Lecture 19. Ancient and Modern Humans

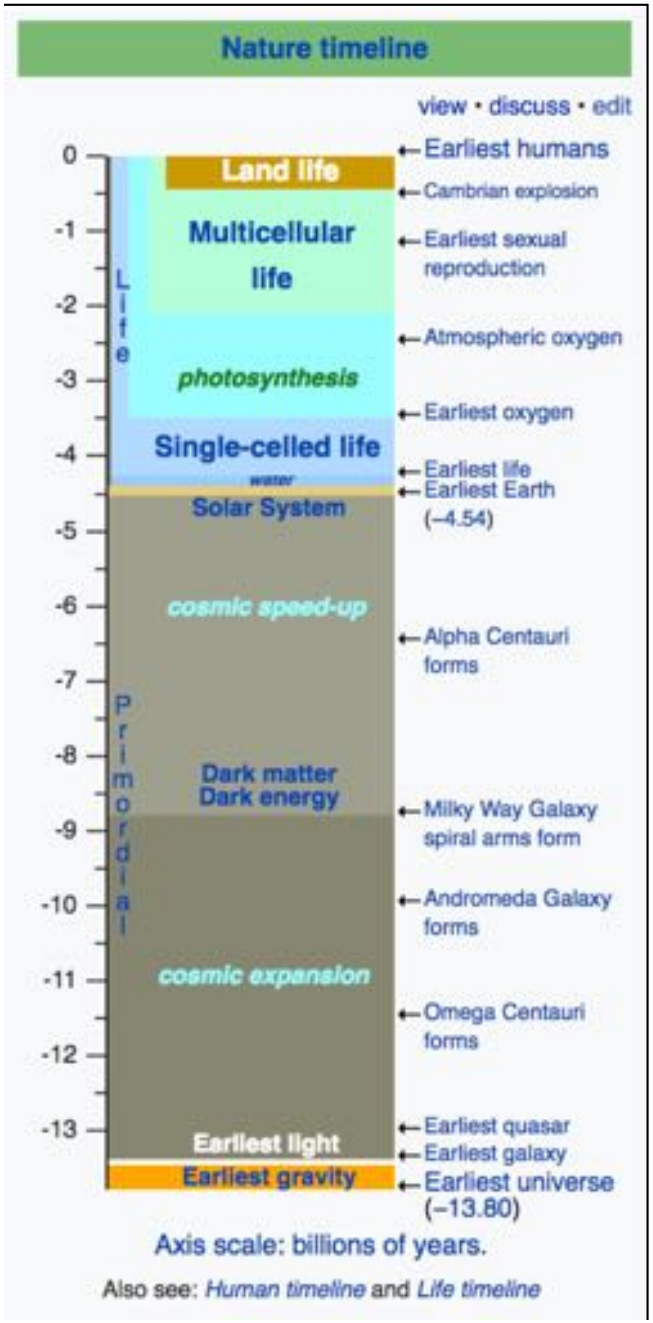
Michael Schatz

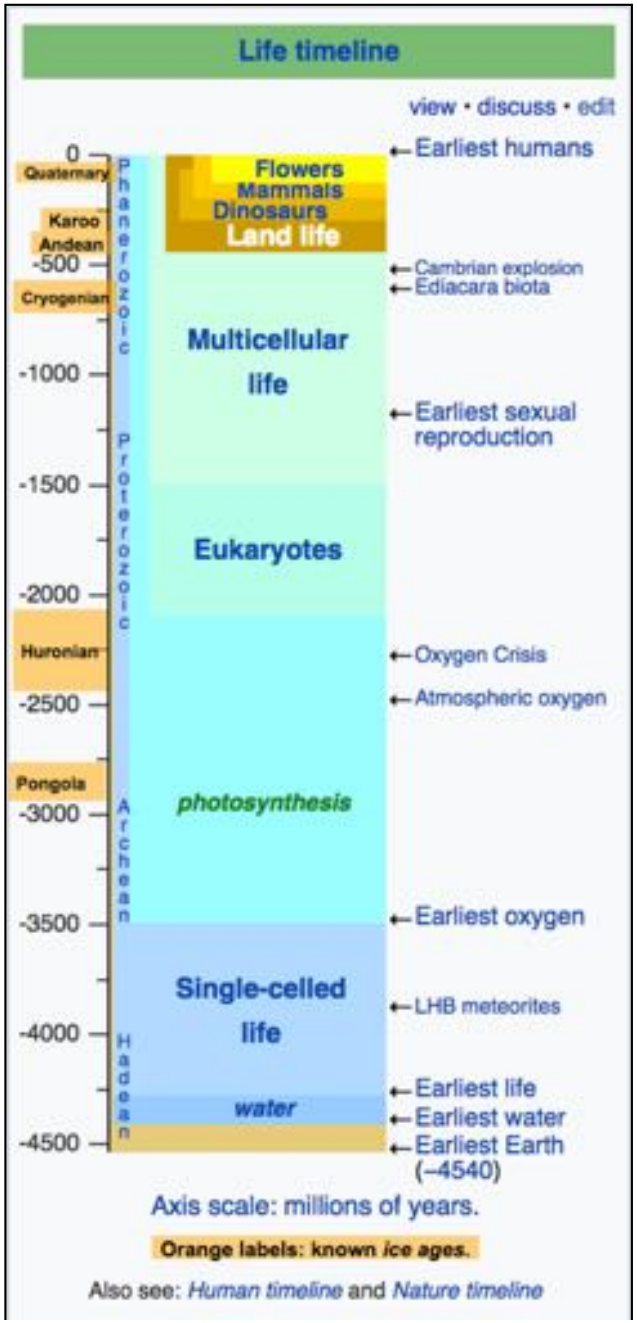
April 10, 2017

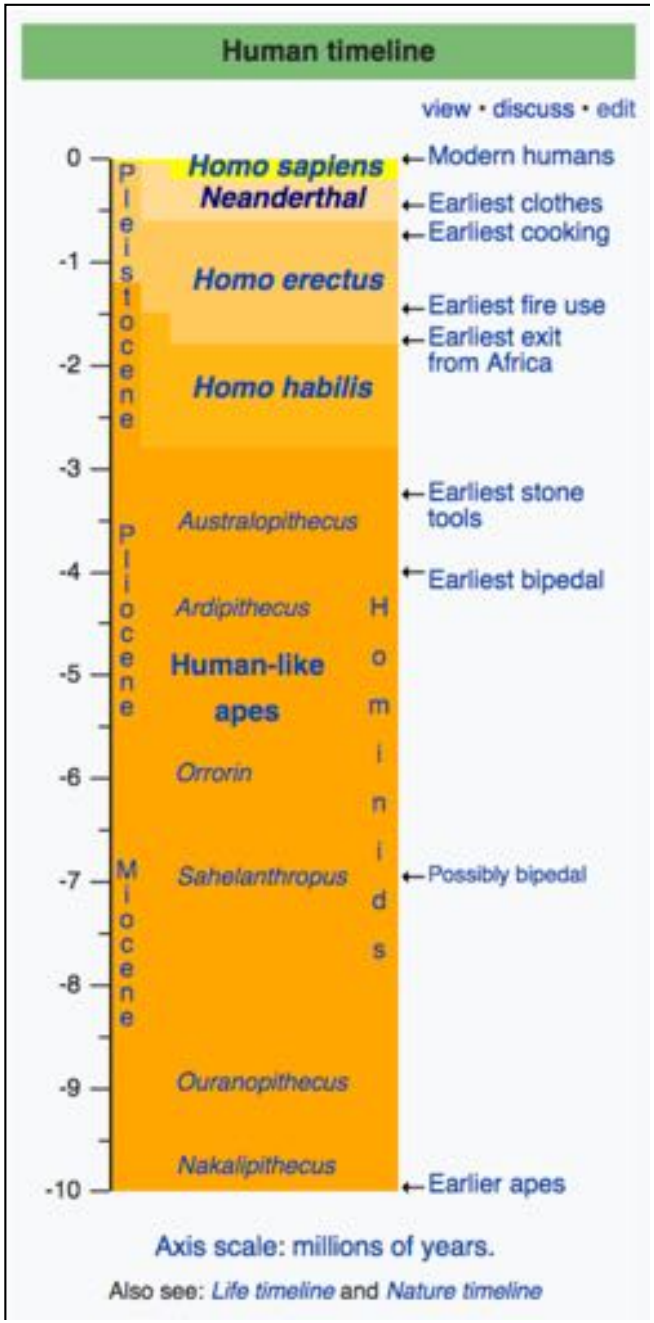
JHU 600.749: Applied Comparative Genomics

Part 1:
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
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

Sequencing ancient genomes

Janet Kelso
Max-Planck Institute


Engis
Neandertal
Spy
Arcy-sur-Cure
La Quina
Saint-Césaire
Le Moustier
La Ferrassie
Maula
Steinheim
Châtelperron
La Chapelle-aux-Saints
Saccopastore
Guattari
Figueira Brava
Zafarraya
Forbe's Quarry
Külna
Sipka
Tata
Erd
Krapina
Vindija
Sukhaya Mechetka
Molodovo
Kiyik-Koba
Starosillya
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 ☺

Science
AAAS

A Draft Sequence of the Neandertal Genome

Richard E. Green, *et al.*

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

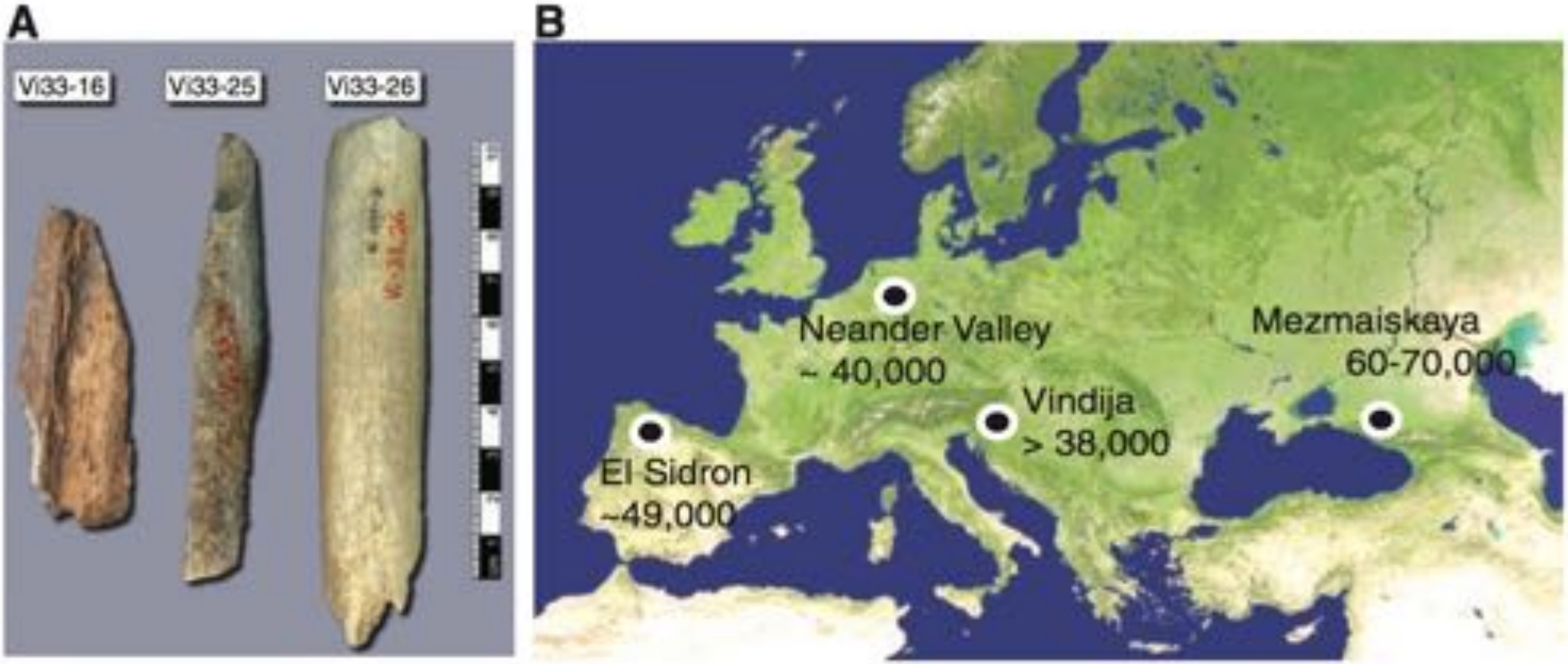
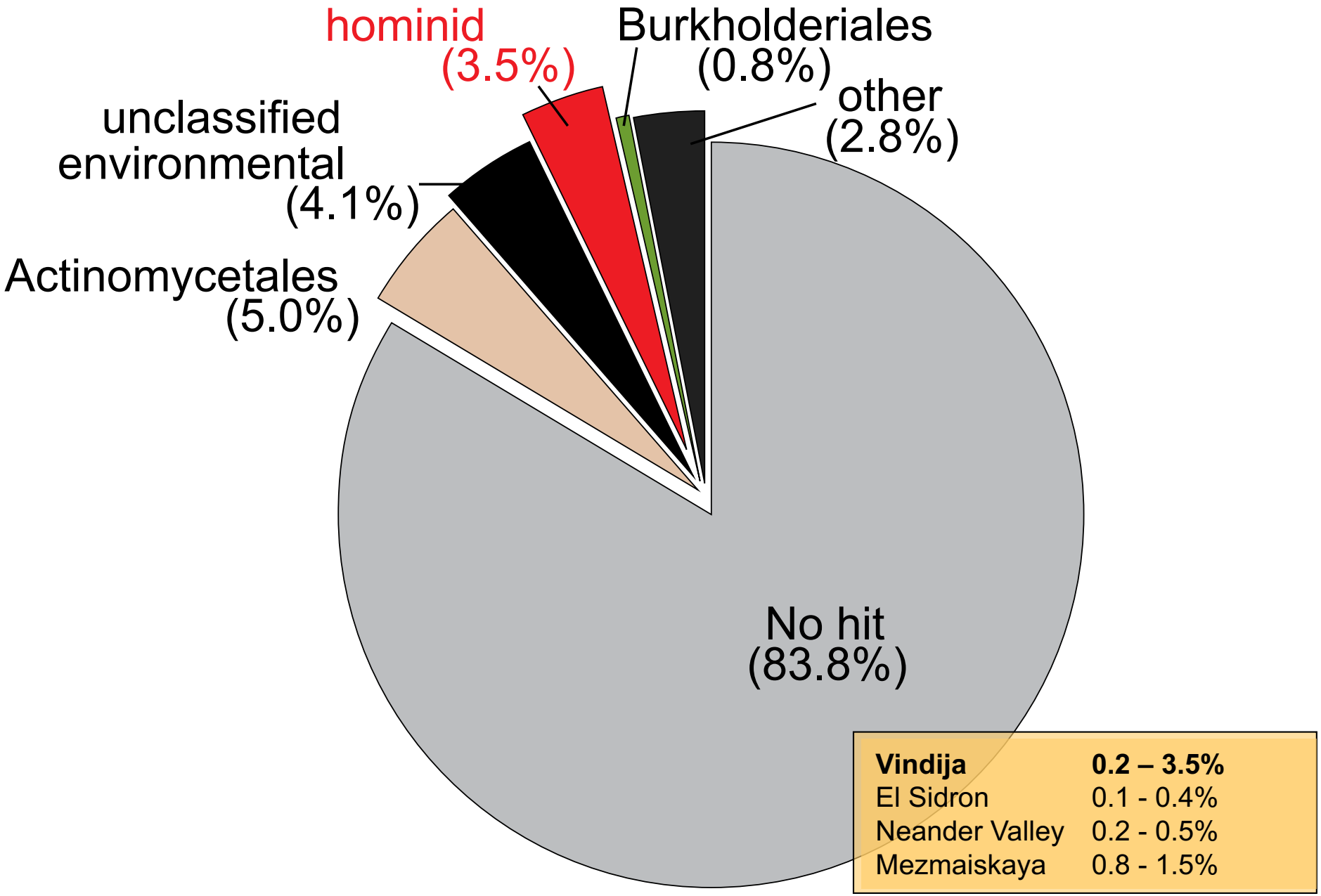


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

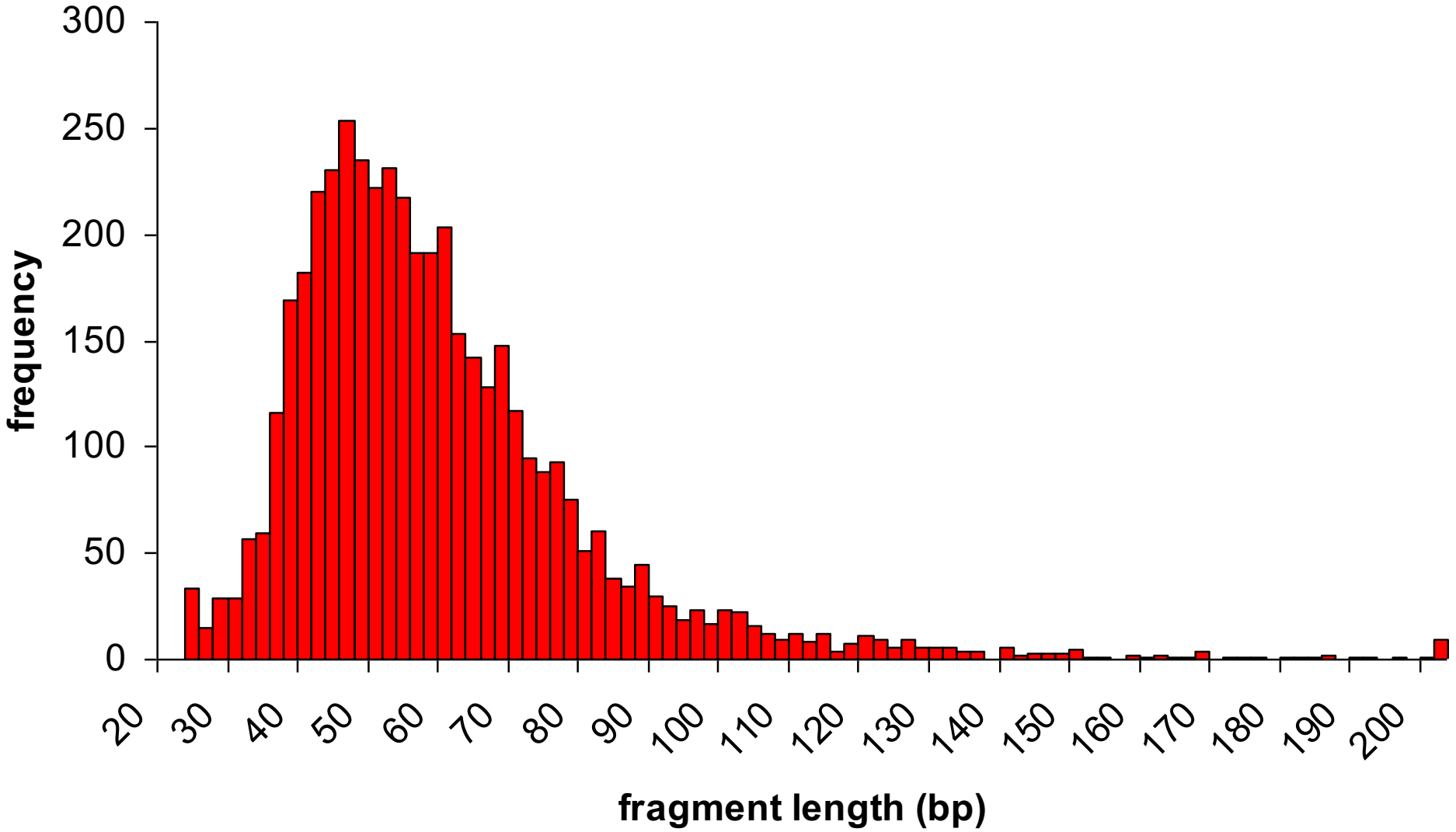
Extracting Ancient DNA

DNA is from mixed sources

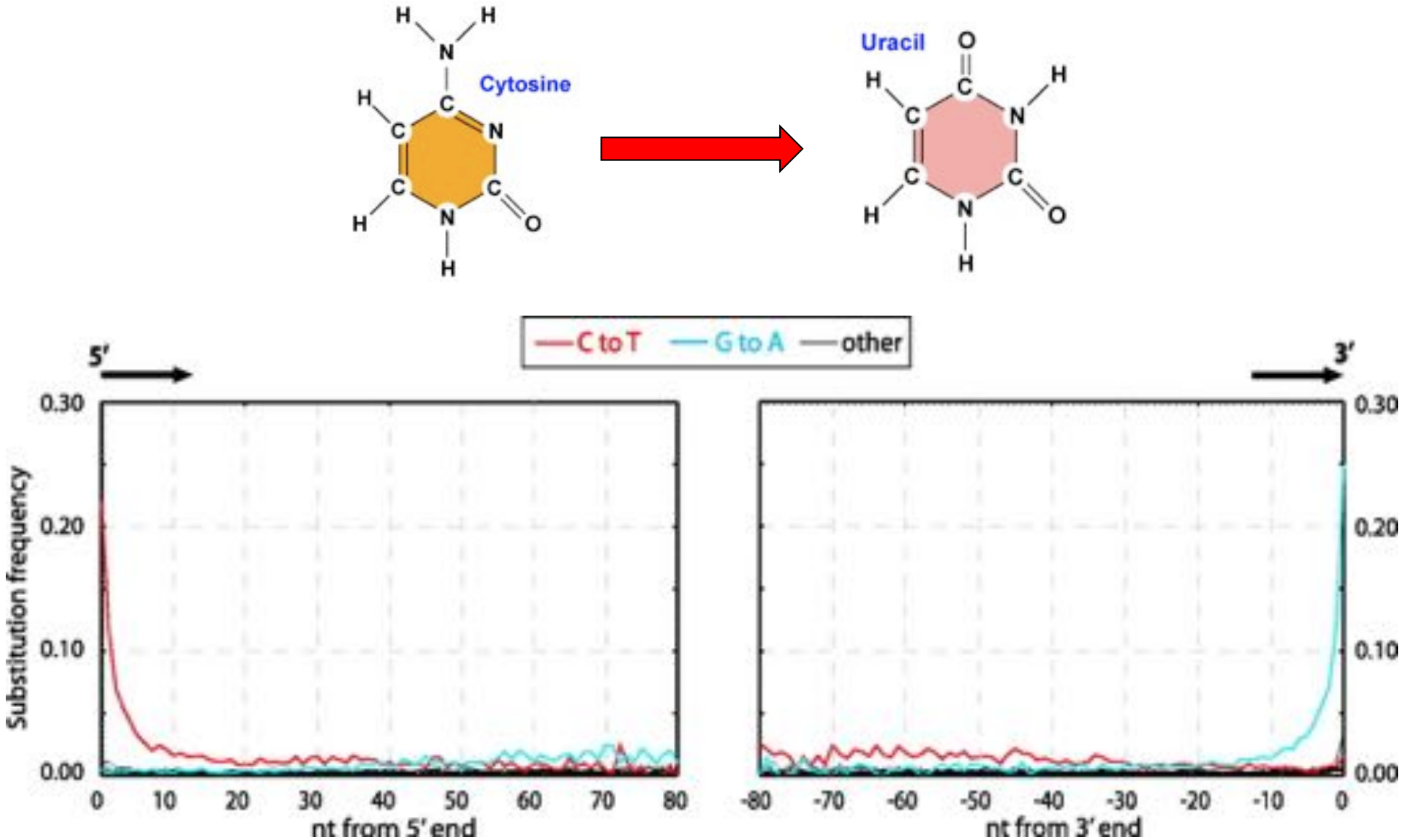


Vindija	**0.2 – 3.5%**
El Sidron	0.1 - 0.4%
Neander Valley	0.2 - 0.5%
Mezmaiskaya	0.8 - 1.5%

DNA is degraded



DNA is chemically damaged



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

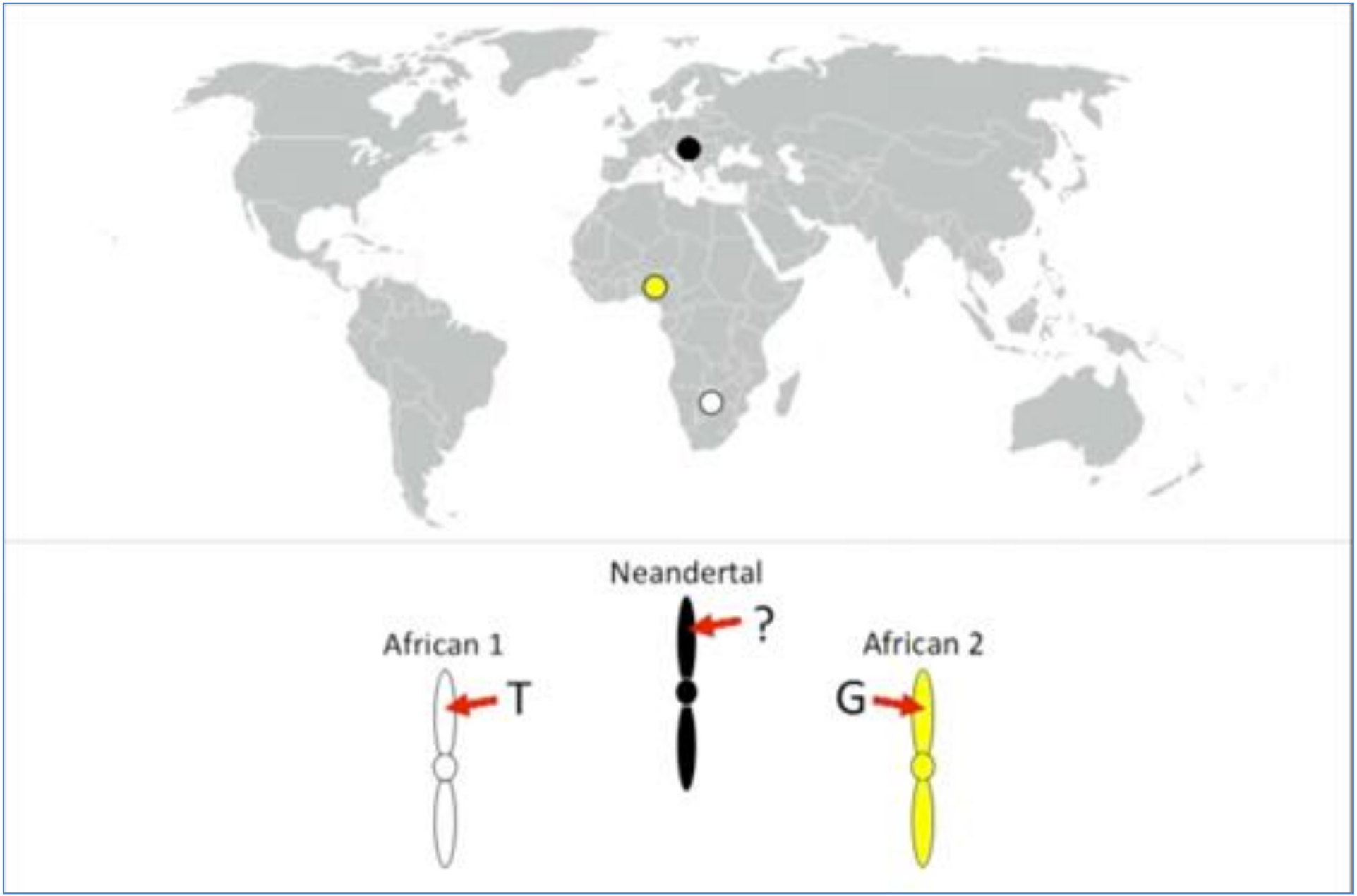
Green et al. 2010

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

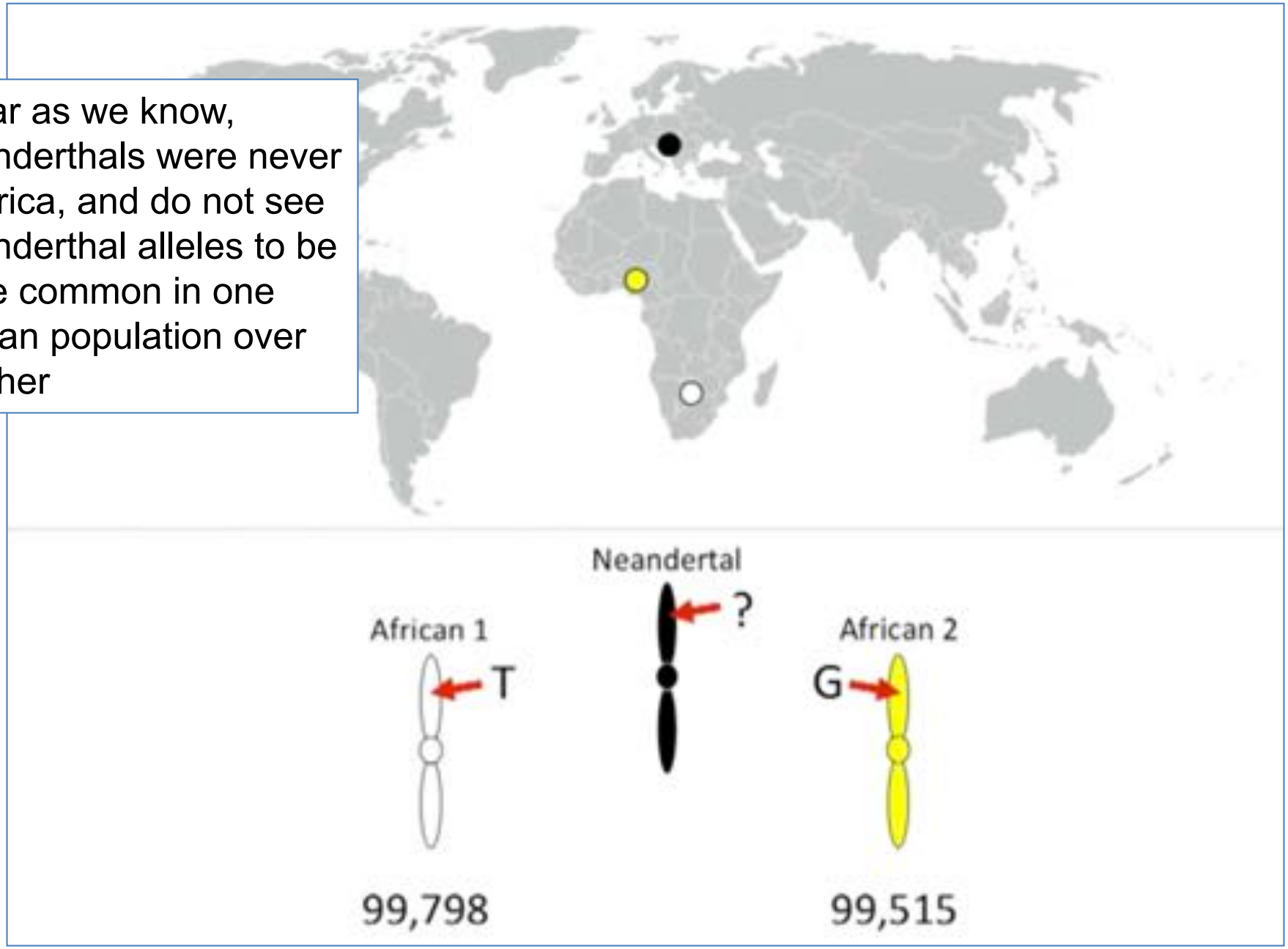
Genome coverage ~1.3 X

Did we mix?



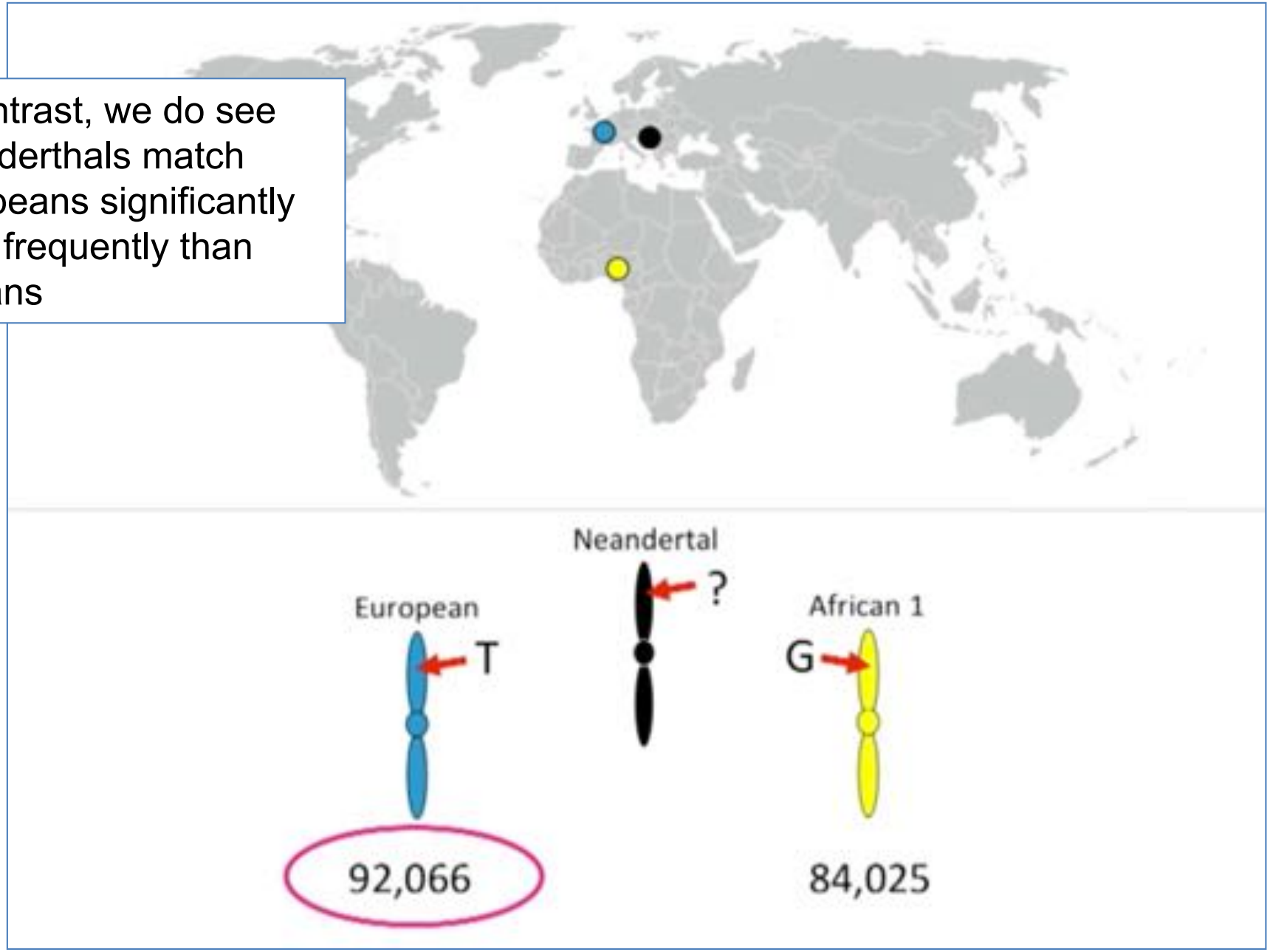
Did we mix?

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?

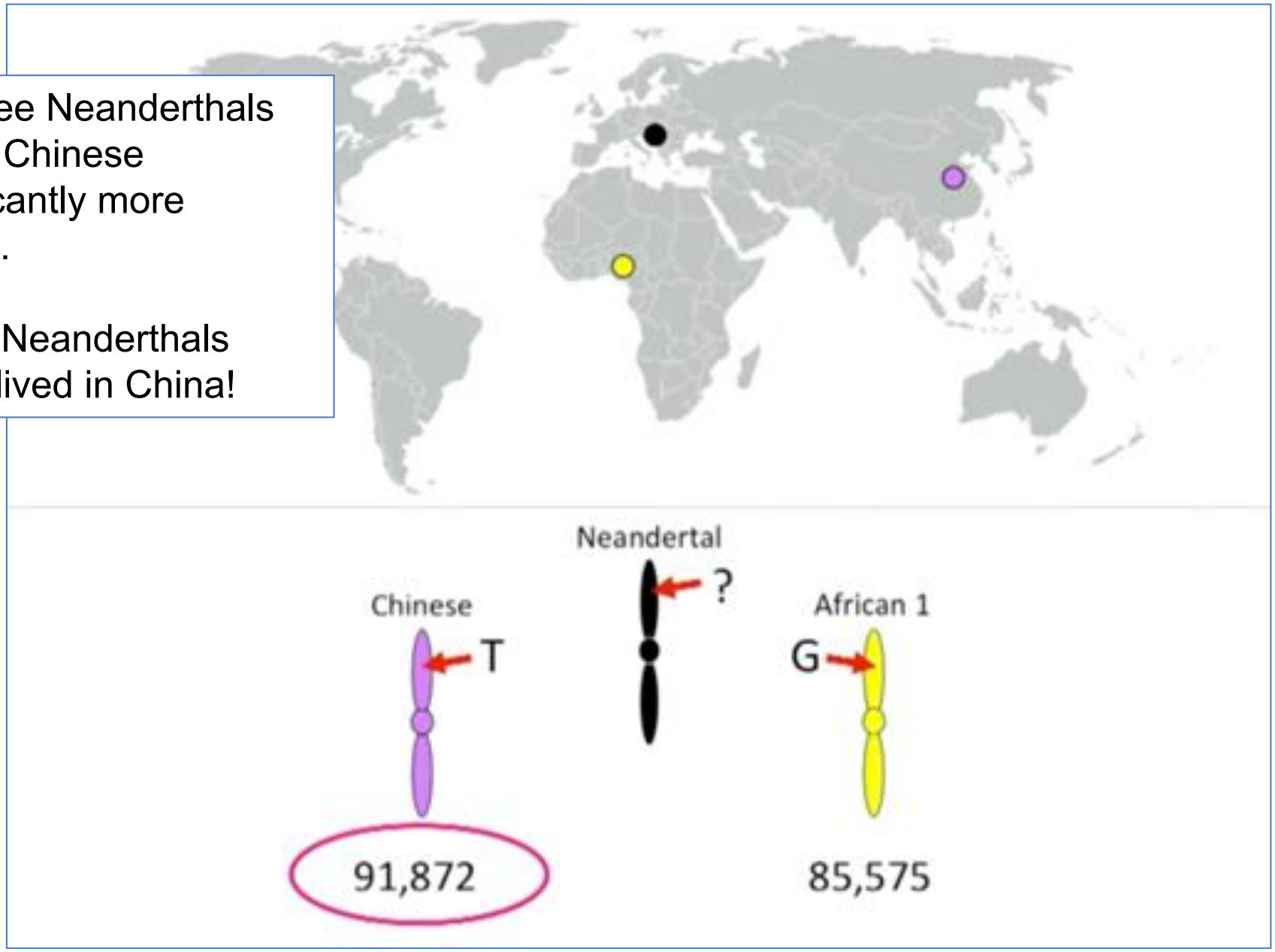
In contrast, we do see Neanderthals match Europeans significantly more frequently than Africans



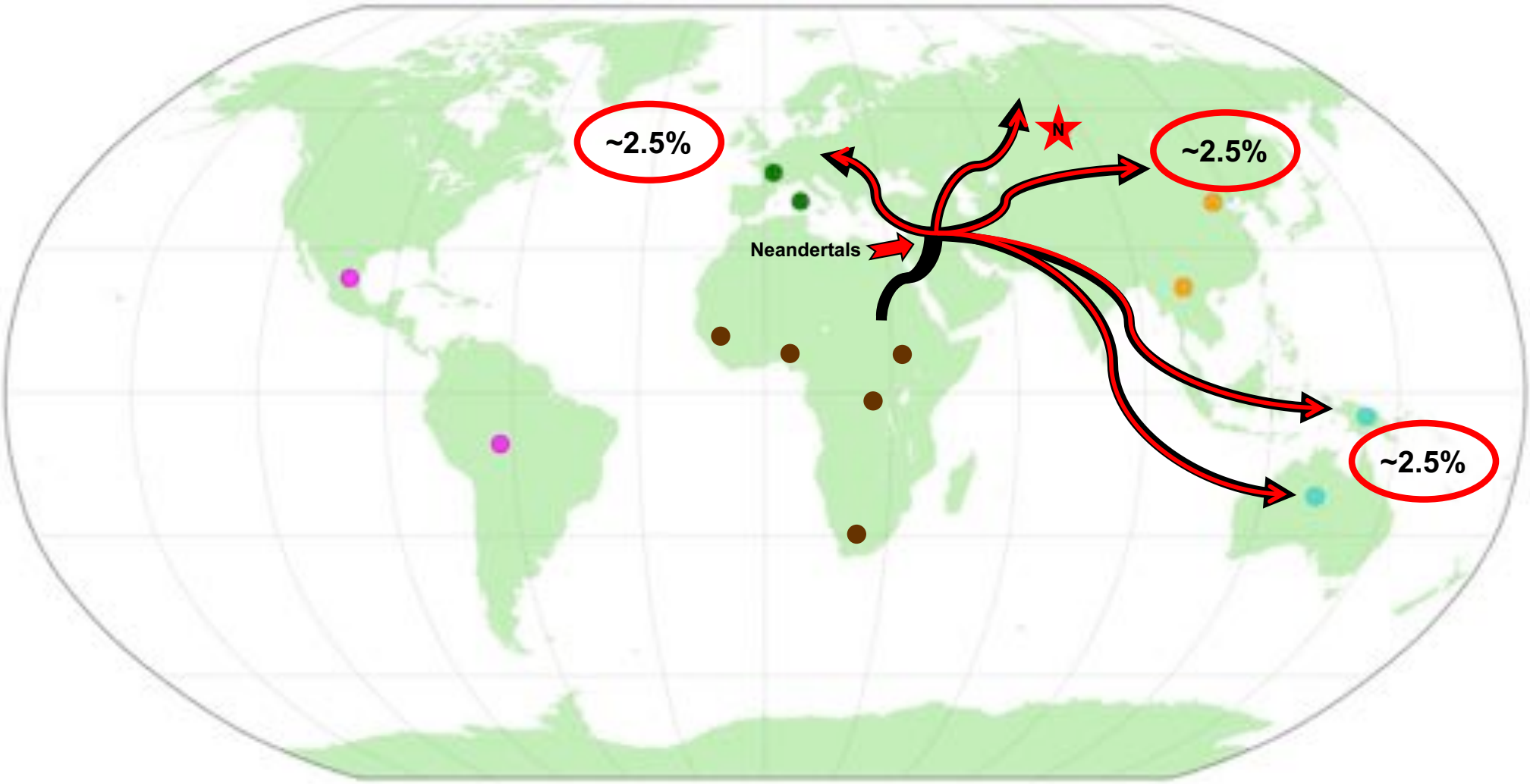
Did we mix?

Also see Neanderthals match Chinese significantly more often…

… but Neanderthals never lived in China!



Neanderthal Interbreeding

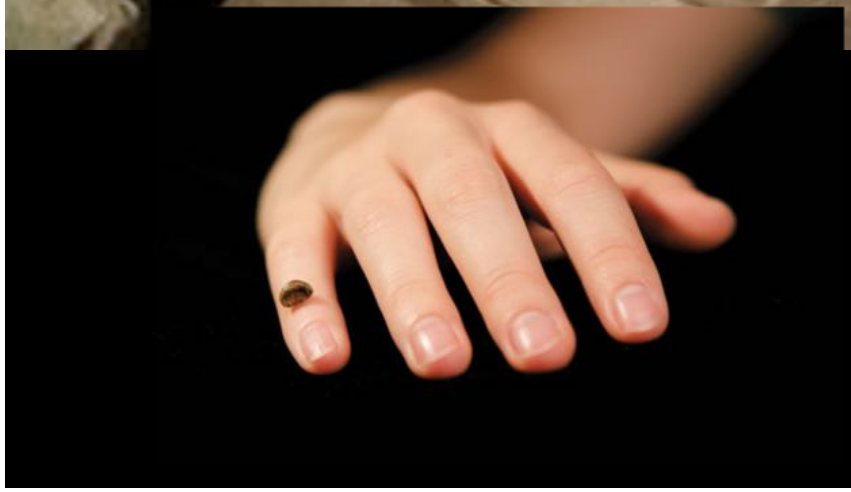


As modern humans migrated out of Africa, they apparently interbred with Neanderthal's so we see their alleles across the rest of the world and carry about 2.5% of their genome with us!

What about other ancient hominids?

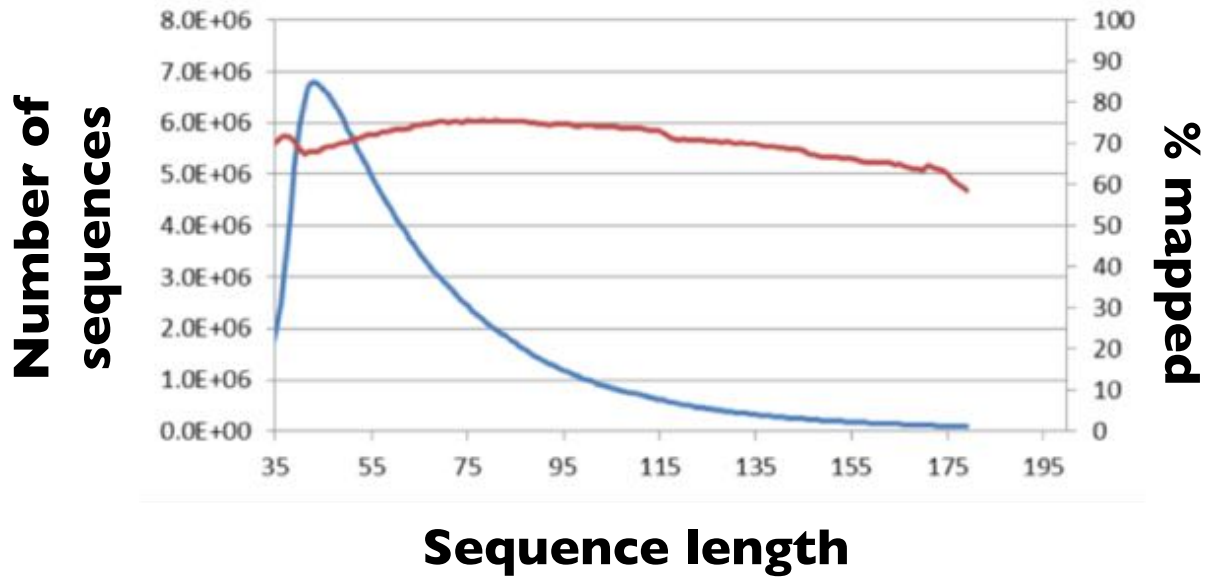
Denisova cave
Altai mountains
Russia

Academician A.P. Derevianko



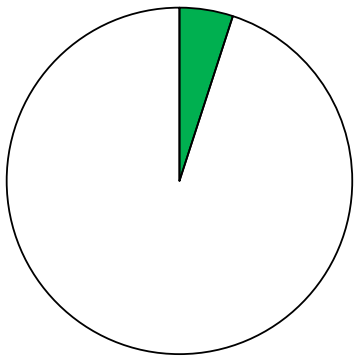

UNITED STATES OF AMERICA
E PLURIBUS UNUM
ONE CENT

Extraordinary preservation

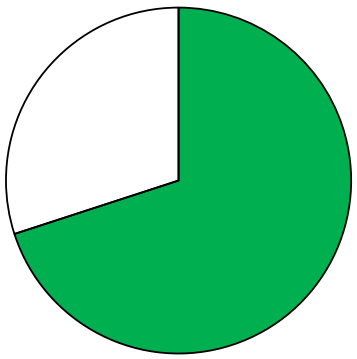


5% endogenous DNA

Best Neandertal bone

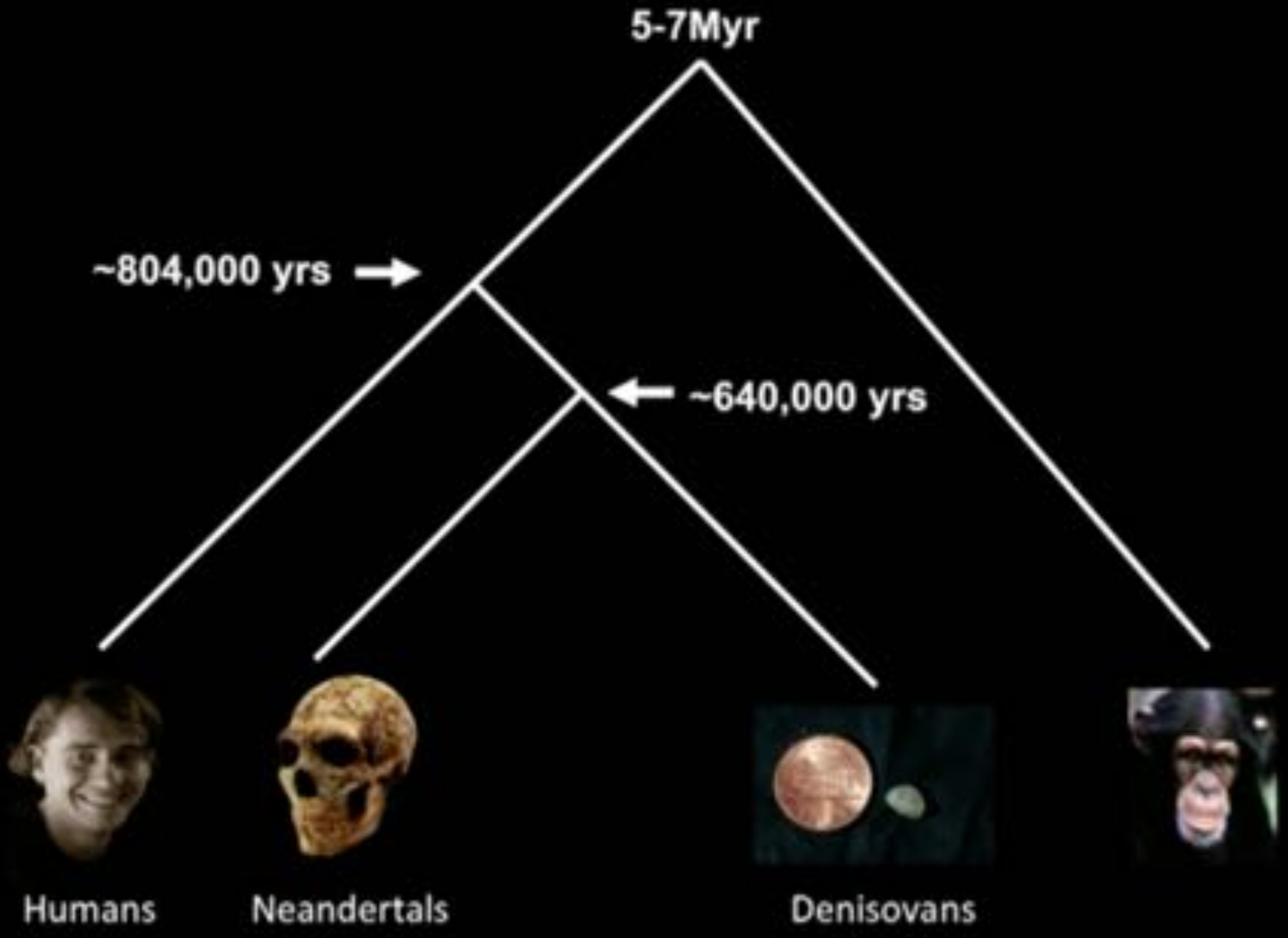


>70% endogenous DNA



Denisova bone

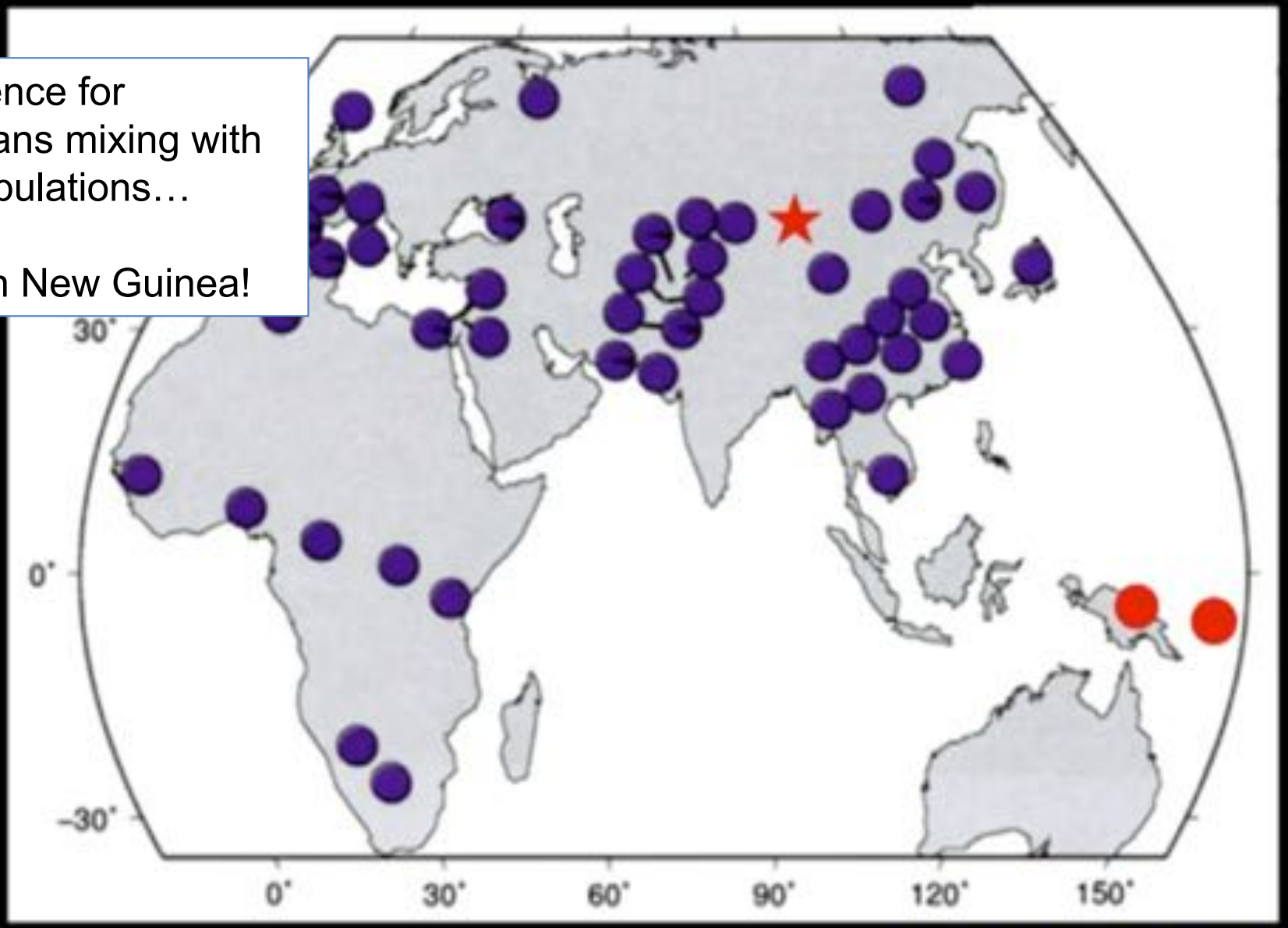
Denisovans & Neandertals



Did we mix?

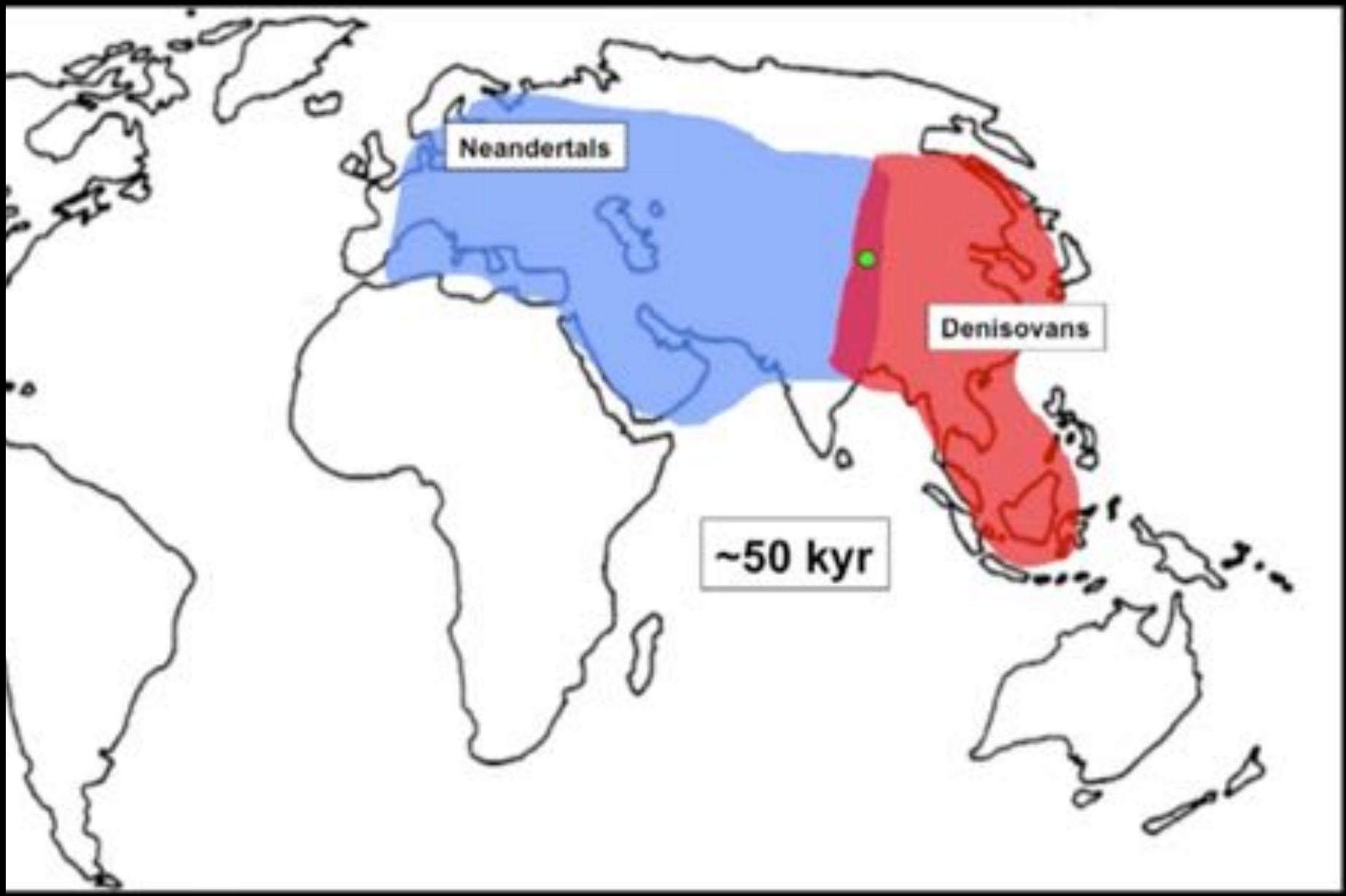
No evidence for Denisovans mixing with other populations…

Except in New Guinea!

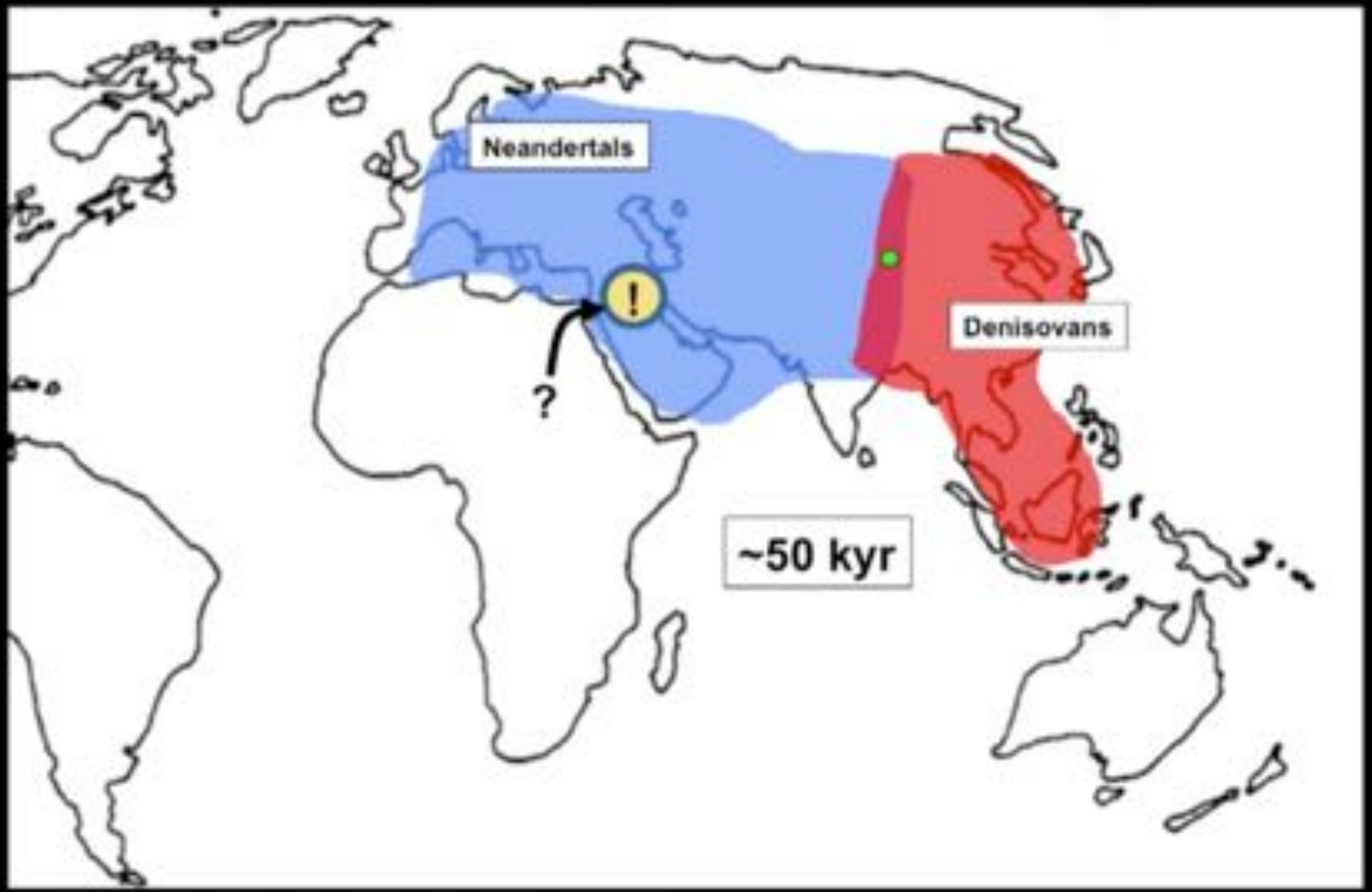


Map after Pickrell et al., 2009

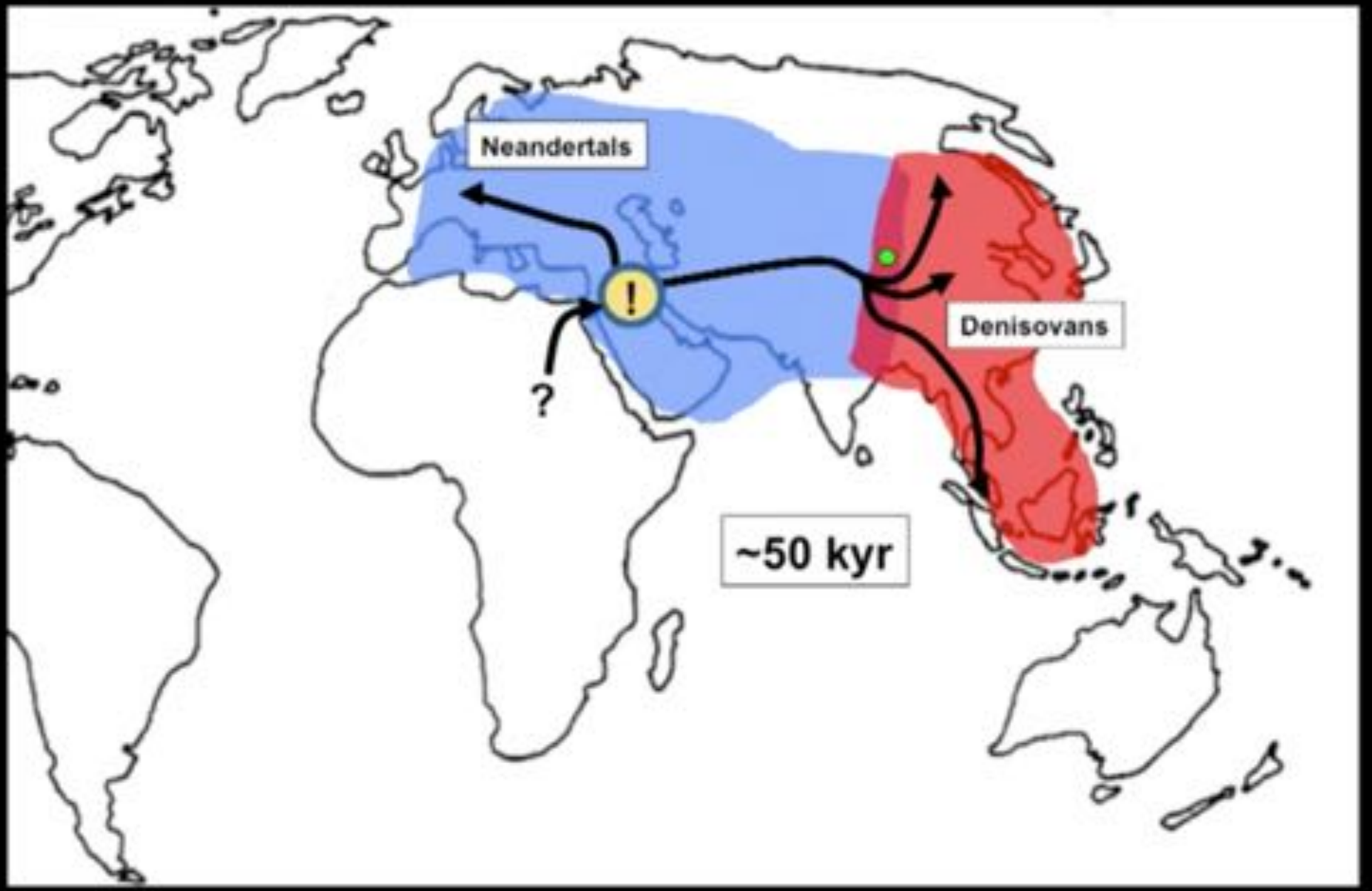
Timeline of ancient hominids



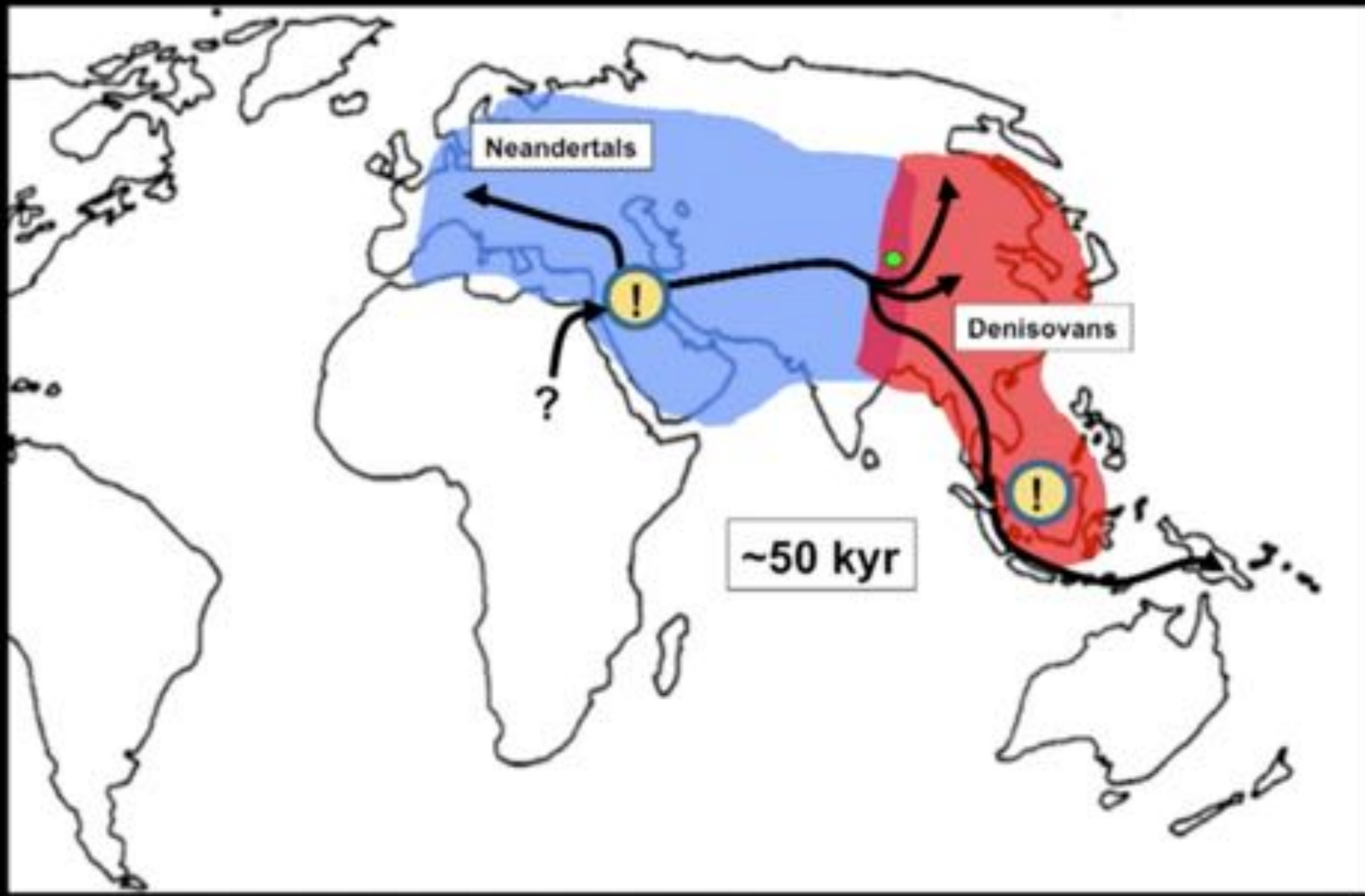
Timeline of ancient hominids



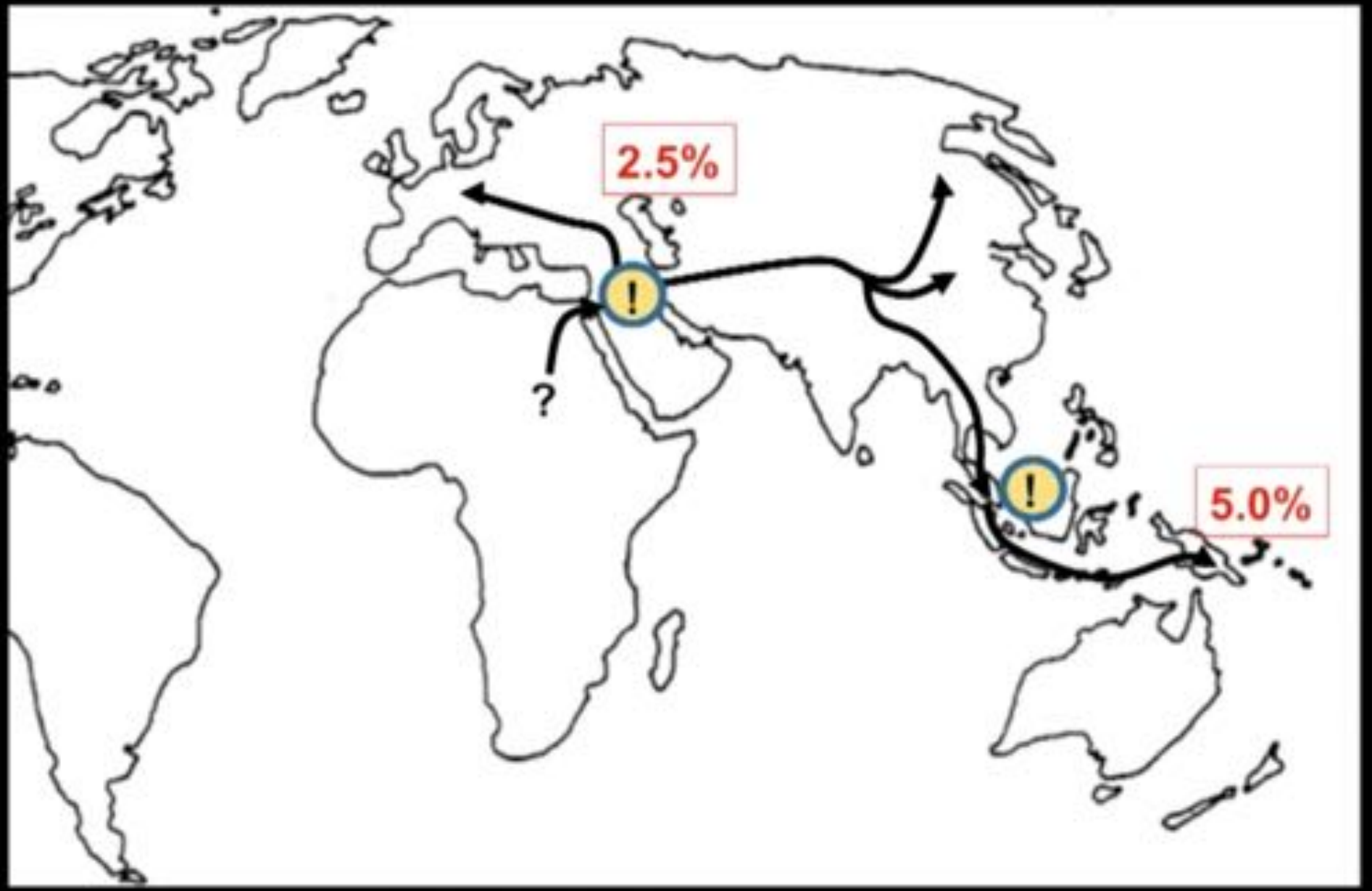
Timeline of ancient hominids



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

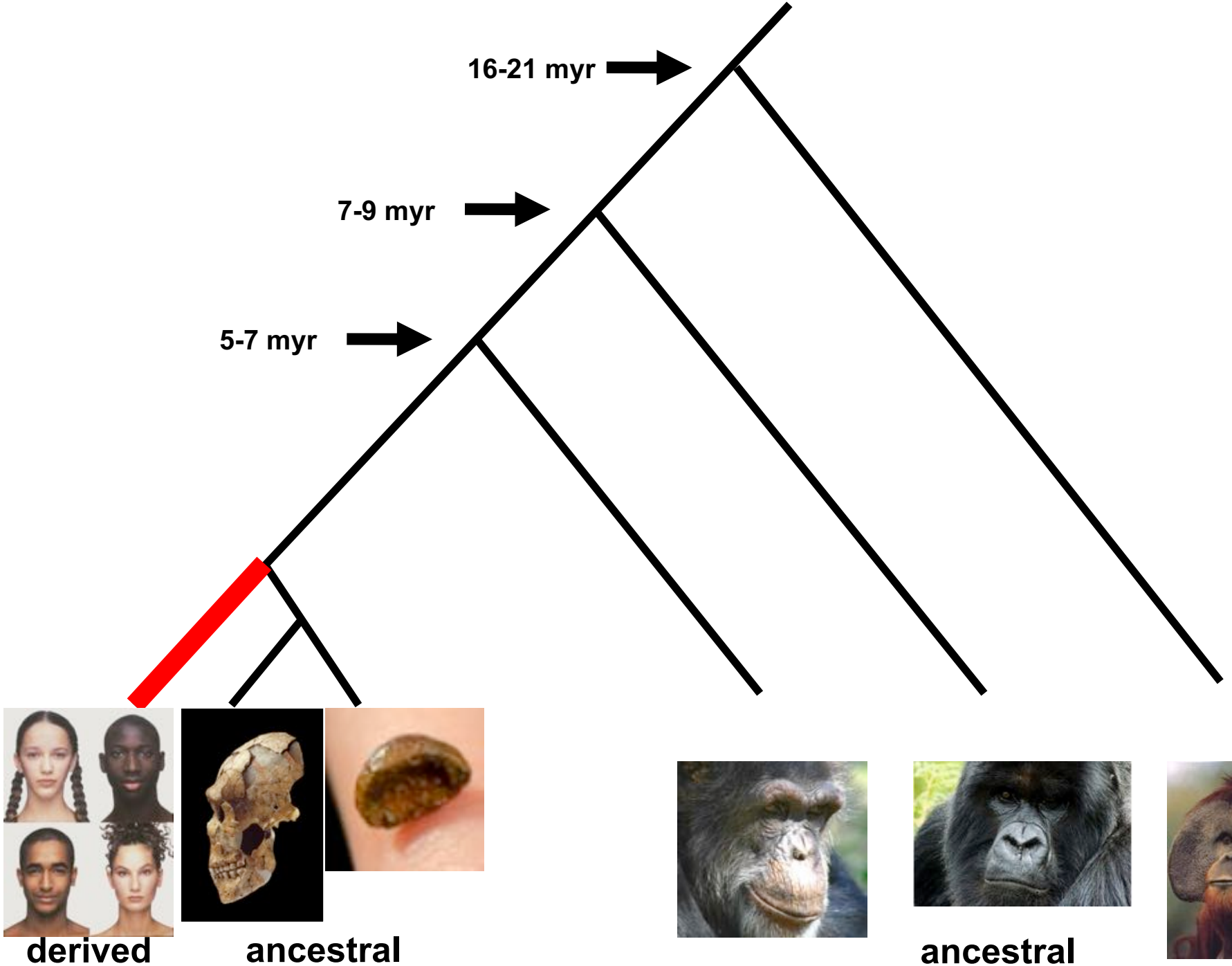
Benjamin Vernot,[1] Serena Tucci,[1,2] Janet Kelso,[3] Joshua G. Schraiber,[1] Aaron B. Wolf,[1] Rachel M. Gittelman,[1] Michael Dannemann,[3] Steffi Grote,[3] Rajiv C. McCoy,[1] Heather Norton,[4] Laura B. Scheinfeldt,[5] David A. Merriwether,[6] George Koki,[7] Jonathan S. Friedlaender,[8] Jon Wakefield,[9]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

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)

FOXP2 Analysis

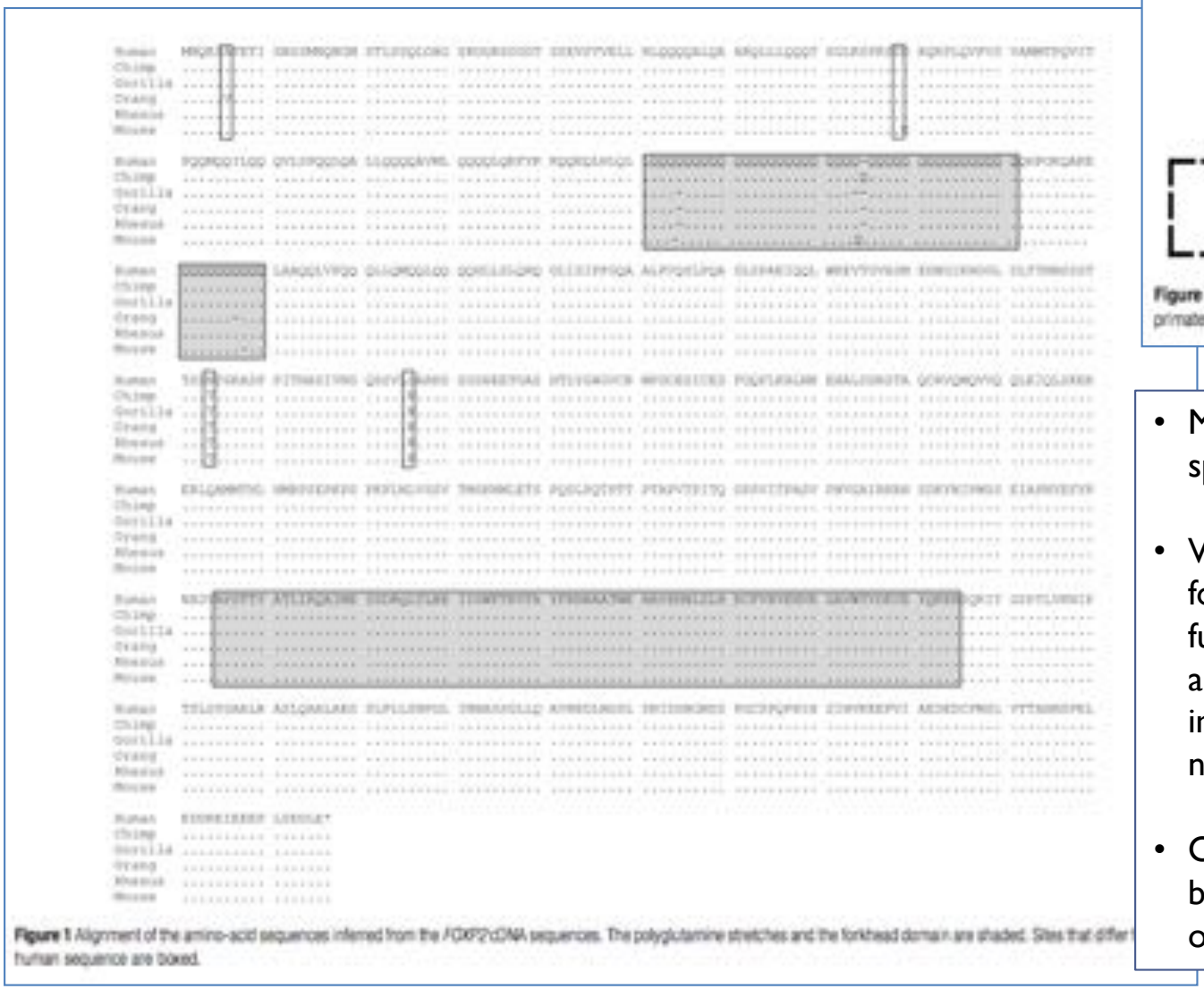


Figure 1 Alignment of the amino-acid sequences inferred from the FOXP2 cDNA sequences. The polyglutamine stretches and the forkhead domain are shaded. Sites that differ from human sequence are boxed.

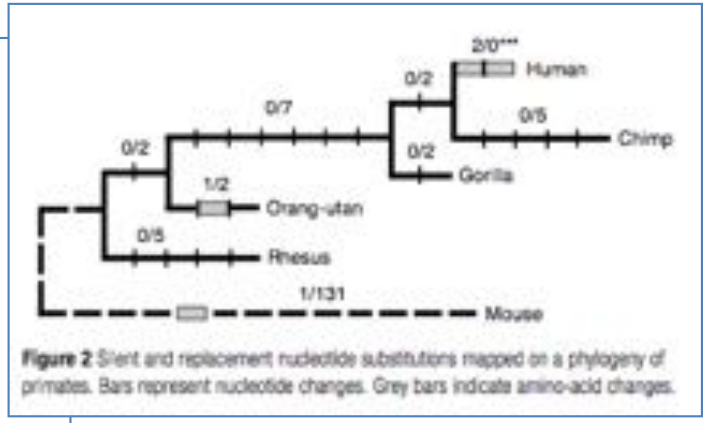


Figure 2 Silent and replacement nucleotide substitutions mapped on a phylogeny of primates. Bars represent nucleotide changes. Grey bars indicate amino-acid changes.

- 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

Part II:
Modern Humans

ARTICLE

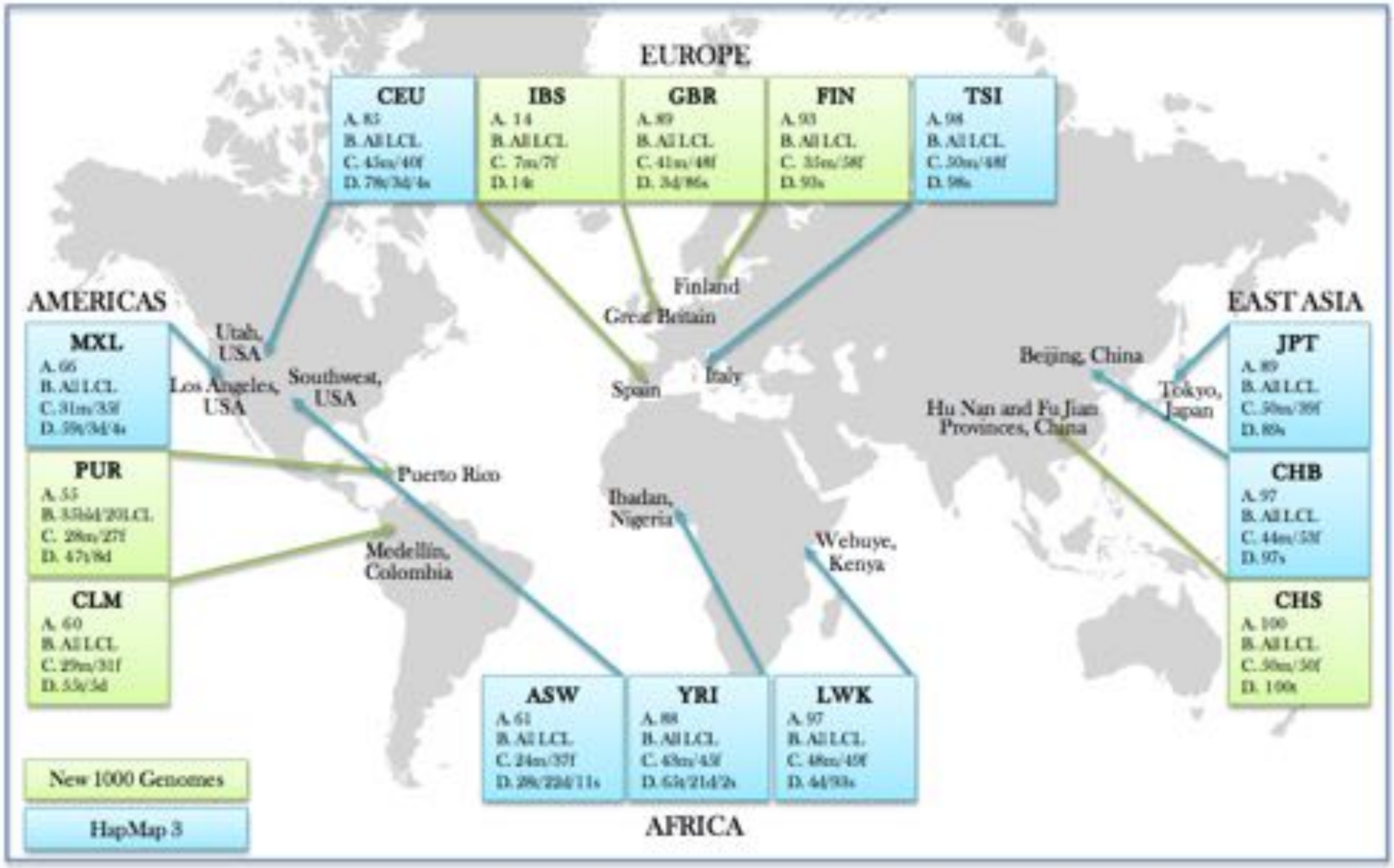
doi:10.1038/nature11632

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



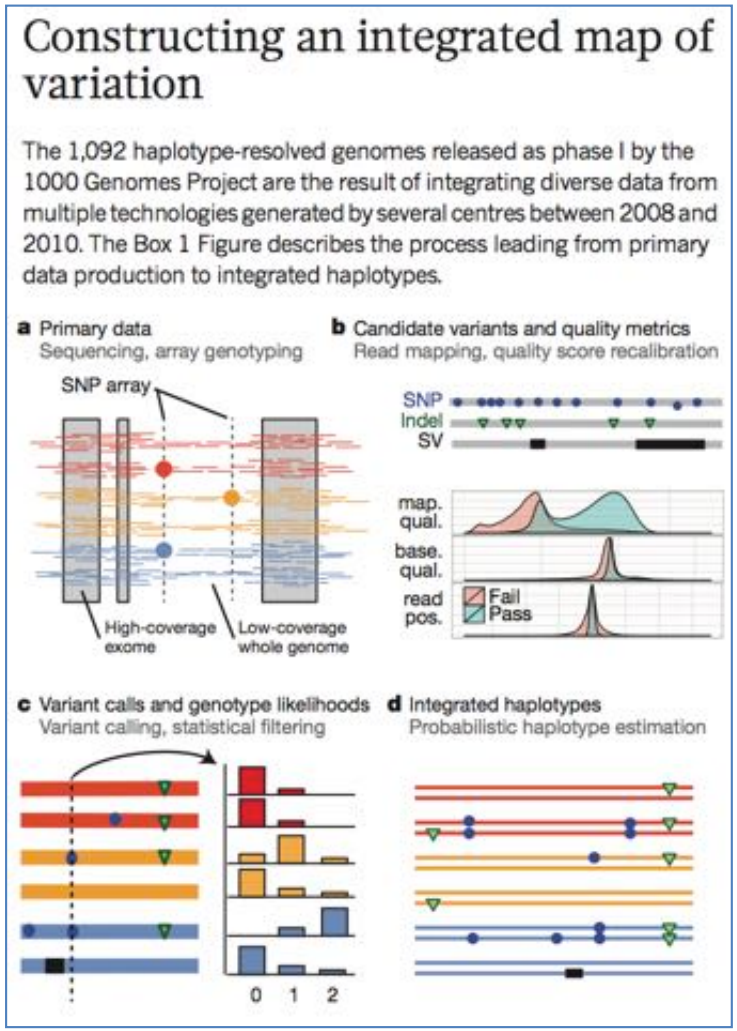
1000 Genomes Populations

Population	DNA sequenced from blood	Offspring Samples from Trios Available	Pilot Samples	Phase 1 Samples	Final Phase Discovery Sample	Final Release Sample	Total
Chinese Dai in Xishuangbanna, China(**CDX**)	no	yes	0	0	99	93	99
Han Chinese in Bejing, China (**CHB**)	no	no	91	97	103	103	106
Japanese in Tokyo, Japan (**JPT**)	no	no	94	89	104	104	105
Kinh in Ho Chi Minh City, Vietnam (**KHV**)	yes	yes	0	0	101	99	101
Southern Han Chinese, China (**CHS**)	no	yes	0	100	108	105	112
Total East Asian Ancestry (EAS)			**185**	**286**	**515**	**504**	**523**
Bengali in Bangladesh (**BEB**)	no	yes	0	0	86	86	86
Gujarati Indian in Houston,TX (**GIH**)	no	yes	0	0	106	103	106
Indian Telugu in the UK (**ITU**)	yes	yes	0	0	103	102	103
Punjabi in Lahore,Pakistan (**PJL**)	yes	yes	0	0	96	96	96
Sri Lankan Tamil in the UK (**STU**)	yes	yes	0	0	103	102	103
Total South Asian Ancestry (SAS)			**0**	**0**	**494**	**489**	**494**
African Ancestry in Southwest US (**ASW**)	no	yes	0	61	66	61	66
African Caribbean in Barbados (**ACB**)	yes	yes	0	0	96	96	96
Esan in Nigeria (**ESN**)	no	yes	0	0	99	99	99
Gambian in Western Division, The Gambia (**GWD**)	no	yes	0	0	113	113	113
Luhya in Webuye, Kenya (**LWK**)	no	yes	102	97	101	99	116
Mende in Sierra Leone (**MSL**)	no	yes	0		85	85	85
Yoruba in Ibadan, Nigeria (**YRI**)	no	yes	106	88	109	108	116
Total African Ancestry (AFR)			**208**	**246**	**669**	**661**	**691**
British in England and Scotland (**GBR**)	no	yes	0	89	92	91	94
Finnish in Finland (**FIN**)	no	no	0	93	99	99	100
Iberian populations in Spain (**IBS**)	no	yes	0	14	107	107	107
Toscani in Italy (**TSI**)	no	no	66	98	108	107	110
Utah residents with Northern and Western European ancestry (**CEU**)	no	yes	94	85	99	98	103
Total European Ancestry (EUR)			**160**	**379**	**505**	**503**	**514**
Colombian in Medellin, Colombia (**CLM**)	no	yes	0	60	94	94	95
Mexican Ancestry in Los Angeles, California (**MXL**)	no	yes	0	66	67	64	69
Peruvian in Lima, Peru (**PEL**)	yes	yes	0	0	86	85	86
Puerto Rican in Puerto Rico (**PUR**)	yes	yes	0	55	105	104	105
Total Americas Ancestry (AMR)				**181**	**352**	**347**	**355**
Total			**853**	**1092**	**2030**	**2004**	**2077**

26 populations from 5 major population groups

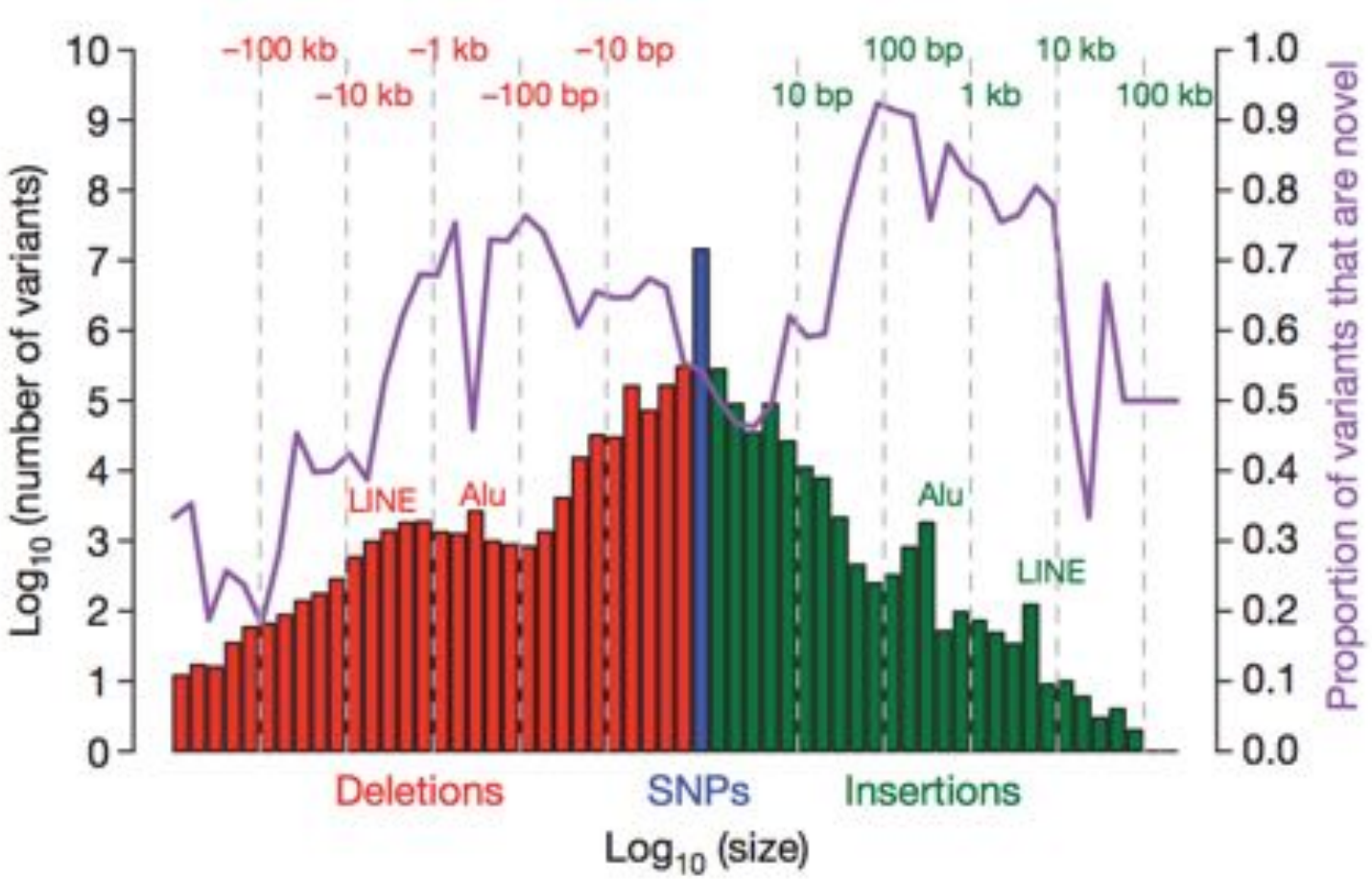
1000 Genomes: Human Mutation Rate

- 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



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

Homozygous LoF Mutations

LETTER

doi:10.1038/nature22034

Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity

Danish Saleheen[1,2]*, Pradeep Natarajan[3,4]*, Irina M. Armean[4,5], Wei Zhao[1], Asif Rasheed[2], Sumeet A. Khetarpal[6], Hong-Hee Won[7], Konrad J. Karczewski[4,5], Anne H. O'Donnell-Luria[4,5,8], Kaitlin E. Samocha[4,5], Benjamin Weisburd[4,5], Namrata Gupta[4], Mozzam Zaidi[2], Maria Samuel[2], Atif Imran[2], Shahid Abbas[9], Faisal Majeed[2], Madiha Ishaq[2], Saba Akhtar[2], Kevin Trindade[6], Megan Mucksavage[6], Nadeem Qamar[10], Khan Shah Zaman[10], Zia Yaqoob[10], Tahir Saghir[10], Syed Nadeem Hasan Rizvi[10], Anis Memon[10], Nadeem Hayyat Mallick[11], Mohammad Ishaq[12], Syed Zahed Rasheed[12], Fazal-ur-Rehman Memon[13], Khalid Mahmood[14], Naveeduddin Ahmed[15], Ron Do[16,17], Ronald M. Krauss[18], Daniel G. MacArthur[4,5], Stacey Gabriel[4], Eric S. Lander[4], Mark J. Daly[4,5], Philippe Frossard[2]§, John Danesh[19,20]§, Daniel J. Rader[6,21]§ & Sekar Kathiresan[3,4]§

A major goal of biomedicine is to understand the function of every gene in the human genome[1]. Loss-of-function mutations can disrupt both copies of a given gene in humans and phenotypic analysis of such 'human knockouts' can provide insight into gene function. Consanguineous unions are more likely to result in offspring carrying homozygous loss-of-function mutations. In Pakistan, consanguinity rates are notably high[2]. Here we sequence the protein-coding regions of 10,503 adult participants in the Pakistan Risk of Myocardial Infarction Study (PROMIS), designed to understand the determinants of cardiometabolic diseases in individuals from South Asia[3]. We identified individuals carrying homozygous predicted loss-of-function (pLoF) mutations, and performed phenotypic analysis involving more than 200 biochemical and disease traits. We enumerated 49,138 rare (<1% minor allele frequency) pLoF mutations. These pLoF mutations are estimated to knock out 1,317 genes, each in at least one participant. Homozygosity for pLoF mutations at *PLA2G7* was associated with absent enzymatic activity of soluble lipoprotein-associated phospholipase A2; at *CYP2F1*, with higher plasma interleukin-8 concentrations; at *TREH*, with lower concentrations of apoB-containing lipoprotein subfractions; at either *A3GALT2* or *NRG4*, with markedly reduced plasma insulin C-peptide concentrations; and at *SLC9A3R1*, with mediators of calcium and phosphate signalling. Heterozygous deficiency of *APOC3* has been shown to protect against coronary heart disease[4,5]; we identified *APOC3* homozygous pLoF carriers in our cohort. We recruited these human knockouts and challenged them with an oral fat load. Compared with family members lacking the mutation, individuals with *APOC3* knocked out displayed marked blunting of the usual post-prandial rise in plasma triglycerides. Overall, these observations provide a roadmap for a 'human knockout project', a systematic effort to understand the phenotypic consequences of complete disruption of genes in humans.

Across all participants (Table 1), exome sequencing yielded 1,639,223 exonic and splice-site sequence variants in 19,026 autosomal genes that passed initial quality control metrics. Of these, 57,137 mutations across 14,345 autosomal genes were annotated as pLoF mutations (that is, nonsense, frameshift, or canonical splice-site mutations predicted to inactivate a gene). To increase the probability that mutations are correctly annotated as pLoF by automated algorithms, we removed nonsense and frameshift mutations occurring within the last 5% of the transcript and within exons flanked by non-canonical splice sites, splice-site mutations at small (<15 bp) introns, at non-canonical splice sites, and where the purported pLoF allele is observed across primates. Common pLoF alleles are less likely to exert strong functional effects as they are less constrained by purifying selection; thus, we define pLoF mutations in the rest of the manuscript as variants with a minor allele frequency (MAF) of <1% and passing the aforementioned bioinformatic filters. Applying these criteria, we generated a set of 49,138 pLoF mutations across 13,074 autosomal genes. The site-frequency spectrum for these pLoF mutations revealed that the majority was seen only in one or a few individuals (Extended Data Fig. 1).

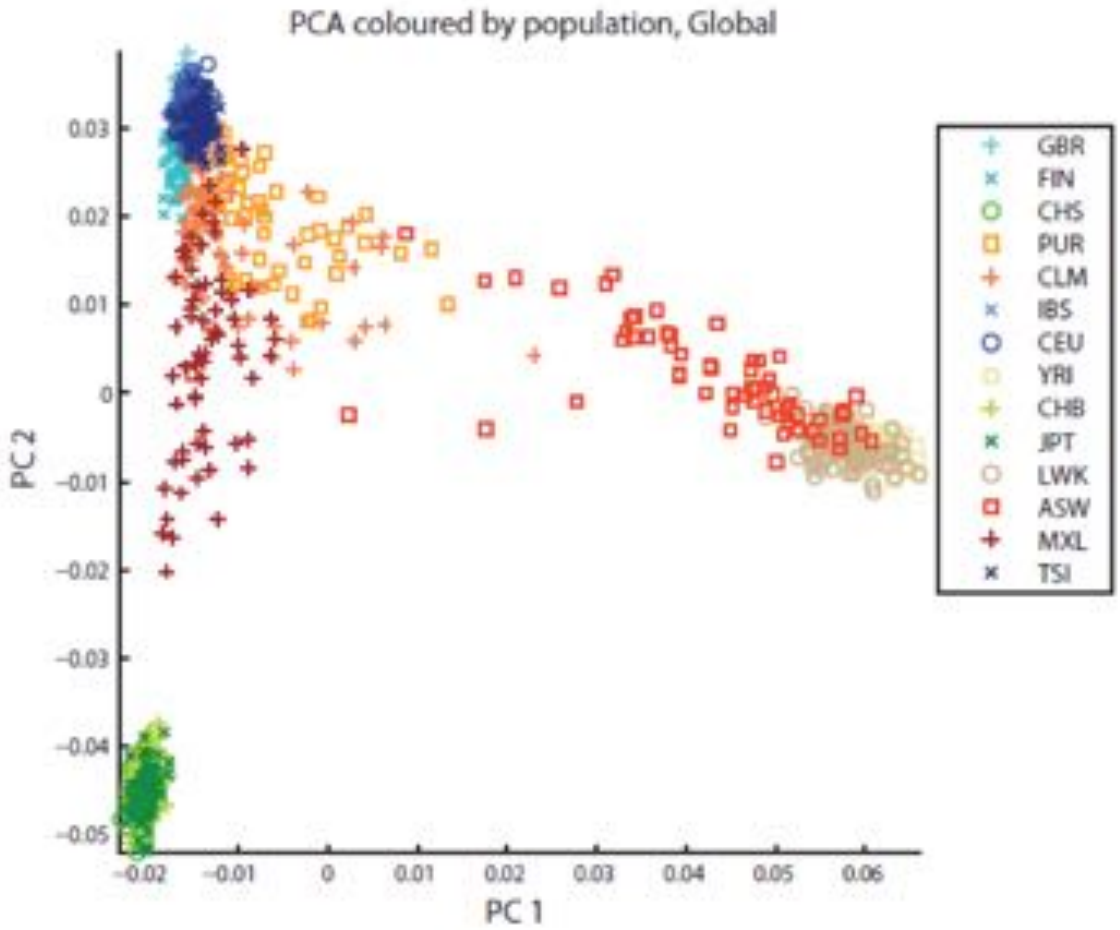
Across all 10,503 PROMIS participants, both copies of 1,317 distinct genes were predicted to be inactivated owing to pLoF mutations. A full listing of all 1,317 genes knocked out, the number of knockout participants for each gene, and the specific pLoF mutation(s) are provided in Supplementary Table 1. 891 (67.7%) of the genes were knocked out only in one participant (Fig. 1a). Nearly 1 in 5 of the participants that were sequenced (1,843 individuals, 17.5%) had at least one gene knocked out by a homozygous pLoF mutation. 1,504 of these 1,843 individuals (81.6%) were homozygous pLoF carriers for just one gene, but the minority of participants had more than one gene knocked out and one participant had six genes with homozygous pLoF genotypes.

We compared the coefficient of inbreeding (*F* coefficient) in PROMIS participants with that of 15,249 individuals from outbred populations of European or African American ancestry. The *F* coefficient estimates the excess homozygosity compared with an outbred ancestor. PROMIS participants had a fourfold higher median inbreeding coefficient compared to outbred populations (0.016 versus 0.0041; $P < 2 \times 10^{-16}$) (Fig. 1b). Additionally, those in PROMIS who reported that their parents were closely related had even higher median inbreeding coefficients than

(2017) Nature. doi:10.1038/nature22034

- Homozygous LoF mutations are rare in most people, but enriched in people born from consanguineous relationships
- Sequence the exomes of many such people, find their homozygous LoFs, relate to 200 biochemical or disease traits
- A "natural" experiment to understand what genes do: people with both copies of APOC3 disabled can clear fat from their bloodstream much faster than others, suggests we should develop compounds to prevent heart attacks

Variation across populations



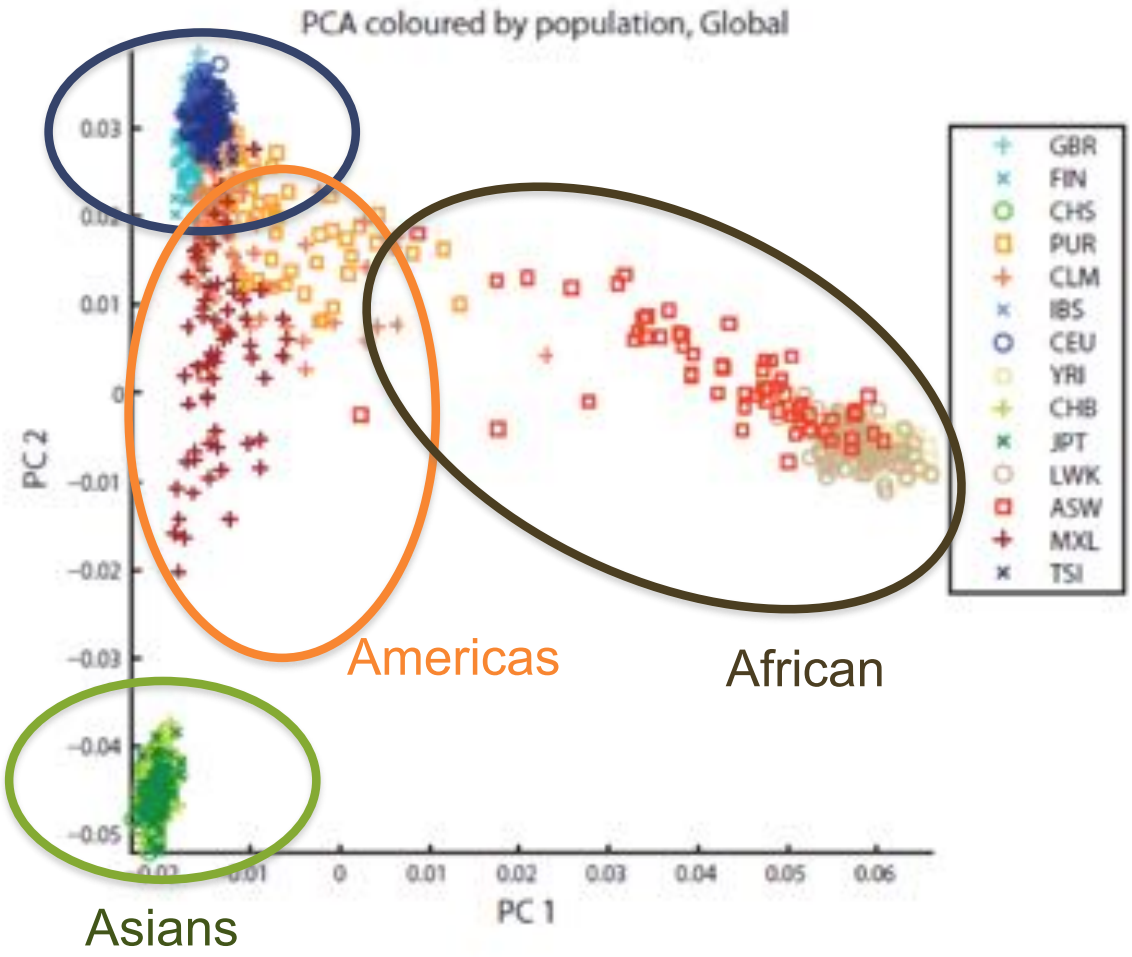
LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

Europeans

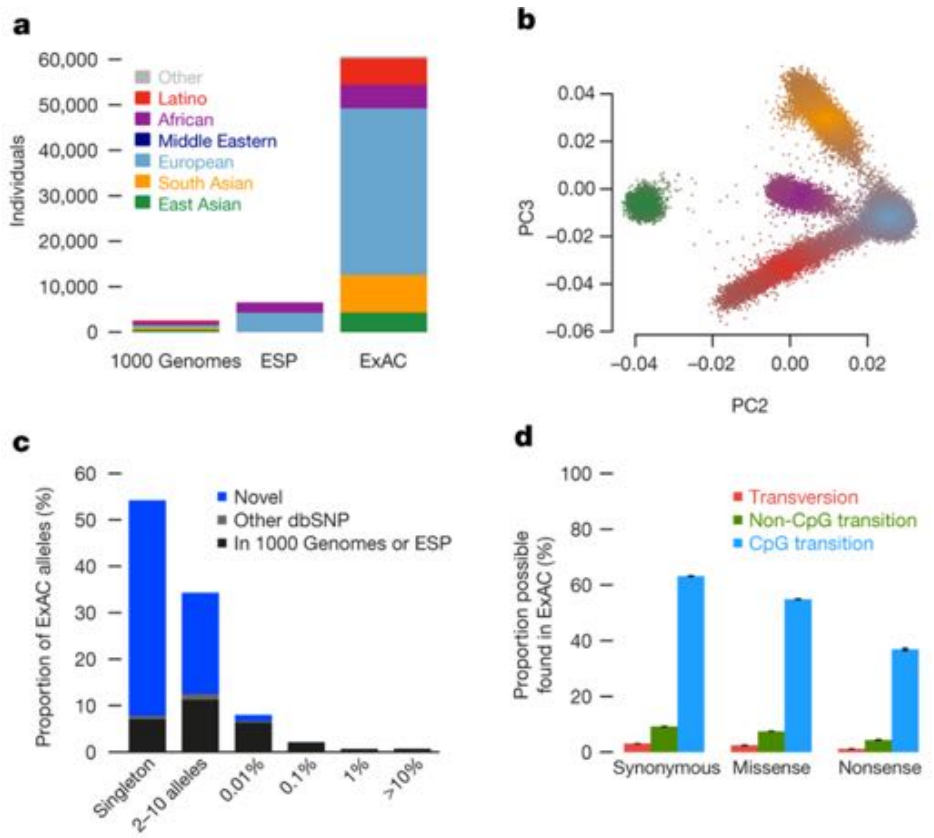


LEVEL	POP_PAIR	# of Highly differentiated SNPs	% in transcribed regions*
AFR	ASW-LWK	258	46.8
AFR	LWK-YRI	251	50.2
AFR	ASW-YRI	213	45.8
ASN	CHS-JPT	275	48.1
ASN	CHB-JPT	176	43.7
ASN	CHB-CHS	79	38.7
EUR	FIN-TSI	343	42.6
EUR	CEU-FIN	201	40.7
EUR	FIN-GBR	197	43.2
EUR	GBR-TSI	100	38.9
EUR	CEU-TSI	57	53.8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

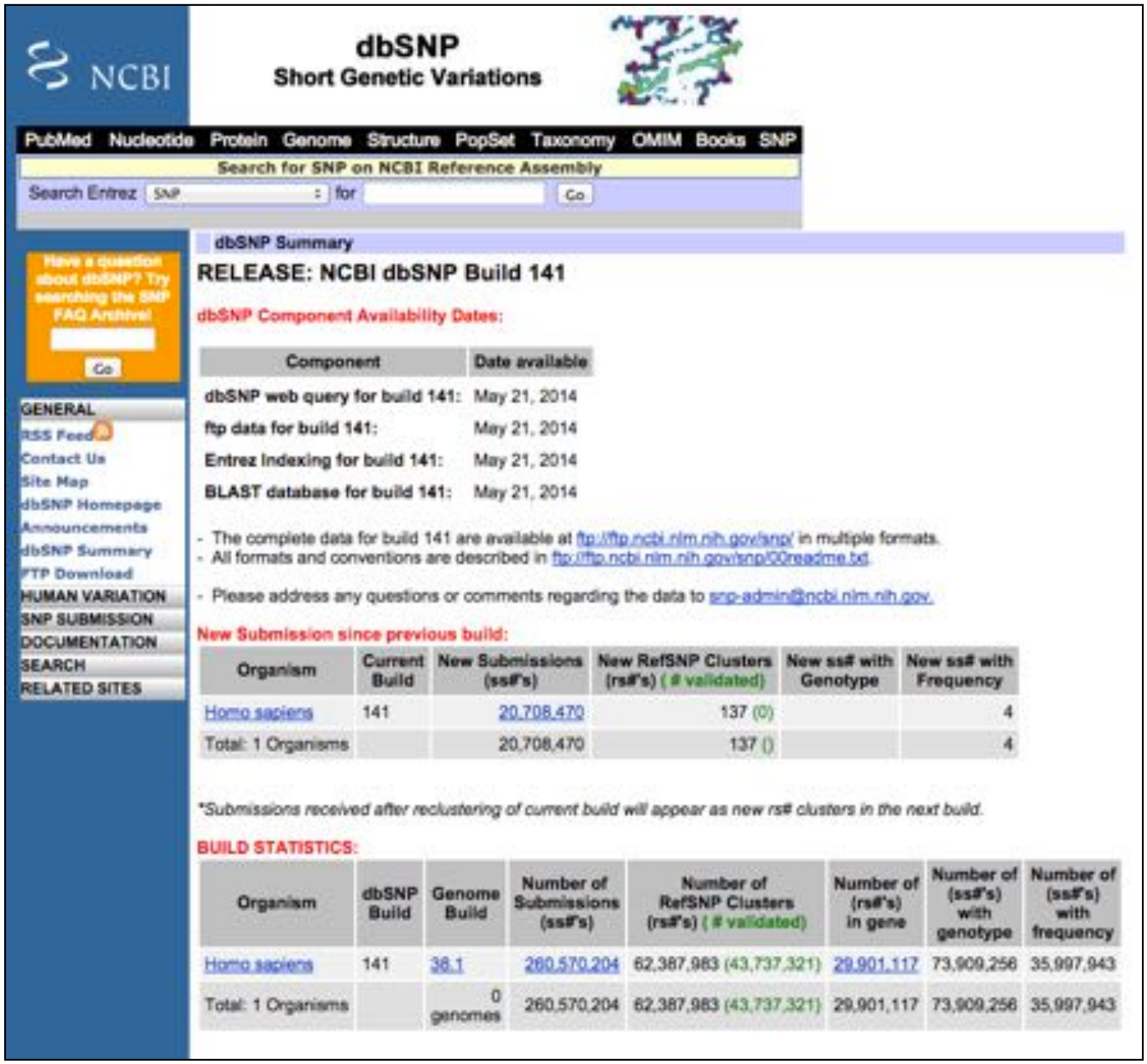
ExAC: Exome Aggregation Consortium



- The aggregation and analysis of high-quality exome (protein-coding region) DNA sequence data for **60,706 individuals**
- This catalogue of human genetic diversity contains an average of **one variant every eight bases of the exome**
- We have used this catalogue to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; **identifying 3,230 genes with near-complete depletion of predicted protein-truncating**

Analysis of protein-coding genetic variation in 60,706 humans
Lek et al (2016) Nature. doi:10.1038/nature19057

dbSNP



NCBI

dbSNP
Short Genetic Variations

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

GENERAL
RSS Feed
Contact Us
Site Map
dbSNP Homepage
Announcements
dbSNP Summary
FTP Download
HUMAN VARIATION
SNP SUBMISSION
DOCUMENTATION
SEARCH
RELATED SITES

dbSNP Summary

RELEASE: NCBI dbSNP Build 141

dbSNP Component Availability Dates:

Component	Date available
dbSNP web query for build 141:	May 21, 2014
ftp data for build 141:	May 21, 2014
Entrez Indexing for build 141:	May 21, 2014
BLAST database for build 141:	May 21, 2014

- The complete data for build 141 are available at ftp://ftp.ncbi.nlm.nih.gov/snp/ in multiple formats.
- All formats and conventions are described in ftp://ftp.ncbi.nlm.nih.gov/snp/00readme.txt.
- Please address any questions or comments regarding the data to snp-admin@ncbi.nlm.nih.gov.

New Submission since previous build:

Organism	Current Build	New Submissions (ss#'s)	New RefSNP Clusters (rs#'s) (# validated)	New ss# with Genotype	New ss# with Frequency
Homo sapiens	141	20,708,470	137 (0)		4
Total: 1 Organisms		20,708,470	137 ()		4

*Submissions received after reclustering of current build will appear as new rs# clusters in the next build.

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	141	38.1	260,570,204	62,387,983 (43,737,321)	29,901,117	73,909,256	35,997,943
Total: 1 Organisms		0 genomes	260,570,204	62,387,983 (43,737,321)	29,901,117	73,909,256	35,997,943

- Periodic release of databases of known variants and their population frequencies
- Generally assumed to be non-disease related
- However, as catalog grows, almost certainly to contain some medically relevant SNPs.

Part III:

Genetic Privacy

Science
AAAS

Identifying Personal Genomes by Surname Inference

Melissa Gymrek *et al.*

Science **339**, 321 (2013);
DOI: 10.1126/science.1229566

What are microsatellites

- **Tandemly repeated sequence motifs**
  - Motifs are 1 – 6 nt long
  - So far, min. 8 nt length, min. 3 tandem repeats for our analyses
- **Ubiquitous in human genome**
  - >5.7 million uninterrupted microsatellites in hg19
- **Extremely unstable**
  - Mutation rate thought to be ~10^{-3} per generation in humans
- **Unique mutation mechanism**
  - Replication slippage during mitosis and meiosis
- **May be under neutral selection**

cCTCTCTCTCTCTCTCTCTCTCTCTCa ➔ $(CT)_{13}$ tCAACAACAACAACAACAACAACAAa ➔ $(CAA)_7$

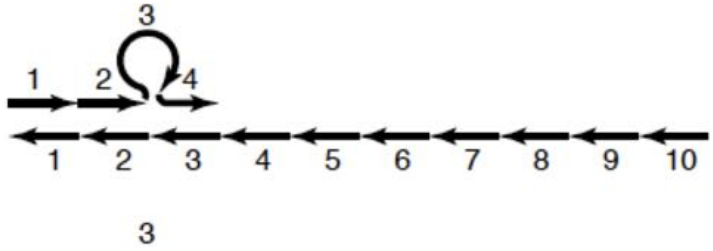
tTTGTCTTGTCTTGTCTTGTCTTGTCTTGTCc ➔ $(TTGTC)_6$ cCATTCATTCATTCATTa ➔ $(CATT)_4$

Microsatellites: Simple Sequences with Complex Evolution
Ellegren (2004) *Nature Reviews Genetics*. doi:10.1038/nrg1348

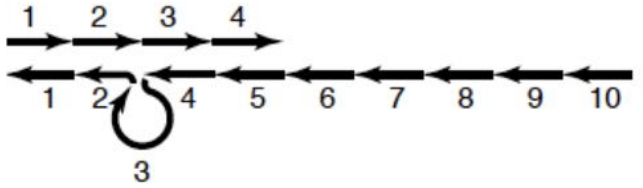
Replication slippage

- **Out-of-phase re-annealing**
  - Nascent and template strands dissociate and re-anneal out-of-phase
- **Loops repaired by mismatch repair machinery (MMR)**
  - Very efficient for small loops
  - Possible strand-specific repair
- **Stepwise process**
  - Nascent strand gains or loses full repeat units
  - Typically single unit mutations
- **Varies by motif length, motif composition, etc.**

Expansion:

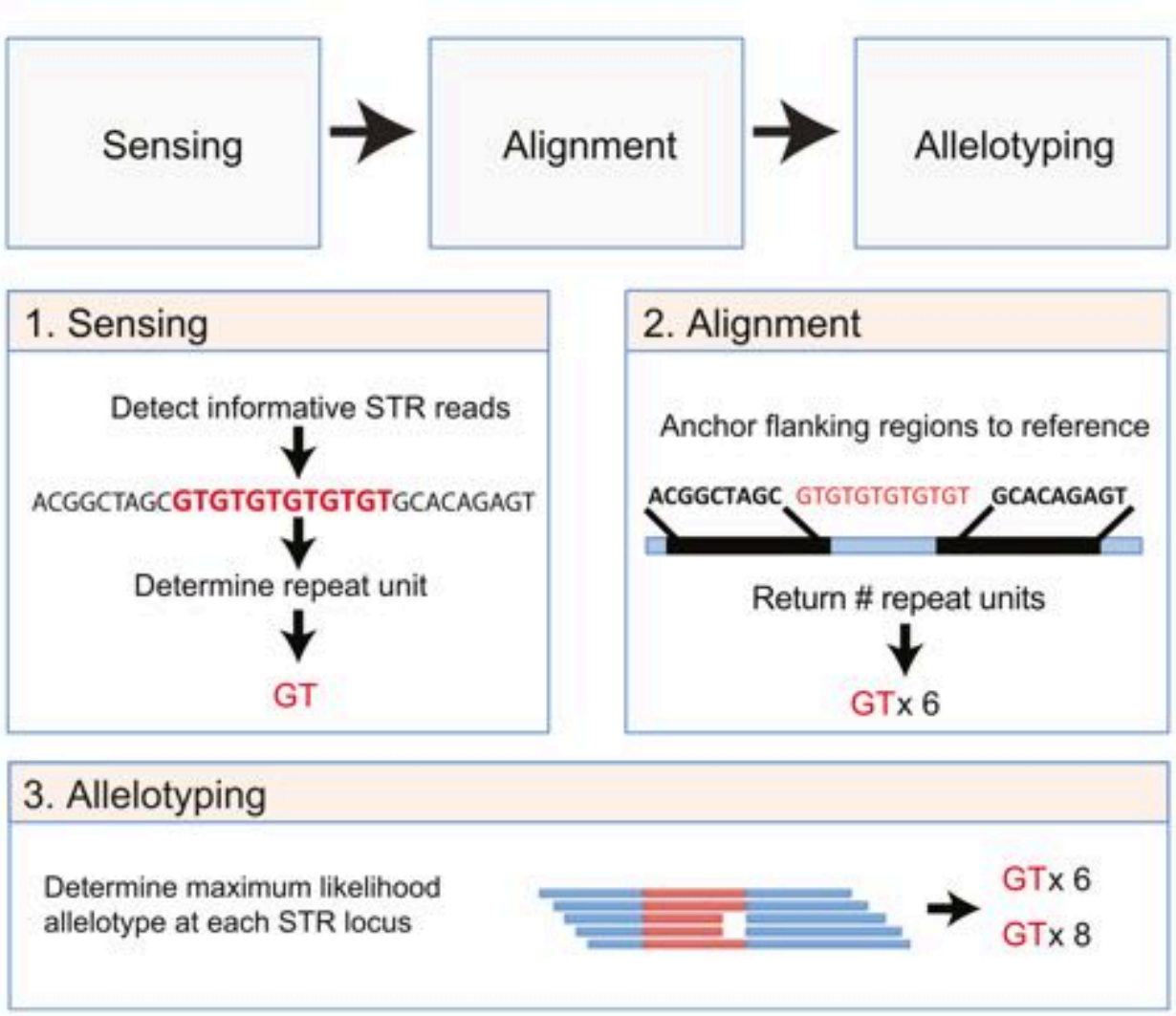


Contraction:



Microsatellites: Simple Sequences with Complex Evolution
Ellegren (2004) *Nature Reviews Genetics*. doi:10.1038/nrg1348

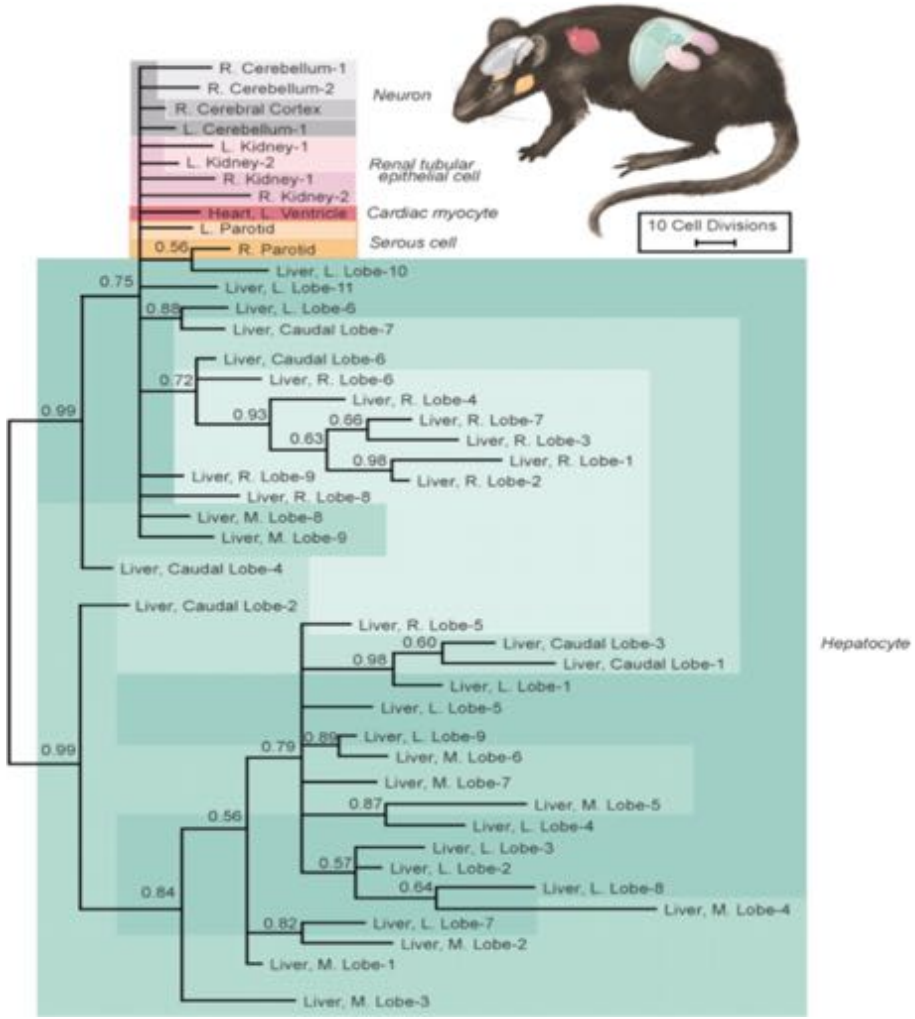
lobSTR Algorithm Overview



lobSTR: A short tandem repeat profiler for personal genomes
Gymrek et al. (2012) *Genome Research.* doi:10.1101/gr.135780.111

Why should we care about microsatellites?

- Polymorphism and mutation rate variation
- Disease
  - Huntington's Disease
  - Fragile X syndrome
  - Friedrich's ataxia
- Mutations as lineage
  - Organogenesis/embryonic development
  - Tumor development



Phylogenetic fate mapping
Salipante (2006) *PNAS*. doi: 10.1073/pnas.0601265103

SERVICES

Combined DNA Index System (CODIS)

The Combined DNA Index System, or CODIS, blends forensic science and computer technology into a tool for linking violent crimes. It enables federal, state, and local forensic laboratories to exchange and compare DNA profiles electronically, thereby linking serial violent crimes to each other and to known offenders. Using the National DNA Index System of CODIS, the National Missing Persons DNA Database also helps identify missing and unidentified individuals.

Overview

CODIS generates investigative leads in cases where biological evidence is recovered from the crime scene. Matches made among profiles in the Forensic Index can link crime scenes together, possibly identifying serial offenders. Based upon a match, police from multiple jurisdictions can coordinate their respective investigations and share the leads they developed independently. Matches made between the Forensic and Offender Indexes provide investigators with the identity of suspected perpetrators. Since names and other personally identifiable information are not stored at NDIS, qualified DNA analysts in the laboratories sharing matching profiles contact each other to confirm the candidate match.

History

The FBI Laboratory's CODIS began as a pilot software project in 1990, serving 14 state and local laboratories. The DNA Identification Act of 1994 formalized the FBI's authority to establish a National DNA Index System (NDIS) for law enforcement purposes. Today, over 190 public law enforcement laboratories participate in NDIS across the United States. Internationally, more than 90 law enforcement laboratories in over 50 countries use the CODIS software for their own database initiatives.

Mission

The CODIS Unit manages CODIS and NDIS. It is responsible for developing, providing, and supporting the CODIS program to federal, state, and local crime laboratories in the United States and selected international law enforcement crime laboratories to foster the exchange and comparison of forensic DNA evidence from violent crime investigations. The CODIS Unit

Genealogy Databases

Genealogy Databases Enable Naming Of Anonymous DNA Donors

Surname Inference

Whose sequence reads are these?

Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

Step 1. Profile Y-STRs from the individual's genome.

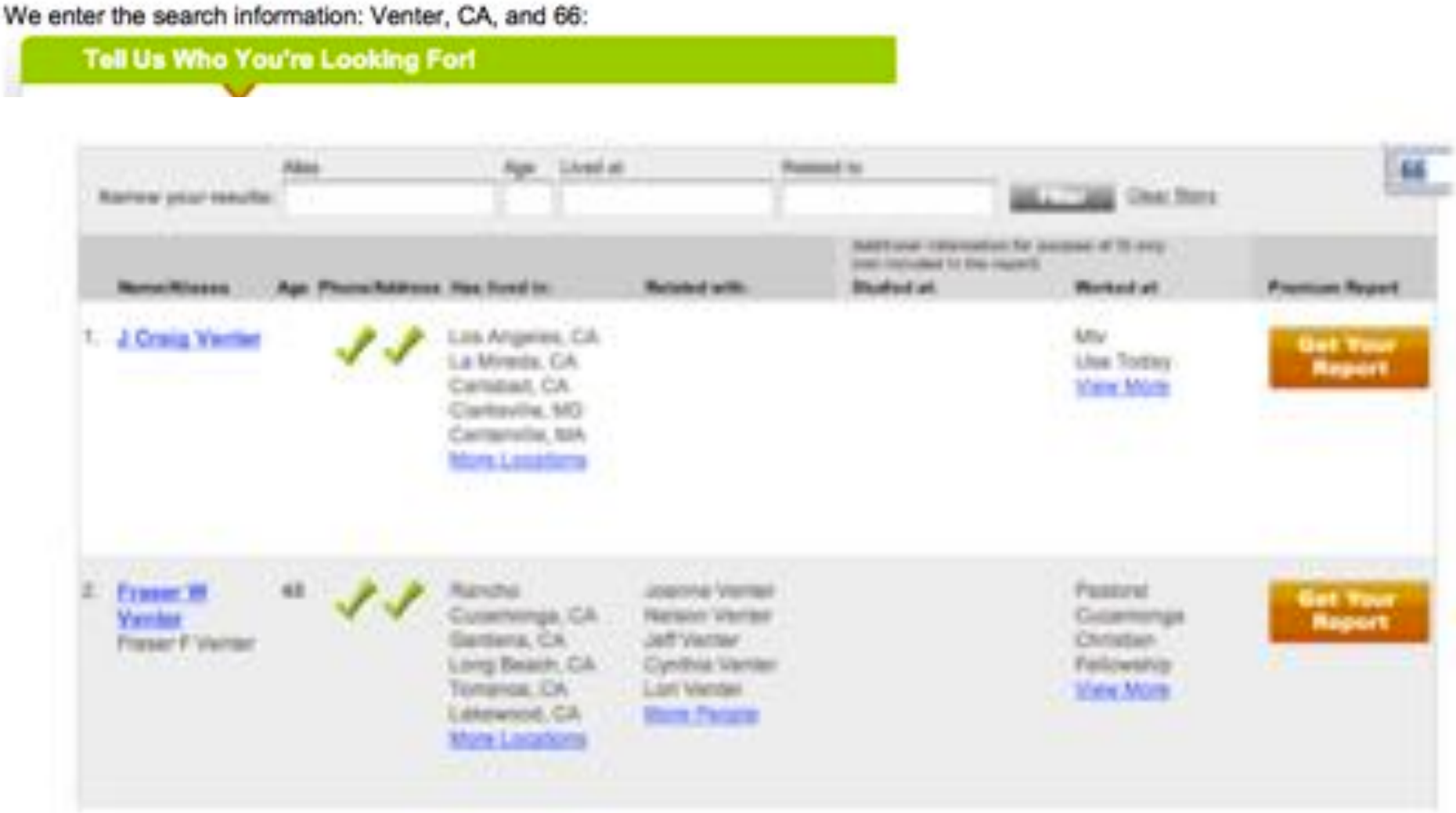
The human reference genome contains 16 copies of "TTTC". Venter has an extra copy of "TTTC", giving him a genotype of "17" at this marker. In a similar way, we can profile all other genealogical STR markers on the Y-chromosome where we know Venter's genome sequence to get the value of a whole panel of these markers.

Step 2. Search for a surname hit in online genetic genealogy databases.

http://www.ysearch.org

Step 3. Search with additional metadata to narrow down the individual.

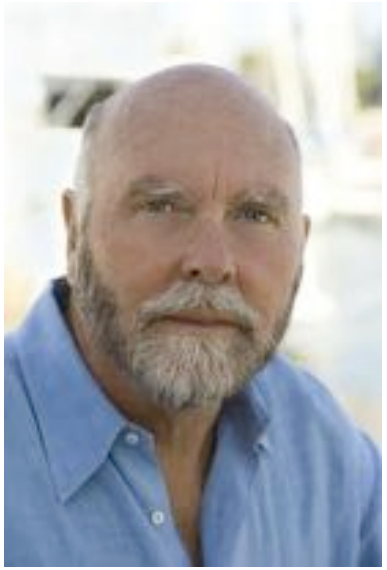
We enter the search information: Venter, CA, and 66:



http://www.ussearch.com

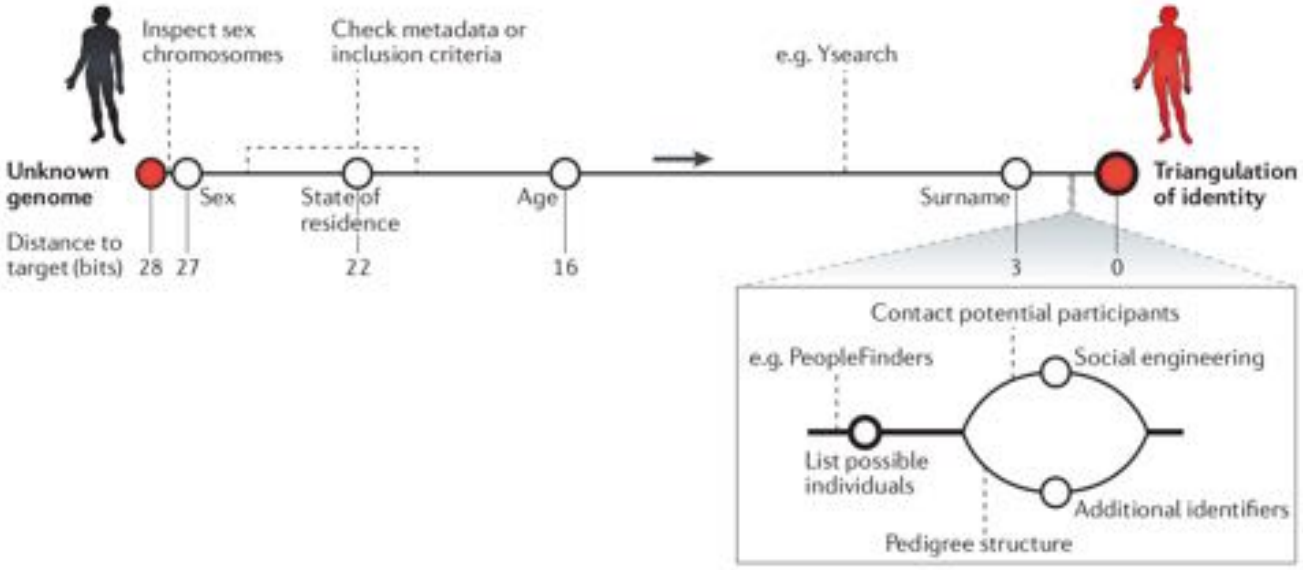
Surname Inference

It's Craig Venter!



Identifying Personal Genomes by Surname Inference
Gymrek et al (2013) *Science*. doi: 10.1126/science.1229566

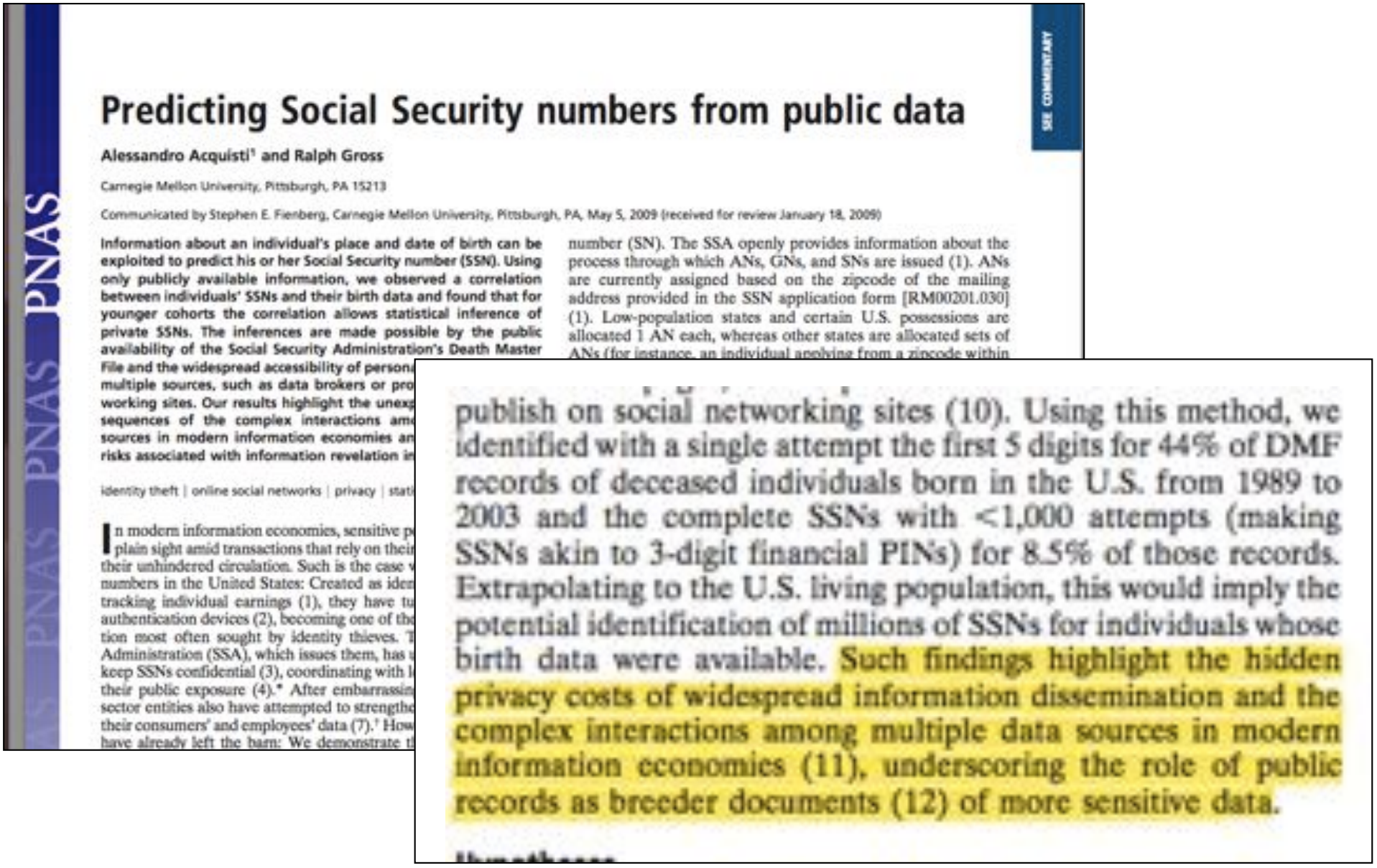
Possible route for identity tracing



- *US population: ~313.9 million individuals*
- *$\log_2$ 313,900,000 = 28.226 bits*
- *Sex ~ 1.0 information bits*
- *$\log_2$ 156,950,000 = 27.226 bits*

- Tracing attacks combine metadata and surname inference to triangulate the identity of an unknown individual.
- With no information, there are roughly 300 million matching individuals in the US, equating to 28.0 bits of entropy.
- Sex reduces entropy by 1 bit, state of residence and age reduces to 16, successful surname inference reduces to ~3 bits.

The risks of big data?



PNAS PNAS PNAS

SEE COMMENTARY

Predicting Social Security numbers from public data

Alessandro Acquisti[1] and Ralph Gross

Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, May 5, 2009 (received for review January 18, 2009)

Information about an individual's place and date of birth can be exploited to predict his or her Social Security number (SSN). Using only publicly available information, we observed a correlation between individuals' SSNs and their birth data and found that for younger cohorts the correlation allows statistical inference of private SSNs. The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of persona multiple sources, such as data brokers or pro working sites. Our results highlight the unexp sequences of the complex interactions am sources in modern information economies an risks associated with information revelation i

identity theft | online social networks | privacy | stati

In modern information economies, sensitive p plain sight amid transactions that rely on thei their unhindered circulation. Such is the case numbers in the United States: Created as ide tracking individual earnings (1), they have tu authentication devices (2), becoming one of th tion most often sought by identity thieves. T Administration (SSA), which issues them, has keep SSNs confidential (3), coordinating with their public exposure (4).* After embarrassi sector entities also have attempted to strengthe their consumers' and employees' data (7).† How have already left the barn: We demonstrate t

number (SN). The SSA openly provides information about the process through which ANs, GNs, and SNs are issued (1). ANs are currently assigned based on the zipcode of the mailing address provided in the SSN application form [RM00201.030] (1). Low-population states and certain U.S. possessions are allocated 1 AN each, whereas other states are allocated sets of ANs (for instance, an individual applying from a zipcode within

publish on social networking sites (10). Using this method, we identified with a single attempt the first 5 digits for 44% of DMF records of deceased individuals born in the U.S. from 1989 to 2003 and the complete SSNs with <1,000 attempts (making SSNs akin to 3-digit financial PINs) for 8.5% of those records. Extrapolating to the U.S. living population, this would imply the potential identification of millions of SSNs for individuals whose birth data were available. Such findings highlight the hidden privacy costs of widespread information dissemination and the complex interactions among multiple data sources in modern information economies (11), underscoring the role of public records as breeder documents (12) of more sensitive data.