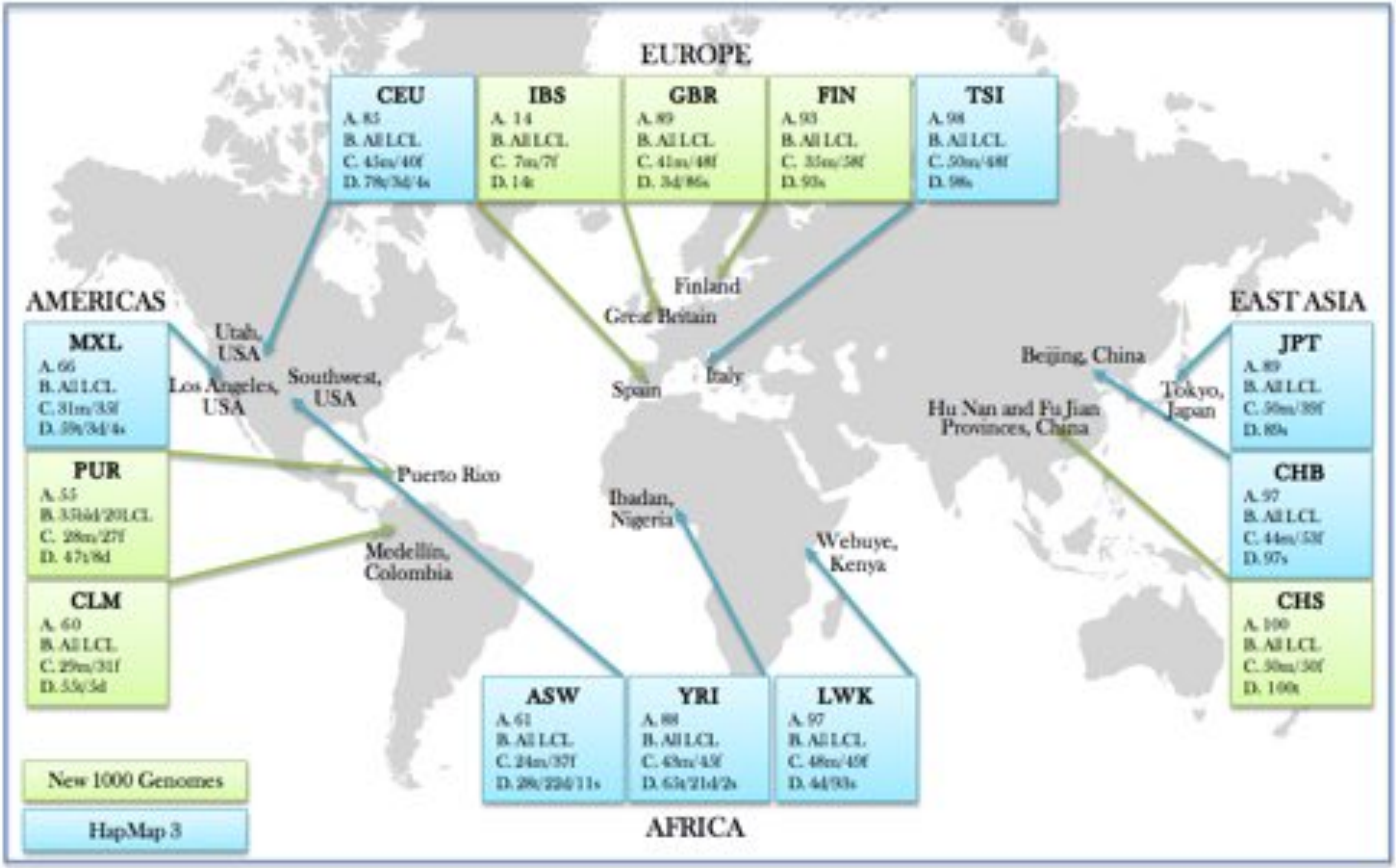
doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

1000 Genomes Populations

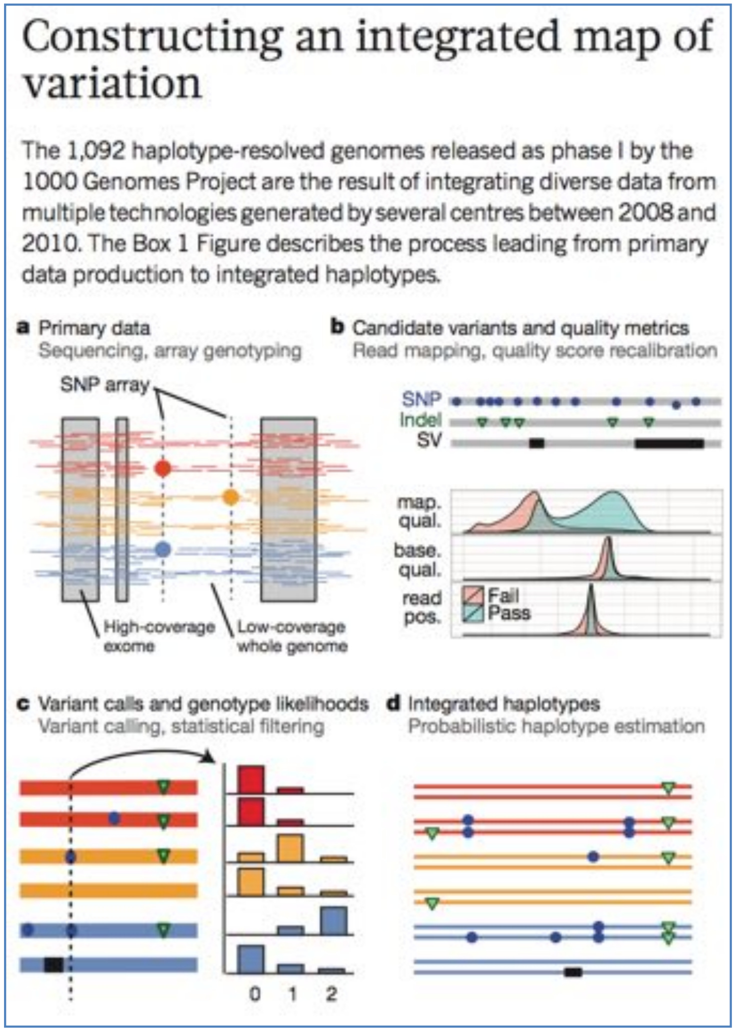


1000 Genomes: Human Mutation Rate

- Phase 1 Release
  - 1092 individuals from 14 populations
  - Combination of low coverage WGS, deep coverage WES, and SNP genotype data

- Overall SNP rate between any two people is ~1/1200bp to ~1/1300
  - ~3M SNPs between me and you (.1%)
  - ~30M SNPs between human to Chimpanzees (1%)

- De novo mutation rate ~1/100,000,000
  - ~100 de novo mutations from generation to generation
  - ~1-2 de novo mutations within the protein coding genes



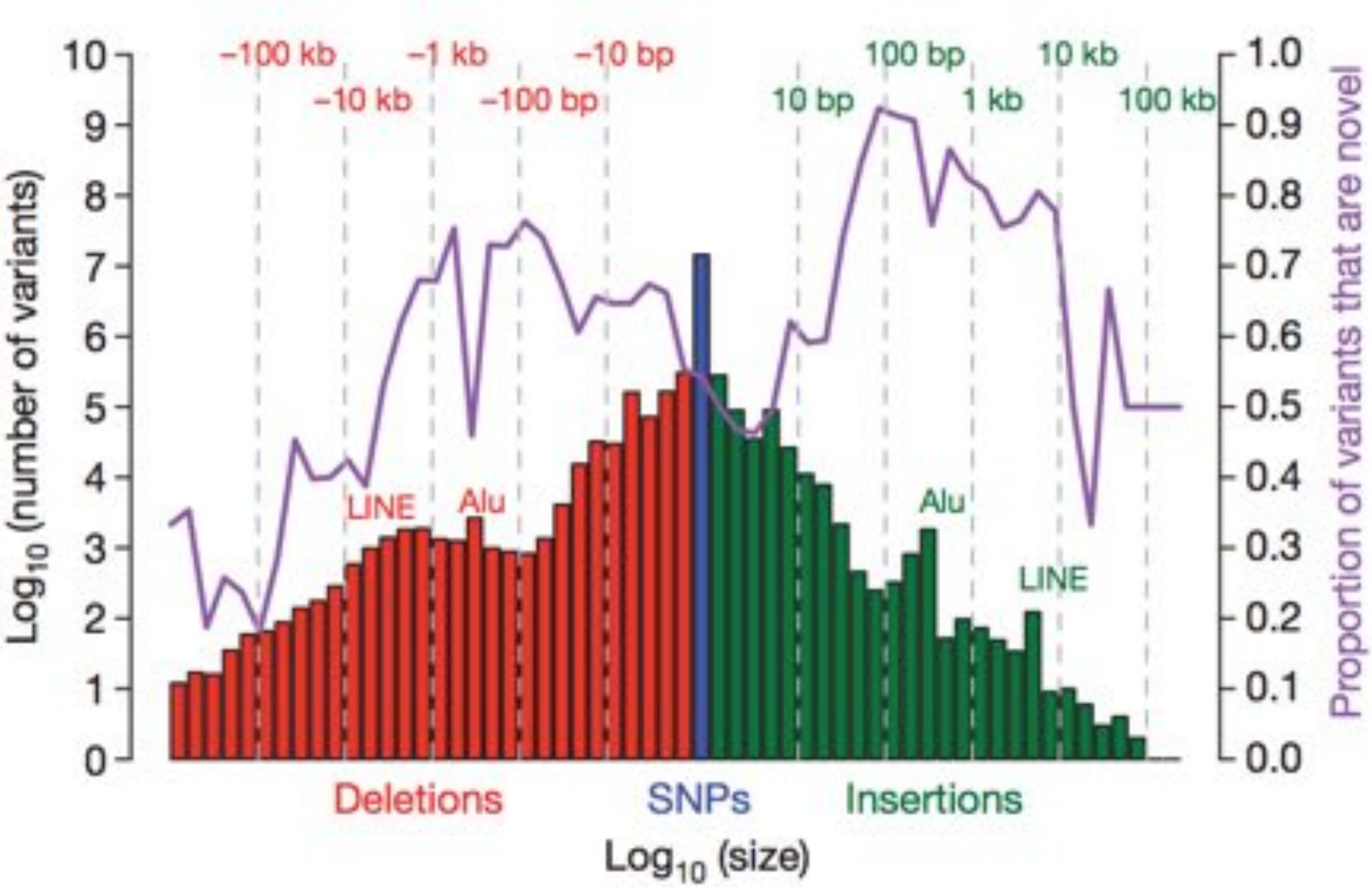
Constructing an integrated map of variation

The 1,092 haplotype-resolved genomes released as phase I by the 1000 Genomes Project are the result of integrating diverse data from multiple technologies generated by several centres between 2008 and 2010. The Box 1 Figure describes the process leading from primary data production to integrated haplotypes.

a Primary data
Sequencing, array genotyping
SNP array
High-coverage exome
Low-coverage whole genome
b Candidate variants and quality metrics
Read mapping, quality score recalibration
SNP
Indel
SV
map. qual.
base. qual.
read pos.
Fail
Pass
c Variant calls and genotype likelihoods
Variant calling, statistical filtering
0
1
2
d Integrated haplotypes
Probabilistic haplotype estimation

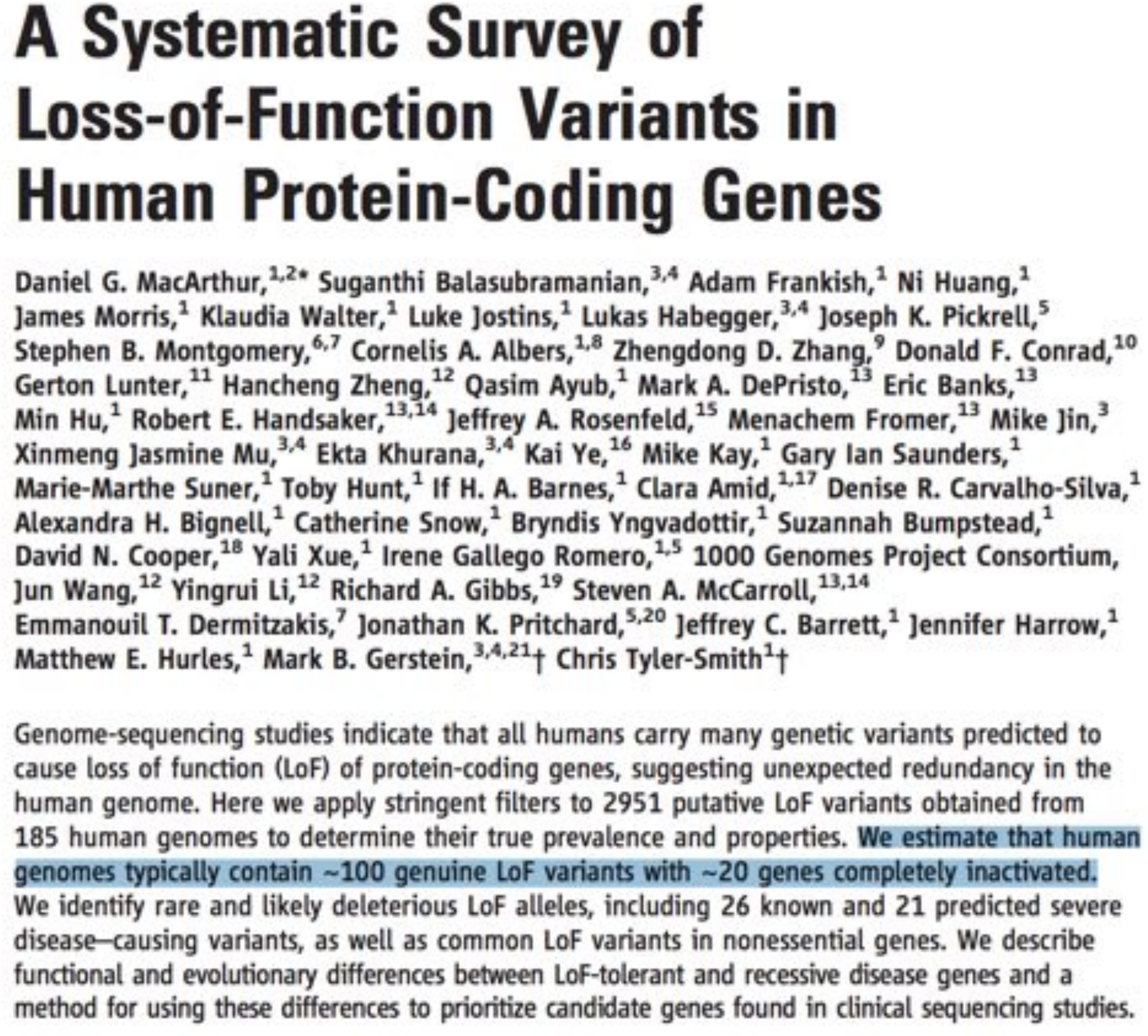
An integrated map of genetic variation from 1,092 human genomes
1000 genomes project (2012) *Nature.* doi:10.1038/nature11632

Human Mutation Types



- Mutations follows a “log-normal” frequency distribution
  - Most mutations are SNPs followed by small indels followed by larger events

A map of human genome variation from population-scale sequencing
1000 genomes project (2010) *Nature*. doi:10.1038/nature09534



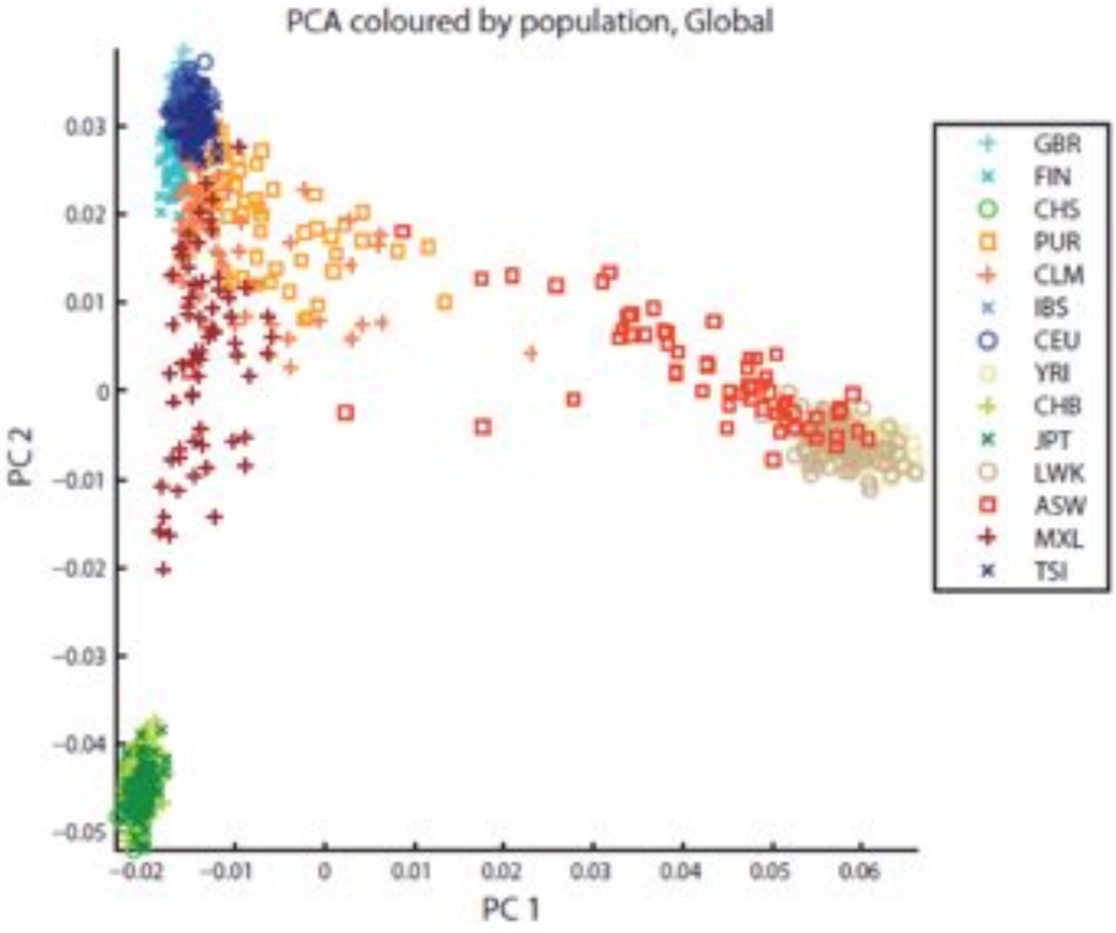
A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes

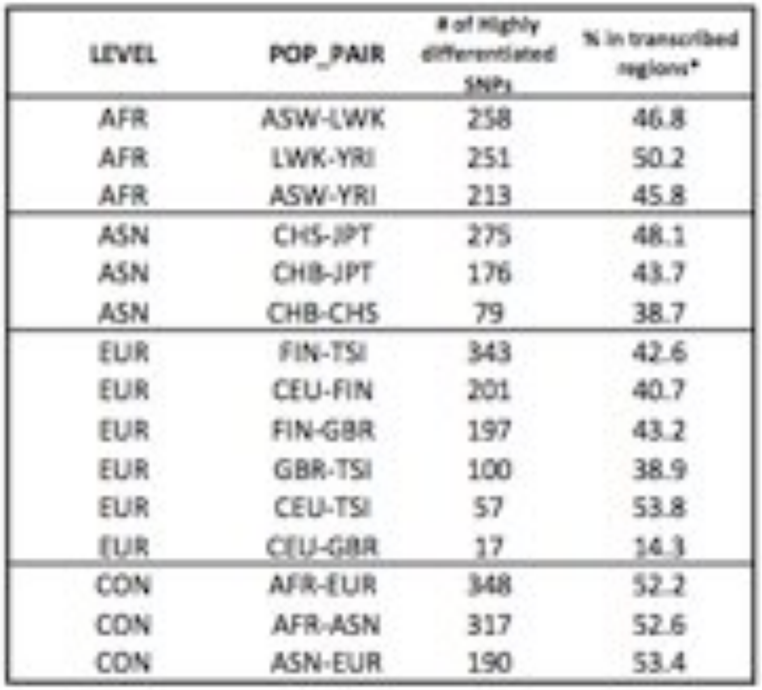
Daniel G. MacArthur,[1,2]* Suganthi Balasubramanian,[3,4] Adam Frankish,[1] Ni Huang,[1] James Morris,[1] Klaudia Walter,[1] Luke Jostins,[1] Lukas Habegger,[3,4] Joseph K. Pickrell,[5] Stephen B. Montgomery,[6,7] Cornelis A. Albers,[1,8] Zhengdong D. Zhang,[9] Donald F. Conrad,[10] Gerton Lunter,[11] Hancheng Zheng,[12] Qasim Ayub,[1] Mark A. DePristo,[13] Eric Banks,[13] Min Hu,[1] Robert E. Handsaker,[13,14] Jeffrey A. Rosenfeld,[15] Menachem Fromer,[13] Mike Jin,[3] Xinmeng Jasmine Mu,[3,4] Ekta Khurana,[3,4] Kai Ye,[16] Mike Kay,[1] Gary Ian Saunders,[1] Marie-Marthe Suner,[1] Toby Hunt,[1] If H. A. Barnes,[1] Clara Amid,[1,17] Denise R. Carvalho-Silva,[1] Alexandra H. Bignell,[1] Catherine Snow,[1] Bryndis Yngvadottir,[1] Suzannah Bumpstead,[1] David N. Cooper,[18] Yali Xue,[1] Irene Gallego Romero,[1,5] 1000 Genomes Project Consortium, Jun Wang,[12] Yingrui Li,[12] Richard A. Gibbs,[19] Steven A. McCarroll,[13,14] Emmanouil T. Dermitzakis,[7] Jonathan K. Pritchard,[5,20] Jeffrey C. Barrett,[1] Jennifer Harrow,[1] Matthew E. Hurles,[1] Mark B. Gerstein,[3,4,21]† Chris Tyler-Smith[1]†

Genome-sequencing studies indicate that all humans carry many genetic variants predicted to cause loss of function (LoF) of protein-coding genes, suggesting unexpected redundancy in the human genome. Here we apply stringent filters to 2951 putative LoF variants obtained from 185 human genomes to determine their true prevalence and properties. We estimate that human genomes typically contain ~100 genuine LoF variants with ~20 genes completely inactivated. We identify rare and likely deleterious LoF alleles, including 26 known and 21 predicted severe disease–causing variants, as well as common LoF variants in nonessential genes. We describe functional and evolutionary differences between LoF-tolerant and recessive disease genes and a method for using these differences to prioritize candidate genes found in clinical sequencing studies.

(2012) Science. doi: 10.1126/science.1215040

Variation across populations



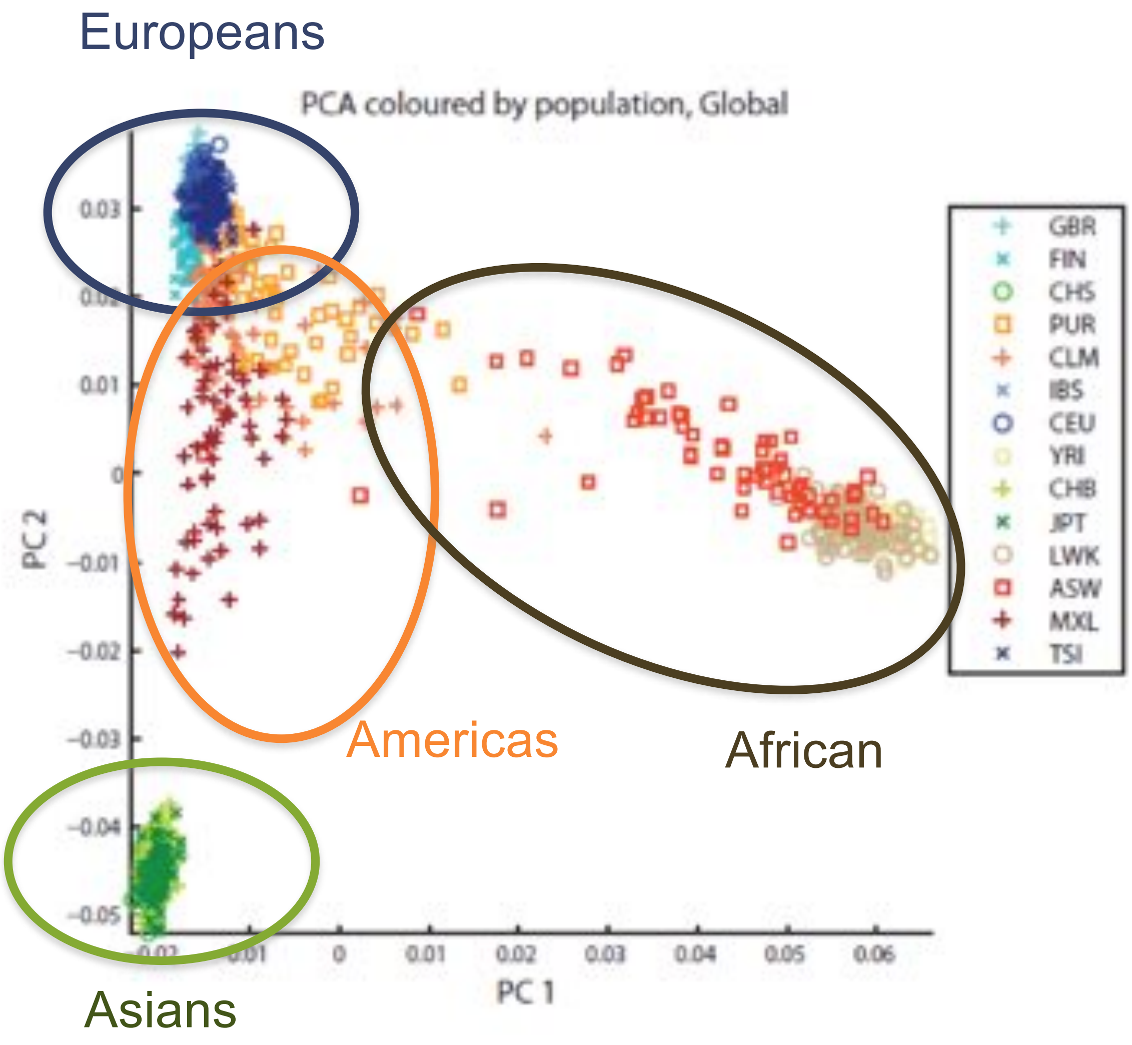


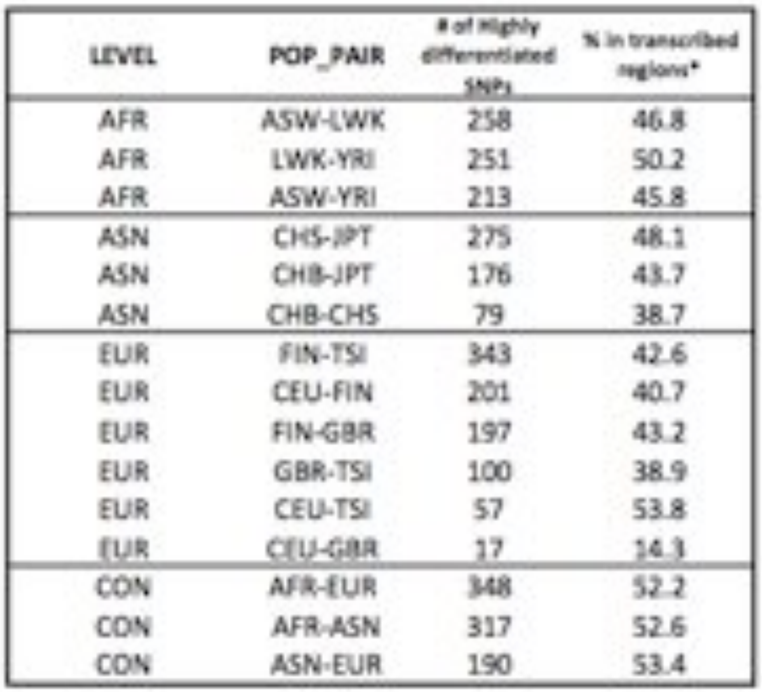
LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8
EUR	CEU-GBR	17	14.3
CON	AFR-EUR	348	52.2
CON	AFR-ASN	317	52.6
CON	ASN-EUR	190	53.4

Table S12A **Summary of sites showing high levels of population differentiation**

- Not a single variant 100% unique to a given population
- 17% of low-frequency variants (.5-5% pop. freq) observed in a single ancestry group
- 50% of rare variants (<.5%) observed in a single population

Variation across populations

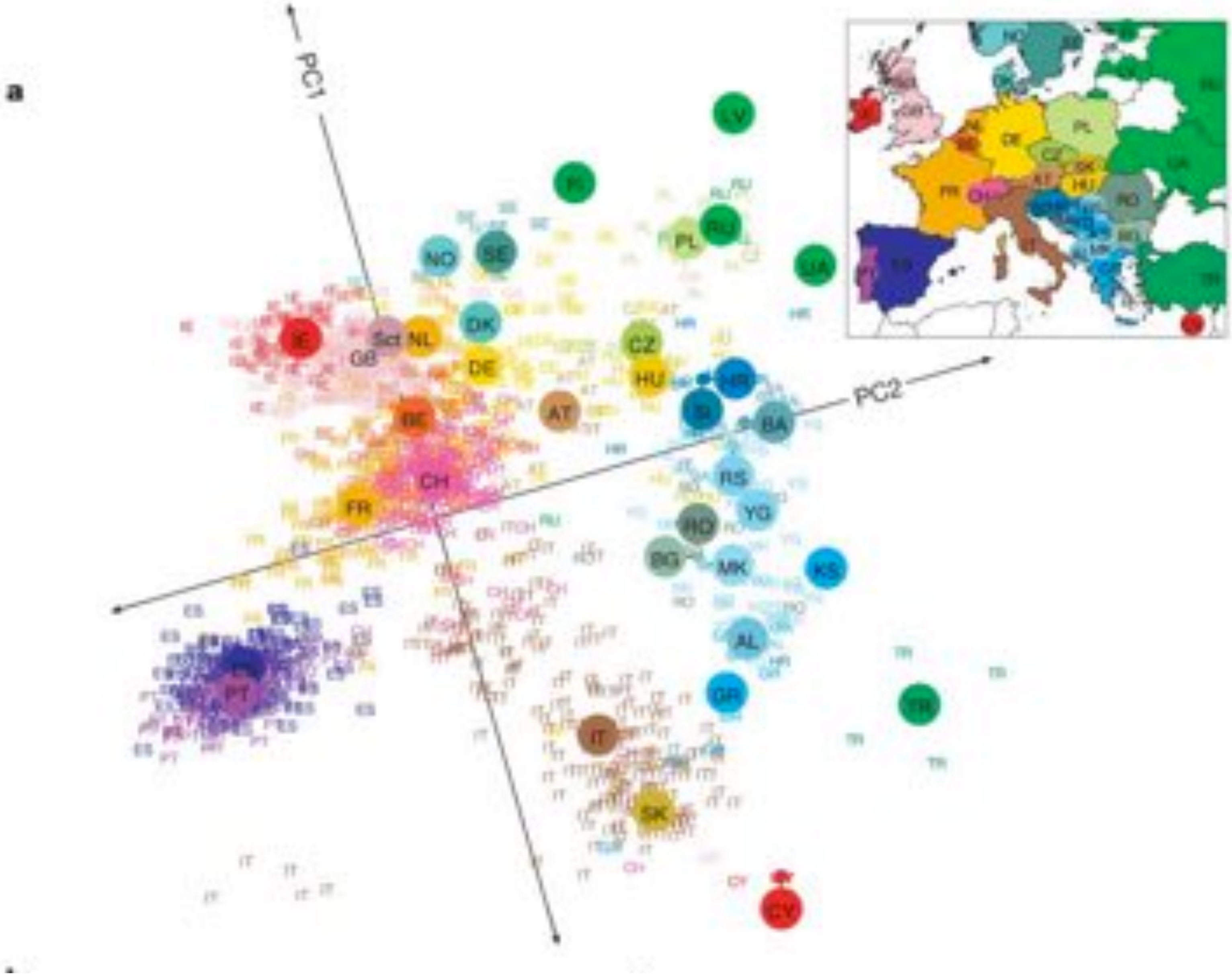




LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8
EUR	CEU-GBR	17	14.3
CON	AFR-EUR	348	52.2
CON	AFR-ASN	317	52.6
CON	ASN-EUR	190	53.4

Table S12A Summary of sites showing high levels of population differentiation

- Not a single variant 100% unique to a given population
- 17% of low-frequency variants (.5-5% pop. freq) observed in a single ancestry group
- 50% of rare variants (<.5%) observed in a single population



Genes mirror geography within Europe

Novembre et al (2008) Nature. doi: 10.1038/nature07331

Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Work on HW3